NCC5010: Data Analytics and Modeling
Spring 2015 Exemption Exam

- Do not look at other pages until instructed to do so.
- The time limit is two hours. This exam consists of 6 problems.
- Do all of your work in the space provided on this exam.
- The exam is closed book. You may not use any materials, including internet resources. You may use a calculator, but not a personal computer.

Name, printed: _____________________________________________________

Cornell ID: _______________________________________________________

Briefly explain below your previous experience in data analytics and modeling:

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

Your signature below signifies that you understand and will abide by the Cornell Code of Academic Integrity.

Sign Here: _________________________________________________________
1. A local bank reviewed its credit card policy with the intention of recalling some of its credit cards. In the past, 5% of cardholders defaulted, leaving the bank unable to collect the outstanding balance. The bank found that the probability of missing a monthly payment is 0.10 for customers who do not default (those customers make the payment eventually). Of course, the probability of missing a monthly payment for those who default is 1.

   a) Are “defaulting” and “missing a monthly payment” two independent events? Show calculations which justify your answer.
   b) Are “defaulting” and “missing a monthly payment” two mutually exclusive events? Explain.
   c) Given that a customer missed a monthly payment, compute the probability that the customer will default.

   a. They are dependent. For example, P(missing | default) = 1.0 ≠ P(missing) < 1.0. P(missing) requires a calculation; it is = P(Miss | Def)*P(Def) + P(Miss | No Def)*P(No Def) = 1.0*0.05 + 0.1*0.95 = 0.145
   b. No; they can both happen at the same time. “Default” and “no default” are mutually exclusive.
   c. P(default | missing) = P(default and missing)/P(Missing) = 0.05/0.145 = 0.345 using joint probabilities from part a.
2. A taxi driver is considering renting snow tires in preparation for a big snow storm. The rental cost is $150 in total, and the tires can be rented for this storm only. If the storm materializes, the driver will make $200 in addition to the $300 he would normally make on regular days. He cannot drive (or make any money) in the storm without snow tires. If he decides to rent the snow tires he must do so several days before the storm hits. What should the probability of stormy weather be to justify renting? Show the calculation.

Rent: He makes either $500-150 or $300-150
Expected payoff = P(storm)*350 + (1 – P(storm))*150

Not rent: He makes either $0 or $300
Expected payoff = P(storm)* 0 + (1 – P(storm))*300

To break even: = P(storm)*350 + (1 – P(storm))*150 = (1 – P(storm))*300
Solving: P(storm) = 150/500 = 0.3. P(storm) must be > 0.3 to justify renting.
3. The average time that an employee works at a call center, before leaving the company, is being studied. Suppose that you have been given the results as a **Confidence Interval**. You have been asked to explain those results to your CEO.

   a) Explain what a confidence interval means (make up numbers as necessary).
   
   b) After you present the results, the CEO is not satisfied, saying that the results are not precise enough. She wants the width of the confidence interval to be cut in half. What would you do to achieve the CEO's required level of precision?

   a. The true value of a parameter (such as the long-run average time spent in a call center) cannot be known from a sample. Suppose the sample average was 100 days. We might be able to say that we are 95% sure that the range from 82 to 118 contains the true long-run average. This actually a loose statement. Officially, we can say that 95% of the intervals constructed as we did this one will contain the true parameter value.
   
   b. The best answer is that if we quadruple the sample size, we expect the standard deviation, and the width of the confidence interval, to become half as large. If we had constructed a 99.5% (or other very high confidence interval, we could construct an 80% confidence interval that is half as wide.
4. A manufactured part is designed to have two holes in it, and the distance between the holes is supposed to be exactly 10mm. However, the machine varies in accuracy, and the standard deviation of the distance between holes is known to be 0.04 mm. Assume that distance between holes does NOT have the normal probability distribution. A sample of 100 parts will be used to test whether the machine is correctly adjusted for a 10mm separation of holes.

a) Briefly explain the Type I and Type II errors that might occur in testing this hypothesis. Your explanation should be in terms of the question above. No numbers or formulas are necessary.

b) Why is it acceptable to use the normal probability distribution to describe the distribution of the sample average in this case?

c) Based on the numbers given above, give a numerical value for the standard error of the sample average distance between holes?

a. The null hypothesis would be that the mean = 10 mm. A Type 1 error would be: believing that the null hypothesis is false, when it is true. A Type 2 error would be: believing the null hypothesis is true when it is false. “Believing that” is more often called “accepting” a hypothesis.

b. Because of the central limit theorem. For most distributions for the data, it is reasonable to assume that the sample means are normally distributed. For n = 100, sample means would be normally distributed no matter what the distribution of the data is.

c. Since we know that σ = 0.04, then \( \sigma(\bar{x}) = \sigma/\sqrt{100} = 0.004 \)
5. A real estate investor has estimated the following relationship for properties in a city of 100,000 people in the Midwest region of the United States:

\[ y = 2890 + 1.1x_1 + 0.9x_2 + 24.0x_3 \]

(0.000) (0.121) (0.002) (0.031) ← p-values given in parentheses

where \( y \) is the sale price,
\( x_1 \) is the appraised land value ($),
\( x_2 \) is the appraised house value ($), and
\( x_3 \) is the area of living space (square feet).

The "R-square" was 0.78, based on a sample of 55 properties.

a) Is there a statistically significant relationship between \( y \) and \( x_1 \) at the 5% significance level? Explain.

b) Give an interpretation of the R-square value that is given above.

c) The investor is interested in properties with appraised land value of $100,000, appraised house value of $200,000, and square footage of 3,000. What sales price does the model predict for such a property?

d) In order for the statistical test in part a) to be valid, a series of "regression assumptions" must be satisfied. State two of those assumptions.

a. No; the p-value given is 0.121 for \( x_1 > 0.05 \).
b. 78% of the variation in \( y \) is ‘explained’ by the relationship. (It is not really explained; the correlation could be spurious.) R-square is also called the multiple coefficient of determination.
c. \[ y = 2890 + 1.1x_1 + 0.9x_2 + 24.0x_3 = 2890 + 1.1 \times 100000 + 0.9 \times 200000 + 24 \times 3000 = 364,890 \]
d. The regression assumptions have to do with the “residuals.” It is assumed that:
   a. The expected value of the residuals is constant at a value of zero
   b. The variance of the residuals is constant
   c. Residuals are not correlated with each other, and
   d. Residuals are normally distributed.
6. DMG, Inc. is considering two different marketing strategies for its latest software product, conveniently denoted by “Strategy A” and “Strategy B.” In order to assess the two strategies, Paula Cooke built and ran a simulation model. Pertinent results are shown on the next page. Using these simulation results, answer the following questions.

(a) What is the probability that Strategy A will be more profitable than Strategy B?

(b) What is a 95% confidence interval for the unknown true mean of the difference in earnings between Strategy A and Strategy B?

(c) Paula Cooke has staked her position that the new software product will earn at least $400,000. What is the probability of achieving this goal under each of the two different marketing strategies?

(d) Notice that the sample mean in the last column of the table is the difference between the sample means of the two different strategies. However, the sample standard deviation of the last column is not the difference between the sample standard deviations of the two strategies. Why is this so?

(e) Which of the two marketing strategies would you recommend? Defend your decision with relevant analysis.

\begin{itemize}
  \item[a.] Between 65% and 70%.
  \item[b.] \(210,923 \pm 1.96 \times \frac{379443}{\sqrt{20000}}\)
  \item[c.] Strategy A: more than 90%. Strategy B: more than 95%.
  \item[d.] The last column is the standard deviation of the difference for each A and B outcome. This calculation is nonlinear in the value of each outcome (i.e. \(\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}\)). As a result, the standard deviation of the difference in the sample means cannot be determined by taking the difference between the two sample standard deviations.
  \item[e.] This is somewhat subjective. Acceptable arguments include:
    \begin{itemize}
      \item[a.] Strategy A – it has a higher mean return than B overall and it will produce a higher return with a 65% to 70% probability.
      \item[b.] Strategy B – although it has a lower mean return than A, the standard deviation is also lower than A. There is a 0% to 5% chance than the earnings are less than $400,000 with B, but this chance increases to 5% to 10% with A.
    \end{itemize}
\end{itemize}
## DMG Simulation Model
### Simulation Results for 20,000 Trials

*(all dollar values are contribution to earnings)*

<table>
<thead>
<tr>
<th>Cumulative Probability</th>
<th>Marketing Strategy A</th>
<th>Marketing Strategy B</th>
<th>A – B</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0001</td>
<td>$47,780</td>
<td>$290,000</td>
<td>($511,723)</td>
</tr>
<tr>
<td>.05</td>
<td>$355,766</td>
<td>$423,101</td>
<td>($259,280)</td>
</tr>
<tr>
<td>.10</td>
<td>$465,761</td>
<td>$485,737</td>
<td>($175,313)</td>
</tr>
<tr>
<td>.15</td>
<td>$531,758</td>
<td>$524,885</td>
<td>($135,163)</td>
</tr>
<tr>
<td>.20</td>
<td>$608,755</td>
<td>$551,729</td>
<td>($93,026)</td>
</tr>
<tr>
<td>.25</td>
<td>$674,752</td>
<td>$579,691</td>
<td>($48,061)</td>
</tr>
<tr>
<td>.30</td>
<td>$762,748</td>
<td>$663,206</td>
<td>($100,072)</td>
</tr>
<tr>
<td>.35</td>
<td>$825,078</td>
<td>$704,963</td>
<td>$22,515</td>
</tr>
<tr>
<td>.40</td>
<td>$889,242</td>
<td>$751,940</td>
<td>$64,402</td>
</tr>
<tr>
<td>.45</td>
<td>$971,738</td>
<td>$832,845</td>
<td>$81,107</td>
</tr>
<tr>
<td>.50</td>
<td>$1,056,068</td>
<td>$888,957</td>
<td>$167,111</td>
</tr>
<tr>
<td>.55</td>
<td>$1,114,732</td>
<td>$924,189</td>
<td>$169,883</td>
</tr>
<tr>
<td>.60</td>
<td>$1,246,726</td>
<td>$981,606</td>
<td>$214,619</td>
</tr>
<tr>
<td>.65</td>
<td>$1,334,722</td>
<td>$1,018,143</td>
<td>$275,622</td>
</tr>
<tr>
<td>.70</td>
<td>$1,417,218</td>
<td>$1,080,779</td>
<td>$338,837</td>
</tr>
<tr>
<td>.75</td>
<td>$1,466,716</td>
<td>$1,166,904</td>
<td>$380,075</td>
</tr>
<tr>
<td>.80</td>
<td>$1,598,710</td>
<td>$1,237,369</td>
<td>$481,646</td>
</tr>
<tr>
<td>.85</td>
<td>$1,697,705</td>
<td>$1,354,812</td>
<td>$590,893</td>
</tr>
<tr>
<td>.90</td>
<td>$1,917,695</td>
<td>$1,566,208</td>
<td>$351,497</td>
</tr>
<tr>
<td>.95</td>
<td>$2,500,669</td>
<td>$1,754,116</td>
<td>$746,553</td>
</tr>
<tr>
<td>1.00</td>
<td>$3,347,631</td>
<td>$2,638,850</td>
<td>$708,781</td>
</tr>
</tbody>
</table>