Statistics for Astronomers
Solutions to Homework #6

Prof. Sundar Srinivasan

January 23, 2021

Note: the solutions below use the script `hw6.py`.

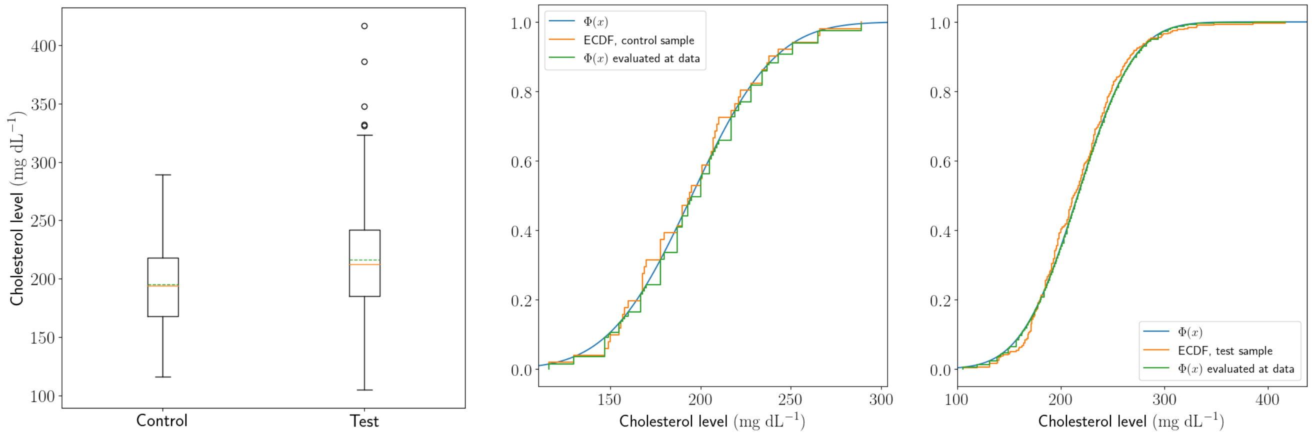

Figure 1: *Left*: box-and-whisker plots for the control and test samples, showing the medians (orange) and means (green) for each sample. *Right*: comparison of the empirical distributions of the samples with normal distributions centred at the sample means and with spreads equal to the sample standard deviations.

1. (a) Figure 1 shows the box-and-whisker plot for the cholesterol levels in the control and test samples. The mean and median of the control population are almost equal, indiicating a symmetric distribution. This is also borne out by the fact that the distance of the top and bottom whiskers from the mean are comparable.
   The mean of the test sample is slightly higher than the median, as is also evidenced by the presence of large-magnitude outliers in this sample (which bias the sample mean to higher values). While both samples are skewed towards larger values, the test sample is definitely more asymmetric than the control sample.

   (b) **This is a trick question!** We do not know the mean and variance of the populations from which the samples are drawn. *Can* we perform a KS test on these sample to test for normality in such a case? Naïvely, we might think that there are two possible ways to do so:

      i. What if we compare the raw data to a normal distribution whose location and scale parameters are set to the sample mean and standard deviation? Unfortunately, **parameters estimated from the data cannot be used to generate the model for a KS test** (see here)!

ii. What if we studentise the raw data and then compare it to a standard normal? The studentised data will be drawn from a Student's t-distribution **by definition**, (at least for large samples), which is similar enough in shape to the Gaussian that a KS test might result in false negatives (being unable to reject the null hypothesis that the sample is drawn from a normal distribution). In general, this won't help us draw correct inferences about the original data. We can demonstrate this for data drawn from an exponential distribution:

```
from scipy.stats import expon, kstest
x = expon.rvs(1.0, size = 10)
_, pvalue1 = kstest(x, 'norm')
xx = (x - x.mean()) / x.std(ddof = 1)
_, pvalue2 = kstest(xx, 'norm')
print(pvalue1, pvalue2)
2.87e-08 9.84e-01
```

The null hypothesis is correctly rejected for the original sample at 5% significance, but can't be rejected for the studentised sample. In fact, the p-value for the latter case is almost one, consistent with a high likelihood that the studentised sample is drawn from a Gaussian.

Therefore, the 1-sample KS test cannot be applied to the datasets in this problem.

We will have to use a different test to test for normality of the two samples. Let's try the Anderson-Darling test. The module `hw6q1b` prints out

```
*****1-sample Anderson-Darling test for control sample.*****
 H0:  sample drawn from a normal distribution.
 AD statistic = 0.22 <= critical value at 5.0% = 0.74, unable to reject H0.
------------------------------------------------------------
******1-sample Anderson-Darling test for test sample.*******
 H0:  sample drawn from a normal distribution.
 AD statistic = 1.3 >= critical value at 5.0% = 0.78, H0 rejected.
------------------------------------------------------------
At least one of the samples is not drawn from a normal distribution.
The t- and F-tests cannot be applied.
```

Since the KS test isn't applicable to this problem and the Anderson-Darling test points to at least one of the two datasets not being drawn from a normal distribution, the t- and F-tests in their original are not applicable to this problem. However, the `scipy.stats.ttest_ind` module is flexible enough that it's worth performing a 2-sample t-test assuming that the samples are independent. We set the keyword `alternative = 'less'` in the call to `ttest_ind` to determine whether the population mean of the control sample is smaller than that of the test sample, as would happen if people with heart disease have higher blood cholesterol. The code prints out

```
***********2-independent-sample t-test************
H0: population means are equal.
Ha: population mean of control sample < population mean of test sample.
p-value < alpha, H0 rejected.
```

```
--------------------------------------------------
```

Interestingly, the 2-sample t-test leads us to believe that there is a correlation between cholesterol level and heart disease. We can reinforce this result with nonparametric tests that do not require the assumption of normality, as is done in the following part of the problem.

(c) We use the `scipy.stats.mannwhitneyu` module with the keyword `alternative = 'less'` to test whether the population mean of the control sample is lower than that of the test sample. The module `hw6q1c` prints out

```
 **********************U-test**********************
 H0: population means are equal.
 Ha: population mean of control sample < population mean of test sample.
 p-value < alpha, H0 rejected.
--------------------------------------------------
```

This result is in agreement with the 2-sample t-test performed in the previous part of the problem. Thus, based on these data, we can establish a connection between cholesterol level and heart disease.

2. (a) The module `hw6q2a` returns

```
*****************2-sample KS test*****************
 H0: samples drawn from the same distribution.
 p-value = 0.09 >= alpha = 0.05, unable to reject H0.
--------------------------------------------------
**********2-sample Anderson-Darling test**********
 H0: samples drawn from the same distribution.
 AD statistic = 3.59 >= critical value at 5% = 1.96, H0 rejected.
--------------------------------------------------
```

There is good agreement in the central parts of the distributions; in fact, the two samples have nearly identical means (Figure 2); since the KS test is not sensitive to disagreement in the wings, it does not reject the null hypothesis. The distribution of absolute magnitudes for M31 globular clusters is tightly peaked around its central location, as demonstrated by the flatness of the distribution at extreme values and its steep rise in the centre. The Milky Way globular clusters are more spread out, as demonstrated by almost constant slope of its empirical distribution curve. The Anderson-Darling test is sensitive to such disagreements in the wings of the distributions, and is therefore able to reject the null hypothesis in this case. Figure 2 also shows box-and-whisker plots and histograms for both samples, confirming the larger spread of magnitudes in the Milky Way sample. Note, also, that the Milky Way distribution is more or less symmetric about its centre; the box plot and histogram show that the M31 data has a longer tail at the faint end.

(b) **Correction:** the question asks to use the U test to determine whether the samples are drawn the same distribution. This is incomplete; the U test is used determine whether two samples

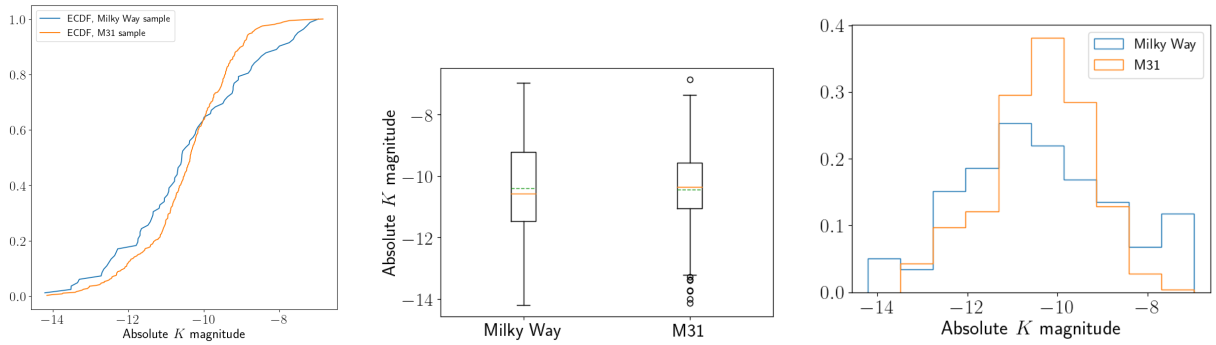

Figure 2: *Left*: empirical distributions for the Milky Way (blue) and M31 (orange) samples. Box-and-whisker plots (*center*) and histograms (*right*) are also shown for the two samples. The distribution of magnitudes is much tighter in the M31 case.

are drawn from distributions **with the same population mean**, regardless (*e.g.*) of whether the population variances are equal. The `hw6q23b` module prints out

```
**********************U test**********************
H0: population means are equal.
p-value = 0.41 >= alpha = 0.05, unable to reject H0.
--------------------------------------------------
```

(c) **Clarification**: formally, the Anderson-Darling test for normality requires studentisation if the population parameters are unknown (see, *e.g.*; here), **however, the `scipy.stats` version doesn't**. We demonstrate this using the module `hw6q2c`, which performs the test for both the original and studentised datasets, obtaining the same result regardless:

```
*********Anderson-Darling test for Milky Way sample*********
 H0:  sample drawn from a normal distribution.
 AD statistic = 0.3 <= critical value at 5.0% = 0.75, unable to reject H0.
------------------------------------------------------------
***Anderson-Darling test for studentised Milky Way sample***
 H0:  sample drawn from a normal distribution.
 AD statistic = 0.3 <= critical value at 5.0% = 0.75, unable to reject H0.
------------------------------------------------------------
************Anderson-Darling test for M31 sample************
 H0:  sample drawn from a normal distribution.
 AD statistic = 1.79 >= critical value at 5.0% = 0.78, H0 rejected.
------------------------------------------------------------
******Anderson-Darling test for studentised M31 sample******
 H0:  sample drawn from a normal distribution.
 AD statistic = 1.79 >= critical value at 5.0% = 0.78, H0 rejected.
------------------------------------------------------------
```

Based on these results, we can say that the M31 distribution (after studentisation) is less likely to be drawn from a standard normal. This is also clear from Figure 2.