Homework 6

5.1.

a. the test rejects the null hypothesis at the 5% significance level for all values (h=1).

h1 = 1×100

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

b. the test rejects the null hypothesis at the 5% significance level only for some values and doesn't reject the null hypothesis at the 5% significance level for the rest (h=0).

h2 = 1×100

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0

c. since the sizes or N values differ for the datasets we cant run full ttests on them.

```
Error using ttest2 (line 156)
The data in a 2-sample t-test must be commensurate.
```

For the first 150 values both the mean 0.2 test and the mean 2 test give similar values and only reject the null hypothesis at the 5% significance level for some values.

h4 = 1×150

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	1	1	1	1	1	1	0	1	1	0	1	0	1	0	0	

2. To compute the table 5.1 we start by forming a normal distribution in the range (0,1) (we can use rnd matlab function for this, its range is (0,1) by default), and set as xp. To calculate p we first find the sample mean using the formula: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. After that we calculate p using the formula, $p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{N(\bar{x}-\hat{\mu})^2}{\sigma^2}\right)$

Table 5.1 Acceptance Intervals $[-x_\rho, x_\rho]$ Corresponding to Various Probabilities for an $\mathcal{N}(0, 1)$ Normal Distribution

$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.282	1.440	1.645	1.967	2.326	2.576	3.090	3.291

3. a. Kernel can be used as a similarity measure in the affinity matrix, and we use it for clustering.

b. we calculate an affinity matrix by calculating the thresholded cosine similarity between documents in our sample space.

c. The data presented for classification comes from treating each row of a matrix X (formed from the k-largest eigenvectors of N from our INPUT data and normalized to unit length) and clustering them using clustering algorithms, we label this points for classification