Statistics Qualifying Exam

June 11, 2012 : 12 - 4 PM

Name: **M-Number:**

1. Consider a population with three kinds of individuals labeled 1, 2, and 3 and occurring in the Hardy-Weinberg proportions $p(1;\theta) = \theta^2$, $p(2;\theta) = 2\theta(1-\theta)$, $p(3;\theta) = (1-\theta)^2$ where $0 < \theta < 1$.

   (a) Assume we observe a sample of three individuals and obtain $x_1 = 1$, $x_2 = 2$, $x_3 = 1$. Find the MLE of θ.

   (b) In general, let n_1, n_2, and n_3 denote the number of $\{x_1, \dots, x_n\}$ equal to 1, 2, and 3, respectively. Assume that $2n_1 + n_2 > 0$ and $n_3 + 2n_2 > 0$. Show that the MLE exists and is given by,

$$\hat{\theta}(\mathbf{x}) = \frac{2n_1 + n_2}{2n} \ .$$

2. Let X be an exponential(1) random variable with pdf $f(x) = e^{-x}$, $x > 0$, and define Y to be the integer part of $X + 1$, that is

$$Y = i + 1 \quad \text{if and only if} \quad i \le X < i + 1 \ , \quad i = 0, 1, 2, \dots$$

   (a) Find the distribution of Y. What well-known distribution does Y have?

   (b) Find the conditional distribution of $X - 4$ given $Y \ge 5$.

3. Suppose that $X_1, \dots, X_m$ are *i.i.d.* Bernoulli(p) where $0 < p < 1$ is the unknown parameter.

   (a) Show that $T = \sum_{i=1}^m X_i$ is a sufficient statistic (SS) for p.

   (b) Find the minimum variance unbiased estimator (MVUE) of p and of $(1-p)^2$.

4. Let $X_1, \dots, X_n$ be a random sample from $N(\theta, \sigma^2)$ population. Consider testing $H_0 : \theta \le \theta_0$ versus $H_0 : \theta > \theta_0$.

   (a) If σ^2 is known, show that the test that rejects H_0 when $\overline{X} > \theta_0 + z_\alpha \sqrt{\sigma^2/n}$ is a test of size α. Show that the test can be derived as an LRT.

   (b) Show that the test in part (a) is a UMP test.

   (c) What happens if σ^2 is unknown?

5. Information about ocean weather can be extracted from radar returns with the aid of a special algorithm. A study is conducted to estimate the difference in wind speed as measured on the ground and via the Seasat satellite. To do so, wind speeds are measured using two methods simultaneously at 6 specified times. Assume the normality of populations. These data result:

Windspeed, m/s

Time	1	2	3	4	5	6
Ground (x)	4.46	3.99	3.73	3.29	4.82	6.71
Satelite (y)	4.08	3.94	5.00	5.20	3.92	6.21

   (a) Find the 95% confidence interval on the mean difference $(\mu_x - \mu_y)$ in measurements taken by these methods.

(b) Test that two population means are equal ($H_0 : \mu_x = \mu_y$) versus ($H_1 : \mu_x < \mu_y$). Use $\alpha = 0.05$.

6. Suppose that X has the normal distribution $N(\mu, \sigma^2)$. A random sample of size $n = 31$ was drawn. The sample mean is 4.5 and the unbiased sample variance is 3.1.

   (a) Suppose both μ and σ^2 unknown. Find the 90% confidence interval for μ.

   (b) Suppose both μ and σ^2 unknown. Find the 90% confidence interval for σ^2 and σ.

   (c) Let $Z = \frac{X-\mu}{\sigma}$. Denote $Y_1 = Z^2$. Find the probability density function of Y_1.

7. An investigation is conducted to study gasoline mileage in automobiles when used exclusively for urban driving. Ten properly tuned and serviced automobiles manufactured during the same year are used in the study. Each automobile is driven for 1000 miles, and the average number of miles per gallon (mi/gal) obtained (Y) and the weight of the car in tons (X) are recorded. These data result:

Car number	1	2	3	4	5	6	7	8	9	10
Miles per gallon(y)	17.9	16.5	16.4	16.8	18.8	15.5	17.5	16.4	15.9	18.3
Weight in tons(x)	1.35	1.90	1.70	1.80	1.30	2.05	1.60	1.80	1.85	1.40

Summary statistics for these data are
$\sum x = 16.75$, $\sum x^2 = 28.6375$, $\sum y = 170.0$, $\sum y^2 = 2900.46$, $\sum xy = 282.405$

   (a) Fill in the ANOVA table below for the linear regression model:

Source	df	Sum of Square (SS)	Mean Squares (MS)	F-ratio
Regression				
Error				
Total (corrected for mean)				

   (b) Give a 95% confidence interval for the slope parameter.

8. A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3 , an index). The administrator randomly selected 36 patients and collected the data, where larger values of Y, X_2 , and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety. Below is the output for all possible linear regression models fitted for this data. Based on the information, answer the following questions.

```
Number
in Model   R-Square     MSE          SSE        Variables in Model
1          0.5656       0.04348      1.47843    x3
1          0.4045       0.05961      2.02660    x1
1          0.3354       0.06652      2.26177    x2
------------------------------------------------------------------------
2          0.6145       0.03976      1.31195    x1 x3
2          0.5917       0.04211      1.38954    x2 x3
2          0.4219       0.05962      1.96735    x1 x2
------------------------------------------------------------------------
3          0.6150       0.04095      1.31025    x1 x2 x3
```

(a) Use the adjusted R-square to compare the model $y = \beta_0 + \beta_2 X_2 + \epsilon$ and the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

(b) For the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, test the following hypothesis. Use $\alpha = 0.05$.
$H_0 : \beta_1 = \beta_3 = 0$
H_1 : not both β_1 and β_3 are equal to 0.

(c) Use the forward selection regression to select the "best" model. Use the significance level of $\alpha = 0.05$.

9. A study is conducted to investigate the effect of temperature (factor A) and humidity (factor B) on the force required to separate an adhesive product from a certain material. Both factors are treated as fixed effects. The following output was obtained from a computer program that performed a two-factor ANOVA on this experiment.

```
                                 Sum of
Source                   DF      Squares   Mean Square    F Value     Pr > F
A                         3      488.83     ___________    _______    < 0.0001
B                       ___      352.67    ___________    _______     < 0.0001
Interaction             ___     ______     ___________    _______      0.9579
Error                    16       157.33        9.83
Corrected Total          23       1001.83
```

   (a) Fill in the ANOVA table.

   (b) Write down the factor effect model with model assumptions and constraints.

   (c) What conclusions would you draw about this experiment based on the information given above? Use nominal significance level $\alpha = 0.05$.

   (d) The following is output from SAS GLM. Examine the factor effects of temperature by pairwise comparison. Use the Tukey multiple comparison procedure with family significance level $\alpha = 0.05$.

```
 The GLM Procedure
Level of             ------------force------------
temperature    N            Mean          Std Dev
1              6      36.3333333       4.84424057
2              6      31.8333333       5.49241902
3              6      28.0000000       4.42718872
4              6      24.1666667       5.41910202

Level of            ------------force------------
humidity      N            Mean          Std Dev
1            12      33.9166667       5.31649805
2            12      26.2500000       5.54526825

 Level of        Level of            ------------force------------
temperature     humidity     N             Mean           Std Dev
1               1            3       39.6666667        3.51188458
1               2            3       33.0000000        3.60555128
2               1            3       36.0000000        3.00000000
2               2            3       27.6666667        3.78593890
3               1            3       31.6666667        2.51661148
3               2            3       24.3333333        1.52752523
4               1            3       28.3333333        4.16333200
4               2            3       20.0000000        2.0000000
```

10. An experiment was performed to investigate the capability of a measurement system. Ten parts were randomly selected, and three randomly selected operators measured each part three times. The tests were made in random order, and the SAS output obtained is listed as below.

```
The GLM Procedure
Dependent Variable: measure
                          Sum of
Source       DF          Squares     Mean Square    F Value    Pr > F
Model        29      4023.733333      138.749425     271.47    <.0001
Error        60        30.666667        0.511111
Corrected    89      4054.400000
total

R-Square     Coeff Var      Root MSE    measure Mean
0.992436      1.996984      0.714920        35.80000

Source                         DF       Type I SS     Mean Square    F Value    Pr > F
part                            9     3935.955556      437.328395     855.64    <.0001
operator                        2       39.266667       19.633333      38.41    <.0001
part*operator                  18       48.511111        2.695062       5.27    <.0001

Source                         DF     Type III SS     Mean Square    F Value    Pr > F
part                            9     3935.955556      437.328395     855.64    <.0001
operator                        2       39.266667       19.633333      38.41    <.0001
part*operator                  18       48.511111        2.695062       5.27    <.0001
```

(a) Write down the appropriate model and its assumptions.

(b) Obtain the point estimation for all the model parameters.

(c) To test whether or not each effect is significantly different, state the appropriate hypotheses, the corresponding F statistics, the decision rules and the conclusions. Use $\alpha = 0.05$.