2DI90 - Probability and Statistics
Final Exam (2DI91)

June 30th, 2017

INSTRUCTIONS:

- This is a CLOSED NOTES exam. You are allowed only a CLEAN copy of the Statistical Compendium and ONE SIDE OF ONE A4 SHEET with HANDWRITTEN notes.
- You may use a calculator (could be a graphical calculator). Cellphones, notebooks or similar devices are not allowed. If you use any non-standard features of the calculator **explain clearly how would you solve the question using only standard features and/or the compendium**, or you might not get full credit for your answer.
- There are 7 pages in the exam questionnaire (including this one) and you have 3 hours (180 minutes) to complete the exam.
- The exam consists of 20 questions of 5 points each (in total 100 points). The final grade of the course will take into account the grades of the homework assignments and electronic test.
- The exam is to be done INDIVIDUALLY. Therefore discussion with your fellow colleagues is strictly forbidden.
- Please **BE ORGANIZED IN YOUR WRITE-UP** – we can't grade what we can't decipher!
- You should clearly and concisely indicate your reasoning and **show all relevant work**. Your grade on each problem will be based on our best assessment of your level of understanding as reflected by what you have written. **JUSTIFY** your answers and be **CRITICAL** of your results.
- The problems are not necessarily in order of difficulty. I recommend that you quickly read through all problems first, then do the problems in whatever order suits you best.

P.I: High-speed stock trading relies on some arbitrage possibilities, arising only within millisecond periods, to make (huge) profits. This requires trading companies to use very fast connections to the stock exchange (often buying buildings that are as close as possible to the exchange building). The trader SnelGeld uses three fiber-optic connections (denoted by A, B and C) in an attempt to ensure a consistently fast connection. The three connections can be either congested or not.

Let A, B and C denote the probabilities of each connection being congested. From historical data you know that $P(A) = 0.1$, $P(B) = 0.05$ and $P(C) = 0.1$. You know also that $P(A \cup B) = 0.11$ and that C is independent from both A and B.

(a) What is the probability connections A and B are simultaneously congested (that is, compute $P(A \cap B)$)?

(b) Are the events A and B independent? Are they mutually exclusive? Carefully justify your answers.

(c) The performance of the trading system is best if all connections are not congested. Compute this probability (that is, compute $P(A' \cap B' \cap C')$).

(d) Compute the probability that no more than two connections are simultaneously congested.

P.II: The number of probing attacks that occur in a certain network router is well described by a Poisson process. On average, the time between two consecutive attacks is 6 hours.

(a) Let Y be the number of attacks that occur in one day (24 hours). What is the distribution of Y? Compute $P(Y \geq 3)$.

(b) Compute the probability that, in one month (30 days) there is at least an entire day without any attacks.

(c) Suppose you just reset the router. Let T be a random variable denoting the time (in hours) when the third attack occurs. What is the probability density function of T? Compute the probability that the third attack will occur within 18 hours, that is, compute $P(T < 18)$.

Hint: there are different ways to answer (c) (some much easier than others).

P.III: You are in charge of designing a MonteCarlo simulation to evaluate the lifetime of a novel design of a lithium battery. Although this is an high-capacity battery the specific design has the drawback that the maximum charge of the battery is highly dependent on the manufacturing process.

From the chemistry of the battery you know that the capacity of the battery (in Ah) is well modeled by a continuous random variable with the following probability density function.

$$f(x) = \begin{cases} \frac{1}{(x+1)^2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} .$$

(a) Check that the cumulative distribution function of X, denoted by $F(x)$, is given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - \frac{1}{x+1} & \text{if } x > 0 \end{cases} .$$

(b) Compute the probability that X is larger than 5.

(c) Suppose you want to simulate the random variable X using the inverse transformation method. Assume you have access to a standard uniform random variable U. Explicitly describe the transformation g such that $g(U)$ is a random variable with the probability density function $f(x)$ described above.

P.IV: When studying the queuing behavior of a computer network router one comes across the following interesting statistical inference problem. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density

$$f(x) = \begin{cases} \frac{2x}{\theta} & \text{if } 0 \leq x \leq \sqrt{\theta} \\ 0 & \text{otherwise} \end{cases} .$$

where $\theta > 0$ is an unknown parameter. Our goal is to estimate θ.

(a) Show that the expected value of X_i is given by $\frac{2}{3}\sqrt{\theta}$ and use that result to derive a moment estimator of θ.

For the rest of the problem consider instead the following estimator of θ:

$$\hat{\theta} = \frac{2(n+1)}{n^2} \sum_{i=1}^{n} X_i^2 .$$

(b) Compute the bias of this estimator. Is the estimator unbiased?

(c) Compute the Mean-Squared-Error (MSE) of $\hat{\theta}$. You may use the fact that

$$\mathbb{E}(X_i^3) = \frac{2}{5}\theta\sqrt{\theta} \quad \text{and} \quad \mathbb{E}(X_i^4) = \frac{1}{3}\theta^2 .$$

P.V: In an important attempt to understand the origins of the universe, scientists have constructed sophisticated experiments to detect and count neutrinos. Because these particles interact very weakly with matter this is not an easy task, and these experiments are generally massive, featuring a large number of sensors in hopes of catching a reasonable number of these elusive particles. A simplified model for the experiment is the following: suppose there are n sensors, each modeled by a random variable X_i, $i \in \{1, \ldots, n\}$. These are assumed to be independent and furthermore it is reasonable to assume them to be Poisson distributed with unknown parameter $\lambda > 0$, namely

$$P(X_i = k) = e^{-\lambda} \frac{\lambda^k}{k!} \ , \quad k \in \{0, 1, 2, \ldots\} \ .$$

The experiment in this question has $n = 10000$ sensors, and current cosmological theories predict that $\lambda = 0.0008$ under normal circumstances. Therefore we would like to test

$$H_0 : \lambda = 0.0008 \quad \text{against} \quad H_1 : \lambda \neq 0.0008 \ .$$

A natural test statistic to consider is total number of neutrinos detected, which is given by $Y = \sum_{i=1}^{n} X_i$. From your knowledge of probability theory you know that Y is distributed as a Poisson random variable with mean $n\lambda$. Under the null hypothesis $n\lambda = 8$, so this observation naturally gives rise to a test that rejects H_0 if $|Y - 8| \geq c_\alpha$ where $c_\alpha > 0$ must be chosen depending on the desired significance level.

(a) Say you take $c_\alpha = 5$. What is the type I error of this test (also known as the false alarm probability)?

(b) During solar flares it is believed we should see a larger number of neutrinos. Suppose actually $\lambda = 0.0015$. What is the power of the test in the previous question (with $c_\alpha = 5$)?

(c) The experiment was conducted and a total of $y = 19$ neutrinos where detected. Compute the p-value of this test. Would you reject the hull hypothesis at significance level $\alpha = 0.01$? Carefully justify your answer.

Hint: To avoid very tedious computations I advise you to use the cumulative distribution tables for the Poisson distribution in the statistical compendium.

P.VI: After a long and hard year of studies you spent a few well deserved weeks on the beach, in a country with consistently good weather. While relaxing on the sand you noticed that the number of swimmers in the water was heavily influenced by the water temperature. This prompted the question: can we use the number of swimmers as a "thermometer"? Over the course of two weeks you counted the number of swimmers entering the water between 10:30 and 11:00 in a selected region of the beach. In addition, you took note of the water temperature (as reported by the meteorological site). Below is the data you collected.

Day	Number of swimmers	Water temperature (°C)
1	57	19.0
2	88	19.5
3	89	23.5
4	73	17.5
5	91	21.0
6	100	20.0
7	70	21.5
8	109	22.0
9	101	25.0
10	91	20.5
11	79	22.0
12	96	23.5
13	82	22.5
14	101	23.5

A simple linear regression analysis was conducted with `R`, where the water temperature is taken as the response variable. The results are summarized in Appendix A.

(a) Write the assumed model for the relation between the number of swimmers and the water temperature. Is it reasonable to assume normally distributed errors?

(b) Test the significance of the regression and give the p-value associated with the test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, where β_1 is the true slope value in the assumed model.

(c) After a quick trip to the country side you returned to the same beach and decided to see if your model was indeed useful in predicting the water temperature. Between 10:30 and 11:00 you counted 93 swimmers. Give an estimate for the water temperature (according to your model).

(d) Your friends were impressed by your model, but a bit doubtful about the quality of your estimate in (c) and wanted to have a better idea of the errors that are involved. Use your knowledge of regression models to give a two sided prediction interval for the water temperature in the scenario in (c) (use $\alpha = 0.05$).

A Beach Data

```
######################################################################
Call:
lm(formula = temp ~ num)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8620 -1.2357  0.4217  1.3660  2.4619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.68833    3.14180   ????  ????
num          0.07772    0.03542   ????  ????

Residual standard error: 1.817 on 12 degrees of freedom
Multiple R-squared: 0.2864,Adjusted R-squared: 0.2269
F-statistic: ???? on 1 and 12 DF,  p-value: ????
######################################################################
Call:
   aov(formula = temp ~ num)

Terms:
                    num Residuals
Sum of Squares  15.8939   39.6061
Deg. of Freedom       1        12

Residual standard error: 1.81673
Estimated effects may be unbalanced
######################################################################
Call:
  mean(num) 87.64286
  sd(num) 14.22677
  mean(temp) 21.5
  sd(temp) 2.066212
######################################################################
Call:
shapiro.test(lm(temp~num)$residuals)

Shapiro-Wilk normality test

data:  lm(temp ~ num)$residuals
W = 0.9227, p-value = 0.2408
######################################################################
```

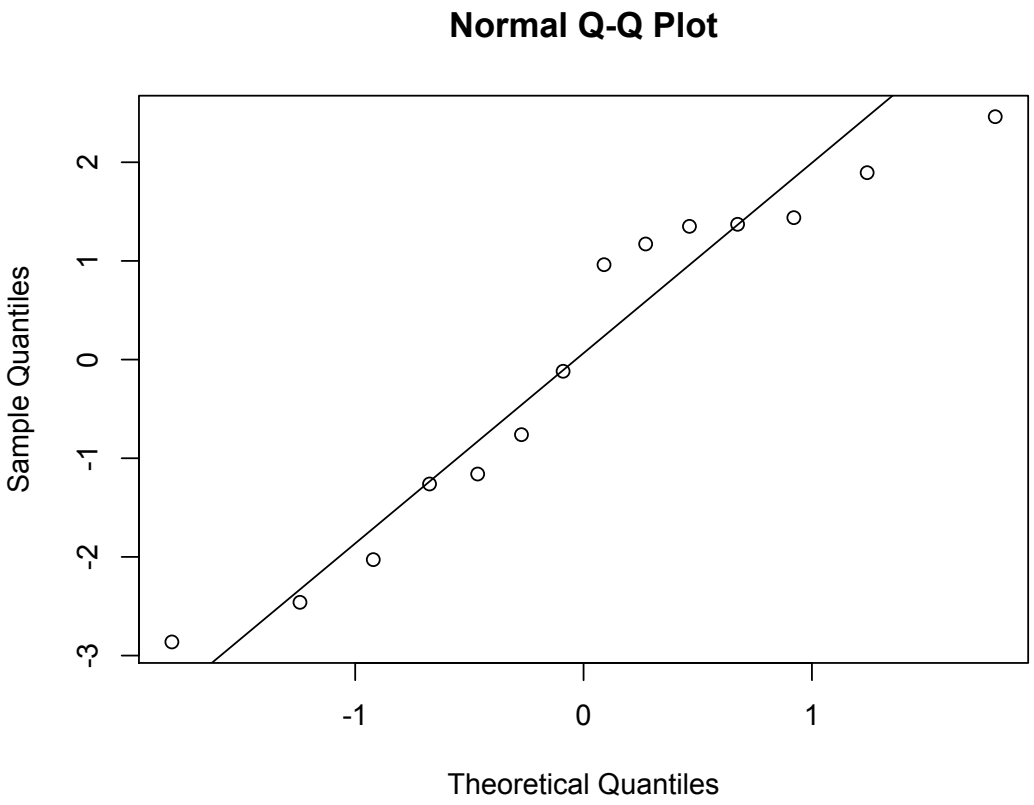

Figure 1: Normal QQ plot of the residuals of the regression model.