The questions below are not meant to be all inclusive but examples of certain type of problems you might expect in the final exam. For a complete list, please also review the lecture notes, practice problems, exam 1 and 2, homework, and all labs.

1. Let X have a uniform distribution on the interval 0 to 3, so the density function of X is:
   \[ f(x) = \begin{cases} 
   \frac{1}{3} & \text{if } 0 \leq x \leq 3 \\
   0 & \text{otherwise} 
   \end{cases} \]

   a) (5pts) Find the expected value E(X).
   \[ \mu_X = E(X) = \int_0^3 f(x)dx = \frac{3}{2} \]

   b) (5pts) Find the variance of X, i.e. V(X). (Hint: one formula is: V(X) = E(X^2) – [E(X)]^2)
   \[ Var(X) = \int_0^3 f(x)(x - \mu_X)^2dx = \frac{3}{4} \]

2. (4pts each) For each part below underline all answers that apply (at least one answer is correct):
   a) Event A occurs with probability 0.3, Event B occurs with probability 0.4. If A and B are disjoint (mutually exclusive), then
   \[ \begin{align*}
   &A. \ P(A \text{ and } B) = 0.12 \\
   &B. \ P(A \text{ or } B) = 0.7 \\
   &C. \ P(A \text{ and } B) = 0 \\
   &D. \ P(A \text{ or } B) = 0.58
   \end{align*} \]

   b) Event A occurs with probability 0.3, Event B with probability 0.4. If A and B independent then
   \[ \begin{align*}
   &A. \ P(A \text{ and } B) = 0.12 \\
   &B. \ P(A \text{ or } B) = 0.7 \\
   &C. \ P(A \text{ and } B) = 0 \\
   &D. \ P(A \text{ or } B) = 0.58
   \end{align*} \]

   c) A statistical test is performed and H_0 is not rejected. This means that:
   \[ \begin{align*}
   &A. \ The \ null \ hypothesis \ is \ true. \\
   &B. \ We \ don’t \ have \ enough \ evidence \ for \ the \ alternative. \\
   &C. \ A \ Type \ I \ error \ may \ have \ been \ committed. \\
   &D. \ A \ Type \ II \ error \ may \ have \ been \ committed.
   \end{align*} \]

3. Do poets die young? Data was collected on the age of death for different types of writers—novel writers, poets and writers of non-fiction. A one-way ANOVA was performed to test for differences among the three types of writers with age at death as the dependent variable. A portion of the SAS output is provided below:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>2744.2</td>
<td>1372.1</td>
<td>6.562</td>
<td>0.001973</td>
</tr>
<tr>
<td>Error</td>
<td>120</td>
<td>25088</td>
<td>209.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>122</td>
<td>27832.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
R-Square Coeff Var Root MSE agedeath Mean
0.098598 20.55092 14.45916 70.35772

<table>
<thead>
<tr>
<th>type</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>nonfic</td>
<td>24</td>
<td>76.875000</td>
<td>14.0969084</td>
</tr>
<tr>
<td>novels</td>
<td>67</td>
<td>71.4477612</td>
<td>13.8515105</td>
</tr>
<tr>
<td>poets</td>
<td>32</td>
<td>63.187500</td>
<td>17.2970956</td>
</tr>
</tbody>
</table>

Tukey's Studentized Range (HSD) Test for agedeath

<table>
<thead>
<tr>
<th>Difference</th>
<th>Comparison</th>
<th>Between Means</th>
<th>Simultaneous 95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nonfic - novels</td>
<td>5.427</td>
<td>-2.924 -13.779</td>
</tr>
<tr>
<td></td>
<td>nonfic - poets</td>
<td>13.688</td>
<td>4.208 -23.167 ***</td>
</tr>
<tr>
<td></td>
<td>novels - nonfic</td>
<td>-5.427</td>
<td>-13.779 2.924</td>
</tr>
<tr>
<td></td>
<td>novels - poets</td>
<td>8.260</td>
<td>0.716 -15.804 ***</td>
</tr>
<tr>
<td></td>
<td>poets - nonfic</td>
<td>-13.688</td>
<td>-23.167 4.208 ***</td>
</tr>
<tr>
<td></td>
<td>poets - novels</td>
<td>-8.260</td>
<td>-15.804 -0.716 ***</td>
</tr>
</tbody>
</table>

a) (12pts) Complete the ANOVA table by filling in the seven missing values.

b) (4pts) State the appropriate null and alternative hypotheses for this situation.

\[ H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vs.} \quad H_a : \text{At least one } \mu_i \text{ is different.} \]

Must Clarify: “1” is the death age for nonfiction-writer, “2” is the death age for novelist, and “3” is the death age for poets.

c) Do you reject \( H_0 \) or fail to reject \( H_0 \)? State your conclusion in layman’s terms.

Use \( \alpha = 5\% \). Since the P-value = 0.001973 < \( \alpha \), we can reject \( H_0 \).

In layman’s terms, our data has sufficient evidence to support the statement that at least the death age for one type of writer is different.

d) Briefly summarize the results of the Tukey multiple comparison output shown. Which pairs:

- Are significantly different?
- Are NOT significantly different?

Nonfic vs. poets
Nonfic vs. Novels
Novel vs. poets

e) Using the confidence intervals provided in the Tukey results, state in layman's terms the results of the ANOVA (Hint: you might say something about poets).

Based on the C.I. from Tukey’s results, we have sufficient evidence to claim that the death age of both novelist and non-fiction writer is significant older than that of the poets, since the corresponding C.I. of their difference only include positive numbers.

f) One assumption for ANOVA is constant or equal variance. Has this assumption been violated? Give a mathematical reason.

No, the constant variance assumption is not violated, based on the rule of thumb: the largest sample s.d. (17.30) is less than twice of the smallest sample s.d. (13.05)
4. For each part below underline the answer that applies (only one option is correct for each question):

a) In an RCBD model (Randomized Block Design) the blocking variable:
   A. Is NOT of interest to the researcher
   B. Eliminates variation that would normally be error (reduces the SSE)
   C. Does not have to be significant to be useful
   D. Needs to be analyzed with multiple comparison if it is significant

b) In 2-way ANOVA we are interested in studying:
   a. Only one Factor
   b. Two Factors and a possible interaction effect
   c. Three Factors
   d. A possible interaction effect

c) In the 2-way ANOVA procedure, assume that there are 5 levels of factor A, 4 levels of factor B, and 3 observations (replications) for each of the 20 combinations of levels of the two factors. Then the number of degrees of freedom of the interaction sum of squares, SS(AB), is
   A. 60
   B. 20
   C. 15
   D. 12
   E. 59

5. For the starting NFL quarterbacks in 2007, data was collected for their passing yards per thousand (yrds) and the number of touchdowns (TD). (On the scatterplot, 1yrds = 1000 passing yards, etc). A fan is interested if he can effectively predict the number of touchdowns based on the passing yards. Below is some SAS output:

```
Source                      DF  Sum of Squares  Mean Square  F Value  Pr > F
Model                        1  3033.89742    3033.89742   123.00    <.0001
Error                       30  739.97758     24.66592
Corrected Total             31  3773.87500

Root MSE                    4.96648
Dependent Mean              17.93750
Adj R-Sq                    0.7974

Parameter Estimates

                      Variable  DF  Estimate  Std Error  t Value  Pr > |t|  95% Confidence Limits
Intercept             1   -8.33928   2.52674    -3.30    0.0025  -13.49957    -3.17899
yrds                   1    9.25636    0.83462    11.15    <.0001    7.55184    10.96088
```

```text
```
a) What is the explanatory variable? ______ passing yards

b) Calculate the value of R-Square and interpret it.

\[
R^2 = \frac{SSM}{SSTotal} = \frac{3033.38742}{3773.87500} = 0.803921\%
\]

the coefficient determination, indicates the proportion of variation that can be explained by the linear regression model.

c) State the estimate of the slope and interpret it.

\[ b = 9.25636 \]

it is the estimate of the slope, indicating the amount of increase in the number of touchdowns (y), when 1 unit is increased in the passing yards. Here the number of touchdowns is the response variable y, and the passing yards is the explanatory/predictor variable.

d) Give an estimate for the error standard deviation \( \sigma \) and interpret it.

\[ se = 4.96648 \]

which also estimates the typical amount by which an observation varies about the regression line.

e) (4pts) Construct a 99% confidence interval for \( \beta \).

\[ \text{df} = n - 2 = 30, \text{ and } P-value = \frac{2}{2} \cdot P(T > |t^*|) = 2 \cdot P(T > 11.09) < 0.0002 \]

Therefore the 99% C.I. for \( \beta \) is \( b \pm t\text{-crit}\cdot s_b = 9.25636 \pm 2.750 \cdot 0.83462 = (6.961, 11.552) \)

f) (12 points) Do a test, with \( \alpha = 0.01 \), to determine if there is a significant linear relationship between passing yards and the number of touchdowns.

i) State your hypotheses.

To test the simple linear relationship between y and x, we can use \( H_0 : \beta = 0 \) \ vs. \ \( H_a : \beta \neq 0 \)

ii) Calculate the test statistic.

\[ t^* = \frac{b - 0}{s_b} = \frac{9.25636}{0.83462} \approx 11.09 \]

iii) State the degrees of freedom and the P-value and if you reject \( H_0 \) or fail to reject \( H_0 \).

\[ \text{df for } t^* = n - 2 = 30, \text{ and } P-value = 2 \cdot P(T > |t^*|) = 2 \cdot P(T > 11.09) < 0.0002 \]

Therefore we can reject \( H_0 \).

iv) Write a conclusion in terms of the problem.

Based on above results, we have enough evidence to claim that there is significant linear relationship between passing yards and the number of touchdowns.

g) (4pts) Explain briefly how you could have used the confidence interval in part e to draw the same conclusion you found in part f.

Since 0 does not fall into the 99% C.I. for \( \beta \) in (e), we can claim that \( \beta \) is significantly different from 0, i.e. we can reject \( H_0 : \beta = 0 \) at the level of 1-99% = 0.01
6. We want to explain a person’s systolic blood pressure (SBP) using two predictors, body size measured by a QUET score and AGE. A multiple linear regression is run with SBP as the response. Here is a portion of the output.

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>4108.59225</td>
<td>2054.29612</td>
<td>25.92</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>29</td>
<td>2305.37650</td>
<td>79.49574</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>31</td>
<td>6425.96875</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimate Standard Error t Value Pr > |t|
Intercept 55.32344 12.53475 4.41 0.0001
AGE 1.04516 0.38606 2.71 0.0113
QUET 9.75073 5.40246 1.80 0.0815

a) (4pts) State the multiple regression model for this problem.

\[ y = \alpha + \beta_1 \text{Age} + \beta_2 \text{Quet} + \epsilon \]

with \( \epsilon \sim \text{N}(0, \sigma) \) independently.

b) Do we have a significant multiple linear regression?

i) State the hypotheses.

\[ H_0: \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_a: \text{At least one } \beta \neq 0. \]

ii) State the test statistic, P-value, if you reject \( H_0 \) or fail to reject \( H_0 \).

F Test statistic = 25.92, P-value < 0.0001, so we should reject \( H_0 \).

iii) Write a conclusion in layman’s terms.

Our data has sufficient evidence to claim that, at least one slope in the multiple linear regression is non-zero, i.e. there is linear relationship between Age, Quet score and the SBP.

c) A person is 50 years old with a QUET score of 3.43. What is this person’s predicted SBP?

By plugging age = 50 and QUET = 3.43 into the equation in (a), we get

\[ \hat{y} = 55.32344 + 1.04516 \times 50 + 9.75073 \times 3.43 \approx 141.026 \]

d) For the person in part c, their actual SBP is 149, what is their residual?

\[ \text{Residual} = \text{Actual Value} - \text{Predicted Value} = 149 - 141.026 = 7.974 \]

e) Using the Backward Elimination procedure and \( \alpha=0.05 \), do we need to remove any predictors from the model? (Indicate numbers from the output you are referencing to determine this.)

Yes, based on Backward Elimination strategy, at first we should remove the predictor QUET from the MLR, since the corresponding P-value = 0.0815 > 0.05. After removing it we should re-run the SLR between x (Age) and y(SBP) to check the updated estimates, test statistic and P-values, etc.
7. **MATCHING:** Suppose we own an ice cream parlor and want to study different aspects of our business. For each question below, write the letter of the most appropriate statistical analysis technique that would answer the question.

*Note: each answer choice may be used once, more than once, or not at all.*

- **D** I. Is the percentage of people who prefer chocolate ice cream higher than 25%?
  - A. 1-sample mean

- **I** II. Can we predict ice cream sales (in dollars) based on the outdoor temperature (in degrees)?
  - B. Matched pairs

- **C** III. Do undergraduate students or graduate students spend more money on average purchasing ice cream?
  - C. 2-sample means

- **E** IV. Is the percentage of women who prefer chocolate ice cream higher than the percentage of men who prefer chocolate ice cream?
  - D. 1-sample proportion

- **B** V. We check the prices for both soft-serve ice cream and hard-packed ice cream at the randomly selected 20 ice cream parlors, is there a significant difference between the average price of both kinds of ice cream?
  - E. 2-sample proportion

- **J** VI. We want to see if a customer's dollar amount spent on ice cream during the year is related to several predictors such as age, income, size of family, and residential proximity to an ice cream parlor (reported in miles)?
  - F. One-way ANOVA

- **G** VII. Is there a significant difference between freshmen, sophomores, juniors, and seniors in how much money they spend on ice cream during the year?
  - G. Randomized Block Design

- **A** VIII. Is the average dollar amount a Purdue student spends on ice cream during the year less than $50?
  - H. Two-way ANOVA

- **H** IX. Do gender and social economic status (low, middle, high), as well as their possible interaction effect, have significant effects on how much a customer spends during the year on ice cream (in dollars)?
  - I. Simple linear regression

- **G** X. We want to compare three different "specials" sale strategies on the sale of ice cream. Because time of year affects ice cream sales, we run each "special" for one week during each of the four seasons. Are there differences between the 3 sale strategies?
  - J. Multiple linear regression