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

July 2nd, 2014

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 8 pages in the exam questionnaire (including this one) and you have 3 hours (180 minutes) to complete the exam.
- The exam consists of 25 questions of 4 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.
- Remember to **IDENTIFY** the materials you give us with you name and student number.

P.I: A RAID system (Redundant Array of Independent/Inexpensive Disks) was built using 3 disks, identified with the names `disk1`, `disk2`, and `disk3`. These disks are known to have probability of failure respectively 0.01, 0.03, and 0.05. The disks are also known to fail independently.

Let A denote the event {`disk1` failure}, and similarly B and C denote the events `disk2` and `disk3` failure, respectively.

(a) The RAID system is such that there is loss of data only if two or more disks fail. Let E denote this event. Write E as a function of events A, B and C, using set notation (**Hint:** you might find helpful to draw a Venn diagram).

(b) Compute the probability there is loss of data, that is, compute $P(E)$ (**Hint:** write the event E as the union of four mutually exclusive events).

(c) Due to a mistake in the configuration of the system there will instead be loss of data if **at least** one of the following happens: (i) `disk1` fails; (ii) `disk2` and `disk3` both fail. What is now the probability that there is a loss of data?

(d) Consider the setting of question (c). Given that `disk 3` has failed, what is now the probability there will be loss of data?

TIP: For the entire problem it might be helpful to recall the following important facts: Let A and B be two independent events. Then event A is also independent of B', and event A' is also independent of B and of B'.

P.II: In the manufacturing of steel cables there is the possibility of creating small localized defects, which are potential points of failure over time. In particular the number of defects is well modeled by a Poisson process, and the average number of defects per meter of cable is 0.15.

Suppose you get a roll of 20 meters of contiguous steel cable:

(a) Let X be the total number of defects in the roll of cable. What is the distribution of X. What is the probability that there are more than two defects in the cable?

(b) The company that sells these steel cables also sells rolls with 200 contiguous meters of cable. Use a suitable approximation to compute the probability that there are more than 40 defects in the cable.

(c) Suppose you start unrolling the cable until you find the first defect. Let T denote the amount of cable you unroll. What is the distribution of T? What is the mean of T? Write the probability density function of T.

(d) Refer to the setting of question (c). Compute the probability that the amount of cable you unrolled is more than 10 meters (in other words, compute $P(T > 10)$).

(e) Actually, for your purposes you need 5 pieces of cable with 4 meters each, and therefore you cut the 20 meters of cable into 5 pieces of 4 meters. What is the probability that exactly three of these pieces of cable have no defects?

P.III: Let X be a continuous random variable with the following probability density function

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

(a) Compute and sketch the cumulative distribution function of X, denoted by $F(x)$.

(b) Compute the mean and variance of X.

(c) Compute $\mathbb{E}[\sqrt{X}]$. Is this equal to $\sqrt{\mathbb{E}[X]}$?

(d) 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: In the study of wireless communication networks one often comes across a process that is well modeled by samples from continuous random variables with density

$$f(x) = \begin{cases} \frac{1}{\sqrt{\lambda}} e^{-\frac{x}{\sqrt{\lambda}}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} ,$$

where $\lambda > 0$ is a parameter we would like to estimate. Suppose you observe two independent samples from the above distribution and consider the following two possible estimators of λ:

$$\hat{\lambda}_A = \frac{X_1^2 + X_2^2}{4} \qquad \text{and} \qquad \hat{\lambda}_B = \frac{X_1^2 + X_2^2 + 2X_1X_2}{6} .$$

(a) Compute the bias of the two estimators. Is there an estimator that is biased?

(b) Compute the variance of $\hat{\lambda}_A$

(c) Compute the MSE of estimator $\hat{\lambda}_A$.

(d) It can be shown that the MSE of $\hat{\lambda}_B$ is given by $\frac{7}{3}\lambda^2$. Given this and your answer to (c) which of the two estimators would you prefer? Justify your answer.

IMPORTANT: you can use the fact that, if X is a random variable with the density above then

$$\mathbb{E}(X) = \sqrt{\lambda} \quad , \quad \mathbb{E}(X^2) = 2\lambda \quad , \quad \mathbb{E}(X^3) = 6\lambda\sqrt{\lambda} \quad , \quad \mathbb{E}(X^4) = 24\lambda^2 .$$

P.V: Computer security companies must always be on the lookout for new threats. Most often than not security breeches are unexpected. For instance, in a *timing attack* the attacker attempts to compromise a cryptosystem by analyzing the time taken to execute cryptographic algorithms. Every logical operation in a computer takes time to execute, and the time can differ based on the input; with precise measurements of the time for each operation, an attacker can work backwards to the input. This information can provide the attacker with information about the CPU running the system, the type of algorithm used, etc...

To check if a certain system is secure 40 login attempts were conducted with randomly chosen passwords of diverse lengths. The amount of time (in ms) the system took to deny access was recorded, and a collection of descriptive statistics is listed in Appendix A.

(a) Complete the table in the appendix filling in the missing values (indicated as ???).

(b) Is it reasonable to assume the collected measurements are samples from a normal distribution? Give both **quantitative** and **qualitative** arguments.

(c) Construct a two-sided 90% confidence interval for the variance of the system's response time. In light of your answer to (b), is it sensible to compute such an interval in this case? Carefully justify your answer.

(d) For this type of cryptosystem, it is known that a high variance of the response time will indicate a weakness of the implementation. Let σ^2 denote the variance of the response time, and test $H_0 : \sigma^2 = 5\text{ms}^2$ against $H_1 : \sigma^2 > 5\text{ms}^2$. Compute (approximately) the corresponding p-value. Will you reject the null hypothesis at a significance level $\alpha = 0.05$?

P.VI: When making beer the mashing step plays a crucial role. In that step the starches of the barley are converted to sugars by various enzymes. Their effectiveness is heavily influenced by the temperature and the pH of the mash.

The brewers making the world famous BuzzBeer systematically record the temperature and pH of the mashes they make, as well as the yield, which can only be measured after the mashing is finished. Yield is measured in terms of original gravity, which has no units. Below is the record of last month's mashes pH and resulting original gravity.

pH	Original Gravity
5.06	1.023
5.21	1.015
5.68	1.018
5.56	1.018
5.55	1.029
5.64	1.020
5.39	1.011
5.51	1.018
5.62	1.017
5.76	1.010
5.52	1.026
5.18	1.033
4.85	1.037
5.29	1.022
5.45	1.010
5.73	1.012
5.42	1.029
5.53	1.011
5.26	1.031
5.18	1.031

A simple data analysis (including a linear regression analysis) was conducted with `R`, and the results are summarized in Appendix B.

(a) Write the assumed model of the original gravity as a function of pH. Do you think the normality assumption of the random errors is reasonable?

(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 model.

(c) Brewer Dan just prepared a mash with pH 5.3. He would like to know what he should expect for the original gravity. Give a point estimate for this quantity.

(d) It turns out that the mash of the previous question resulted in a yield of 1.023. To check if this value is within reasonable limits compute a 90% predition interval for the original gravity of this mash.

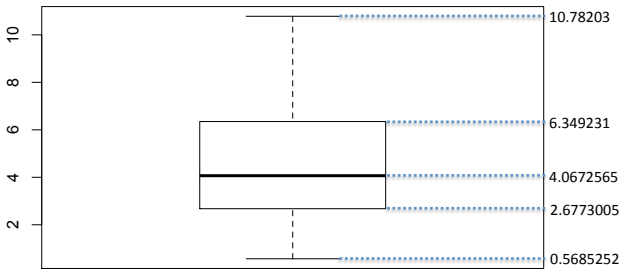
A Timing Attacks

System response time (in miliseconds):

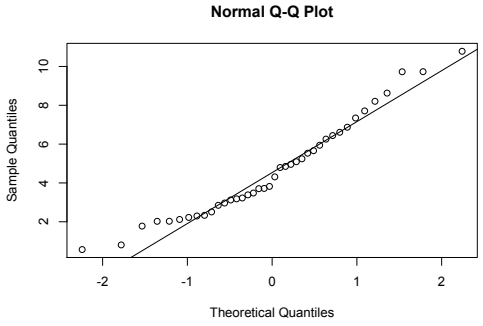
1.777763 ; 2.223587 ; 6.869387 ; 10.78203 ; 2.332443 ; 4.312676 ; 6.440998 ; 2.023269 ; 5.531647 ; 3.481276 ; 2.965045 ; 2.84759 ; 7.710503 ; 3.821837 ; 9.726404 ; 2.507011 ; 0.5685252 ; 3.123111 ; 0.8075089 ; 4.79541 ; 4.850344 ; 2.119353 ; 8.628205 ; 7.345244 ; 3.226172 ; 6.608948 ; 2.298478 ; 3.180011 ; 5.660042 ; 3.385601 ; 5.243938 ; 9.7317 ; 3.716626 ; 5.093568 ; 2.026606 ; 4.956999 ; 8.205676 ; 5.941724 ; 3.704429 ; 6.257464

Sample size=n	40
Sample Mean	???
Sample Median	???
Sample Variance	???
Minimum	0.5685252
Maximum	10.78203
Range	???
Shapiro-Wilk test statistic	0.9529
Shapiro-Wilk test p-value	0.09574
$\sum_{i=1}^{n} x_i$	186.8291
$\sum_{i=1}^{n} x_i^2$	1125.009

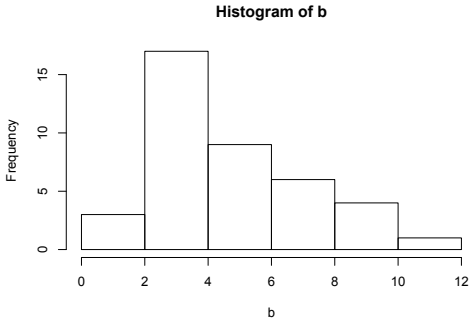
(a)



(b)



(c)



(d)

Figure 1: (a) Descriptive Statistics table; (b) Box and whisker plot ; (c) Normal QQ plot; (d) Histogram.

B Brewery Data Analysis

```
######################################################################
Call:
lm(formula = y ~ x)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0107034 -0.0041341  0.0002267  0.0050879  0.0108407

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.141096   0.035133   ??? ???
x           -0.022151   0.006477   ??? ???

Residual standard error: 0.00677 on 18 degrees of freedom
Multiple R-squared: 0.3939,Adjusted R-squared: 0.3602
F-statistic:  11.7 on 1 and 18 DF,  p-value: ???
######################################################################
Call:
   aov(formula = y ~ x)

Terms:
                           x    Residuals
Sum of Squares  0.0005360352 0.0008249148
Deg. of Freedom            1           18

Residual standard error: 0.006769682
Estimated effects may be unbalanced
######################################################################
Call:
  mean(x) 5.4195
  var(x) 0.05749974
  mean(y) 1.02105
  var(y) 7.162895e-05
######################################################################
Call:
  shapiro.test(lm(y ~ x)$residuals)

Shapiro-Wilk normality test

data:  lm(y ~ x)$residuals
W = 0.9495, p-value = 0.3602
######################################################################
```

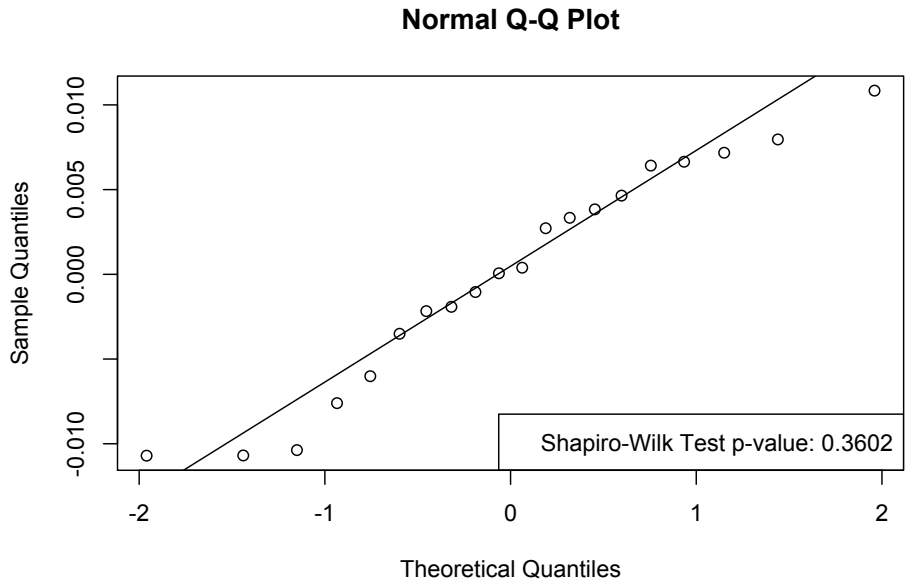


Figure 2: Normal QQ plot of the residuals.