MATH 322 Probability & Statistics II
Spring 2011
Final Exam

1. Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from the distribution with the p.d.f

\[
f(x; \theta) = \begin{cases} \theta^2 xe^{-\theta x} & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad \theta > 0
\]

(a) Find the Method of Moments estimators for this distribution.

\[
E(X) = \int_{\theta}^{\infty} x \theta^2 e^{-\theta x} dx
\]

\[
= \theta^2 \left[ \frac{-x}{\theta} e^{-\theta x} \right]_{0}^{\infty} = \theta \left[ \frac{-\infty}{\theta} e^{-\theta x} \right]_{0}^{\infty} + \int_{0}^{\infty} \frac{1}{\theta} x e^{-\theta x} dx
\]

Again integrating by parts

\[
u = x \Rightarrow du = dx, \quad dv = e^{-\theta x} dx, \quad v = -\frac{1}{\theta} e^{-\theta x}
\]

\[
\int_{0}^{\infty} x e^{-\theta x} dx = \frac{1}{\theta^2}
\]

\[
E(X) = x \Rightarrow \frac{2}{\theta} = \bar{x} \Rightarrow \theta = \frac{2}{\bar{x}}
\]

(b) Find the Maximum likelihood estimators (MLE) for this distribution.

\[
L(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta)
\]

\[
= \left( \theta x_1 e^{-\theta x_1} \right) \cdot \left( \theta x_2 e^{-\theta x_2} \right) \cdots \left( \theta x_n e^{-\theta x_n} \right)
\]

\[
= \theta^n \cdot \prod_{i=1}^{n} x_i e^{-\theta \sum_{i=1}^{n} x_i}
\]

The log likelihood is

\[
\ln(L(x; \theta)) = 2n \ln(\theta) + \sum_{i=1}^{n} \ln(x_i)
\]

\[
- \theta \sum_{i=1}^{n} x_i
\]

\[
\frac{d \ln(L(x; \theta))}{d\theta} = \frac{2n}{\theta} - \sum_{i=1}^{n} x_i
\]

\[
= 0 \Rightarrow \frac{2n}{\theta} = \sum_{i=1}^{n} x_i
\]

\[
\Rightarrow \theta = \frac{2n}{\sum_{i=1}^{n} x_i}
\]

\[
= \frac{2}{\bar{x}}
\]
2. In 1995, John Wayne played Genghis Khan in a movie called The Conqueror. Unfortunately the movie was filmed downwind of the site of 11 above-ground nuclear bomb tests. Of the 220 people who worked on this movie, 91 had been diagnosed with cancer by the early 1980s, including Wayne, his co-stars and the director. According to large-scale epidemiological data, only about 14% of people of this age group, on average, should have been stricken with cancer within this time frame. We want to know whether there is evidence for an increased cancer risk of people associated with this film.

(a) What is the best estimate of the probability of a member of the cast or crew getting cancer within the study interval? Assume that this probability is the same for each member of the cast.

\[ \hat{p} = \frac{X}{n} = \frac{91}{220} = 0.4136 \]

(b) What is the standard error of your estimate?

\[ s.e = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.4136)(0.5864)}{220}} = 0.033 \]

(c) What is the 95% confidence interval for this probability estimate? Does this interval bracket the typical cancer rate of 14% for people of the same age group? Interpret the result.

\[ \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

\[ 0.4136 \pm 1.96 \times 0.033 \]

\[ [0.3485, 0.4787] \]

The interval does NOT bracket the typical cancer rate of 14%.

3. A publishing wants to know what percent of the population might be interested in a new magazine on making the most of your retirement. Secondary data (that is several years old) indicates that 22% of the population is retired. They are willing to accept an maximum error rate of 5% and they want to be 95% certain that their finding does not differ from the true rate by more than 5%. What is the required sample size?

\[ n = \left[ \frac{2 \times Z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})}}{M.E} \right]^2 = \frac{4 \times (1.96) \times (0.22)(0.78)}{(0.1)^2} \]

\[ = 263.687 \]

\[ \approx 264 \]
4. A member of Congress wants to determine her popularity in a certain part of the state. She indicates that the proportion of voters who will vote for her must be estimated within plus or minus 2 percent of the population proportion. Further, a 95% confidence level will be used. In past elections, the representative received 40 percent of the popular vote in that area of the state. She doubts whether it has changed much. How many registered voters should be sampled?

\[
\hat{p} = 0.4, \quad \omega = 2(0.02) = 0.04, \quad \alpha = 0.05, \quad Z_{\alpha/2} = 1.96
\]

\[
n = \left[ \frac{2 \cdot Z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})}}{\omega} \right]^2 = \frac{4 \cdot (1.96)^2 \cdot (0.4)(0.6)}{(0.04)^2} = 2304.96 \approx 2305
\]

5. Six healthy three-year-old female Suffolk sheep were injected with the antibiotic Gentamicin, at a dosage of 10 mg/kg body weight. Their blood serum concentration of Gentamicin 1.5 hours after injection was as follows:

\[
33 \quad 26 \quad 34 \quad 31 \quad 23 \quad 25
\]

Construct a 90% confidence interval for the population mean.

\[
\bar{x} = 28.67, \quad s = 4.59, \quad n = 6, \quad \alpha = 0.1, \quad t_{\alpha/2, n-1} = t_{0.05, 5}
\]

\[
\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} = 28.67 \pm (2.015) \frac{4.59}{\sqrt{6}} = (24.89, 32.45)
\]

6. Consider the lengths (in minutes) of the 58 nine-inning games from the first week of the 2010 major league baseball season.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game length</td>
<td>58</td>
<td>178.24</td>
<td>19.53</td>
<td>136.00</td>
<td>165.00</td>
<td>177.00</td>
<td>188.50</td>
<td>218.00</td>
</tr>
</tbody>
</table>

(in min.)

Construct a 90% prediction interval about the game length of the next nine-inning game, not about the average length of all games.

\[
\bar{x} \pm t_{\alpha/2, n-1} \cdot s \sqrt{1 + \frac{1}{n}}
\]

\[
178.24 \pm 1.671 \cdot 19.53 \sqrt{1 + \frac{1}{58}}
\]

\[
(145.325, 211.155)
\]

Since n is large, we can also use \( Z_{\alpha/2} \) instead of \( t_{\alpha/2, n-1} \); we get almost the same interval.
7. A plan for an executive traveler’s club has been developed by an airline on the premise that 5% of its current customers would qualify for membership. A random sample of 500 customers yielded 40 who would qualify. Using this data, (a). Test at level \( \alpha = 0.01 \) the null hypothesis that the company’s premise is correct against the alternative that it is not correct.

\[
H_0 : \hat{p} = P_0 \quad \hat{p} = \frac{40}{500} = 0.08, \quad P_0 = 0.05
\]

\[
H_1 : \hat{p} \neq P_0
\]

Test statistic:

\[
Z = \frac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} = \frac{0.08 - 0.05}{\sqrt{(0.05)(0.95)/500}} = 3.08
\]

\[
P-value = 2 \times [1 - P(Z \leq 3.08)] = 2 \times [1 - P(Z \leq 3.08)] = 2 \times [1 - 0.9990] = 2 \times 0.001
\]

\[
\text{Rejection region } Z \geq Z_{\alpha/2} \text{ or } Z \leq -Z_{\alpha/2}
\]

The critical values are:

\[
Z_{0.005}, -Z_{0.005}
\]

(b). What is the probability that when the test of part (a) is used, the company’s premise will be judged correct when in fact 10% of all current customers qualify?

Type II error = \( P \{ \text{fail to reject } H_0 \text{ when } H_0 \text{ is false} \} = P \{ \text{accept } H_0 \text{ when } P_0 = 0.1 \} \)

Using the equation in book:

\[
\beta = \Phi \left[ \frac{0.05 - 0.1 + 2.575 \sqrt{(0.05)(0.95)/500}}{\sqrt{(0.01)(0.9)/500}} \right] - \Phi \left[ \frac{0.05 - 0.1 - 2.575 \times \sqrt{(0.01)(0.9)/500}}{\sqrt{(0.01)(0.9)/500}} \right]
\]

\[
= \Phi (-1.80) - \Phi (-5.6)
\]

\[
= 0.0314 - 0 \approx 0.0314
\]

8. An advertisement for a particular brand of automobile states that it accelerates from 0 to 60 mph in an average of 5.0 seconds. Makers of a competing automobile feel that the true average number of seconds it takes to reach 60 mph from zero is above 5.0. Suppose the population standard deviation is believed to be 0.43 seconds. We want to test the hypothesis

\( H_0 : \mu = 5.0 \text{ vs } H_a : \mu > 5.0 \)

(a). For what values of the sample mean \( \bar{X} \) should we reject \( H_0 \), if a 5% level of significance will be used, and 50 automobiles will be tested.

\[
\alpha = 0.05 = P \{ \text{reject } H_0 \text{ when } H_0 \text{ is true} \}
\]

\[
= P \{ \bar{X} > \bar{X} \text{ when } \mu = 5, \sigma = 0.43, n = 50 \}
\]

\[
0.05 = P \{ Z > \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \}
\]

\[
1.645 = \frac{\bar{X} - 5}{0.43 / \sqrt{50}} \Rightarrow \bar{X} > 5.1
\]
(b). Find the power of the test if the true average is 5.15 seconds, a 5% level of significance will be used, and 50 automobiles will be tested.

\[
\beta = P\{\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}\} \\
= P\{\bar{x} < 5.1 \text{ when } \mu = 5.15, \sigma = 0.43, n = 50\} \\
= P\left\{ \frac{Z < \frac{5.1 - 5.15}{\sqrt{0.43^2/50}}} \right\} \\
= P\{Z < -0.822\} \\
= 0.2061
\]

\[
\text{Power} = 1 - \beta = 1 - 0.2061 = 0.7939
\]

(c). Accept \( H_0 \) if \( \bar{X} \) (\( \bar{X} \)) is between 4.9 and 5.1, otherwise reject \( H_0 \). If the sample size 50 is used and sigma = 0.43, then compute alpha (\( \alpha \))

\[
\alpha = P\{\text{Reject } H_0 \text{ when } H_0 \text{ is true}\} \\
= P\{\bar{x} < 4.9 \text{ or } \bar{x} > 5.1 \text{ when } \mu = 5, \sigma = 0.43, n = 50\} \\
= P\{\bar{x} < 4.9\} + P\{\bar{x} > 5.1\} \\
= P(Z < -1.644) + P(Z > 1.644) \\
= 0.0505 + (1 - P(Z \leq 1.644)) \\
= 0.0505 + (1 - 0.9495) = 0.0505 + 0.0505 = 0.101
\]

9. Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n=19 \) from the normal distribution \( N(\mu, \sigma) \).

(a). Find the rejection region of size \( \alpha = 0.05 \) for testing \( H_0: \sigma^2 = 30 \text{ vs } H_a: \sigma^2 > 30 \)

Test statistic: \( \chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{18.82}{30} \)

Reject \( H_0 \) if \( \chi^2 > \chi^2_{\alpha, (n-1)} \) where \( \chi^2_{0.05, 18} = 28.87 \)

\[
\Rightarrow \frac{18.82}{30} > 28.87 \\
\Rightarrow s^2 > 48.116
\]
(b) What is the probability of type II Error if true \( \sigma^2 = 80. \)

\[
\text{Type II error} = P\{\text{Accept } H_0 \text{ when } H_0 \text{ is false}\} \\
= P\{S^2 < 48.116 \text{ when } \sigma^2 = 80\} \\
= P\left\{ \frac{(n-1)s^2}{\sigma^2} < \frac{(48.116) \times 18}{80} \right\} \\
= P\left\{ \chi^2_{(18)} < 10.826 \right\} \\
= 0.10 \\
\text{Since } \chi^2_{0.90, 18} = 10.86
\]

10. Madison Department of Transportation needs to analyze the average speed of cars passing downtown area, in order to see whether more traffic lights are needed for safety. Suppose, it is believed that an average speed of 15mph (or 24.14 kmph: kilometers per hour) is ideal for the balance of safety and smooth traffic. The velocity-detecting radar records 10 cars speed as shown in the following table.

Use a two-sided test to analyze this dataset and represent your conclusions using P-Value. (choose \( \alpha = .05 \))

<table>
<thead>
<tr>
<th>Cars</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>( \bar{X} )</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed (kmph):</td>
<td>21.42</td>
<td>21.87</td>
<td>26.10</td>
<td>30.78</td>
<td>28.03</td>
<td>35.80</td>
<td>29.44</td>
<td>35.91</td>
<td>27.90</td>
<td>27.30</td>
<td>28.455</td>
<td>4.9</td>
</tr>
</tbody>
</table>

\( H_0 : \mu = 15 \)

\( H_a : \mu \neq 15 \)

Test Statistic: \( T^* = \frac{\bar{X} - \mu}{S/\sqrt{n}} = 2.79 \)

\[ d.f = 9 \]

\[ \text{P-Value} = 2 \left[ 1 - P(T^* > 2.79) \right] \]

\[ = 2 \left[ 1 - P(T \leq 2.79) \right] \]

\[ = 2 \left[ 1 - .01 < \alpha < .005 \right] \]

\[ = .02 < \alpha < .005 \]
11. Each year in Britain there is a No Smoking Day, where many people voluntarily stop smoking for a day. This No Smoking Day occurs on the second Wednesday of March each year. Data are collected about nonfatal injuries on the job, which allows a test of the hypothesis that stopping smoking affects the injury rate. Many factors affect injury rate, though, such as year, time of work, etc., so we would like to be able to control some of these factors, one way to do this is to compare the injury rate on the Wednesday of the No Smoking Day to the rate for the previous Wednesday in the same years. Those data for 1987 to 1996 are listed in the following table:

<table>
<thead>
<tr>
<th>Year</th>
<th>87</th>
<th>88</th>
<th>89</th>
<th>90</th>
<th>91</th>
<th>92</th>
<th>93</th>
<th>94</th>
<th>95</th>
<th>96</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>516</td>
<td>610</td>
<td>581</td>
<td>586</td>
<td>554</td>
<td>632</td>
<td>479</td>
<td>583</td>
<td>445</td>
<td>522</td>
</tr>
<tr>
<td>on</td>
<td>540</td>
<td>620</td>
<td>599</td>
<td>639</td>
<td>607</td>
<td>603</td>
<td>519</td>
<td>560</td>
<td>515</td>
<td>556</td>
</tr>
</tbody>
</table>

(a) How many more or fewer injuries are there on No Smoking Day, on average, compared with the normal day?

(b) Test whether the accident rate increases on No Smoking Day.

\[ H_0 : \mu_D = 0 \]
\[ H_a : \mu_D < 0 \]

\[ T = \frac{-25}{3.231/\sqrt{10}} = -2.447 \]

\[ P\text{-value} = P(T_{10} < -2.447) = 0.01 \]

Reject \( H_0 \) if \( T \leq t_{0.05, 9} \)

Critical value = -1.833

12. Government environmental regulations specify a PCB limit of 5 parts per million in water. A major manufacturing firm, producing PCB's for electrical insulation, discharges small amounts from its plant. The company management, attempting to control the PCB level in its discharge, has given instructions to halt production if the mean amount of PCB in the effluent exceeds 3 parts per million. A random sample of water specimens will be taken regularly to determine if the effluent meets the company's specifications. The hypothesis to be tested is

\[ H_0 : \mu = 3 \text{ ppm} \quad H_a : \mu > 3 \text{ ppm} \]
From past experience, the plant manager believes the standard deviation in the PCB found in their effluent is $\sigma = 0.5$ part per million. If he is willing to risk a 5% probability of Type I error and a 10% probability of Type II error if the true mean is 3.38 parts per million, how large a sample should be taken to test the hypothesis?

$$P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = 0.05$$

$$= P(Z > Z_{0.05} \text{ when } \mu = 3, \sigma = 0.5) = 0.05$$

$$= P(Z > \frac{\bar{x} - 3}{0.5/\sqrt{n}}) = 0.05$$

$$\Rightarrow \frac{\bar{x} - 3}{0.5/\sqrt{n}} = 1.645 \Rightarrow \bar{x} = 3 + (1.645) \frac{(0.5)}{0.5/\sqrt{n}} \text{ ......(1)}$$

$$P(\text{accept } H_0 \text{ when } H_0 \text{ is false}) = 0.1$$

$$P(Z < Z_{0.05} \text{ when } \mu = 3.38, \sigma = 0.5) = 0.1$$

$$\Rightarrow \frac{\bar{x} - 3.38}{0.5/\sqrt{n}} = -1.282 \Rightarrow \bar{x} = 3.38 - (0.5) \frac{(1.282)}{0.5/\sqrt{n}} \text{ ......(2)}$$

Solving equations (1) and (2) we get $n = 14.8 \approx 15$

13. Each day the major stock markets have a group of leading gainers in price (stocks that go up the most). On one day the standard deviation in the percent change for a sample of 10 NASDAQ leading gainers was 15.8. On the same day, the standard deviation in the percent change for a sample of 10 NYSE leading gainers was 7.9 (USA Today, September 14, 2000). Conduct a test for equal population variances to see whether it can be concluded that there is a difference in the volatility of the leading gainers on the two exchanges. Use $\alpha = 0.10$. What is your conclusion?

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2$$

Test statistic: $F = \frac{S_1^2}{S_2^2} = \frac{(15.8)^2}{(7.9)^2} = 4$

Reject $H_0$ if $F > f_{\alpha/2, v_1, v_2}$ (or) $F < f_{1-\alpha/2, v_1, v_2}$

$$f_{0.05, 99} = 3.18$$

$$f_{0.05, 99} = \frac{1}{f_{0.05, 99}}$$

$$f_{0.05, 99} = f_{0.95, 99} = \frac{1}{3.18} = 0.314$$

Since $F > f_{\alpha/2, v_1, v_2} = 4 > 3.18$ reject $H_0$ at $0.10$. 8

So these two variances are not equal.
14. Let $P_1$ and $p_2$ be the probabilities of a child getting paralytic polio for the control and treatment conditions, respectively. The objective was to test $H_0: P_1 - p_2 = 0$ versus $H_a: P_1 - p_2 \neq 0$. Supposing the true value of $P_1$ is 0.0003 (an incidence rate of 30 per 100,000), the vaccine would be a significant improvement if the incidence rate was halved—that is $P_2 = 0.00015$. Using a level $\alpha = 0.05$ test, it would then be reasonable to ask for sample sizes for which $\beta = 0.1$ when $P_1 = 0.0003$ and $P_2 = 0.00015$. What should be the required sample size to achieve 90% power? (Assume equal sample size)

$$n = \frac{\left(\frac{Z_{\alpha/2} \sqrt{(P_1 + P_2)(q_1 + q_2)/2} + Z_{\beta} \sqrt{P_1 q_1 + P_2 q_2}}{P_1 - P_2}\right)^2}{(P_1 - P_2)^2}$$

$$n = 209,903.$$

15. A researcher is concerned about the level of knowledge possessed by university students regarding United States history. Students completed a high school senior level standardized U.S. history exam. Major for students was also recorded (Education, Business/Management, Behavioral/Social Science, Fine Arts). Data in terms of percent correct is recorded below for 32 students from each group. Is there a significant difference between the four testing conditions? Complete the ANOVA table and carry out the $F$ test at level $\alpha = 0.05$. Write your conclusions by finding the $P$-value.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>63.25</td>
<td>3</td>
<td>21.083</td>
<td>0.048</td>
</tr>
<tr>
<td>Within</td>
<td>12298.25</td>
<td>28</td>
<td>439.2232143</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12361.5</td>
<td>31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$F_{0.05, 3, 28} = 2.95$$

$$F_{0.05, 3, 28} = \frac{1}{F_{0.05, 28, 3}} = \frac{1}{8.63} = 0.1159.$$