Example sheet 3

Frequentist inference

Data Science—DJW—2023/2024

Before attempting this example sheet, it is a good idea to work through the exercises in lecture notes sections 9.2 (confidence intervals), 9.3 (hypothesis testing), and 9.6 (non-parametric resampling).

Questions labelled * are more challenging. For some questions you can test your answers using the online tester; there is a notebook with templates for answers and instructions for submission on the course materials webpage.

Following this example sheet is a page with hints for each question. There is also a set of more advanced supplementary questions. These are not intended for supervision (unless your supervisor directs you otherwise).

IF AT FIRST YOU DON'T SUCCEED
TRY TRY AGAIN UNTIL YOUR FAILURE IS STATISTICALLY SIGNIFICANT

Question 1. Sketch the cumulative distribution function, and calculate the density function, for this random variable:

```
def rx():
    u1 = random.random()
    u2 = random.random()
    return min(u1,u2)
```

Question 2*. The dataset at `https://www.cl.cam.ac.uk/teaching/current/DataSci/data/responsetime_ms.txt` is a list of web server response times, measured in milliseconds.

(a) Plot the empirical cumulative distribution function (ecdf) of this sample.

(b) Plot the empirical tail distribution function (etdf = 1 − ecdf), on a log-log plot.

(c) You should see that, for large enough response times, the etdf looks roughly like a (noisy) straight line on a log-log plot. Using this observation, estimate the 99.9%ile and 99.99%ile of response time.

Question 3. We are given a dataset $x = [x_1, \dots, x_n]$ which we believe is drawn from $\text{Normal}(\mu, \sigma^2)$ 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.

[Optional: To test your code using the online tester, fill in the answer template for `sd_confint_parametric` and `sd_confint_nonparametric`.]

Question 4. We are given data $x = [x_1, \ldots, x_m]$ which we believe is sampled from $\mathrm{Exp}(\mu)$, and further data $y = [y_1, \ldots, y_n]$ which we believe is sampled from $\mathrm{Exp}(\nu)$. We wish to test the hypothesis that $\mu = \nu$.

(a) Under this hypothesis, all the datapoints in x and y are sampled from a commmon distribution $\mathrm{Exp}(\lambda)$, where λ is the common value of μ and ν. Find the maximum likelihood estimator $\hat{\lambda}$.

(b) Explain how to compute the p-value for this test, using the test statistic $\hat{\nu} - \hat{\mu}$, with parametric resampling.

[Optional: To test your code using the online tester, fill in the answer template for `exp_equality_test`.]

Question 5. We are given a dataset of (g_i, x_i, y_i) records, $i = 1, \ldots, n$, where $g_i \in \{1, 2, 3\}$ is the group that record i belongs to, $x_i \in \mathbb{R}$ is a predictor variable, and $y_i \in \mathbb{R}$ is the response. We are interested in the model

$$Y_i \sim \alpha_{g_i} + \beta_{g_i} x_i + N(0, \sigma^2).$$

(a) Explain how to fit this model to the dataset.

(b) We wish to test the hypothesis that $\beta_1 = \beta_2 = \beta_3$. Suggest a test statistic, and describe how to conduct the test.

Question 6*. (a) I toss a coin n times and get x heads. My model is that the number of heads is $\mathrm{Bin}(n, \theta)$ and I wish to test the null hypothesis that $\theta = 1/2$. Explain how to find the p-value for this test.

(b) I make many attempts at a task, and I have no successes at all, just a string of failures. Modelling my attempts as independent random variables with success probability θ and failure probability $1 - \theta$, how many failures does it take for me to reject $\theta = 1/2$ at p-value 5%?

Question 7. We have a climate dataset of $(\mathtt{t}, \mathtt{temp})$ pairs. Considered a model in which temperatures increase linearly,

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

Let $\hat{\gamma}$ be the maximum likelihood estimator for the rate of temperature increase. Explain how to find a 95% confidence interval for $\hat{\gamma}$.

Question 8. I have computed the maximum likelihood estimators for all the parameters in the model in question 7, and I have used them to define a temperature prediction function

$$\texttt{def pred(}\mathtt{t}_{\mathtt{new}}\texttt{): return } \hat{\alpha} \texttt{ + } \hat{\beta}_1 \texttt{ sin(}2\pi\mathtt{t}_{\mathtt{new}}\texttt{) + } \hat{\beta}_2 \texttt{ cos(}2\pi\mathtt{t}_{\mathtt{new}}\texttt{) + } \hat{\gamma}\texttt{(}\mathtt{t}_{\mathtt{new}}\texttt{-2000)}$$

Modify this code so that in addition to predicting the temperature it also produces a 95% confidence interval for its prediction.

Question 9. 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 $\mathrm{Poisson}(\mu)$ under the old chief and $\mathrm{Poisson}(\nu)$ under the new chief.

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

(b) Explain how to test the hypothesis that $\mu = \nu$, using parametric sampling.

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

Hints and comments

Question 1. Work through exercise 5.3.4 in lecture notes, then apply the same strategy to this question.

Question 2. The code for plotting the ecdf is in lecture notes section 7.1.

When there isn't enough data to get a reliable percentile from the dataset itself, in other words when the ecdf is very noisy, it's a good idea to sketch a continuous cdf that approximates the ecdf and read off percentiles from our sketch. When we're looking for extreme events, such as the 99.99%ile which has very low probability, it's a good idea to transform the y-axis of the ecdf plot to give more resolution at extreme events, by plotting log(1 − cdf(x)). To get the 99.9%ile, we simply want

$$\text{cdf}(x) = 0.999 \quad \Rightarrow \quad \log_{10}(1 - \text{cdf(x)}) = -3.$$

For a more thoughtful answer, you might think about how much noise there is in your etdf plot. How might you use non-parameteric resampling to get a sense of this?

Question 3. 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 9.2.1 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 9.6.1.

Question 4. Check the definition of the Exponential distribution in lecture notes section 1.2, and watch out for difference in convention between maths (which refers to the rate parameter μ) and numpy (which refers to the scale parameter $1/\mu$). You should find the mles to be $\hat{\mu} = 1/\bar{x}$, $\hat{\nu} = 1/\bar{y}$, $\hat{\lambda} = 1/\bar{z}$ where z is the concatenation of x and y. For the test, look at exercise 9.3.1 in lecture notes.

Question 5. 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. Look at exercise 9.3.2 for inspiration.

This question tells us a general hypothesis H_1, namely that $Y \sim \alpha_g + \beta_g x + N(0, \sigma^2)$; and it proposes a null hypothesis H_0 that is a restriction on the parameters of H_1, namely that $\beta_0 = \beta_1 = \beta_2$. Can you think up a test statistic using the $\hat{\beta}_g$ parameters from H_1 and $\hat{\beta}$ from H_0?

Question 6. For part (a), we only have a single datapoint namely the number of heads x, so we might as well use x itself as the test statistic. For this question, we can do much better than just giving pseudocode: we know the distribution that this test statistic will have under H_0, so we can write out the p-value exactly in terms of the cdf of the Binomial distribution.

For part (b), just use your expression for the p-value from part (a), applied to data $x = 0$. Your expression will depend on n. Find the smallest n such that $p \leq 0.05$.

Question 7. Follow the general strategy from section 8.2 of lecture notes. 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. You can generate a synthetic dataset with `np.random.normal(loc=pred, scale=`$\hat{\sigma}$`)`, as in exercise 8.2.4 lines 14–15, and you can compute the predicted temperatures `pred` as in section 2.1 line 13.

Question 8. We want to generate a multiverse of synthetic datasets, and canvas the opinion of data scientists across this multiverse. If a parallel-universe data scientist sees dataset X^*, what value would they produce for `pred(t`$_\text{new}$`=2050)`? You just need to assemble a large collection of these predictions, then find a 95% confidence interval in the usual way.

For an extra challenge, write your code so that it accepts a vector-valued `t`$_\text{new}$.

Question 9. The Poisson distribution is a common choice for counts of events, for example the number of buses passing a bus-stop over a given time period, or the number of radioactive particles emitted by a source, or the number of murders in a quiet English village.

For part (a), follow example 9.2.3 from lecture notes. For the maximum likelihood calculation, see your answers to Example Sheet 1. For part (b), follow example 9.3.1 (though you need to think about what test statistic to use; a sensible choice is $\hat{\nu} - \hat{\mu}$).

Supplementary questions

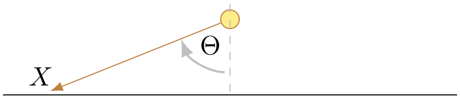
These questions are not intended for supervision (unless your supervisor directs you otherwise). Some require careful maths, some are best answered with coding, some are philosophical.

Question 10. Sketch the cumulative distribution function, and calculate the density function, for this random variable:

```
def rx():
    u = random.random()
    return u * (1-u)
```

Question 11. A point lightsource at coordinates $(0,1)$ sends out a ray of light at an angle Θ chosen uniformly in $(-\pi/2, \pi/2)$. Let X be the point where the ray intersects the horizontal line through the origin. What is the density of X?

Note: This random variable is known as the Cauchy distribution. It is unusual in that it has no mean.



Question 12. 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 13. I implement the two resamplers from question 12. 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.

> *Naive resampling (based on mle parameter estimates or on empirical distributions) 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).*
>
> *The idea of resampling is that we want to simulate novel unseen versions of the dataset. The best way to do this is to use a model that we think is a good description for novel unseen data—in other words, to use a model that fits a holdout dataset well. (See section 9 of lecture notes for a longer discussion of generalization. That section of notes is non-examinable.) One ad hoc way to get better generalization in this case is to use an unbiased estimator for θ rather than a maximum likelihood estimator; though this is happenstance, not a general principle!*

Question 14. Test the hypothesis that temperatures in Cambridge have not been changing, using a non-parametric test.

> *In lectures we looked at several examples of tests using parametric resampling. We also looked at one example of a test with non-parametric resampling, namely Fisher's permutation test. Example 8.6.2 in lecture notes gives another illustration of non-parametric sampling for hypothesis tests.*
>
> *For this dataset, it's blindingly obvious that there is an annual cycle in temperatures, so your resampling strategy must respect this. If there were no global warming, and you wanted to simulate a January, how could you simulate it using the data in this dataset?*
>
> *Second, the test statistic. You are at liberty to use any test statistic at all; it doesn't have to be linked to the resampling strategy. You might as well use $\hat{\gamma}$ from question 7.*

Question 15. We have a dataset $x_1, x_2, \ldots, x_n$, and we wish to model it as Normal(μ, σ^2) where μ and σ are unknown. How different are Bayesianist and frequentist confidence intervals for the mean? To be concrete, let's work with the first 10 values for `temp` in the climate dataset.

(a) Plot the log likelihood function $\log \Pr(x_1, \ldots, x_n | \mu, \sigma)$ as a function of μ and σ. (A code skeleton is provided in `https://github.com/damonjw/datasci/blob/master/ex/ex3.ipynb`.)

(b) Using frequentist resampling, generate 50 resampled datasets, find the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ for each, and show these 50 points on your plot.

(c) Using computational Bayesian methods, with priors $\mu \sim \text{Normal}(0, 10^2)$ and $\sigma \sim \Gamma(k = 2, \theta = 1)$ (where k and θ are as in the numpy documentation), sample 500 pairs from the prior distribution and show them on your plot. Then compute the posterior weights of these sampled pairs, and show the weighted pairs on your plot by setting the size of the plot marker in proportion to weight.

(d) Find the 95% confidence interval (for $\hat{\mu}$ in the frequentist case, and for $(\mu \mid \text{data})$ in the Bayesianist case), and show them on your plot.

(e) Repeat the exercise, using the first 100 values from the climate dataset.

> *You should see broadly similar outcomes, whether you're plotting frequentist samples of $(\hat{\mu}, \hat{\sigma})$ or whether you're plotting the Bayesianist samples that get non-negligible weight. When there are more datapoints, then the results are even more similar: there's a very narrow peak in the log likelihood plot, and the samples from both Bayesianist and frequentist approaches are heavily concentrated arount this peak. (Though the naive computational Bayesian procedure we learnt in this course doesn't work very well when the log likelihood has such a sharp spike.)*

Question 16. In hypothesis testing, what p-value would you expect if H_0 is true?

> *This is a mindbender! At first glance it's surprising that this question even has an answer that applies to any sort of hypothesis testing. And it's tricky to even work out what it's asking us to prove. Think of it this way ...*
>
> *In frequentist inference, we decide on a sampling distribution X^* that tells us what the dataset might have been if H_0 were true. We then compute the p-value by an operation on $t(x)$ and on the histogram of $t(X^*)$.*
>
> *Now, if H_0 were true, then the actual dataset x will look like a sample from X^*. If we perform the p-value operation not on the actual value $t(x)$ but on a typical value $t(X^*)$, what's the distribution we'll get for the p-value?*
>
> *You can find the answer at `https://en.wikipedia.org/wiki/Fisher's_method`. The page also describes how the answer can be used to combine the results of several independent tests.*

Question 17. We are given a dataset $x_1, \ldots, x_n$. Our null hypothesis is that these values are drawn from $\text{Normal}(0, \sigma^2)$, 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 $\text{Normal}(0, 1)$, for all x. Suggest how to test the hypothesis that the dataset is indeed drawn from $\text{Normal}(0, \sigma^2)$, 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 on Wikipedia.*
>
> *When we fit a linear model, there's an assumption that the residuals are normally distributed (as discussed in section 2.4). After fitting a linear model, it's always worth testing whether the residuals are indeed normally distributed, and this question gives you a way to do this test.*

Question 18 (Cardinality estimation).

(a) Let T be the maximum of m independent Uniform$[0, 1]$ random variables. Show that $\mathbb{P}(T \leq t) = t^m$. Find the density function $\Pr_T(t)$. *Hint. For two independent random variables U and V,*

$$\mathbb{P}\big(\max(U, V) \leq x\big) = \mathbb{P}(U \leq x \text{ and } V \leq x) = \mathbb{P}(U \leq x)\,\mathbb{P}(V \leq x).$$

(b) A common task in data processing is counting the number of unique items in a collection. When the collection is too large to hold in memory, we may wish to use fast approximation methods, such as the following: Given a collection of items $a_1, a_2, \ldots$, compute the hash of each item $x_1 = h(a_1), x_2 = h(a_2), \ldots$, then compute $t = \max_i x_i$.

If the hash function is well designed, then each x_i can be treated as if it were sampled from $\text{Uniform}[0, 1]$, and unequal items will yield independent samples..

The more unique items there are, the larger we expect t to be. Given an observed value t, find the maximum likelihood estimator for the number of unique items. *[Hint. This is about finding the mle from a single observation, as in lecture notes example 1.3.1.]*

`http://blog.notdot.net/2012/09/Dam-Cool-Algorithms-Cardinality-Estimation`

Question 19. 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.

> *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*
>
> $$y_i = \left[y_{i,1855}, \ldots, y_{i,2020}\right]$$
>
> *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.*

Question 20. 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 $\hat{\gamma}_{\text{2010s}} - \hat{\gamma}_{\text{1980s}}$. Conduct a hypothesis test of whether $\gamma_{\text{1980s}} = \gamma_{\text{2010s}}$.

Question 21. 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.