Example sheet 3

Frequentist inference

Data Science—DJW—2021/2022

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 $x_1, \ldots, x_n$ which we believe is drawn from Normal(μ, σ^2) where μ and σ are unknown.

(a) Find the maximum likelihood 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. 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. Model the numbers as Poisson(μ) under the old chief and Poisson(ν) under the new chief.

(a) Report a 95% confidence interval for $\hat{\nu} - \hat{\mu}$, using parametric sampling.

(b) Conduct a hypothesis test of the hypothesis $\mu = \nu$, using parametric sampling, and using the test statistic $\hat{\nu} - \hat{\mu}$. Explain your choice between a one-sided and a two-sided test.

(c) Explain carefully the difference in sampling methods between parts (a) and (b).

[Note. The Poisson(λ) *distribution takes values in* $\{0, 1, \ldots\}$ *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 3. (a) In section 2.2 we considered a climate model in which temperatures increase linearly. The probabilistic version of the model is

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

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

(b) To allow for non-linear temperature increase, example sheet 1 suggested a model with a step function,

$$\texttt{temp} \sim \beta_1 \sin(2\pi \texttt{t}) + \beta_2 \cos(2\pi \texttt{t}) + \gamma_{\texttt{decade}} + \text{Normal}(0, \sigma^2).$$

Find a 95% confidence interval for $\hat{\gamma}_{2010\text{s}} - \hat{\gamma}_{1980\text{s}}$. Conduct a hypothesis test of whether $\gamma_{1980\text{s}} = \gamma_{2010\text{s}}$.

[Note. You are encouraged to implement your solution. A code skeleton is provided at `https://github.com/damonjw/datasci/blob/master/ex3.ipynb`*]*

Question 4. I toss a coin n times and get the answers $x_1, \ldots, x_n$. My model is that each toss is $X_i \sim \text{Bin}(1, \theta)$, and I wish to test the null hypothesis that $\theta \geq 1/2$.

(a) Find an expression for $\Pr(x_1, \ldots, x_n \,;\, \theta)$. Give your expression as a function of $y = \sum_i x_i$.

(b) Sketch $\log \Pr(x_1, \ldots, x_n \,;\, \theta)$ as a function of θ, for two cases: $y < n/2$, and $y > n/2$.

(c) Assuming H_0 is true, what is the maximum likelihood estimator for θ?

(d) Let the test statistic be y. What is the distribution of this test statistic, when θ is equal to your value from part (c)?

(e) Explain why a one-sided hypothesis test is appropriate. Give an expression for the p-value of the test.

Question 5. Your attempts at a task succeed with probability θ, and fail with probability $1 - \theta$. How long an unbroken list of failures does it take, for you to reject "$\theta \geq 1/2$" at p-value 5%?

Question 6. A recent paper *Historical language records reveal a surge of cognitive distortions in recent decades* by Bollen et al., `https://www.pnas.org/content/118/30/e2102061118.full`, claims that depression-linked turns of phrase have become more prevalent in recent decades. This paper reports both confidence intervals and null hypotheses. Explain how it is computes them, in particular (1) the readout statistic, (2) the sampling method.

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), use the general method of example 8.1.1 from lecture notes, but remember this question is asking you for a confidence interval for $\hat{\sigma}$ not for $\hat{\mu}$. For part (c), see example 8.3.1.

Question 2. For part (a), follow example 8.1.3 from lecture notes. For the maximum likelihood calculation, see your answers to example sheet 1. For part (b), follow example 8.2.1 (though, as this question is asking you to use a different parameterization of the model, you need to think about what test statistic to use; a sensible choice is $\hat{\nu} - \hat{\mu}$).

In questions where you're given a parametric model, and asked to test a hypothesis that restricts the parameters, and it's left to you to choose a test statistic, it's a good strategy to (i) find the maximum likelihood estimators under the general model, (ii) invent some plausible-looking function based on those maximum likelihood estimators. Ask yourself how your statistic would differ between the scenario where H_0 is true, and the scenario where H_0 isn't true. This will tell you what "more extreme" means, in the definition of p-value, and hence whether to use a one-sided or two-sided test.

Question 3. Follow the general strategies from sections 8.1 and 8.2 of lecture notes. Note that the model in part (b) doesn't have an intercept term; an intercept term is redundant since the one-hot encoded $\gamma_{\texttt{decade}}$ terms sum to 1.

In your answers for this question, it's a good idea to use sklearn wherever reasonable. There's no point going through lots of algebra, when there are fast easy routines that you can use. Use `LinearRegression.predict`, then look at exercise 8.1.4 lines 14–15.

For the hypothesis test, you'll need to fit the model under the assumption that $\gamma_{1980\text{s}} = \gamma_{2010\text{s}}$. This restricted model can be written as

$$\begin{aligned}\texttt{temp} \sim{}& \beta_1 \sin(2\pi \texttt{t}) + \beta_2 \cos(2\pi \texttt{t}) \\ &+ \gamma_{1990\text{s}} 1_{\texttt{decade}=1990\text{s}} + \gamma_{2000\text{s}} 1_{\texttt{decade}=2000\text{s}} + \gamma_{2020\text{s}} 1_{\texttt{decade}=2020\text{s}} + \alpha 1_{\texttt{decade}\in\{1980\text{s},2010\text{s}}.\end{aligned}$$

Question 4. Parts (a)–(c) are plain old maximum likelihood estimation. The only novelty here is that we need to find the value of θ that maximizes the likelihood function, *under the restriction* $\theta \geq 1/2$. You should get the answer

$$\hat{\theta} = \max\Big(\frac{y}{n}, \frac{1}{2}\Big).$$

In lecture notes we only went through numerical computation of p-values. In this question, the distribution of $y(X)$ is so simple that you can write out an explicit formula for the p-value.

Question 5. Use your expression for the p-value from question 4 part (e), with the observed data $y = 0$. Let the value of this expression be ≤ 0.05 and solve for n.

Question 6. Skim-read the whole paper, and read the *Materials and Methods* section closely. Note that the word 'bootstrapping' is another name for 'non-parametric resampling'. You can find a definition of z-score on Wikipedia, but it doesn't add anything to the explanation given in the paper.

In the notation used in this course, the dataset used in the paper is $(x_1, y_1), \ldots, (x_k, y_k)$ where y_k is a vector

$$y_i = \big[y_{i,1855}, \ldots, y_{i,2020}\big]$$

giving the prevalence of n-gram i in each year, and $x_i \in \{1, 2, 3, 4, 5\}$ is the number of words in that n-gram.

The readout statistic $t(x_1, \ldots, x_k)$ is well hidden, and you will have to dig through the whole paper to find it.

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 7. We are given a dataset $x_1, \ldots, x_n$ 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.

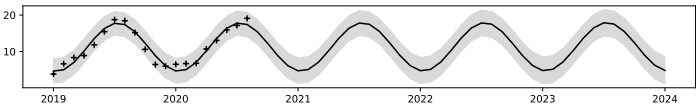
Question 8. I implement the two resamplers from question 7. To test them, I generate 1000 values from Uniform$[0, \theta]$ with $\theta = 2$, and find a 95% confidence interval for $\hat{\theta}$. I repeat this 20 times. Not once does my confidence interval include the true value, $\theta = 2$, for either 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 9 of the extended lecture notes, on overfitting, frames the discussion; that section is not examinable material.*

Question 9. In the setting of question 3, I have defined a function for computing the fitted temperature at an arbitrary future timepoint,

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

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

> *Remember that the readout statistic must be a function **of the dataset**. In this case, think of the predictor function as*
> `def pred(x,tnew): return` $\hat{\alpha}(x)$ `+` $\hat{\beta}_1(x)$ `sin(2π tnew) +` $\hat{\beta}_2(x)$ `cos(2π tnew) +` $\hat{\gamma}(x)$`(tnew−2000)`
> *where x is a vector of temperatures (either from the original dataset, or from the synthetic dataset), and t_{new} is an arbitrary new timepoint, nothing to do with the `t` vector in the dataset. For the confidence interval at t_{new}, you need to sample many values of X, and find the spread of `pred(`X`,`t_{new}`)`.*
>
> *You should organize your code so that it accepts a vector of t_{new} values, and it doesn't resample for every value in the t_{new} vector. That way, it's easy to plot a 'confidence ribbon' which shows, for many values of t_{new}, both the predicted value and the confidence range. (In this illustration, I have drawn the confidence ribbon artificially wide. You should get a confidence interval that's barely visible.)*
>
> 
>

Question 10. 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. Let

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1[x_i/\hat{\sigma} \leq x]$$

where $\hat{\sigma} = \sqrt{n^{-1} \sum_i x_i^2}$ is the maximum likelihood estimator for σ. If the null hypothesis is true, we'd expect $\hat{F}(x)$ to be reasonably close to $\Phi(x)$, the cumulative distribution function for Normal$(0, 1)$, for all x. Suggest how to test the hypothesis that the dataset is indeed drawn from Normal$(0, \sigma^2)$, using a test statistic based on $\hat{F}$ and Φ.

> *This question is asing you to be creative in inventing a test statistic. If you don't feel creative, look up the Kolmogorov-Smirnov test.*
>
> *When we fit a linear model, there's an assumption that the residuals are normally distributed (as discussed in sections 2.4 and 2.7). After fitting a linear model, it's worth testing whether the residuals are indeed normally distributed.*

Question 11. While driving along the motorway, you pass as many cars as pass you. Does this mean you are driving at the median speed?

There are many ways to answer this question. Here's a plodding methodical method, based on the spirit of empirical distributions.

First, imagine a reference line at one location along the motorway. Suppose that each vehicle's speed is one of a finite set of speeds $v_1, \ldots, v_n$. Let t be the total number of vehicles per hour that cross the reference line, and let p_i be the fraction of them that have speed v_i. From this, you can derive the distance x_i between cars in class i, assuming that cars in each class are evenly spaced. This is a roundabout way to get at what we mean by 'median over all vehicles'.

Now, suppose you're driving at speed v_0. Given all the v_i and x_i, find how many vehicles per hour pass you, and how many are passed by you. Let these numbers be equal, and solve for v_0.

This answer is based on the empirical distribution in that, rather than grappling with abstract distributions, we have imagined a very concrete tangible dataset of vehicles, and we've reasoned about this dataset.