NAME Key

MATH 245 - PROBABILITY AND STATISTICS

DECEMBER 9, 2014

LONG EXAM IV

Instructions: *Include all relevant work to get full credit. Encircle your final answers.*

1. A market analyst wished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 120 people provide these data:

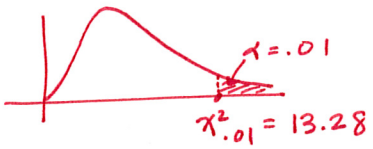
Cherry	Strawberry	Orange	Lime	Grape
32	36	20	18	14

a. Using appropriate mathematical symbols, formulate the most appropriate null and alternative hypotheses to test if consumers have equal preference among the five flavors. [Hint: *The sum of the proportions should equal 1.*] [2]

$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = .2$ vs. H_1: At least one is not equal to .2

b. Using $\alpha = 0.01$, define your rejection rule. [2]

$df = k-1 = 5-1 = 4$



Reject H_0 if $X^2_{obs} > 13.28$

c. Compute the observed value of the appropriate test statistic. [6]

$$X^2_{obs} = \frac{(32-24)^2}{24} + \frac{(36-24)^2}{24} + \frac{(20-24)^2}{24} + \frac{(18-24)^2}{24} + \frac{(14-24)^2}{24}$$

$$= 2.67 + 6 + 0.67 + 1.5 + 4.17$$

$$= 15.01$$

d. Do you reject the null hypothesis? Write a practical conclusion. [3]

Since $X^2_{obs} = 15.01 > 13.28$, we reject H_0. Therefore, we found sufficient evidence that consumers do have a preference among the 5 flavors.

2. A sociologist wishes to see whether the number of years of college a person has completed is related to her or his place of residence. A sample of 400 people is selected and classified as shown in the table below:

Location	*No college*	*4-year degree*	*Advanced degree*	Total
Urban	60	52	40	152
Suburban	32	??	??	140
Rural	24	??	??	108
Total	116	164	120	400

a. In the context of the problem, formulate the appropriate null and alternative hypotheses. [2]

H_0: The number of years of college a person has completed is not related to his/her place of residence.

H_1: The number of years of college completed is related to place of residence

b. Using $\alpha = 0.01$, define your rejection rule. [2]

$df = (r-1)(c-1) = (3-1)(3-1) = 4$

→ Reject H_0 if $X^2_{obs} > 13.28$.

c. Compute the missing expected counts. *Use 2 decimal places.* [4]

Location	*No college*	*4-year degree*	*Advanced degree*	Total
Urban	44.08	62.32	45.60	152
Suburban	40.60	57.4	42	140
Rural	31.32	44.28	32	108
Total	116	164	120	400

d. The table below shows the contribution of each cell to the value of the test statistic. Fill up the missing contributions then compute the observed value of the test statistic. [6]

Location	*No college*	*4-year degree*	*Advanced degree*
Urban	5.74	1.71	0.69
Suburban	1.82	1.54	7.72
Rural	1.71	8.78	4.74

$$X^2_{obs} = 5.74 + 1.71 + \cdots + 4.74$$

$$= 34.45$$

e. Do you reject the null hypothesis? Write a practical conclusion. [3]

Yes. Therefore, we have enough evidence to conclude that the number of years of college completed by a person is related to his/her place of residence.

3. Income and road rage. Is a driver's propensity to engage in road rage related to his or her income? Researchers at Mississippi State University attempted to answer this question by conducting a survey of a representative sample of U.S. adult drivers. Based on how often each driver engaged in certain road rage behaviors (e.g., making obscene gestures, tailgating, and thinking about physically hurting another driver), a road rage score was assigned. (Higher scores indicate a greater pattern of road rage behavior.) The drivers were grouped according to their annual income: under \$30,000, between \$30,000 and \$60,000, and over \$60,000. The data are summarized in the table below.

Income Group	n_i	Average ($\bar{x}_i$)	Sample S.D. (s_i)
Under \$30,000	48	4.60	1.25
\$30,000 to \$60,000	45	5.08	1.32
Over \$60,000	30	5.32	1.42

$n = 123$ $\quad \hookrightarrow \bar{x} = \dfrac{48(4.6) + 45(5.08) + 30(5.32)}{123} = \dfrac{609}{123} \approx 4.95$

a. Identify the following:

i. Experimental units: Drivers [1]

ii. Response Variable: Road rage score [1]

iii. Factor: Income [1]

iv. Treatments: < \$30K ; \$30K - \$60K ; > \$60K [1]

b. Using mathematical symbols, formulate the appropriate null and alternative hypotheses for this problem. Clearly define one of the parameters you used in the null hypothesis. [3]

$H_0: \mu_1 = \mu_2 = \mu_3$ vs H_1: At least one is different.

μ_1 = mean road rage score of people earning < \$30 k a year.

c. Construct the ANOVA Table. **ANOVA Table** [10]

Source	DF	Sum of Squares	Mean Square	F
Treatment	3-1=2	10.75	5.375	3.0926
Error	123-3=120	208.58	1.738	xxxxx
Total	123-1=122	219.33	xxxxx	xxxxx

$$SSTr = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 = 48(4.6-4.95)^2 + 45(5.08-4.95)^2 + 30(5.32-4.95)^2$$
$$= 10.75$$

$$SSE = \sum_{i=1}^{3} (n_i - 1)s_i^2 = (48-1)(1.25^2) + (45-1)(1.32^2) + (30-1)(1.42^2)$$
$$= 208.58$$

d. Using $\alpha = 0.01$, define your rejection rule. [2]

Reject H_0 if $F_{obs} > 4.79$

e. Do you reject the null hypothesis? Write a practical conclusion. [3]

Since $F_{obs} = 3.09 \not> 4.79$, we do not reject H_0.
Hence, we did not find enough evidence to conclude that the mean road rage score of people differ by income level.

f. What are the assumptions for the ANOVA model? [2]

i. The samples came from populations with equal standard deviations.
ii. The samples came from normal populations.
iii. The samples are independent.

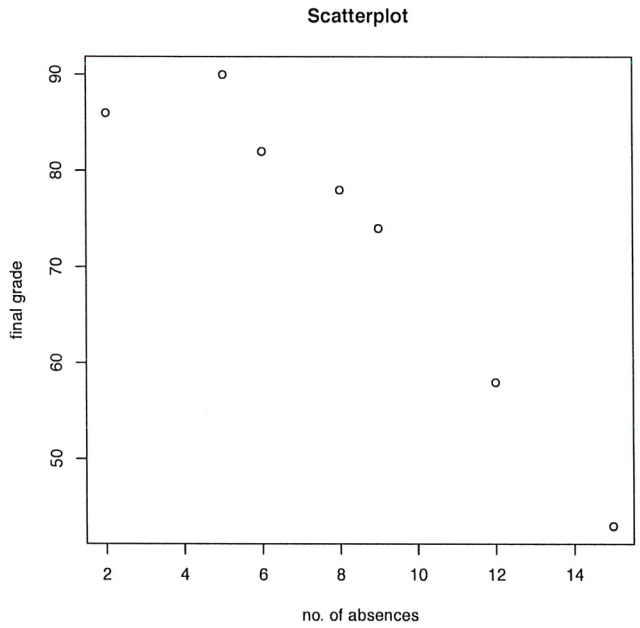
g. Based on our data, was it reasonable to assume equal standard deviations? Explain. [2]

Yes, because the largest s.d. (1.42) is not more than twice the smallest s.d. (1.25).

4. Data from a random sample of 7 students are used to examine the relation between the number of absences(x) and final grade (y) in an elementary statistics course.

The summary statistics for the available data are given below.

$$\sum x = 57 \quad \sum x^2 = 579$$
$$\sum y = 511 \quad \sum y^2 = 38,993$$
$$n = 7 \quad \sum xy = 3,745$$



a. Compute the value of SS_{xx}, SS_{yy}, and SS_{xy}. [6]

$SS_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y) = 3745 - \frac{1}{7}(57)(511) \approx -416$

$SS_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2 = 579 - \frac{1}{7}(57^2) \approx 114.86$

$SS_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2 = 38993 - \frac{1}{7}(511^2) \approx 1690$

b. Determine the correlation coefficient r and the coefficient of determination r^2. Explain the meaning of r^2 in the context of this problem. [5]

$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}} = \frac{-416}{\sqrt{(114.86)(1690)}} \approx -.944 \Rightarrow r^2 = (-.944)^2 = .891$

Hence, about 89.1% of the variability of students' final grade can be explained by a linear relationship with the number of absences.

c. Determine the regression line. [5]

$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{-416}{114.86} \approx -3.62$ and $\hat{a} = \bar{y} - \hat{b}\bar{x} = \left(\frac{511}{7}\right) - (-3.62)\left(\frac{57}{7}\right) = 102.48$

$\Rightarrow \hat{y} = 102.48 - 3.62x$

d. On average, what happens to a student's final grade in an elementary statistics course if he/she misses an additional class? [2]

The student's final grade will decrease by 3.62 points.

e. Calculate the value of $s = \hat{\sigma}$, the estimate of σ_ϵ. [3]

$$SSE = SS_{yy} - \hat{b}\, SS_{xy} = (1690) - (-3.62)(-416) = 184.08$$

$$\Rightarrow \hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{184.08}{7-2}} \approx 6.07$$

f. If the person with 2 absences had a final grade of 86 (see plot), determine the residual for this person. Use the $\hat{\sigma}$ that you obtained in the previous problem to check if this point is an outlier. [5]

$$e_1 = y_1 - \hat{y}_1 = 86 - (102.48 - 3.62(2))$$
$$= 86 - 95.24 = -9.24$$

This residual is not an outlier because it is still within 2 standard deviation of 0.

g. Construct and interpret a 90% confidence interval for b. [7]

$$\hat{b} \pm t_{\frac{\alpha}{2}} \cdot SE_{\hat{b}}, \quad \text{where } SE_{\hat{b}} = \frac{\hat{b}\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{(-3.62)\sqrt{1-.891}}{-.944\sqrt{7-2}} \approx .57$$

$$= -3.62 \pm 2.015\,(.57)$$
$$= -3.62 \pm 1.15 = [-4.77, -2.47]$$

We are 90% confident that b (slope) is between -4.77 and -2.47

h. Using the confidence interval for b that you obtained in the previous problem, test $H_0 : b = 0$ vs. $H_1 : b \neq 0$. Explain how you arrived at your conclusion. What is the significance level of your test? [3]

Since 0 is not in the c.I., we reject $H_0: b = 0$.

The level of significance for this test is $\alpha = .10$.

5. In the simple linear model, $Y_i = a + bx_i + \epsilon_i$, where ϵ_i are independent and identically distributed with mean $E(\epsilon_i) = 0$ and variance $V(\epsilon_i) = \sigma^2$, show that $\hat{b} = \dfrac{SS_{xy}}{SS_{xx}}$ is an unbiased estimator of b. [Note: In this setup, x_i is considered constant and so $E(Y_i) = a + bx_i$.] [10]

Proof:

$$E(\hat{b}) = E\left(\frac{SS_{xy}}{SS_{xx}}\right) = E\left[\frac{\sum x_i Y_i - \frac{1}{n}(\sum x_i)(\sum Y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\right]$$

Since x_i's are considered constants

$$\Rightarrow E(\hat{b}) = \frac{1}{\sum (x_i^2) - \frac{1}{n}(\sum x_i)^2} E\left[\sum x_i Y_i - \frac{1}{n}(\sum x_i)(\sum Y_i)\right]$$
$$= \frac{1}{SS_{xx}} E(SS_{xy})$$

But $E[SS_{xy}] = \sum x_i E(Y_i) - \frac{1}{n}(\sum x_i)(\sum E(Y_i))$

$$= \sum x_i (a + bx_i) - \frac{1}{n}(\sum x_i)(\sum (a + bx_i))$$
$$= a\sum x_i + b\sum x_i^2 - \frac{1}{n}(\sum x_i)(na + b\sum x_i)$$
$$= a\sum x_i + b\sum x_i^2 - a\sum x_i - \frac{b}{n}(\sum x_i)^2$$
$$= b\left(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right)$$

Therefore,

$$E(\hat{b}) = \frac{1}{\sum (x_i^2) - \frac{1}{n}(\sum x_i)^2} * b\left(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right)$$

$$= b. \quad \blacksquare$$