X_400004 - Statistics
Mock Final Exam

9 December 2020

I recommend that you try to solve this mock Final exam before you watch the final lecture on the 9th of December where we will solve it together. The instructions for the Final Exam follow.

Instructions:

- The exam is to be solved **individually**.
- Please **write clearly and in an organised way**: we can't grade illegible answers.
- Please change pages when starting a new question.
- This is an exam on a mathematical subject, so support your answers with **computations**, rather than words, whenever possible.
- You should report **all relevant computations** and **justify** non-trivial steps.
- This is a **closed book exam**; you are only allowed to have one A4 sheet with **handwritten** notes with you.
- You may use a calculator.
- There are 5 pages in the exam questionnaire (including this one) and you have two hours (120 minutes) to complete the exam.
- The exam consists of 14 questions spread throughout 3 problems.
- The number of points per question is indicated next to it for a total of 100 points.
- Your final grade is $\max(1, \text{score}/10)$, where "score" is the number you points you get.
- The problems are not necessarily ordered in term 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** at least one of your answer sheets with you name and student number.

After completing the actual Midterm exam, digitalise your answer sheets (with the correct order and orientation) and submit them as **a single PDF** on Canvas for grading. **You have 10 minutes following the end of the exam to do this** after which any submission will be marked as late. These instructions do not replace the VU's *Protocol of online examination for 2020-2021* that you can find on Canvas.

Prob.I: Consider a random sample $X_1, \ldots, X_n$ of size n from the $\mathrm{Ber}(p)$ distribution, where $p \in [0,1]$ and suppose that you put a $\mathrm{Beta}(\alpha, \beta)$ prior on p. This problem concerns the choice of the hyperparameters α and β.

A few facts that you *may* need to know in order to solve this questions are:

- If $X \sim \mathrm{Ber}(p)$, then the probability mass function of X is $f(x) = p^x(1-p)^{1-x}$;
- If $Y \sim \mathrm{Beta}(\alpha, \beta)$, $\alpha, \beta > 0$, then the probability density function of Y is $f(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}$;
- If $Y \sim \mathrm{Beta}(\alpha, \beta)$, then $\mathbb{E}Y = \alpha/(\alpha+\beta)$, and $\mathbb{V}Y = \alpha\beta/\big\{(\alpha+\beta)^2(\alpha+\beta+1)\big\}$;

(a) Compute the posterior distribution of p.

(b) Compute the expectation of the posterior distribution and call is $\hat{p}$; this is your estimator of p.

(c) Compute the bias and the variance of $\hat{p}$. (Note that these will depend on α and β.)

(d) Is it possible to pick hyperparameters α, β (not depending on p) so that the resulting $\hat{p}$ is unbiased? If so, what is the resulting estimator?

(e) The estimator $\bar{X}$ is the Maximum Likelihood estimator and as such has the smallest variance among any estimator of p. Is it possible to pick α, β so that the corresponding $\hat{p}$ has a **smaller Mean Squared Error** than the Maximum Likelihood estimator? (Justify your answer.)

Prob.II: One of the important features of the power supply unit of a computer server is to deliver power to the server at a steady voltage, regardless of the demand of the server. You are considering placing a large order of power units for a cluster of servers that your company is setting up but, since these are quite expensive, you want to know if they work as advertised.

To do this, you purchased one power unit and carefully measured the output voltage every hour, for one day. Let $X_1, \dots X_{24}$ be independent random variables modelling the voltage output (in volt) for each of the 24 hourly measurements. It can safely be assumed these have a Normal distribution with unknown mean μ, and standard deviation $\sigma = 0.25$ volt.

Suppose that these units are advertised to have an output of 5 volt; you are quite confident that the output is not above 5 volt. You are not sure, however, if it might happen that the voltage might drop below 5 volt – if this were to happen, then the server might shut down, with dramatic consequences.

(To answer the following questions you may need one or more of the following quantiles: $z_{0.01} = -2.33$, $z_{0.0125} = -2.24$, $z_{0.025} = -1.96$, $z_{0.05} = -1.64$.)

(a) You decide to base your purchase decision on the outcome of a statistical test. You want to make sure that you set up the test such that there is a small chance α that the test will tell you not to buy the power units in the case where the units work as advertised. Formulate the appropriate null hypothesis and alternative hypothesis for such a test.

(b) Report an appropriate **test statistic** to be used and the **rejection rule** that ensures that the test has the right significance level. **Calibrate the test** to have significance level $\alpha = 0.05$.

(c) From the measurements that you took, you computed a voltage sample mean of $\bar{x} = 4.91$ volt. Show that the p-value of the test is $\Phi(-1.7636) = 0.0389$. Should you reject the null hypothesis at significance level $\alpha = 0.05$?

(d) Repeat question (c), for a significance level $\alpha = 0.01$. In other words, compute the p-value and make a decision when $\alpha = 0.01$.

(e) Suppose that actually $\mu = 4.95$ volt. Is the power of the test in (b) at least 0.5? (Justify your answer by computing the power.)

Prob.III: A common background sound in many places during Summer when you go for a walk in nature is the sound of field crickets. You may have noticed before that crickets tend to chirp (sing) faster when it's warmer. In Table 1 you can see some data (x_i, Y_i), $i = 1, \ldots, 42$: the x_i represents the number of times that a cricket chirped in 60 seconds, and Y_i represents the respective temperature (in degrees Celsius) at the time that you counted the chirps. We plot the data in Figure 1.

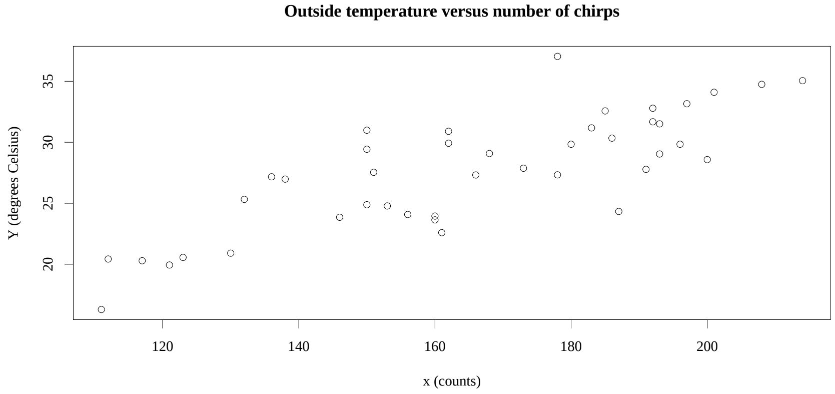

Figure 1: Plot of the cricket dataset.

The data in Figure 1 seems to suggest that the relation between the number of times that a cricket chirps and the outside temperature might be linear. The actual data follows in Table 1.

Variables	Values
$(x_1, \ldots, x_n)$	111 112 117 121 123 130 132 136 138 146 150 150 150 151 153 156 160 160 161 162 162 166 168 173 178 178 180 183 185 186 187 191 192 192 193 193 196 197 200 201 208 214
$(y_1, \ldots, y_n)$	16.29 20.43 20.29 19.94 20.56 20.91 25.32 27.17 26.98 23.85 29.43 24.88 30.99 27.54 24.78 24.08 23.64 23.95 22.59 30.90 29.92 27.32 29.08 27.87 37.04 27.33 29.84 31.18 32.57 30.34 24.33 27.78 32.79 31.68 29.04 31.51 29.84 33.16 28.58 34.10 34.75 35.05

Table 1: The cricket dataset.

From the observations in Table 1 we see that $n\bar{x} = 6942.00$, $n\bar{y} = 1159.60$, $SS_{xx} = 31894.57$, and $SS_{xy} = 4355.36$. There are 42 measurements in total.

(a) Suppose that you would like to use a Simple Linear Regression model to derive a formula that allows you to predict the outside temperature (Y) based on the number of times that a cricket chirps in 60 seconds (x). In a linear regression model you assume that

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\alpha, \beta \in \mathbb{R}$ are unknown, and the ϵ_i are random error terms. In order for Simple Linear Regression to be an adequate model here, what should you assume about: (i) the relation between (x_i, Y_i) and (x_j, Y_j), $i \neq j$; (ii) the expectation of the noise terms ϵ_i; (iii) the variance of the noise terms ϵ_i.

(b) Consider the data from Table 1 and assume that the assumption from (a) hold. Based on the data, what are your estimates of the intercept and the slope of the line in your model? (If you do not manage to compute the estimates, assume in the following that your prediction formula is $\hat{Y} = 7.82 + 0.11\,x$.)

(c) It might just be that there is actually no relation between x and Y. In this case you would expect x not to help explain Y or, in other words, you would expect the slope β to be 0. To test this out assume that the noise terms have a Normal distribution with variance $\sigma = 2$. Based on the fact that you know that in this case

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{SS_{xx}}\right),$$

derive an exact, 95% confidence interval for β. Based on this, what do you conclude about the possibility that the slope might be 0? (You may need one or more of the following Normal quantiles: $z_{0.01} = -2.33$, $z_{0.0125} = -2.24$, $z_{0.025} = -1.96$, $z_{0.05} = -1.64$.)

(d) Say that you go outside and hear a cricket chirp 181 times in 60 seconds. What would be your prediction of the outside temperature?