Problem Set 8

Prof. Oke

CEE 260/MIE 273: Probability & Statistics in Civil Engineering *12.06.2024*

Due Tuesday, December 1, 2026 at 12:59 PM as PDF uploaded via Gradescope. ***Show as much work as possible in order to get FULL credit.*** *There are 8 problems with a total of 55 points available.*

Problem 1 *Chi-square hypothesis test conclusions (9 points)*

What conclusions would be appropriate for an upper-tailed chi-square test in each of the following situations (where χ^2 is the test statistic)? (In each case, show explicitly how you compute and compare the critical value $\chi^2_{1-\alpha,\nu}$. Then circle the correct option (**i.**) or (**ii.**).)

(a) $\alpha = 0.01$, $\nu = 3$, $\chi^2 = 8.54$ *[3]*

i. Fail to reject H_0

ii. Reject H_0

(b) $\alpha = 0.10$, $\nu = 2$, $\chi^2 = 4.36$ *[3]*

i. Fail to reject H_0

ii. Reject H_0

(c) $\alpha = 0.01$, $k = 6$, $\chi^2 = 10.20$ *[3]*

i. Fail to reject H_0

ii. Reject H_0

Problem 2 *p-value of chi-square statistic (6 points)*

Calculate the p-value for an upper-tailed chi-square test in each of the following situations (show the Matlab/Python/calculator function you use in each case):

[2] **(a)** $\chi^2 = 13.0$, $\nu = 6$

[2] **(b)** $\chi^2 = 18.0$, $\nu = 9$

[2] **(c)** $\chi^2 = 5.0$, $k = 4$

Problem 3 *ANOVA (8 points)*

The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments[1]. It, for example, collects information on number of hours parents spend taking care of their children under age 6. Provided below is the partial ANOVA output for comparing average hours across educational attainment categories.

	d.o.f.	SS	MS	f	$P(>F)$
Education	4	4142.09	1035.52	?	?
Residuals	794	653047.83	822.48		

(a) Write the hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels. *[2]*

(b) Compute the test statistic. *[2]*

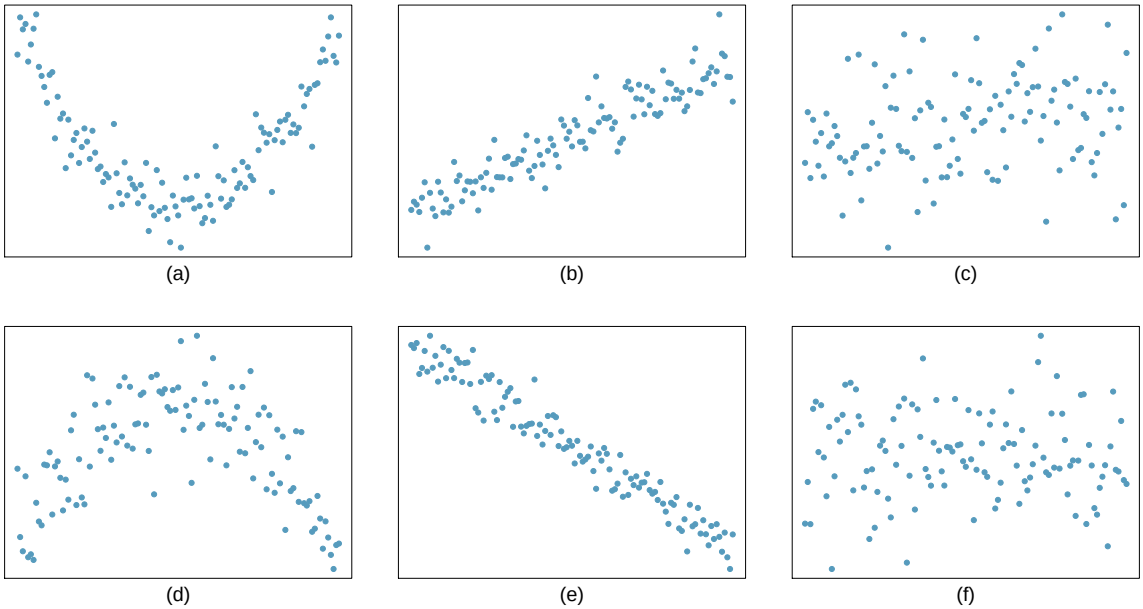
(c) Compute the p-value. *[2]*

(d) What is the conclusion of the hypothesis test and why? *[2]*

[1] UNC Carolina Population Center, https://www.cpc.unc.edu/projects/china

Problem 4 *Identifying relationships (12 points)*

For each of the six plots, identify the strength of the **relationship** (whether linear or not) (e.g. weak, moderate, or strong) in the data **and** whether fitting a linear model would be reasonable.



(a)

(b)

(c)

(d)

(e)

(f)

Problem 5 *Understanding correlation: True/False (4 points)*

Determine if the following statements are absolutely true or false. Briefly explain your reasoning.

(a) A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

(b) Correlation is a measure of the relationship (linear or otherwise) between any two variables.

Problem 6 *Working with a regression equation (7 points)*

Suppose that in a certain chemical process the reaction time y (hr) is related to the temperature (°F) in the chamber in which the reaction takes place according to the simple linear regression model with equation:

$$y = 5.00 - 0.01x$$

and $\sigma = 0.075$.

[1] **(a)** What is the expected change in reaction time for a 1°F increase in temperature?

[2] **(b)** What is the expected reaction time when the temperature is 250°F?

[4] **(c)** Suppose five observations are made independently on reaction time, each one for a temperature of 250°F, what is the probability that all five times are between 2.4 and 2.6 hr?

Problem 7 *Interpreting least-squares linear regression output in MATLAB (9 points)*

The table below shows MATLAB output from fitting the least squares regression line to the number of fatal accidents in each U.S. state to the population of the state.[2]

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	142.71	64.88	2.200	0.0326
x1	0.00013	7.8854	15.933	5.1357×10^{-21}

Table 1: Output from least-squares regression of accidents on population in MATLAB

Let y be the number of accidents in a state and x its population. The "Estimate" column of the table shows the values of the regression coefficients.

(a) Write the regression equation in the form: $y = \hat{\beta}_0 + \hat{\beta}_1 x$ *[2]*

The "SE" column shows the standard errors of the points estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The "tStat" column indicates the respective test statistics (T-scores). The "pValue" column shows the p-values of the test statistics. The hypotheses for β_0 (first row of table) are:

$$H_0 : \beta_0 = 0 \quad \text{(The true intercept is zero)}$$
$$H_1 : \beta_0 \neq 0 \quad \text{(The true intercept is not zero)}$$

(b) Based on the p-value of the intercept, determine whether the intercept is significant (i.e. non-zero) at $\alpha = 0.05$. Would your conclusion change if $\alpha = 0.01$? *[2]*

[2] I would strongly encourage you run the script `least_squares_regression.m` in the M6a folder on Moodle to gain further perspective. More will be said on inference in linear regression during M6b.

The hypotheses for β_1 (second row of table) are:

$$H_0 : \beta_1 = 0 \quad \text{(The true slope is zero)}$$
$$H_1 : \beta_1 \neq 0 \quad \text{(The true slope is not zero)}$$

[2] (c) Based on the size of the p-value for $\hat{\beta}_1$, what is the conclusion of the hypothesis test for β_1? Explain your answer briefly.

[1] (d) Would you agree that the size of of the p-value is a strong indication of the linear relationship between x and y?

[2] (e) If the population of a state were 10,000,000, what would be the expected number of accidents?

Problem 8 *Regression and ANOVA (8 points)*

The following script was run in MATLAB to regress the number of accidents on the population of a state.

```
load accidents
x = hwydata(:,14); %Population of states
y = hwydata(:,4);  %Accidents per state
mdl = fitlm(x,y);
mdl.anova('summary')
```

Below is the output of running the ANOVA method on the model estimate:

	SumSq	DF	MeanSq	F	pValue
Total	3.5776e+07	50	7.1553e+05		
Model	2.9988e+07	1	2.9988e+07	253.86	5.1357e-21
Residual	5.7883e+06	49	1.1813e+05		

(a) What are the values of SSE, SSR and SST? *[3]*

(b) Use the above output to compute the coefficient of determination R^2. *[2]*

(c) How many observations are in the data sample? *[1]*

(d) The null hypothesis of the ANOVA test for a linear regression statest that the true slope is equal to zero: $H_0 : \beta_1 = 0$. Given the p-value in the above output, do you reject or fail to reject the null hypothesis? Briefly explain the reason behind your answer. *[2]*