Exercise 1

- H1 The probability that a student gets question 1 right is 50%
- H2 The average total that students obtain is less than 50 points.
- H3 BLA students get better exam results than CRU students
- H4 Question 5 is more difficult than question 8

A. For each of the four hypotheses:

Determine whether it is a one sample performance hypothesis, or a two-sample comparative hypothesis

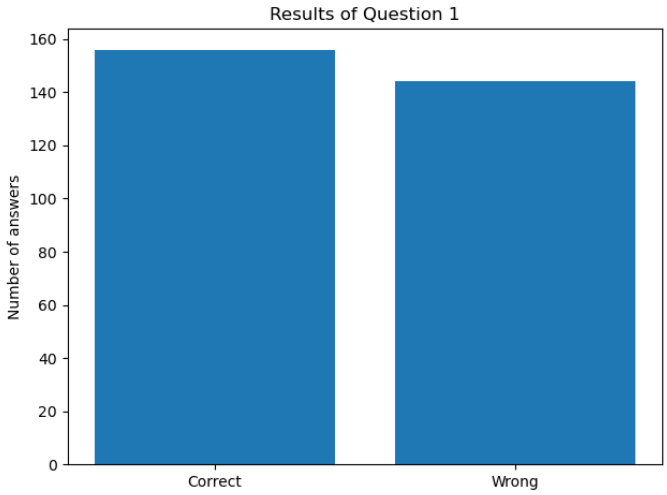
H1: one-sample: it's a hypothesis about performance.

H2: one-sample: it's a hypothesis about performance.

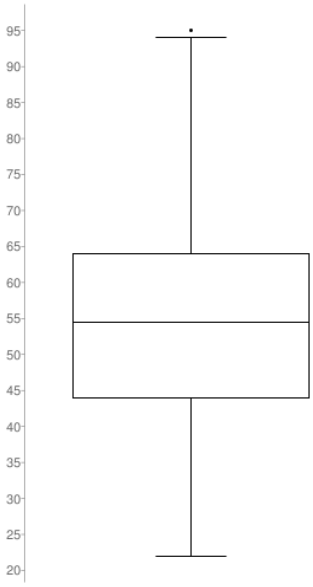
H3: two-sample: the hypothesis makes a comparison between BLA students and CRU students.

H4: two-sample: the hypothesis makes a comparison between two questions.

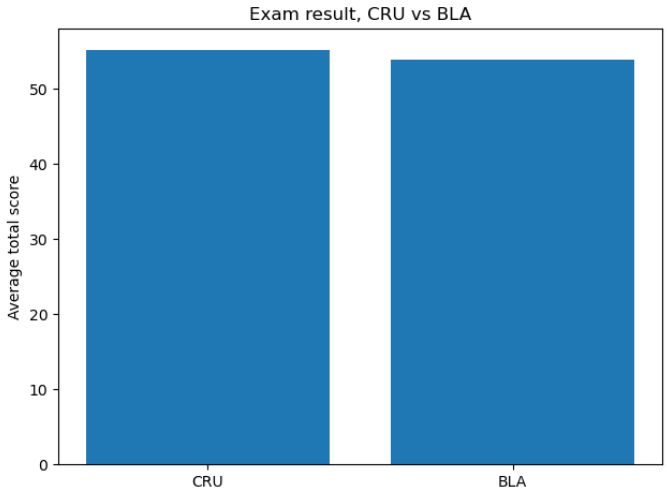
Perform a preliminary investigation of the plausibility of the hypothesis using descriptive analysis tools (boxplots, histograms, biplots, . . .)



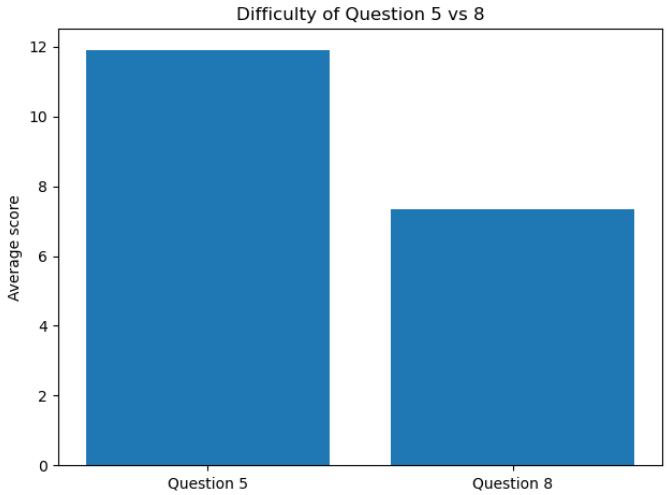
H1: Based on the bar diagram above - it seems like the hypothesis is plausible. The probability that a student gets question 1 right is around 50%.



H2: The box plot above is based on the total score of all students. We can see that the median score is higher than the 50 points which the hypothesis claims that students have less than, so it seems like it is not plausible.



H3: Based on the bar diagram above - the hypothesis does not seem to be plausible. The above diagram is based on the total average score of each set of students, and in this case, they are pretty even. And if anything, based on the data, CRU gets a slightly higher score on average.



H4: Based on the bar diagram above - the hypothesis seems to not be plausible at all. Student, on average, gets a much higher score in question 5 then question 8, which leads to the assumption that the harder question of the two is question 8.

B. For H1 and H2

Notes:

- Binomial Test (17-24)
  - A binomial test examines if some population proportion is likely to be x.
  - parameters: N = number of cases, p = probability of the desired outcome
    - Ex: N = 20, p = 0.5:
  - You scale the number of test cases up until you with the equation get a very small p-value. At that point you know you have a reasonable about of data to tests the hypothesis.
  - Set at what value you will reject the hypothesis before starting the data analysis: $\leq \alpha$
  - One-sided: "upper/lower bound" hypothesis, such as: $p \leq 0.5$
  - Two-sided: "point" hypothesis, such as: $p = 0.5$
- t-Test (25-31)
  - Use to compare two given samples
  - A t-test is used when the population parameters (mean and standard deviation) are not known.

Identify a statistical test that is suitable to test the hypothesis

H1: I would like to apply a Binomial test for this hypothesis. This test has the purpose of estimating the probability of a specific outcome, which in this case is, that the chance of a student answering question one correctly is 50%.

H2: I would like to apply a t-Test for this hypothesis.

Apply the test to the vitdata and determine whether the hypothesis is rejected at a significance level of 0.01.

H1: Using the calculator at link, and entering the following values;

- n, which is the number of cases, is it to 300.
- K, which is the number of occasions the actual outcome occured, which in this case is the number of times the answer to the question was correct, is 156.
- p, being the probability the outcome will occur on any particular occasion, is initially set to $50\% = 0.5$. This value is only correct if we assume that every student answers at random.

n	K	*p*	*q*
300	156	0.5	0.5

The probability of exactly 156 (K) out of 300 (*n*) is *p* = .03623314.

The probability of exactly, or fewer than, 156 (K) out of 300 (*n*) is *p* = .77351214.

The probability of exactly, or more than, 156 (K) out of 300 (*n*) is *p* = .262721.

Using the before mentioned numbers we get the result at the above picture. The probability values renders the data unusable. If we re-evaluate the entered values, we find that the value p is only accurate if each student answers completely random. If we assume that most students has studied for the test, and that each student will make a qualified answer, we can increase the

chance of a correct answer, the value p. If we increase the value to $60\% = 0.6$, we get the following result:

n	K	*p*	*q*
300	156	0.6	0.3

The probability of exactly 156 (K) out of 300 (*n*) is *p* < .00000001.

The probability of exactly, or fewer than, 156 (K) out of 300 (*n*) is *p* < .00000001.

The probability of exactly, or more than, 156 (K) out of 300 (*n*) is *p* > .99999999.

This result makes us not reject the hypothesis because the significance level is $p < 0.01$

H2: I use the calculator at link. The *population mean* is the 50 points in the hypothesis and *sample X* is each students total score.

Single Sample T-Test Calculator

The value of *t* is 5.501147.

Population mean (μ)
50

Sample X
48,54,42,66,71,39,66,58,61,59,45,35,41,46,38,58,80,51,52,35,53,55,36,48,53,56,48,62,55,56,43,63,62,50,50,54,70,39,50,62,60,71,68,47,72,42,73,74,43,65,65,52,54,95,40,35,53,34,69,53,44,56,62,48,29,67,73,60,25,35,59,43,62,61,48,42,60,55,91,61,53,49,29,70,43,71,48,45,57,31,53,55,61,27,50,38,64,44,51,56,53,69,45,52,80,84,46,79,45,57,37,32,58,57,82,59,59,66,40,64,57,50,71,48,48,40,37,59,58,67,34,51,27,27,35,55,55,37,55,47,55,58,74,64,52,70,64,35,70,49,25,48,34,67,46,66,41,64,57,93,63,56,42,53,56,47,66,52,74,55,65,37,57,71,68,33,62,67,34,28,40,62,56,59,38,54,58,60,64,29,53,46,33,38,61,45,50,40,81,37,36,62,64,66,94,81,71,59,81,75,46,45,72,43,53,69,87,67,75,54,47,59,68,83,76,36,61,61,88,80,22,54,37,30,58,55,48,50,55,46,85,32,49,51,41,50,71,80,45,42,67,68,50,31,61,42,63,52,63,60,58,38,54,64,62,48,35,49,40,53,40,70,42,63,42,73,52,44,60,30,61,68,41,64,75,42,37,34,41,59,65,75,51,54,37,76,85,69,55,52

Significance Level:
- (•) 0.01
- ( ) 0.05
- ( ) 0.10

One-tailed or two-tailed hypothesis?:
- (•) One-tailed
- ( ) Two-tailed

The *t*-value is 5.501147. The value of *p* is < .00001. The result is significant at *p* < .01.

The significance level is at 0.01 and the hypothesis is therefore rejected.

Exercise 2

C. For hypotheses H3 and H4 from Exercise 1:

- H3 BLA students get better exam results than CRU students
- H4 Question 5 is more difficult than question 8

Identify a statistical test that is suitable to test the hypothesis

H3: I would apply a T-Test to test the probability that BLA students gets better results than CRU students.

H4: I would apply a Binomial test.

Apply the test to the vitdata and determine whether the hypothesis is rejected at a significance level of 0.01.

H3: Using the calculator at link, and entering the following values:

Treatment 1 (*X*)

```
48,71,66,58,61,45,38,58,80,36,48,53,56,62,55,50,50,54,50,62,60,68,47,42,74,43,65,95,35,69,53,67,2
5,59,61,42,55,61,53,29,70,43,45,57,53,55,61,56,80,84,46,79,57,37,58,82,66,40,57,48,48,40,37,58,34
,27,55,37,74,35,70,25,67,66,57,93,56,42,53,56,47,66,74,65,37,33,67,34,40,62,59,54,58,46,61,50,81,
37,62,94,69,87,54,59,68,76,36,61,61,88,80,22,54,48,46,32,51,41,71,80,67,50,42,63,52,63,60,58,54,6
4,62,35,49,40,42,63,44,30,61,75,42,37,34,65,37,55
```

Treatment 2 (*X*)

```
54,42,66,39,59,35,41,46,51,52,35,53,55,48,56,43,63,62,70,39,71,72,73,65,52,54,40,53,34,44,56,62,4
8,29,73,60,35,43,62,48,60,91,49,71,48,31,27,50,38,64,44,51,53,69,45,52,45,32,57,59,59,64,50,71,59
,67,51,27,35,55,55,47,55,58,64,52,70,64,49,48,34,46,41,64,63,52,55,57,71,68,62,28,56,38,60,64,29,
53,33,38,45,40,36,64,66,81,71,59,81,75,46,45,72,43,53,67,75,47,83,37,30,58,55,50,55,85,49,50,45,4
2,68,31,61,38,48,40,53,70,42,73,52,60,68,41,64,41,59,75,51,54,76,85,69,52
```

Treatment 1 is the total scores of all CRU students, and Treatment 2 is the total scores of all BLA student.

The result of the calculation is:

> *The t-value is 0.77286. The p-value is .440218. The result is not significant at p < .01.*

Based on the **p** value, I would not reject the hypothesis.

H4: Using the calculator at link, and entering the following values;

- n, which is the number of cases, is it to 300.
- K, which is the number of occasions the actual outcome occurred, which in this case is the number of times a student scores less points in question 5 compared to question 8. This happened 77 times.
- p, being the probability the outcome will occur on any particular occasion, is set to $50\% = 0.5$. This value is only correct if we assume that each question is equality hard.

n	K	*p*	*q*
300	77	0.5	0.5

The probability of exactly 77 (K) out of 300 (*n*) is $p < .00000001$.

The probability of exactly, or fewer than, 77 (K) out of 300 (*n*) is $p < .00000001$.

The probability of exactly, or more than, 77 (K) out of 300 (*n*) is $p > .99999999$.

Using the above numbers results in a significance level below 0.01 and the hypothesis is therefore not rejected.

D. Define yourself 4 new hypotheses about the students and exam questions in vitdata, such that you have one hypothesis each with the following characteristics:

Define 4 hypotheses

One sample, and can be tested with Binomial test:

- M1 The probability that a student gets question 1-4 right is 50%

One sample, and can be tested with the one sample t-test

- M2 The average score of question 5-8 is more than 15 points

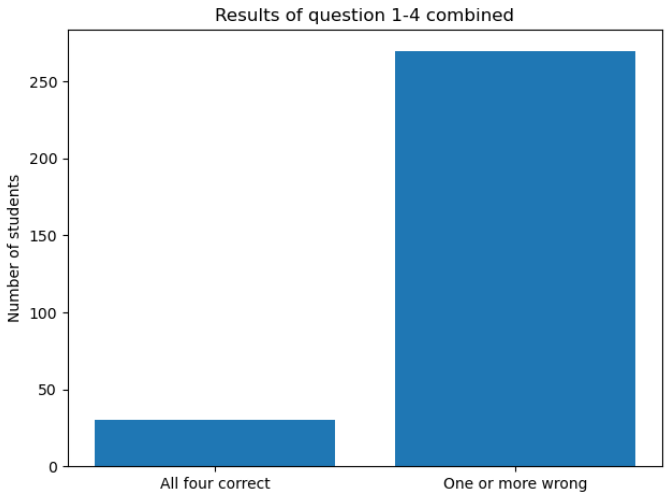
two sample, and can be tested with the two sample t-test

- M3 The average score of question 5 and 6 is higher than the average score of question 7 and 8

two sample, and can be tested with the Wilcoxon test

- M4 Given 100 students from each program which has the highest total score, CRU will have the highest average score.

For each hypothesis, perform a preliminary analysis using descriptive statistics, and based on this analysis, decide whether you think the hypothesis should be A: not rejected, B: rejected, or C: very clearly rejected (with a p-value < 0.00001)



M1: Based on the above bar chart I think the hypothesis should be: C: very clearly rejected.

M2: If we for each student take the average of their scores on question 5-8, and calculate the average of all this for all students combined, we get the value: **10**. The value is significant lower then 15, so I think it should be: B: rejected.

M3: From the given vitdata we can calculate the following:

- The average score for question 5 and 6 is **10.985**
- The average score for question 7 and 8 is **9.053333333333333**

The average score for question 5 and 6 is higher, but the difference is too small to make any conclusions. I would think that B: the hypothesis would be reject.

M4: From the given vitdata we can calculate the following:

- The average total score of top 100 students from CRU is **47.41**
- The average total score of top 100 students from BLA is **45.94**

Based on these calculations I think the hypothesis will be: B: rejected.

Apply a suitable statistical test to your hypotheses, and check whether the result of the test corresponds to your expectation.

M1:

n	K	*p*	*q*
300	30	0.063	0.937

The probability of exactly 30 (K) out of 300 (*n*) is *p* = .00387647.

The probability of exactly, or fewer than, 30 (K) out of 300 (*n*) is *p* = .99501148.

The probability of exactly, or more than, 30 (K) out of 300 (*n*) is *p* = .00886499.

Based on the above values we should reject as my expectations were.

M2:

Population mean (μ)

15

Sample X

```
8.5,8.25,8.75,11.25,10.75,8.0,11.25,11.0,10.0,11.25,9.5,7.0,8.5,6.25,7.75,11.0,13.0,7.5,9.5,8.75,9
.75,10.25,7.25,10.25,11.5,8.75,12.0,12.0,10.25,8.75,7.25,14.0,12.0,7.25,10.75,13.5,12.25,8.0,7.25,
12.0,9.75,12.5,10.0,6.5,12.75,7.0,14.75,11.5,5.5,12.75,11.0,9.5,8.25,16.75,8.25,7.0,9.75,8.5,12.0,
9.75,9.25,8.75,12.0,8.5,7.25,9.75,11.25,9.75,4.5,7.0,11.25,7.25,12.0,13.5,10.25,7.0,9.75,6.75,19.2
5,11.75,11.5,8.75,5.5,12.25,7.25,12.5,10.25,11.25,12.5,6.0,9.75,6.75,10.0,6.75,9.0,6.0,12.5,7.5,11
.0,8.75,9.75,12.0,11.25,9.5,13.0,15.75,8.0,14.5,9.5,10.75,7.5,8.0,11.0,9.0,15.25,9.5,9.5,9.5,8.25,
10.75,10.75,10.75,14.25,10.25,10.25,6.5,5.75,7.75,9.25,15.0,6.75,7.5,6.75,5.0,5.25,10.25,8.5,9.25,
8.5,8.25,10.25,9.25,13.25,10.75,11.25,12.25,12.5,7.0,10.5,8.75,6.25,10.25,6.75,9.75,8.0,9.5,5.0,10
.75,12.5,16.25,12.25,8.75,5.25,8.0,10.5,8.25,11.25,9.5,15.0,10.25,14.5,7.5,9.0,12.5,11.75,8.25,12.
0,11.5,8.5,7.0,8.25,13.75,12.25,9.5,6.0,8.25,7.5,11.5,12.5,5.5,9.75,9.75,8.25,7.75,13.5,11.25,9.0,
10.0,13.25,5.75,9.0,12.0,14.25,11.25,16.5,15.0,12.5,9.5,13.25,17.0,9.75,7.75,12.75,7.25,13.25,10.2
5,14.75,11.5,11.75,6.5,8.25,11.25,11.75,13.75,13.75,9.0,11.75,10.0,18.5,16.5,5.5,8.25,7.5,5.75,12.
75,10.25,6.75,9.0,8.5,8.0,14.25,6.25,10.5,9.25,6.75,10.75,14.25,16.5,6.0,7.0,11.5,11.75,9.0,6.0,11
.75,7.0,10.5,11.25,10.5,9.75,9.25,6.0,11.75,14.25,13.75,12.0,5.25,7.0,8.25,9.75,4.75,14.0,7.0,8.75
,7.0,11.25,9.5,9.25,11.5,7.5,13.5,13.5,8.5,10.75,13.5,7.0,5.75,6.75,8.5,11.25,12.75,11.75,11.0,8.2
5,7.5,12.0,16.0,12.0,10.25,11.25
```

With the above number the result is: $t - value = -31.65222$ and $p < .00001$. So the hypothesis should be rejected. This was my expectation.

M3:

Treatment 1 (*X*)

```
9.0,13.5,12.0,12.5,15.0,6.0,17.5,9.0,9.5,11.5,10.5,6.0,11.0,7.0,11.5,13.0,9.0,7.5,5.0,11.0,12.0,10
.0,10.0,18.0,10.5,4.5,20.0,15.0,12.5,5.5,5.5,11.0,15.5,13.5,11.5,17.5,17.0,13.0,6.5,12.5,10.0,7.5,
9.5,12.0,8.5,10.5,14.0,10.0,5.5,18.5,8.5,10.0,14.5,17.0,7.0,8.5,6.5,7.0,13.0,13.5,13.0,10.5,15.0,8
.5,11.5,6.5,6.5,8.5,3.5,11.0,13.5,7.0,6.5,16.5,8.0,9.0,15.0,9.5,20.0,10.0,10.0,6.5,8.0,10.0,8.5,12
.0,8.5,16.0,15.0,8.5,14.0,7.0,12.5,6.5,12.0,7.0,11.0,13.0,17.5,9.5,12.0,16.0,14.0,10.0,15.0,12.0,1
2.0,21.0,8.0,11.0,11.0,11.0,17.0,12.0,14.5,16.5,12.5,10.5,9.0,8.5,11.0,9.5,16.5,14.5,15.5,9.5,3.0,
5.5,9.0,13.0,4.5,7.5,6.5,7.0,5.0,8.5,11.0,8.5,11.5,11.0,11.5,13.0,15.0,20.0,14.0,9.5,11.0,11.5,13.
5,17.0,8.5,10.0,8.0,11.5,8.5,9.0,7.5,12.0,15.5,18.5,14.0,7.0,8.0,6.5,13.0,11.5,8.5,6.0,13.5,14.5,1
4.0,9.0,5.0,15.0,10.0,12.5,11.5,9.5,10.0,12.0,10.0,11.5,14.0,9.0,7.0,10.0,8.5,20.0,16.5,7.5,12.0,8
.5,7.0,8.5,10.5,14.5,11.5,11.5,7.5,9.0,8.0,14.0,18.0,4.5,21.0,15.5,6.5,9.5,15.5,15.5,14.5,3.5,14.5
,5.0,16.0,13.5,15.5,5.0,14.5,6.5,5.5,10.0,11.0,13.0,14.5,11.5,16.0,11.0,22.0,19.0,5.5,4.5,11.5,8.5
,17.0,9.0,3.0,10.0,8.5,10.5,11.5,9.0,9.0,7.5,5.5,8.0,13.0,20.0,6.0,8.0,10.5,13.5,10.0,4.0,13.0,7.5
,10.5,16.0,12.0,12.5,7.5,8.5,18.0,15.5,10.0,15.0,8.5,5.5,9.5,5.0,3.0,13.5,6.0,10.5,11.0,12.5,11.0,
6.5,14.5,9.0,17.0,12.0,8.0,11.0,13.0,10.5,4.5,6.0,7.5,16.0,7.0,14.0,17.5,11.0,11.5,11.5,20.0,11.5,
11.5,14.5
```

Treatment 2 (*X*)

```
8.0,3.0,5.5,10.0,6.5,10.0,5.0,13.0,10.5,11.0,8.5,8.0,6.0,5.5,4.0,9.0,17.0,7.5,14.0,6.5,7.5,10.5,4.
5,2.5,12.5,13.0,4.0,9.0,8.0,12.0,9.0,17.0,8.5,1.0,10.0,9.5,7.5,3.0,8.0,11.5,9.5,17.5,10.5,1.0,17.0
,3.5,15.5,13.0,5.5,7.0,13.5,9.0,2.0,16.5,9.5,5.5,13.0,10.0,11.0,6.0,5.5,7.0,9.0,8.5,3.0,13.0,16.0,
11.0,5.5,3.0,9.0,7.5,17.5,10.5,12.5,5.0,4.5,4.0,18.5,13.5,13.0,11.0,3.0,14.5,6.0,13.0,12.0,6.5,10.
0,3.5,5.5,6.5,7.5,7.0,6.0,5.0,14.0,2.0,4.5,8.0,7.5,8.0,8.5,9.0,11.0,19.5,4.0,8.0,11.0,10.5,4.0,5.0
,5.0,6.0,16.0,2.5,6.5,8.5,7.5,13.0,10.5,12.0,12.0,6.0,5.0,3.5,8.5,10.0,9.5,17.0,9.0,7.5,7.0,3.0,5.
5,12.0,6.0,10.0,5.5,5.5,9.0,5.5,11.5,1.5,8.5,15.0,14.0,2.5,7.5,0.5,4.0,10.5,5.5,8.0,7.5,10.0,2.5,9
.5,9.5,14.0,10.5,10.5,2.5,9.5,8.0,5.0,14.0,13.0,16.5,6.0,15.0,6.0,13.0,10.0,13.5,4.0,12.5,13.5,7.0
,2.0,6.5,16.0,10.5,10.0,5.0,6.5,6.5,3.0,8.5,3.5,7.5,11.0,9.5,7.0,16.5,8.0,6.5,8.5,19.0,2.5,10.0,10
.0,10.5,18.0,12.0,14.5,18.5,9.5,11.0,18.5,5.0,12.0,11.0,9.5,10.5,7.0,14.0,18.0,9.0,6.5,11.0,12.5,1
2.5,14.5,13.0,6.5,7.5,9.0,15.0,14.0,5.5,12.0,3.5,3.0,8.5,11.5,10.5,8.0,8.5,5.5,17.0,3.5,12.0,11.0,
8.0,13.5,15.5,13.0,6.0,6.0,12.5,10.0,8.0,8.0,10.5,6.5,10.5,6.5,9.0,7.0,11.0,3.5,5.5,13.0,17.5,9.0,
2.0,8.5,7.0,14.5,6.5,14.5,8.0,7.0,3.0,10.0,8.0,12.0,8.5,6.0,10.0,15.0,9.0,10.5,14.0,3.5,7.0,7.5,9.
5,6.5,18.5,9.5,4.5,5.5,3.5,12.5,12.0,12.5,9.0,8.0
```

Entering the above numbers into the T-Test calculator gives the following results:
$t - value = 5.94806$ and $p - value < .00001$. So the hypothesis should be rejected. This was my expectation.

M4:

Treatment 1

```
22, 25, 25, 27, 29, 30, 32, 33, 34, 34, 34,
35, 35, 35, 36, 36, 37, 37, 37, 37, 37, 37,
37, 38, 40, 40, 40, 40, 41, 42, 42, 42, 42,
42, 42, 43, 43, 44, 45, 45, 46, 46, 46, 47,
47, 48, 48, 48, 48, 48, 49, 50, 50, 50, 50,
50, 51, 52, 53, 53, 53, 53, 53, 54, 54, 54,
54, 54, 55, 55, 55, 55, 55, 56, 56, 56, 56,
57, 57, 57, 57, 58, 58, 58, 58, 58, 58, 59,
59, 59, 60, 60, 61, 61, 61, 61, 61, 61, 61
```

Treatment 2

```
27, 27, 28, 29, 29, 30, 31, 31, 32, 33,
34, 34, 35, 35, 35, 35, 36, 37, 38, 38,
38, 38, 39, 39, 40, 40, 40, 41, 41, 41,
41, 42, 42, 42, 43, 43, 43, 44, 44, 45,
45, 45, 45, 45, 46, 46, 46, 47, 47, 48,
48, 48, 48, 48, 48, 49, 49, 49, 50, 50,
50, 50, 51, 51, 51, 51, 52, 52, 52, 52,
52, 52, 52, 53, 53, 53, 53, 53, 53, 54,
54, 54, 55, 55, 55, 55, 55, 55, 55, 56,
56, 56, 57, 57, 58, 58, 59, 59, 59, 59
```

The above numbers results in the following: $z = -6.3334$, $p < .00001$ and $W = 331.5$. So the hypothesis should be rejected. This was my expectation.

Exercise 3

I have chosen a project which I am currently working on in my spare time. The idea came from playing the game *Escape from Tarkov* which contains a large number of quests that the player has to complete. Each quest has some objectives which are descripted using a mix of lore and dialog. This makes it very hard to understand the specifics required to complete the quest and often the player finds himself browsing the Wikipedia and other external sources to make sense of the in-game quest description. The process is time consuming and often leaves you with a lot of confusing notes and time wasted not playing the game.

Purpose of the system and performance measures

The purpose of the application is to track the players quest progression throughout the game, and display all objectives in each quest as specific as possible, with the purpose of making each quest as easy as possible to complete.

The best performance measure for measuring the quality of my system could be how many quests can be completed without using external resources. To be specific: given a user has completed all quests in the game - what percentage of quests were completed without using any other information than what is available in the game and in my application.

Would there be an alternative, existing solution to which you could compare your solution?

There is not any other applications which does this, but it could be compared to the Wikipedia for the game. So what percentage of the quests can be completed without using any other information then what is in the game.

How do you set up an experiment to collect empirical data for the comparison?

I would get two groups of people to play the game trying to complete each quest. One group would only use the information in the game, and the other group would use my application and the information provided by the game. I would then record how many quests each person in each group was able to complete with the given information.

State precisely a null hypothesis you would like to test. Which of the tests described in the lecture would be suitable for testing your hypothesis based on your data? If none of them seems applicable, in what respect do these tests not fit your data or your hypothesis?

The null hypothesis that I would like to test is:

There is no statistically significant relationship between the amount of quests completed by a player using my application and a player that does not.

I would use a two sample t-test, to test the hypothesis.