Exercise 4.3

a)

i) No, we think this is not the case. In short: this test is no longer distribution free since it does make assumptions about the distribution of the data (the test assumes the data is normally distributed by including , the standard normal cdf (cumulative distribution function)).

A somewhat longer and more formal answer:

Note: since the assignment format document asks for brief answers, we will only recap a part of the original proof shortly, without restating the entire proof.

Recall the simple KS (Kolmogorov-Smirnov) test statistic:

Consider a random sample (i.i.d.) from an unknown distribution .

When testing:

versus

The KS test statistic is defined as:

The fact that the (simple) KS test is distribution free hinges (amongst other things) on the following:

(\*) Take the order statistics (i.i.d.) from the unknown distribution . Define the random variables for (note: **where is the cdf of each** ). It follows that these () variables are distributed. Since one can as such rewrite such that is no longer part of the equation, it follows that , and therefore the (simple) KS test is distribution free.

Is the above the case for the modified statistic ? In this case, is replaced with (a transformation of) , the standard normal cdf. Since we want to test normality specifically, but as the random variable X does not have distribution necessarily after the transformation (with the sample mean and variance, which are additional random variables), (\*) no longer follows. Therefore, the arguments for the proof in the lecture cannot be used to show the same for the adjusted Kolmogorov-Smirnov test.

ii) The adjusted KS statistic applies a transformation to the data, which is assumed to be normally distributed with location (expectation) and scale (variance) . By applying the transformation (inside the formula), the location and scale parameters of the data are corrected to a location of 0 and scale of 1 regardless of the values of the parameters. This also ensures that the standard normal cdf can be used. Therefore, this adjusted KS test statistic is independent of the location and scale parameters of the data.

b)

**1879 dataset**

Denote the random variable (and part of the random sample) : the *i*th measurement error, measured by Michelson in 1879. Denote also : the number of measurements in 1879. According to the dataset: . As such: .

Denote this random sample as sampled from the unknown distribution .

Informally:

*is a normal distribution*

*is not a normal distribution*

More formally:

The (adjusted Kolmogorov-Smirnov) **test statistic** is

We need to simulate under .

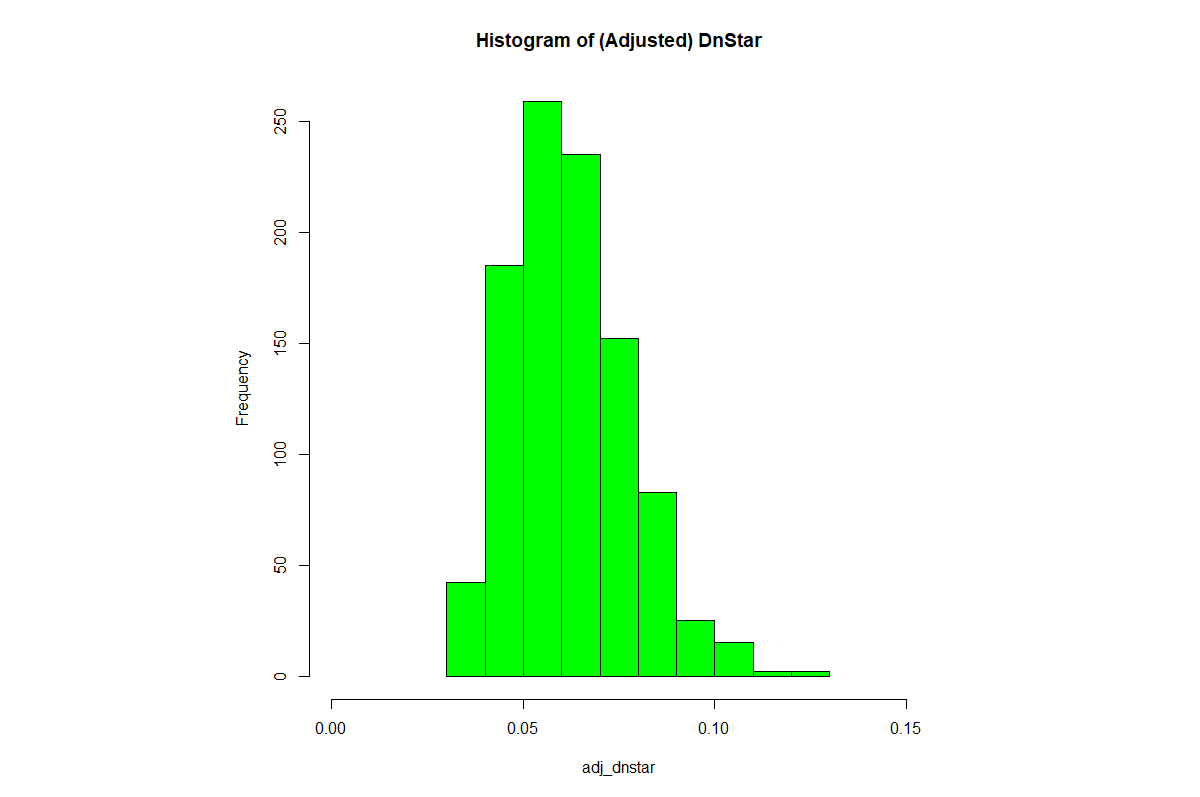
Since is independent of the location and scale parameters of the data, we can use the parametric bootstrap with random samples from the standard normal distribution.

For each =:

We generate , (i.i.d) from .

We generate with , for .

We take .



The test score for the sample from 1879 is 0.083. This gives an **p-value** of 0.091 for the (right-tailed adjusted bootstrapped) KS test (see the R code).

**Conclusion:** Since this p-value is greater than the significance level, we do not reject the null hypothesis. Therefore we conclude that the measurement errors from 1879 could possibly be normally distributed.

**1882 dataset**

NB: For the sake of brevity, please see the above steps again (until the histogram). The only change is that the dataset is now the set of measurement errors from 1882 (instead of the dataset for 1879), note that therefore this time around.

The test score for the sample from 1882 is 0.145. Since this exceeds all the values, the corresponding (right-tailed) **p-value** is 0.000 (for the right-tailed adjusted bootstrapped KS test).

**Conclusion:** Since this p-value is lower than the significance level, we reject the null hypothesis. Therefore we do not have any reasoning to say that the measurement errors from 1879 could possibly be normally distributed.

c)

Yes, the p-values obtained directly from the output of ks.test (when one uses as input for par: the estimated mean and standard deviation) are much higher than the p-values in part b. This has multiple reasons:

1. The p-values are wrongly estimated, ks.test assumes the parameters are not derived from the sample, as this would affect the sampling distribution of . Moreover, in that case, the changes from simple, to composite.
2. The standard way of calculating ks.test in R is done with a two-sided alternative hypothesis, whilst our test in part b was right-tailed. An right-tailed test would have provided (somewhat) lower p-values. However, reason 1 is of a larger influence on the p-values.