Example sheet 3

Frequentist inference

Foundations of Data Science—DJW—2019/2020

For the questions that ask "find …', you may give either a formula, or pseudocode. Or, if the question gives you numerical data, you are encouraged to give actual code and a numerical answer.

Question 1. We are given a dataset $x_1, \dots, x_n$ which we believe is drawn from Normal(μ, σ^2) where μ and σ are unknown.

(a) Find the maximum likeihood estimators $\hat{\mu}$ and $\hat{\sigma}$.

(b) Find a 95% confidence interval for $\hat{\sigma}$, using parametric resampling.

(c) Repeat, but using non-parametric resampling.

Question 2. We are given a dataset $x_1, \dots, x_n$ which we believe is drawn from Uniform$[0, \theta]$ where θ is unknown. Recall from example sheet 1 that the maximum likelihood estimator is $\hat{\theta} = \max_i x_i$. Find a 95% confidence interval for $\hat{\theta}$, both using parametric resampling and using non-parametric resampling.

Question 3. The linear model for temperature increase from lecture notes section 2.2.5 invites the probability model

$$\mathsf{temp}_i = \alpha + \beta_1 \sin(2\pi \mathsf{t}_i) + \beta_2 \cos(2\pi \mathsf{t}_i) + \gamma(\mathsf{t}_i - 2000) + \text{Normal}(0, \sigma^2).$$

Find a 95% confidence interval for $\hat{\gamma}$, the maximum likelihood estimator for the rate of temperature increase.

Question 4. The number of unsolved murders in Kembleford over three successive years was 3, 1, 5. The police chief was then replaced, and the numbers over the following two years were 2, 3. We know from general policing knowledge that the number of unsolved murders in a given year follows the Poisson distribution. Assuming that the numbers come from Poisson(λ) under the old chief and from Poisson(μ) under the new chief, estimate the mean change $\mu - \lambda$ and report a 95% confidence interval.

Note. The Poisson(λ) *distribution takes values in* $\{0, 1, \dots\}$ *and has probability mass function* $\Pr(x) = \lambda^x e^{-\lambda}/x!$. *Its cdf can be found using* `scipy.stats.poisson.cdf(x, mu=λ)`.

Question 5. I tossed a coin $n = 100$ times and got $x = 65$ heads. My null hypothesis is that the coin is unbiased, and my alternative hypothesis is that the probability of heads is some arbitrary value $p \in [0, 1]$. Consider testing H_0 using the test statistic $t = x$ itself.

(a) If H_0 is true, what is the distribution of the test statistic?

(b) Should I use a one-sided or a two-sided test?

(c) Find the p-value.

[Hint. `scipy.stats.binom.cdf(x, n, p)` *gives* $\mathbb{P}(\text{Binom}(n, p) \le x)$.*]*

Hints and comments

Question 1. For part (a) you should learn these formulae by heart, and be able to derive them without thinking: $\hat{\mu}$ is the sample mean $\bar{x}$, and $\hat{\sigma}$ is $\sqrt{n^{-1} \sum_i (x_i - \bar{x})^2}$. For part (b), see example 7.1 from lecture notes. For part (c), see example 7.2 from lecture notes for the resampling method; the statistic we're interested in is the same as in part (b), namely $\hat{\sigma}$.

Question 2. The maths is very similar to question 1. However, the result is unhelpful—see the supplementary question sheet.

Question 3. Use parametric resampling. You have to find maximum likelihood estimators for all the parameters, including σ.

Question 4. See example 7.6 from lecture notes.

Question 5. See section 7.3 of lecture notes. Part (a) is just the same as every single illustration in that section, except it's simpler because there are no parameters to estimate, and we know the distribution precisely. It's just a fancy way of asking "What is the distribution of the number of heads, when I toss a fair coin?", and the answer is of course Binom$(100, 1/2)$. For parts (b) and (c), read the notes about the Neyman–Pearson approach, and follow example 7.13.

Supplementary questions

These questions are not intended for supervision (unless your supervisor directs you otherwise). Some require careful maths, some are best answered with coding, some are philosophical.

Question 6. Given a dataset $x_1, \ldots, x_n$, let X^* be the random variable produced by selecting one of the datapoints at random. This has distribution

$$\Pr_{X^*}(x) = \frac{1}{n}\begin{pmatrix}\text{num. datapoints}\\ \text{that are equal to } x\end{pmatrix}.$$

Show that $\mathbb{E}\, X^* = \bar{x}$ and $\operatorname{Var} X^* = n^{-1}\sum_i (x_i - \bar{x})^2$, where $\bar{x} = n^{-1}\sum_i x_i$.

> *Don't be put off by the fancy setting. This is just a question asking you for the mean and variance of a random variable, the sort of question you answered in IA Maths.*
>
> *Try to write out your answer formally, using the Law of Total Expectation, which is a fancy form of the Law of Total Probability on example sheet 0. It says that for any pair of random variables Y and Z,*
>
> $$\mathbb{E}\, Y = \sum_z \mathbb{E}(Y \mid Z = z)\, \mathbb{P}(Z = z).$$
>
> *You could let Z be the index of the selected datapoint, an integer in $\{1, \ldots, n\}$. Or you could let Z be the value of the selected datapoint.*

Question 7. I implement the two resamplers from question 2. To test them, I generate 1000 values from Uniform$[0, a]$ with $a = 2$, and find a 95% confidence interval for $\hat{a}$. I repeat this 20 times. Not once does my confidence interval include the true value, $a = 2$, for either resampler. Explain.

> *Resampling is an heuristic, not a perfect procedure. It works well for 'central' statistics like averages or sums. It doesn't work well for certain types of extreme statistics (like the maximum of a dataset) nor for certain types of distribution (like the uniform). Section 8 of the extended lecture notes, on overfitting, frames the discussion; that section is not examinable material.*

Question 8. In the setting of question 3, I have defined a function for computing the fitted temperature at an arbitrary future timepoint,

`def temp(t): return` $\hat{\alpha} + \hat{\beta}_1$ `sin(2πt)` $+ \hat{\beta}_2$ `cos(2πt)` $+ \hat{\gamma}$`(t−2000)`

(a) Modify the code to also return a 95% confidence interval.

(b) This code returns the *fitted* temperature at `t`. Modify it to return a 95% confidence interval for the *actual* temperature, which is fitted temperature plus noise.

> *For part (a), you should organize your code so that it accepts a vector of `t` values, and it doesn't resample for every value in the vector. This is how we typically want to use it, to plot nice smooth confidence ribbons.*
>
> *For part (b), here's one way to think about it. From part (a), you can generate a resampled set of values for $X^* =$ `temp(t)`, and this is what you used to generate a confidence interval. For part (b), we want to report the distribution of $Y = X^* + \text{Normal}(0, \sigma^2)$. (Let's pretend that $\hat{\sigma}$ is known, for now.) We know the joint distribution of X^* and the noise term, so we can find the distribution of Y using marginalization. See section 3.4 of lecture notes, or the slides for lecture 8 (slides are on Moodle). Next: how should you deal with the fact that $\hat{\sigma}$ is estimated?*

Question 9. We are given a dataset $x_1, \ldots, x_n$. Our null hypothesis is that these values are drawn from Normal$(0, \sigma^2)$, where σ is an unknown parameter. Alternatively, we suspect that values might come from distributions with unequal variances. Consider the test statistic

$$t = \sum_{i=1}^{n} \left(\frac{x_i}{\hat{\sigma}}\right)^2.$$

This would tend to be bigger under the alternative hypothesis than under the null hypothesis.

(a) Find the distribution of t that we'd expect to see, if the null hypothesis holds.

(b) Find the p-value.