# X_400004 - Statistics
# Solutions to the Final Exam

## 15 December 2020

**Below are answers to the exam questions. Some of these are slightly abbreviated, while others include extra comments. These are for your reference only but should inform the level of detail that is expected from your answers in the exam. Also keep in mind that there might be different ways to approach each question. If you find typos and or omissions, please report them to the lecturer so they can be corrected.**

**Prob.I:** Consider a random sample $X_1, \ldots, X_n$ of size $n$ from the Poisson($\lambda$) distribution, where $\lambda > 0$ and suppose that you put a Gamma($\alpha, \beta$) prior on $\lambda$. This question concerns the choice of the hyper-parameters $\alpha$ and $\beta$.

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

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

**8 pts** (a) Compute the posterior distribution of $\lambda$ corresponding to a Gamma prior. **Solution: The posterior distribution of $\lambda$ is proportional to the likelihood times the prior. The likelihood is**

$$L(\lambda) = e^{-\lambda}\lambda^{X_1}/X_1! \times \cdots \times e^{-\lambda}\lambda^{X_n}/X_n! \propto e^{-n\lambda}\lambda^{n\bar{X}}.$$

**The prior density is $\pi(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$. Combining the two we get that the posterior density satisfies**

$$\pi(\lambda \mid X_1, \ldots, X_n) \propto L(\lambda)\pi(\lambda) \propto e^{-n\lambda}\lambda^{n\bar{X}}\lambda^{\alpha-1}e^{-\beta\lambda} = \lambda^{n\bar{X}+\alpha-1}e^{-(n+\beta)\lambda}.$$

**We identify this as being proportional to the density of a Gamma$(n\bar{X}+\alpha, n+\beta)$ distribution. We conclude that the posterior is Gamma$(n\bar{X}+\alpha, n+\beta)$.**

**4 pts** (b) Compute the expectation of the posterior distribution and call it $\hat{\lambda}$; this is your estimator of $\lambda$. **Solution: Using the expression for the expectation of a Gamma distribution, we have $\hat{\lambda} = (n\bar{X}+\alpha)/(n+\beta)$.**

**10 pts** (c) Compute the bias and the variance of $\hat{\lambda}$. (Note that these will depend on $\alpha$ and $\beta$.) **Solution: The expectation and variance of $\bar{X}$ are respectively $\lambda$ and $\lambda/n$ , so**

$$Bias_{\alpha,\beta}(\lambda) = \mathbb{E}\hat{\lambda} - \lambda = \frac{n\mathbb{E}\bar{X}+\alpha}{n+\beta} - \lambda = \frac{n\lambda+\alpha-(n+\beta)\lambda}{n+\beta} = \frac{\alpha-\beta\lambda}{n+\beta}.$$

**The variance of $\hat{p}$ is**

$$Var_{\alpha,\beta}(\lambda) = \mathbb{V}\hat{\lambda} = \frac{\mathbb{V}(n\bar{X}+\alpha)}{(n+\beta)^2} = \frac{n^2\mathbb{V}(\bar{X})}{(n+\beta)^2} = \frac{n^2\lambda/n}{(n+\beta)^2} = \frac{n\lambda}{(n+\beta)^2}.$$

**2 pts** (d) Is it possible to pick some $\alpha$, $\beta$ not depending on $\lambda$ so that the resulting $\hat{\lambda}$ is unbiased? **Solution: For any $\beta$, setting $\alpha = \beta\lambda$ leads to the respective $\hat{\lambda}$ being unbiased. The only choice of this type that does not depend of $\lambda$ is to set $\alpha = \beta = 0$. Although technically this is not allowed for the prior, the resulting posterior is well defined when $\alpha = \beta = 0$ and has expectation $\bar{X}$.**

**8 pts** (e) The estimator $\bar{X}$ is the Maximum Likelihood estimator and has the smallest variance among all estimators of $\lambda$. What is the optimal choice of $\alpha$, $\beta$ that you can make in practice so that the corresponding $\hat{\lambda}$ has the smallest possible Mean Squared Error? Justify your answer. **Solution: As we just saw, setting $\alpha = \beta = 0$ leads to the posterior expectation being the Maximum Likelihood estimator which is unbiased and has the smallest variance of any estimator of $\lambda$. The bias-variance decomposition then tells us that the Maximum Likelihood estimator has the smallest Mean Squared Error of any estimator of $p$ (no bias and smallest possible variance.) So any other choice of $\alpha$ and $\beta$ leads to an estimator with at least as much bias and at least as much variance as the Maximum Likelihood Estimator. So to get the estimator with the smallest Mean Squared Error we should set $\alpha = \beta = 0$.**

**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. **Solution: The sample mean is just** $\bar{x} = \sum_{i=1}^n x_i/n = 43.66/12 = 3.6383$ **. The sample variance can be computed as** $S^2 = n(\bar{x^2} - \bar{x}^2)/(n-1) = (159.1936 - 43.66^2/12)/11 = 0.0313$ **.**

**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 $\alpha = 0.1$? **Solution: In this situation we clearly want to test** $H_0 : \mu = 3.6$ **vs** $H_1 : \mu \neq 3.6$ **. We can take as test statistic** $T = \bar{X}$ **. Considering the alternative, clearly we want to reject** $H_0$ **if** $\bar{X}$ **is too different from** $3.6$ **so our rejection rule is that we reject** $H_0$ **if** $|\bar{X} - 3.6| > c^*$ **for some appropriate** $c^*$ **. To calibrate the test we need to ensure that we pick** $c^*$ **in such a way that the probability of rejecting** $H_0$ **when it is true is** $\alpha$ **(which we later want to take as** $0.1$**.). Under the null,** $\bar{X} \sim N(3.6, \sigma^2/n)$ **so that, under the null,** $(\bar{X} - 3.6)/(S/\sqrt{n}) \sim t_{n-1}$ **. This means that we want**

$$\alpha = \mathbb{P}_{\mu=3.6}(|\bar{X} - 3.6| > c^*) = 1 - \mathbb{P}_{\mu=3.6}(|\bar{X} - 3.6| \leq c^*),$$

**so that**

$$\alpha = 1 - \mathbb{P}_{\mu=3.6}(-c^* \leq \bar{X} - 3.6 \leq c^*) = 1 - \mathbb{P}_{\mu=3.6}(-\sqrt{n}c^*/S \leq \sqrt{n}(\bar{X} - 3.6)/S \leq \sqrt{n}c^*/S).$$

**If** $F_{t_{n-1}}$ **denotes the CDF of a** $t_{n-1}$ **distribution, then we want**

$$\alpha = 1 - \{F_{t_{n-1}}(\sqrt{n}c^*/S) - F_{t_{n-1}}(-\sqrt{n}c^*/S)\}.$$

**Since** $F_{t_{n-1}}(z) = 1 - F_{t_{n-1}}(-z)$**, this is the same as** $\alpha = 1 - \{1 - 2F_{t_{n-1}}(-\sqrt{n}c^*/S)\}$ **or** $\alpha = 2F_{t_{n-1}}(-\sqrt{n}c^*/S)$**. We can now easily solve for** $c^*$ **to get** $F_{t_{n-1}}(-\sqrt{n}c^*/S) = \alpha/2$ **or**

$-\sqrt{n}c^*/S = F^{-1}_{t_{n-1}}(\alpha/2) = t_{n-1;\alpha/2}$ **or** $c^* = -S\,t_{n-1;\alpha/2}/\sqrt{n}$**. Plugging everything in, we get** $c^* = -\sqrt{0.0313/12}\,t_{11;0.1/2} = \sqrt{0.0313/12}\,1.7959 = 0.0917$ **. Since it is not true that** $|\bar{x} - 3.6| = |3.6383 - 0.36| = 0.0383$ **is larger than** $0.0917$ **the conclusion is that we do not reject** $H_0$ **at level 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? **Solution: We saw in the previous question that we should reject** $H_0$ **at level** $\alpha$ **if** $|\bar{x} - 3.6| = 0.0383 > c^* = -S\,t_{n-1;\alpha/2}/\sqrt{n}$ **. So the values of** $\alpha$ **for which we reject satisfy** $-0.0383 \leq -S\,t_{n-1;\alpha/2}/\sqrt{n}$ **or** $-\sqrt{n}0.0383/S \leq t_{n-1;\alpha/2}$**. This means that all** $\alpha$ **such that** $\alpha \geq 2F_{t_{n-1}}(-\sqrt{n}0.0383/S) = 2F_{t_{n-1}}(-\sqrt{12}0.0383/0.1768) = 2F_{t_{n-1}}(-0.7509)$ **lead to rejection. As such, since the p-value is the smallest** $\alpha$ **for which we reject, we conclude that the p-value is** $2F_{t_{n-1}}(-0.7509)$ **. This is actually** $0.4684$ **which is larger than 0.1 (no rejection) but you already know that at level** $0.1$ **you don't reject since the p-value just gives you an alternative way of conduction the test but gives you the same answer at the same significance level .**

**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. **Solution: We know that in this setting** $\sqrt{n}(\bar{X} - \mu)/S$ **has a** $t_{n-1}$ **distribution and is thus a pivot for** $\mu$**. We then know that**

$$\mathbb{P}\big(t_{n-1;0.05} \leq \sqrt{n}(\bar{X} - \mu)/S \leq -t_{n-1;0.05}\big) = 0.9.$$

**Solving the above for** $\mu$ **we get that** $[\bar{X} + St_{n-1;0.05}/\sqrt{n},\ \bar{X} - St_{n-1;0.05}/\sqrt{n}]$ **is a confidence interval of level** $0.9$ **for** $\mu$**. Plugging everything in, we get that this confidence interval is** $[3.5467, 3.7300]$ **. The answer to the question is yes: since 3.6 belongs to this 0.9 level confidence interval, we know that we cannot reject** $H_0$ **at level** $0.1$**. As we saw in class, a test that rejects** $H_0$ **if 3.6 belongs to a 0.9 level confidence interval for** $\mu$ **has level** $1 - 0.9 = 0.1$**.**

**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.

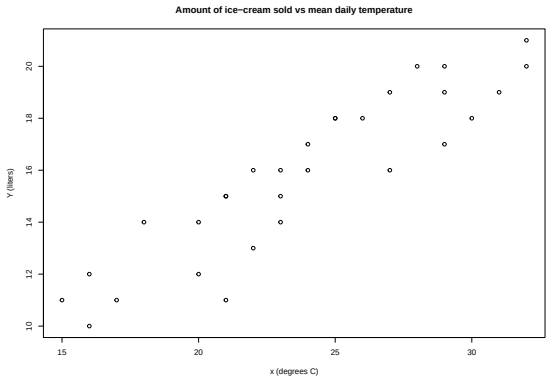


Figure 1: Plot of the ice-cream dataset.

| 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 $\beta$ and $\alpha$, respectively. **Solution: We model each response as $Y_i = \alpha + \beta x_i + \sigma\epsilon_i$ . In the Simple Linear Regression model we assume that the $\epsilon_i$ terms are i.i.d., with expectation 0 and variance 1 . In this particular case you can also just say that the noise terms are i.i.d. $N(0,1)$ .**

**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. **Solution: We can use $\hat{\alpha} + 27 \times \hat{\beta}$ to estimate the expected amount of ice-cream that will be sold in a day with temperature forecast of 27 degrees Celsius . We have that $\hat{\beta} = S_{xy}/SS_{xx} = 403.3333/691.4667 = 0.5833$ and $\hat{\alpha} = \bar{y} - \bar{x} \times SS_{xy}/SS_{xx} = \bar{y} - \bar{x} \times \hat{\beta} = 15.8333 - 23.8667 \times 0.5833 = 1.9119$ . So our point estimate for the expected quantity of ice-cream that will be sold is $\hat{\alpha} + 27\hat{\beta} = 1.9119 + 27 \times 0.5833 = 17.6610$ litres .**

**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 $\beta$ 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

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

**Solution: We can use $T = \hat{\beta}$ (or just the pivot above) as test statistic and should reject if $|T| > c^*$ . Under the null we know that $\sqrt{SS_{xx}}(\hat{\beta} - 0)/\hat{\sigma} \sim t_{n-2}$ so for a test of level $\alpha$ we need**

$$\alpha = \mathbb{P}_{\beta=0}(|\hat{\beta}| > c^*) = \mathbb{P}_{\beta=0}(|\sqrt{SS_{xx}}\hat{\beta}/\hat{\sigma}| > \sqrt{SS_{xx}}c^*/\hat{\sigma}).$$

**Using that $F_{t_{n-2}}(z) = 1 - F_{t_{n-2}}(-z)$, this is the same as**

$$\alpha = 1 - \left\{F_{t_{n-2}}(\sqrt{SS_{xx}}c^*/\hat{\sigma}) - F_{t_{n-2}}(-\sqrt{SS_{xx}}c^*/\hat{\sigma})\right\} = 2F_{t_{n-2}}(-\sqrt{SS_{xx}}c^*/\hat{\sigma}).$$

**From this we conclude that using the critical value $c^* = -\hat{\sigma}t_{n-2;\alpha/2}/\sqrt{SS_{xx}}$ leads to a test of level $\alpha$ . The $\alpha$ that lead to rejection are the ones for which $|\hat{\beta}| > -\hat{\sigma}t_{n-2;\alpha/2}/\sqrt{SS_{xx}}$. Solving for $\alpha$ gives $t_{n-2;\alpha/2} = F^{-1}_{t_{n-2}}(\alpha/2) > -\sqrt{SS_{xx}}|\hat{\beta}|/\hat{\sigma}$, or $\alpha > 2F_{t_{n-2}}(-\sqrt{SS_{xx}}|\hat{\beta}|/\hat{\sigma})$. So the smallest $\alpha$ that leads to rejection is $2F_{t_{n-2}}(-\sqrt{SS_{xx}}|\hat{\beta}|/\hat{\sigma})$, so that is the p-value . We already know that $\hat{\beta} = 0.5833$ and that $SS_{xx} = 691.4667$, so we just need to compute $\hat{\sigma}$. This is $\hat{\sigma} = \sqrt{SS_{yy}/n - \hat{\beta}^2 SS_{xx}/n} = \sqrt{284.1667/30 - 0.5833^2 \times 691.4667/30} = 1.2767$ . Plugging everything in we see that p-value$= 2F_{t_{n-2}}(-\sqrt{691.4667}|0.5833|/1.2767) = 2F_{t_{n-2}}(-\sqrt{691.4667}|0.5833|/1.2767) = 2F_{t_{n-2}}(-12.0137) \approx 0$. Under these circumstances we can conclude that that we reject the null at pretty much any significance level and so conclude that $\beta \neq 0$.**