DS-GA 1002 - Homework 9

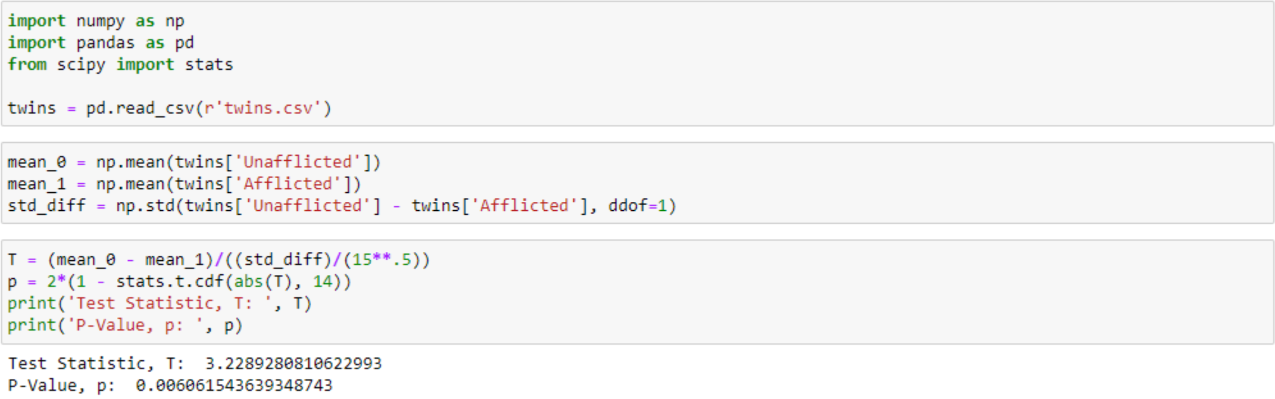
Eric Niblock

November 7th, 2020

1. **A twins study is performed on 15 sets of monozygotic twins where one twin is afflicted with schizophrenia and the other is not. For each twin, a measurment is taken of the volume of their left hippocampus using an MRI machine. The goal is to determine if there is a link between the affliction and left hippocampal volume.**

   (a) **Perform a 2-sided t-test on the data in *twins.csv* to see if there is a significant difference in left hippocampal volume. Report your p-value.**

   Below is the code we used to preform the 2-sided t-test on the twins data,



```
import numpy as np
import pandas as pd
from scipy import stats

twins = pd.read_csv(r'twins.csv')
```

```
mean_0 = np.mean(twins['Unafflicted'])
mean_1 = np.mean(twins['Afflicted'])
std_diff = np.std(twins['Unafflicted'] - twins['Afflicted'], ddof=1)
```

```
T = (mean_0 - mean_1)/((std_diff)/(15**.5))
p = 2*(1 - stats.t.cdf(abs(T), 14))
print('Test Statistic, T: ', T)
print('P-Value, p: ', p)
```

```
Test Statistic, T:  3.2289280810622993
P-Value, p:  0.006061543639348743
```

   As you can see, the resulting test statistic and p-value have been reported in the print statement.

   (b) **For the test you just performed:**

   i. **Give a model for the data.**

   If we call $X_1, ..., X_{15}$ the measurements of hippocampal volume concerning the unafflicted twins, and $Y_1, ..., Y_{15}$ the measurements of hippocampal volume concerning the afflicted twins. We assume these data are approximately normally distributed. Furthermore, it makes sense to use the paired t-test, since it is likely that the hippocampal volume of identical twins is not independent.

We therefore use the following test statistic,

$$T = \frac{\sum_{i=1}^{n} X_i - Y_i}{\frac{s_{X-Y}}{\sqrt{n}}} \tag{1}$$

Where $n = 15$, and s_{X-Y} is the sample standard deviation of the differences.

ii. State the null hypothesis.

The null hypothesis is,

$$\Theta_0 = \{(\mu_X, \mu_Y) : \mu_X - \mu_Y = 0\}$$

iii. State the alternative hypothesis.

The alternative hypothesis is,

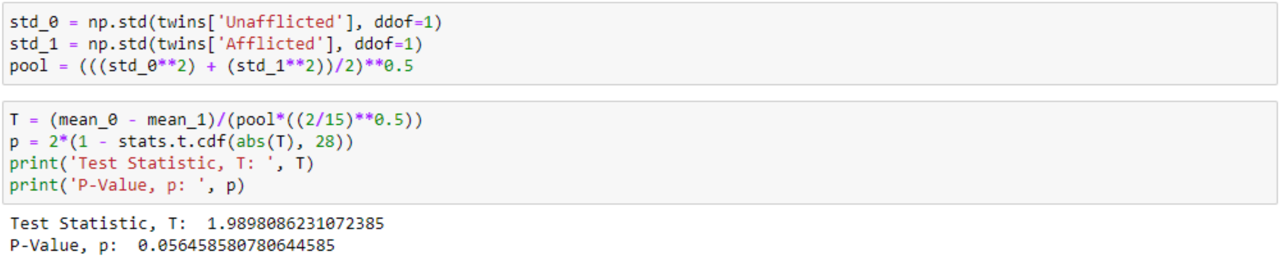
$$\Theta_1 = \{(\mu_X, \mu_Y) : \mu_X - \mu_Y \neq 0\}$$

iv. Do you reject at the 5% level?

Given the p-value we should reject the null hypothesis since it falls below the significance level.

(c) Suppose the 30 people had no familial relationships (i.e., just 15 afflicted and 15 unafflicted, but no twins). Perform a 2-sided t-test under these circumstances and report your p-value. [Note: Use the pooled version.]

Assuming that the variance between the two samples is approximately equal, we generated a pooled variance for our 2-sided t-test done below,



```
std_0 = np.std(twins['Unafflicted'], ddof=1)
std_1 = np.std(twins['Afflicted'], ddof=1)
pool = (((std_0**2) + (std_1**2))/2)**0.5
```

```
T = (mean_0 - mean_1)/(pool*((2/15)**0.5))
p = 2*(1 - stats.t.cdf(abs(T), 28))
print('Test Statistic, T: ', T)
print('P-Value, p: ', p)
```

```
Test Statistic, T:  1.9898086231072385
P-Value, p:  0.056458580780644585
```

As you can see, the resulting test statistic and p-value have been reported in the print statement.

Since we no longer can conclude that the data is paired, we no longer simply use the differences when calculating the standard error. Instead, we use the pooled method,

$$T = \frac{\sum_{i=1}^{n} X_i - Y_i}{s_d\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} \tag{2}$$

$$s_d = \frac{(N_X - 1)s_X^2 + (N_Y - 1)s_Y^2}{N_X + N_Y - 2} \tag{3}$$

Where we are again working with the same null and alternative hypotheses.

(d) Suppose you want to conclude that people with schizophrenia tend to have smaller left hippocampal volume. What assumptions must be made to draw this conclusion?

We would need to make a plethora of assumptions. We would first need to assume that our sample population is representative of the population of schizophrenics. We would also need to assume that the distribution of sample mean differences does not include zero.

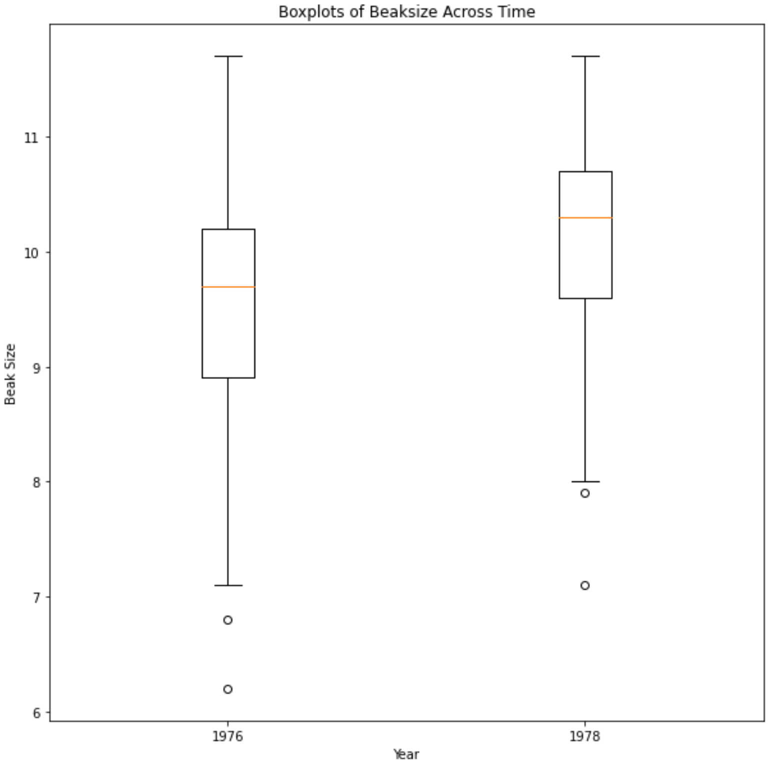
(e) Give a short explanation why the data does not give evidence that differences in left hippocampal volume cause schizophrenia.

No scientific experiment has been preformed here, and causation cannot be presumed. It appears that there is a correlation between left hippocampal volume and schizophrenia, however this may mean that differences in left hippocampal volume causes schizophrenia, schizophrenia causes differences in left hippocampal volume, some third variable causes both differences in left hippocampal volume and schizophrenia, or there may be no causal link between them at all!

2. **In 1977 a severe drought caused vegetation to whither on one of the Galapagos islands. The remaining food sources were harder to eat. Biologists wanted to test if the remaining finches on the island had larger beaks – a demonstration of natural selection. In *finches.csv* you have beak depth data from 178 finches. Of those 178 measurements, 89 are from finches selected before the drought, and 89 are from finches selected after the drought.**

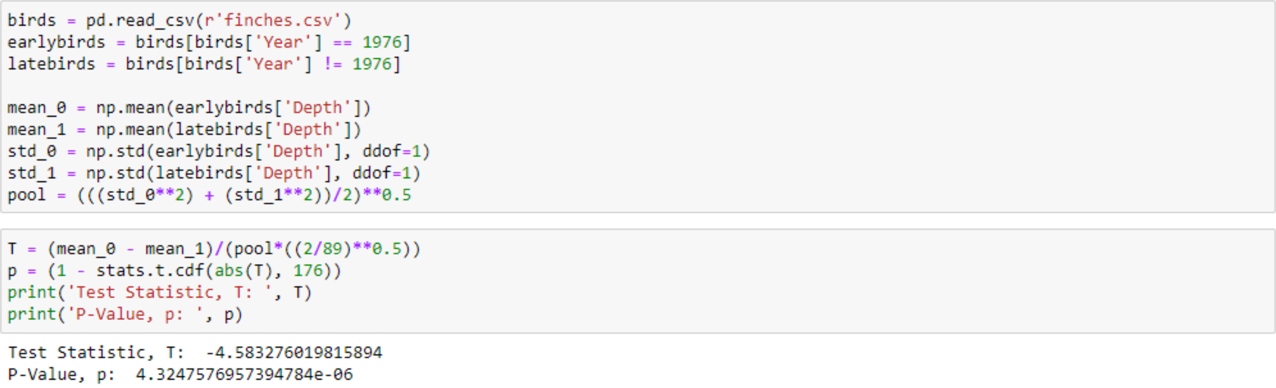
   (a) **Give a box-plot of the data separated by year.**

   The following is a boxplot of the beak-size data for the two years in question,



   (b) **Perform a one-sided t-test to compare the beak depths. [Note: Use the pooled version.]**

   The following is the code was used to generate the one-sided t-test and associated test statistic and p-value,



```
birds = pd.read_csv(r'finches.csv')
earlybirds = birds[birds['Year'] == 1976]
latebirds = birds[birds['Year'] != 1976]

mean_0 = np.mean(earlybirds['Depth'])
mean_1 = np.mean(latebirds['Depth'])
std_0 = np.std(earlybirds['Depth'], ddof=1)
std_1 = np.std(latebirds['Depth'], ddof=1)
pool = (((std_0**2) + (std_1**2))/2)**0.5
```

```
T = (mean_0 - mean_1)/(pool*((2/89)**0.5))
p = (1 - stats.t.cdf(abs(T), 176))
print('Test Statistic, T: ', T)
print('P-Value, p: ', p)
```

```
Test Statistic, T:  -4.583276019815894
P-Value, p:  4.3247576957394784e-06
```

i. Give a model for the data.

If we note the beak lengths of the finches from 1976 as $X_1, ..., X_{89}$ and the beak lengths of the finches from 1978 as $Y_1, ..., Y_{89}$, we then can construct a one-sided t-test. We assume $X_1, ..., X_{89}$ and $Y_1, ..., Y_{89}$ are approximately normally distributed. Therefore the differences, $X_1 - Y_1, ..., X_{89} - Y_{89}$ will be approximately normally distributed as well.

Again we use the following test statistic,

$$T = \frac{\sum_{i=1}^{n} X_i - Y_i}{s_d\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} \tag{4}$$

$$s_d = \frac{(N_X - 1)s_X^2 + (N_Y - 1)s_Y^2}{N_X + N_Y - 2} \tag{5}$$

ii. State the null hypothesis.

The null hypothesis is,

$$\Theta_0 = \{(\mu_X, \mu_Y) : \mu_X - \mu_Y \geq 0\}$$

iii. State the alternative hypothesis.

The alternative hypothesis is,

$$\Theta_1 = \{(\mu_X, \mu_Y) : \mu_X - \mu_Y < 0\}$$

iv. **Do you reject at the 5% level?**

We do reject the null hypothesis at the 5% level, since our p-value is less than 0.05.

(c) **Propose a reason why the beak depth data from the finches before and after the drought may not be independent. [Hint: The lifespan of a finch is typically 5-10 years.]**

It is entirely possible that our two groups of finches are not independent. There may be finches from 1976 that appear again in our sample from 1978. Obviously using the same finch twice would introduce dependency.

3. Suppose $X_1, ..., X_n \overset{iid}{\sim} F$, where F is the CDF of a discrete distribution on the values $1, ..., k$ assigning probability p_i to the value i. That is, if $X \sim F$ then $P(X = i) = p_i$ for $i = 1, ..., k$. We want to test if $(p_1, ..., p_k) \neq (\theta_1, ..., \theta_k)$ for given values $\theta_1, ..., \theta_k \in [0, 1]$ with $\sum_{i=1}^{k} \theta_i = 1$. If we assume that $(p_1, ..., p_k) = (\theta_1, ...\theta_k)$ and n is large then the test statistic

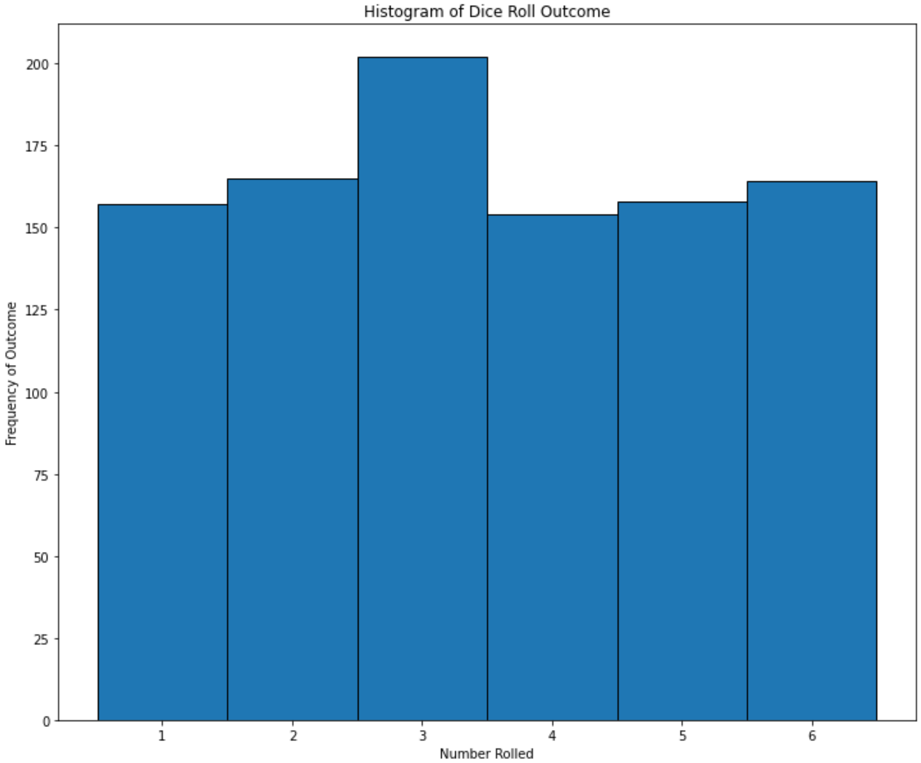
$$T = \sum_{i=1}^{k} \frac{(n_i - n\theta_i)^2}{n\theta_i} \tag{6}$$

approximately has a χ^2 (chi-squared) distribution with $k - 1$ degrees of freedom, where n_i is the number of data points taking the value i:

$$n_i = \sum_{j=1}^{n} I_{\{i\}}(X_j) \tag{7}$$

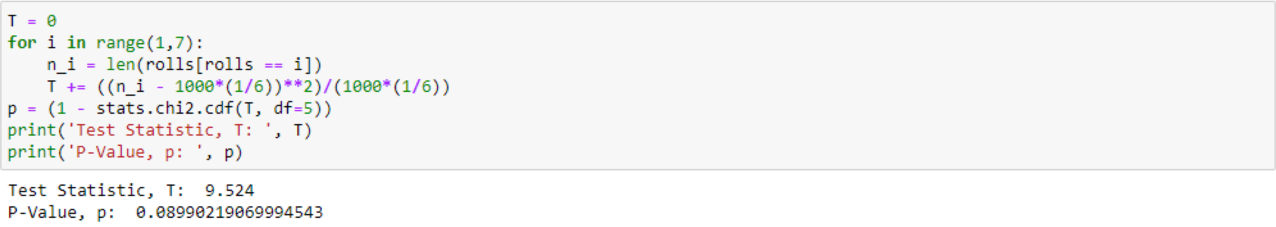
In *rolls.txt* you have data from 1000 rolls of a 6-sided die.

(a) Give a histogram of the data. [Hint: Use edgecolor='black' and set bins appropriately.]



(b) Test whether the die is not fair, and report your p-value.

In order to evaluate whether the dice was fair, we calculated the approximately χ^2 (chi-squared) test statistic, and used the CDF to calculate the p-value



```
T = 0
for i in range(1,7):
    n_i = len(rolls[rolls == i])
    T += ((n_i - 1000*(1/6))**2)/(1000*(1/6))
p = (1 - stats.chi2.cdf(T, df=5))
print('Test Statistic, T: ', T)
print('P-Value, p: ', p)
```

```
Test Statistic, T:  9.524
P-Value, p:  0.08990219069994543
```

If we use a significance level of $\alpha = 0.05$, we can not conclude that the dice is rigged.

4. **In this problem we will investigate a flawed testing paradigm using simulations.**

   You work at a firm that conducts many experiments with their website. To monitor ongoing experiments, there is a webpage (called a dashboard) that contains the number of samples you have collected thus far, and the current value of the test statistic. Pressured for time, you constantly refresh the dashboard, and decide to stop the experiment and reject the moment it passes your threshold, or fail to reject otherwise. Below we examine this type of testing procedure.

   Suppose our data obeys the model $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ with $\mu \in \mathbb{R}$ unknown. We want to test if $\mu > 0$.

   (a) **Determine the smallest n and the corresponding rejection region such that the following criteria are simultaneously satisfied by a one-sided test on the full data $X_1, ..., X_n$.**

   - **The test is significant at the 1% level.**
   - **The chance of rejection is at least 80% assuming $\mu = 0.5$.**

   [Hint: First solve for τ in terms of n, then find n. If you are worried about your calculation, you can check using a simulation.]

   For the test to be significant at the 1% level, we must have,

$$P(\bar{X}_n > \gamma) = 0.01 \tag{8}$$

For some threshold γ, which represents the lower end of the rejection region. Then,

$$P\left(\frac{\bar{X}_n - \mu_0}{\frac{1}{\sqrt{n}}} > \frac{\gamma - \mu_0}{\frac{1}{\sqrt{n}}}\right) = 0.01 \tag{9}$$

And,

$$1 - \Phi\left(\frac{\gamma - \mu_0}{\frac{1}{\sqrt{n}}}\right) = 0.01 \tag{10}$$

$$\frac{\gamma - \mu_0}{\frac{1}{\sqrt{n}}} = 2.326 \tag{11}$$

$$\gamma = \frac{2.326}{\sqrt{n}} \tag{12}$$

Then, for the chance of rejection to be at least 80% assuming $\mu = 0.5$, we have,

$$P(\bar{X}_n > \gamma; \mu = 0.5) \geq 0.80 \tag{13}$$

$$P\left(\frac{\bar{X}_n - 0.5}{\frac{1}{\sqrt{n}}} > \frac{\gamma - 0.5}{\frac{1}{\sqrt{n}}}\right) \geq 0.80 \tag{14}$$

$$1 - \Phi\left(\frac{\gamma - 0.5}{\frac{1}{\sqrt{n}}}\right) \geq 0.80 \tag{15}$$

$$\gamma \leq -\frac{0.842}{\sqrt{n}} + 0.5 \tag{16}$$

Combining both conditions yields,

$$n \geq 40.145 \tag{17}$$

$$n \approx 41 \tag{18}$$

(b) Assuming the null hypothesis, generate 2000 datasets consisting of i.i.d. draws of size n (the n you computed in the previous part) from a $\mathcal{N}(0,1)$ distribution. For $i = 1, ..., 2000$, do the following:

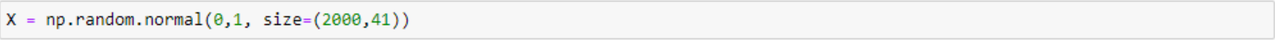
i. Consider the i-th generated dataset $X_{i,1}, ..., X_{i,n}$.

ii. For $k = 15, ..., n$ treat $X_{i,1}, ..., X_{i,k}$ as the full dataset, and determine if you reject the null hypothesis at the 1% level (i.e., perform a hypothesis test for $\mu > 0$ at the 1% level with a sample size of k).

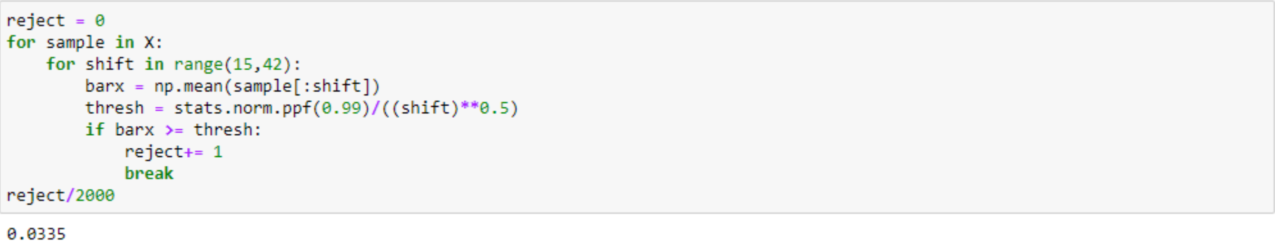
iii. Declare that you reject the null hypothesis if you reject for at least one $k \in \{15, ..., n\}$ in the previous part.

What proportion of your simulated datasets are rejected by the above procedure?

The above procedure was preformed using Python, and the code is shown below,



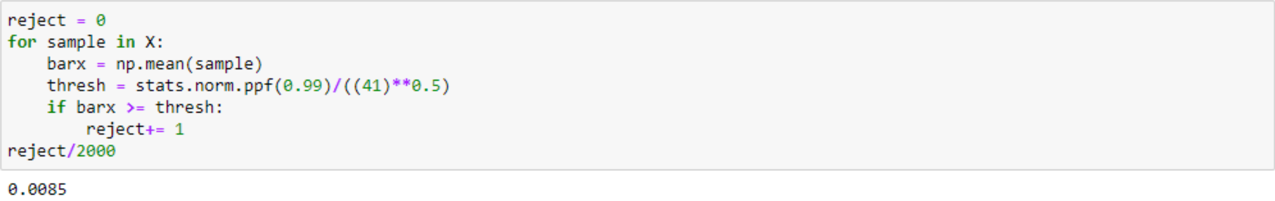
```
X = np.random.normal(0,1, size=(2000,41))
```



```
reject = 0
for sample in X:
    for shift in range(15,42):
        barx = np.mean(sample[:shift])
        thresh = stats.norm.ppf(0.99)/((shift)**0.5)
        if barx >= thresh:
            reject+= 1
            break
reject/2000
```

```
0.0335
```

Where the printed result represents the proportion of datasets which would result in the rejection of the null hypothesis. Below, we contrast this with the number of datasets which would be rejected using the standard procedure (preforming the test once per dataset),



```
reject = 0
for sample in X:
    barx = np.mean(sample)
    thresh = stats.norm.ppf(0.99)/((41)**0.5)
    if barx >= thresh:
        reject+= 1
reject/2000
```

```
0.0085
```

This is expected, because the result is close to a 1% rejection rate, whereas the other result yields a rate which is over three times greater.