X_400004 - Statistics
Final Exam

15 December 2020

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 12 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 your 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 Poisson(λ) distribution, where $\lambda > 0$ and suppose that you put a Gamma(α, β) prior on λ. This question concerns the choice of the hyper-parameters α and β.

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

- If $X \sim \text{Poisson}(\lambda)$, then its probability mass function is $f(x) = e^{-\lambda}\lambda^x/x!$;
- If $Y \sim \text{Gamma}(\alpha, \beta)$, $\alpha, \beta > 0$, then its probability density function is $f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$;
- If $Y \sim \text{Gamma}(\alpha, \beta)$, then $\mathbb{E}Y = \alpha/\beta$, and $\mathbb{V}Y = \alpha/\beta^2$;

8 pts (a) Compute the posterior distribution of λ corresponding to a Gamma prior.

4 pts (b) Compute the expectation of the posterior distribution and call it $\hat{\lambda}$; this is your estimator of λ.

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

2 pts (d) Is it possible to pick some α, β not depending on λ so that the resulting $\hat{\lambda}$ is unbiased?

8 pts (e) The estimator $\bar{X}$ is the Maximum Likelihood estimator and has the smallest variance among all estimators of λ. What is the optimal choice of α, β that you can make in practice so that the corresponding $\hat{\lambda}$ has the smallest possible Mean Squared Error? Justify your answer.

Prob.II: One of the steps in the production of a certain drug involves a salt precipitation reaction in an acid solution. The pH of the solution heavily affects the effectiveness of this step, and in turn the production cost. The technical staff at your company tells you that a pH of 3.6 is ideal for such a reaction. However, obtaining this exact value is difficult, since the solution uses some plant ingredients and there is natural variation from batch to batch. Your technicians have been working really hard at refining their procedures and provided you with pH measurements of 12 batches of solution.

```
3.50 3.60 3.85 3.34 3.69 3.81 3.64 3.72 3.80 3.33 3.80 3.58
```

The way that the measurements were made ensures these are statistically independent pH samples. The technicians also tell you that the measurements have a normal distribution (although they have no idea what the expectation and variance would be.)

Your goal here is to decide, based on the data above, if it seems like your technicians have finally refined their technique enough to ensure that the expected pH of the solution is 3.6. If this is not the case you need to know so you can tell them to further calibrate their procedure.

For the data above we have $n = 12$, $\sum_{i=1}^{n} x_i = 43.66$, $\sum_{i=1}^{n} x_i^2 = 159.1936$. You may also need one or more of the following quantiles: $t_{12;0.1} = -1.3562$, $t_{12;0.05} = -1.7823$, $t_{12;0.025} = -2.1788$, $t_{12;0.01} = -2.681$, $t_{11;0.1} = -1.3634$, $t_{11;0.05} = -1.7959$, $t_{11;0.025} = -2.201$, $t_{11;0.01} = -2.7181$, $t_{10;0.1} = -1.3722$, $t_{10;0.05} = -1.8125$, $t_{10;0.025} = -2.2281$, $t_{10;0.01} = -2.7638$.

4 pts (a) Compute the sample mean and the sample variance.

18 pts (b) Suppose you want to test if the pH is really 3.6 with significance level $\alpha = 0.1$. Conduct the appropriate hypothesis test by clearly stating **the null- and alternative-hypotheses**, an appropriate **test statistic**, and the **rejection rule**. **Calibrate the test**. What is your conclusion at significance level $\alpha = 0.1$?

10 pts (c) Show that the p-value of the test that you designed in (b) is (approximately) $2F_{t_{n-1}}(-0.7509)$. What is your conclusion at significance level 0.1?

8 pts (d) Compute a two-sided 90% confidence interval for the mean pH. Could you have reached the conclusion of your answer to (c) using this interval instead? Justify your answer.

Prob.III: Imagine that you are managing an ice-cream parlour. One of the secrets behind the success of any ice-cream store is the freshness of the ice-cream. Because of this, it is important for you to be able to predict how much ice-cream needs to be made each day so that there are no left-overs. If you over-produce then you have to sell left-overs and if you under-produce then you miss out on sales.

Over the years you've noticed a trend that might be useful: every evening you check the forecast of the temperature for the following day and you've noticed that on days for which the weather forecast is higher, you tend to sell more ice-cream. In the table below is a list of the forecasted highest temperature and the respective amount of ice-cream sold in thirty randomly selected summer days (only weekdays are considered). Figure 1 below shows a plot of the data.

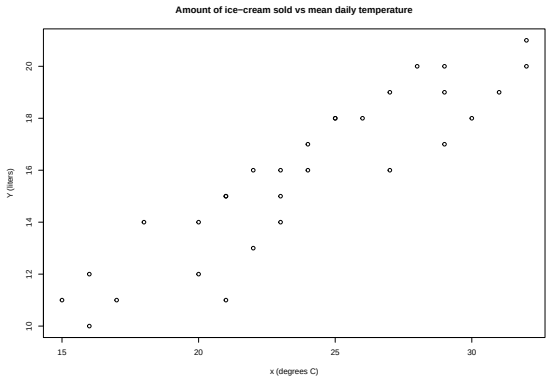

Figure 1: Plot of the ice-cream dataset.

forecasted temperature (°C)	ice-cream sold (liters)	forecasted temperature (°C)	ice-cream sold (liters)
25	18	15	11
21	15	30	18
18	14	31	19
16	12	29	19
22	13	32	21
16	10	23	16
17	11	23	15
20	14	22	16
26	18	20	12
25	18	21	15
24	16	23	14
24	17	29	17
27	19	32	20
28	20	21	11
29	20	27	16

Denote the measurements as (x_i, y_i) where x_i represents the temperature forecast and y_i the respective amount of ice-cream sold. From the table: $n = 30$, $\bar{x} = 23.8667$, $\bar{y} = 15.8333$, $SS_{xx} = 691.4667$, $SS_{xy} = 403.3333$, and $SS_{yy} = 284.1667$.

A strategy starts forming in your mind: if you had a way of predicting ice-cream demand based on the temperature forecast you could use that to decide how much ice-cream to make in advance...

6 pts (a) Describe a Simple Linear Regression model to explain the amount of ice-cream sold (Y) as a function of the temperature forecast (x). Make sure to write down the **model equation**, and **list the necessary assumptions** on the noise term. You can assume that the noise is normal; denote the slope and intercept as β and α, respectively.

10 pts (b) The temperature forecast for tomorrow is 27 degrees. Give a point estimate of the expected amount of ice-cream that will be sold.

12 pts (c) Test the significance of the regression by showing that the p-value associated with the test of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$, where β is the true slope value in the model, is $2F_{t_{28}}(-12.0137) \approx 0$. What do you conclude? You should use the fact that you know that

$$\sqrt{SS_{xx}}\frac{\hat{\beta} - \beta}{\hat{\sigma}} \sim t_{n-2}.$$