Homework 4: Due Tuesday April 5, 2019

In what follows $\Phi(x)$ is the error function, i.e. the CDF of a standard normal. Φ^{-1} is its inverse function.

Given $X_1, \ldots, X_n$ and a CDF F define

$$D_n = \max_{1 \le i \le n} \left| \frac{i}{n} - F(X_{i:n}) \right|.$$

We have seen that this is a good statistic for testing the hypothesis that the data has CDF F. Furthermore, we know that D_n has the same distribution for all continuous F.

Problem 1a. Write a code that takes as input an integer n (sample size) and a number $\alpha \in (0,1)$ (significance level). It then generates a sample of n Uniform(0,1) random variables, computes D_n (which now simply has $F(X_{i:n}) = X_{i:n}$), and repeats this $10,000$ times (you can add another input parameter M and set it to $10,000$, if you wish). Then, given these $10,000$ values of D_n it finds a cut-off $\delta > 0$ (the smallest possible) such that $P(D_n > \delta) \le \alpha$. The output of the code is this value δ.

Write also a variant of the code that takes as input an integer n and a number $\delta > 0$ and gives as output (an estimate of) the value of $P(D_n > \delta)$ (i.e. a p-value).

Problem 1b. Use your code to create a table that shows the values of δ for $\alpha \in \{0.01, 0.05, 0.1, 0.15, 0.2\}$ and $n \in \{5, 10, 20, 30, 40, 50, 100, 500, 1000\}$.

Problem 1c. Create a similar table using the formula that comes from the Brownian bridge limiting distribution, i.e.

$$\delta \approx \sqrt{\frac{\ln(2/\alpha)}{2n}}.$$

Compare to the table from Problem 1b.

Problem 1d. Adjust the code from Problem 1a to use Normal(0,1) instead of Uniform(0,1). (You need then to use $F(X_{i:n}) = \Phi(X_{i:n})$ when you calculate D_n.) Then create a table as in Problem 1b. Compare the two tables.

Problem 1e. Adjust the code again to use Bernoulli(1/2). (Now, $F(X_{i:n})$ needs to again be adjusted appropriately. What should it be?) Create a table and compare again with the table from Problem 1b.

Now, we would like to test a few data sets for normality. The data sets are created as follows:

1) 1000 i.i.d. computer-generated Normal(0,1) random variables.

2) 1000 i.i.d. computer-generated Normal(8,4) random variables.

3) First, let the computer pick a random mean μ (say uniformly from $(-2, 2)$) and a random variance σ^2 (say uniformly from $(1, 2)$). Then the sample is given by 1000 i.i.d. computer-generated Normal(μ, σ^2) random variables. Do not have the code tell you the values of μ and σ. Now you have a sample from a normally distributed population, but with unknown mean and variance.

4) 1000 computer-generated Uniform(0,1) random variables.

5) 1000 computer-generated Cauchy random variables.

6) 1000 computer-generated Exponential(1) random variables.

7) 1000 computer-generated Bernoulli(1/2) random variables.

8) Download the file `pi.txt` from the course website. This gives you the first million digits of π (not including the leading integer 3). The numbers are conveniently grouped into groups of 10. From every 10 consecutive digits create one data point from $(0, 1)$. For instance, the first data point would be 0.1415926535 and the second one is 0.8979323846. Create this way 1000 data points.

Problem 2. Plot a histogram and a Q-Q plot for each of the above data sets. For the Q-Q plot do not use a built-in function. Just plot $(X_{i:n}, \Phi^{-1}(i/n))$ for $i = 1, \ldots, 1000$. You can use a built-in function to compute Φ^{-1}. Which of the above data sets seem like they come from a normal distribution? (The most interesting one should be the last data set.)

The Q-Q plot is only a visual test. There are several ways to turn it into a concrete test. We will next explore two of these ways.

If we are testing for normality with a given known mean and variance, then we know the hypothesized CDF exactly (simply $F(x) = \Phi((x-\mu)/\sigma))$ and we can compute D_n.

Problem 3a. Test at significance level 0.01 the hypothesis that the data from 2) is Normal(0,2). Note that here you have to use $\mu = 0$, $\sigma^2 = 2$, and the CDF to use in the formula for D_n is $F(x) = \Phi(x/\sqrt{2})$. Once you compute D_n, you can use the table you created in Problem 1b (or in Problem 1d).

Problem 3b. Test at significance level 0.01 the hypothesis that the data from 2) is Normal(8,4). Now, $\mu = 8$, $\sigma^2 = 4$, and the CDF to use in the formula for D_n is $F(x) = \Phi((x-8)/2)$. You can again use the tables from Problems 1b or 1d.

If, on the other hand, we do not know the mean and variance (which is usually the case), then we cannot compute D_n (because we do not know F). But we can approximate it if we use

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}$$

in place of μ and σ^2. So then we define

$$\hat{D}_n = \max_{1 \le i \le n} \left| \frac{i}{n} - \Phi\Big(\frac{X_{i:n} - \bar{X}_n}{S_n}\Big) \right|$$

and use it as a measure of how far our data is from normality.

Observe that if we replace each data point X_i by $(X_i - \mu)/\sigma$, then the formula for $\hat{D}_n$ remains the same. This tells us that if we are after the distribution of $\hat{D}_n$, when the data is normal, then we can assume the data to be standard normal.

Problem 3c. Write a code similar to the one from Problem 1d but that computes $\hat{D}_n$ instead of D_n. (So in place of $\Phi(X_{i:n})$ you would use $\Phi((X_{i:n} - \bar{X}_n)/S_n)$. Note that it is S_n here, not S_n^2.) Use the code to create a table similar to Problem 1b. Now, use that table to test (at significance level say 0.01) for normality for all eight data sets.

Again, the most interesting case is the last one. Hence, provide also the p-value for the last data set. (This means you use the variant of the code where you give a cut-off δ and it gives you back an estimate of $P(\hat{D}_n > \delta)$. If you feed the code as δ the value of $\hat{D}_n$ from your data set, then the value it spits back would be the p-value.)

Another way is to use linear correlation.

Define R_n to be the correlation of the pairs $(X_{i:n}, \Phi^{-1}(i/n))$, $i \in \{1, \dots, n\}$. Observe again that if we replace each data point X_i with $(X_i - \mu)/\sigma$, then the formula for R_n does not change. This means that, under the assumption that the data is normal (with unknown mean and variance), the distribution of R_n is the same regardless of the mean and variance.

Problem 4a. Write a code that takes as input an integer n (sample size) and a number $\alpha \in (0, 1)$ (significance level). It then generates a sample of n standard normal random variables, computes R_n, and repeats this $10,000$ times (you can add another input parameter M and set it to $10,000$, if you wish). Then, given these $10,000$ values of R_n it finds a cut-off $r_0 \in (0, 1)$ (the largest possible) such that $P(R_n < r_0) \leq \alpha$. The output of the code is this value r_0.

Write another variant of the code that takes as input n and $r_0 \in (0, 1)$ and gives as output (an approximation of) the value of $P(R_n < r_0)$ (i.e. a p-value).

Problem 4b. Use the code from Problem 4a to test at significance level 0.01 the hypothesis that the above data sets are normal. Also, provide the p-value for the last data set.

Problem 5a. Use the data from 8) to plot a Q-Q plot for this data set, but against the Uniform(0,1) distribution. What does the plot suggest?

Problem 5b. Now, use your code from Problem 1b to test the hypothesis that the dataset in 8) is coming from i.i.d. Uniform(0,1) random variables. Give a p-value.

Problem 5c. Adjust your code from Problem 4a to test the same hypothesis as in 5b. Give a p-value.