Research Track II
Fundaments of Statistics

Carmine Tommaso Recchiuto

Exercises

Exercise 1:

1) A random variable X follows a normal distribution with μ = 6 and σ = 1.24. Calculate the following probabilities:

a) $P(X \leq 8)$
b) $P(X \leq 3)$
c) $P(X \geq 7)$
d) $P(7 \leq X \leq 10)$

Z Score Table - Z Table and Z score calculation (z-table.com)

We have here the mean and the standard deviation of the population, hence we can compute the probabilities with the Z-score

$$Z = \frac{x-\mu}{\sigma}$$

a) (8-6)/1.24=1.613 -> 94.6%

b) (3-6)/1.24=-2.42 -> 0.8%

c)(7-6)/1.24=0.806 -> 79.10%
But since it is x>=7, the probability to be considered is 1-p = 20.90%

d) p(x<=7)=79.10% p(x>=10)=0.06%
Hence p(7<=x<=10) = 20.84%

Exercises

Exercise 2:

A machine in a factory is responsible for filling boxes of cereal. The weight of cereal in each box is has a mean of 500g and a variance of 20g. A box is picked at random. Calculate the probability that it contains less than 490g of cereal.

Again we have here the mean and the standard deviation (variance) of the population, hence we can compute the probabilities with the Z-score

$P(X \leq 490) = P(Z \leq (490-500)/\sqrt{20}) = P(Z \leq -2.236) = 12.7\%$.

In other words, we would expect roughly 12.7% of boxes to weigh less than 490g.

Exercises

Exercise 3:

3) A visual research lab has purchased a digital colour blindness test from a company. Before they can use the test in their research, they must ensure it is properly calibrated. To do this they must check that they get the same results as the company when testing participants with no colour deficiencies. The company states that participants with healthy colour vision will score 15 on the test on average. The research lab tests 13 participants with healthy colour vision. On average they score 12 with a standard deviation of 3.6. Is their machine properly calibrated?

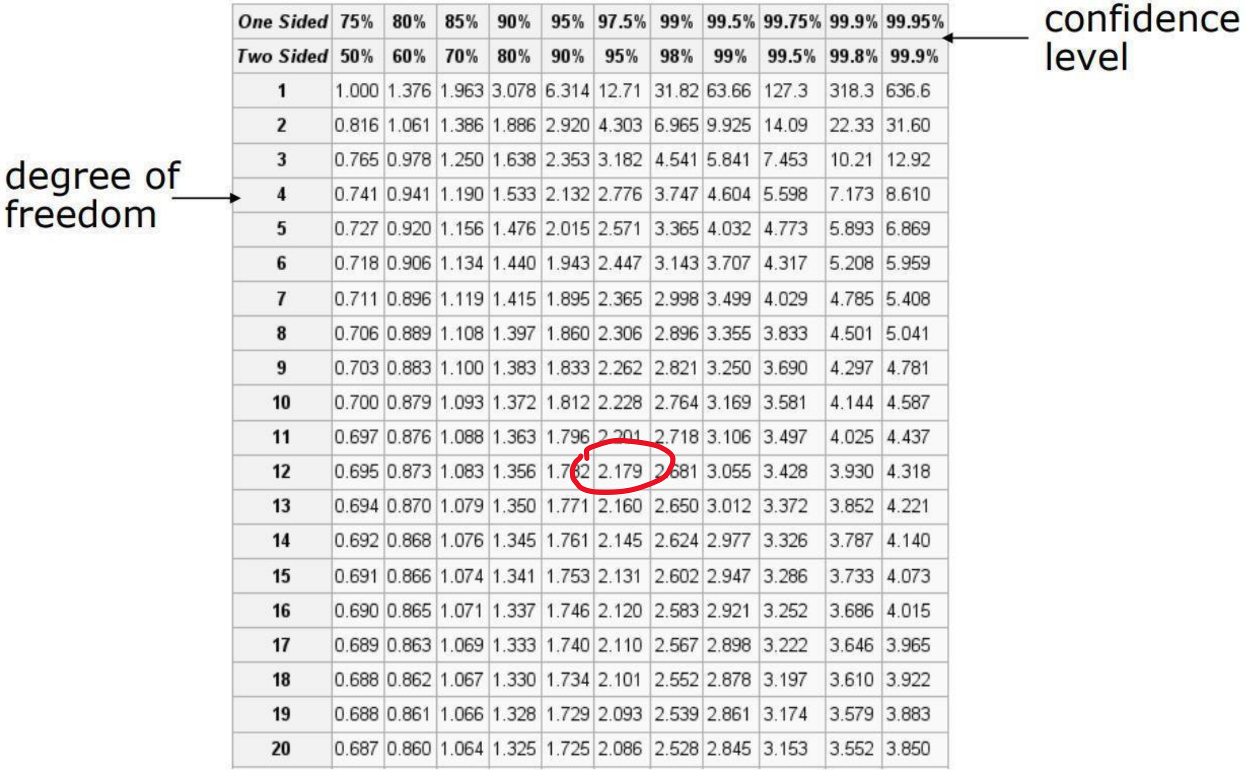
We need to use a t-test as we don’t have the population standard deviation. It will be 2-tailed as we are checking to see if the machine gives different values, rather than specifically lower or higher values.

$$p = (12-15)/(3.6/\sqrt{13}) = -3.0046$$

Exercises

Exercise 3:

3) A visual research lab has purchased a digital colou
test in their research, they must ensure it is properly ca
results as the company when testing participants with n
with healthy colour vision will score 15 on the test on a
colour vision. On average they score 12 with a standarc



degree of freedom

confidence level

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

We need to use a t-test as we don't have the population sta
the machine gives different values, rather than specifically lc

$$p = (12-15)/(3.6/\sqrt{13}) = -3.0046$$

The critical value for a 2-tailed t-test at the 95% level with 12 (N-1) degrees of freedom is $2.179 < 3.0046$, therefore our result is significant and we can conclude that the machine is not calibrated properly.

Exercises

Exercise 4:

4) The authors of a paper about path planning had tested 20 different environments and found that on average the robot takes 39s to navigate in the environment, with a standard deviation of 4.3 s. They have compared the implemented algorithm with a state-of-the-art approach. The average score in that case is 51s. Is there significant evidence at the 5 % level to support the researcher's claim?

We need to use a t-test as we don't have the population standard deviation. In this case, the usage of a 1-tailed or 2-tailed test depends on the researchers' claim. If the claim is that their algorithm is faster ($H_1 = \mu_1 < \mu_2$), then the 1-tailed test should be used.

$$\mathrm{p} = (39-51)/(4.3/\sqrt{}20)\text{=-}12.48$$

The critical value for a 1-tailed t-test at the 95% level with 19 degrees of freedom is $1.729 < 12.48$, therefore the result is significant and we conclude that there is evidence to suggest the proposed approach is faster than classical other approaches.

Two sample T-Test

T-test

There is a quite strong limitation in the exercises / examples seen so far.

We are assuming that we know something about the population (either the mean, or the standard deviation), which is however not true in many cases (e.g., exercise 4).

In many real scenarios, we don’t know much about the population, and we want to compare two approaches based on the results that we obtain

Eg, in the case of exercise 4, we run the system 20 times with two different approaches, and we want in the end to compare the results, i.e., check if the data obtained belong to the same population (null hypothesis) or not (alternative hypothesis).

In this case we use the **two-sample T-Test**!

Pooled Standard Deviation

We have seen how the standard deviation of the mean can be computed as:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$$

For the two sample t-Test, we have two variances. The Pooled variance is the weighted average of the variances, using as weights -> N-1

The pooled, estimated variance of the sampling distribution of the difference of means is then:

$$\hat{\sigma}^2_{pooled} = \frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}$$

where s1 and s2 are the two standard deviations

Pooled Standard Deviation

Which leads to the pooled, estimated SE of the sampling distribution of the difference of means

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\hat{\sigma}^2_{pooled}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

We are interested in the differences, thus the t-statistics turns into

$$t_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

Another example:

- Two planning algorithms A and B
- Evaluate A and B, each in 25 randomly generated scenarios ($N_A = N_B = 25$)

Two Sample T-Test

Situation:

$$H_0 \;:\; \mu_A = \mu_B \;\leftrightarrow\; \mu_A - \mu_B = 0$$
$$H_1 \;:\; \mu_A \neq \mu_B \;\leftrightarrow\; \mu_A - \mu_B \neq 0$$
$$\bar{x}_A = 127 \;\; s_A = 33; \;\; \bar{x}_B = 131, \;\; s_B = 28$$

$\hat{\sigma}^2_{pooled} =$ $(24*33^2 + 24*28^2) / 48 =$ 936.5

$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} =$ sqrt ($\hat{\sigma}^2_{pooled}$ * 2 / 25) = 8.6556

Finally, the t-value of the two sample t-test may be computed as:

$t_{\bar{x}_1 - \bar{x}_2} = \dfrac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$ => 4 / 8.66 = 0.46

DoF is $N_A + N_B - 2 = 48$ -> We cannot reject H_0 since |t| should be > 2.01

Paired T-Test

A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample. Examples of where this might occur are:

- Before-and-after observations on the same subjects (e.g. students’ diagnostic test results before and after a particular module or course).
- A comparison of two different methods of measurement or two different treatments where the measurements/treatments are applied to the same subjects
- A comparison of two different approaches applied to the same scenarios

To test the null hypothesis that the true mean difference is zero, the procedure is as follows:

1. Calculate the difference ($d_i = y_i − x_i$) between the two observations on each pair, making sure you distinguish between positive and negative differences
2. Calculate the mean difference $\bar{d}.$
3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of

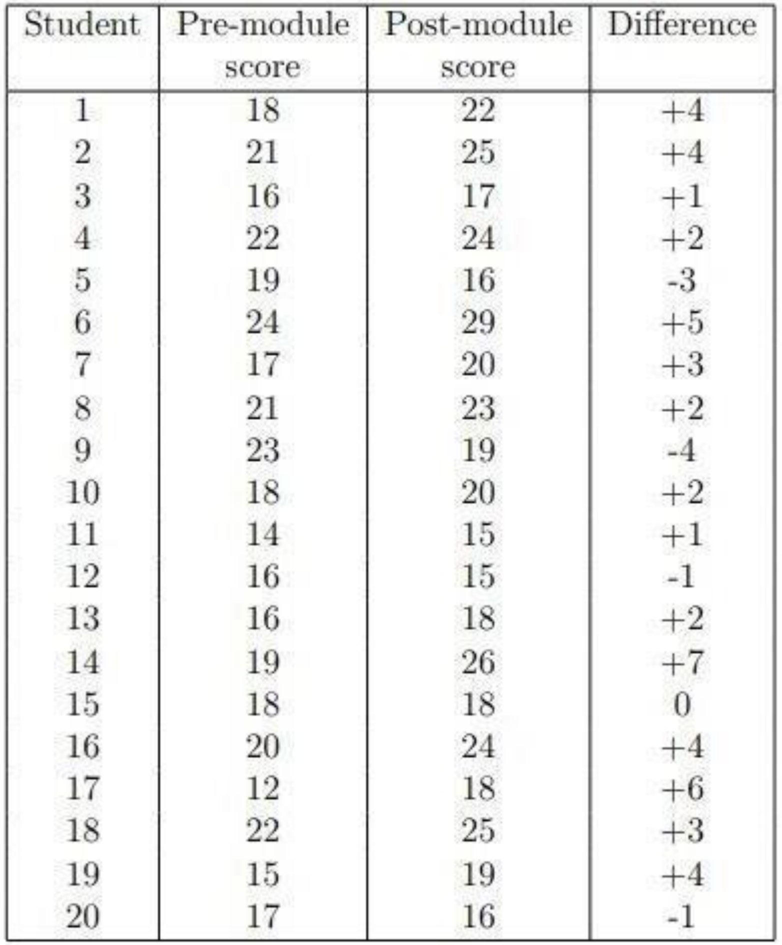
the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$

4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(\bar{d})}$. Under the null hypothesis, this statistic follows a t-distribution with n − 1 degrees of freedom.

5. Use tables of the t-distribution to compare your value for T to the $t_{n−1}$ distribution. This will give the p- value for the paired t-test.

Paired T-Test

Example. With n = 20 students, the following results were obtained ->



Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Calculating the mean and standard deviation of the differences gives:

$\bar{d}$ = 2.05 and s_d = 2.837. Therefore,

$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$ = 2.837 / sqrt(20) = 0.634.

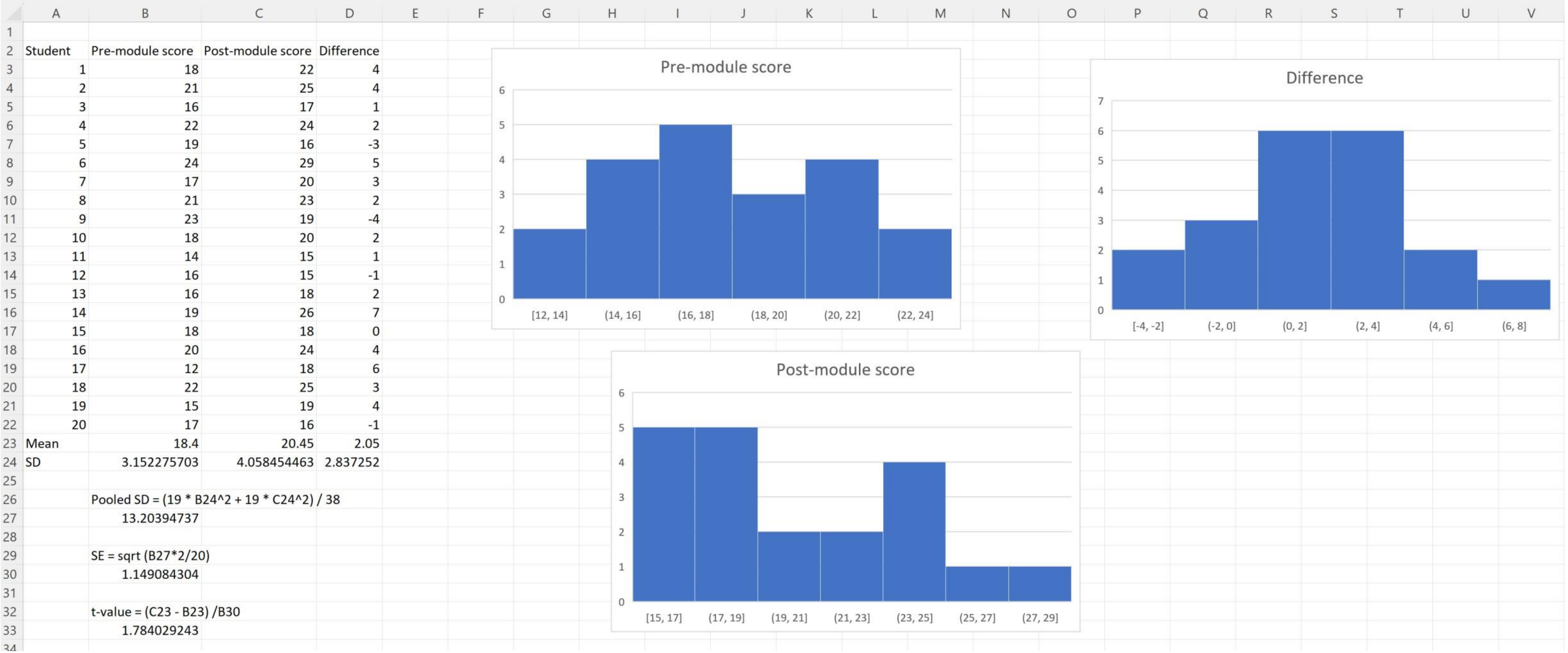
Finally, t = $\frac{\bar{d}}{SE(\bar{d})}$ = 2.05 / 0.634 = 3.231 (on 19 DoF)

This means that we can reject the null hypothesis, with a confidence of 99.5%

Two sample test: Test if the differences of the means differs from zero
Paired sample test: Test if the means computed over the individual differences is differ from zero

Try to repeat the same exercise with a two sample t-test -> Whenever possible, use the paired sample t-Test!

Paired T-Test



	A	B	C	D
1				
2	Student	Pre-module score	Post-module score	Difference
3	1	18	22	4
4	2	21	25	4
5	3	16	17	1
6	4	22	24	2
7	5	19	16	-3
8	6	24	29	5
9	7	17	20	3
10	8	21	23	2
11	9	23	19	-4
12	10	18	20	2
13	11	14	15	1
14	12	16	15	-1
15	13	16	18	2
16	14	19	26	7
17	15	18	18	0
18	16	20	24	4
19	17	12	18	6
20	18	22	25	3
21	19	15	19	4
22	20	17	16	-1
23	Mean	18.4	20.45	2.05
24	SD	3.152275703	4.058454463	2.837252

Pooled SD = (19 * B24^2 + 19 * C24^2) / 38

13.20394737

SE = sqrt (B27*2/20)

1.149084304

t-value = (C23 - B23) /B30

1.784029243

A General Comment

From the formulas and examples seen before, you can see how the probability of rejecting the null hypothesis when it is false is greater as the sample size increases. The power of the test is the ability to detect a false null hypothesis, and power increases as sample size increases.

It is generally recommended for a test to have high power. However, in an attempt to raise the power of a test, the *t*-value can be made as large as one wishes. In other words, to guarantee a rejection of the null hypothesis, one only needs a large enough sample. For example, while computing the t-value, as the sample size increases, x¯ and *S* will approximately stabilize at the true parameter values. Hence, a large value of *n* translates into a large value of *t*

$$t = \frac{\bar{x}-\mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x}-\mu}{s/\sqrt{N}}$$

std. dev estimated form the sample

sample size

While minimum sample sizes are strictly adhered to in choosing an appropriate test statistic, maximum sample sizes are not set. Alternatively, α is not usually adjusted according to the sample size. For example, for larger sample sizes (e.g., *n* > 1000), the set significance should be adjusted to a smaller value (e.g., α = 0.01 or 0.001).

A General Comment

Even more arbitrary is the tendency to adhere to a standard set significance of 0.05 or 0.01. ***T*-values which imply a statistical significance less than or equal to α are deemed significant; those implying a statistical significance greater than α are nonsignificant**.

This rule promotes the nonsensical distinction between a statistical significant finding if *P* = 0.04, and a nonsignificant finding of *P* = 0.06. Such minor differences are not so relevant actually, as they derive from tests whose assumptions often are only approximately met (e.g., a random sample).

Statistical significance is often an imperfect method for determining the reliability of a statistic. However, if the assumption of the tests are met, attention is paid to sample size, and if care is taken to interpret *t*-values in relation to confidence intervals, statistical significance tests can be the starting point for further analysis.

Non-parametric Tests

What does it happen if I have a small sample, which does not come from a normally distributed population (i.e., the Lilliefors test rejects the null hypothesis)?

I can’t use the t-test, and obviously the z-score, but in this case I can apply some **non-parametric tests.**

Wilcoxon-Mann-Whitney test (or U-test): This test is used to determine whether two independent samples have been drawn from the same population (hence, it’s the equivalent of the two stample t-test, for generic sample distributions).

- To perform this test, we first of all rank the data jointly, taking them as belonging to a single sample in either an increasing or decreasing order of magnitude. We usually adopt low to high ranking process which means we assign rank 1 to an item with lowest value, rank 2 to the next higher item and so on. In case there are ties, then we would assign each of the tied observation the mean of the ranks which they jointly occupy. For example, if sixth, seventh and eighth values are identical, we would assign each the rank (6 + 7 + 8)/3 = 7.

After this we find the sum of the ranks assigned to the values of the first sample (and call it $R1$) and also the sum of the ranks assigned to the values of the second sample (and call it $R2$). Then we work out the test statistic i.e., U, which is a measurement of the difference between the ranked observations

Non-parametric Tests

$$U = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

where *n*1, and *n*2 are the sample sizes and *R*1 is the sum of ranks assigned to the values of the first sample. (In practice, whichever rank sum can be conveniently obtained can be taken as *R*1, since both samples could be the first sample)

In applying *U*-test we take the null hypothesis that the two samples come from identical populations. If this hypothesis is true, it seems reasonable to suppose that the means of the ranks assigned to the values of the two samples should be more or less the same. Under the alternative hypothesis, the means of the two populations are not equal and if this is so, then most of the smaller ranks will go to the values of one sample while most of the higher ranks will go to those of the other sample.

If the null hypothesis that the *n*1 + *n*2 observations came from identical populations is true, the said '*U*' statistic has a sampling distribution with

$$\mu_U = \frac{n_1 \cdot n_2}{2} \qquad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Non-parametric Tests

If n_1 and n_2 are sufficiently large (i.e., both greater than 8), the sampling distribution of U can be approximated closely with normal distribution and the limits of the acceptance region can be determined in the usual way at a given level of significance. But if either n_1 or n_2 is so small that the normal curve approximation to the sampling distribution of U cannot be used, then exact tests may be based on special tables, showing selected values of Wilcoxon's (unpaired) distribution.*

In case we need to compare more than 2 samples, the **Kruskal-Wallis** test (or H test) can be adopted: this test is used to test the null hypothesis that '*k*' independent random samples come from identical universes against the alternative hypothesis that the means of these universes are not equal. In this test, like the *U* test, the data are ranked jointly from low to high or high to low as if they constituted a single sample; after that, the test statistic H is computed.

* N.B. It also exist a Wilcoxon paired test: Wilcoxon Matched-pairs Test (or Signed Rank Test)

Chi – Square Test

Finally, a chi-square statistic is one way to show a relationship between two categorical variables. In statistics, there are two types of variables: numerical (countable) variables and non-numerical (categorical) variables. The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.

However, as usual, some conditions should be satisfied to apply it:

(i) Observations recorded and used are collected on a random basis.
(ii) All the items in the sample must be independent.
(iii) No group should contain very few items, say less than 10.
(iv) The overall number of items must also be reasonably large. It should normally be at least 50.

Chi – Square Test

How to apply it?

- First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis (e.g.: divide the overall number of data by the number of groups)
- Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(Oij – Eij)^{2.}$
- Divide the quantity $(Oij – Eij)^2$ obtained as stated above by the corresponding expected frequency to get $(Oij – Eij)^2/Eij$ and this should be done for all the group frequencies
- Sum the $(Oij – Eij)^2/Eij$ values. This is the required ξ^2 value.

The ξ^2 value obtained as such should be compared with relevant table value of ξ^2 and then inference be drawn

E.g.: Two different path planning algorithms are applied to 60 different scenarios. The robot reaches the target in 48 of the 60 runs when using the first algorithm, and in 44 runs using the second one. Are these differences due to chance?

Chi – Square Test

I can build the following table:

	Success	Failure
Algorithm 1	48	12
Algorithm 2	44	16
Total	92	28

If I consider the null hypothesis, Algorithm1 has the same performance of Algorithm2, I expect my algorithm to succeed (it does not depend if alg1 or alg2 is used): 92/2 = 46 runs out of 60.

	Success	Failure
Algorithm 1	48 (46)	12 (14)
Algorithm 2	44 (46)	16 (14)
Total	92	28

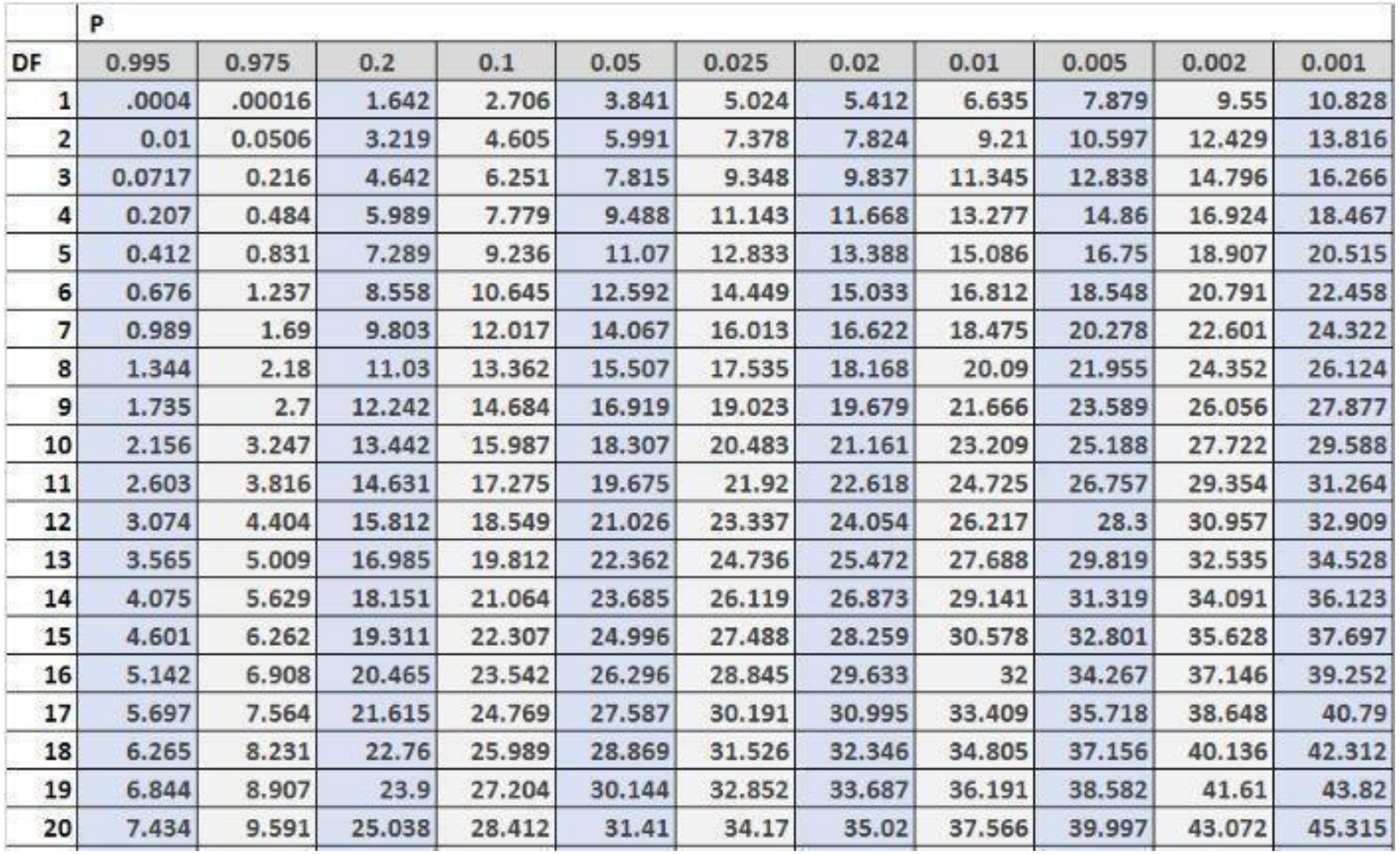
Chi – Square Test

$\xi^2 = (2^2 / 46 + 2^2 / 14) = 0.37$

Looking at the table (DoF = $(n_r - 1)(n_c - 1)$), I can see that the probability of committing an error by rejecting the null hypothesis is > 0.2; hence, it is quite significant.

In this case, I can't reject the null hypothesis.

To reject the null hypothesis with a confidence level of 95%, ξ^2 should have been bigger than 3.841



	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Assignment (continuous evaluation – RT2)

✓ Properly comment (sphinx and/or doxygen) the 2rd RT1 – assignment

✓ Create a jupyter notebook to interact with the simulation of the 2rd assignment

❑ Perform a statistical analysis on the first assignment, considering two different implementations (yours, and a solution of one of your colleagues) and testing which one performs "better", when tokens are randomly placed in the environment.

As performance evaluators you may possibly consider:

- the average time required to finish the task
- the number of success / failures
- …

Possibly, you can also vary the number of boxes to see how the algorithm behaves.

You can possibly modify the file: robot-sim/sr/robot/arenas/two_colours_assignment_arena.py

Assignment (continuous evaluation – RT2)

```
token.location = (cos(angle) * radius, sin(angle) * radius)
```

- ❑ It’s very important here to clearly define what is the hypothesis that you want to test. Do you want to test which algorithm does perform better when the token are placed on the two circles? Or when the token are randomly placed in the environment?

  - Based on the hypothesis, you need to design the experiments accordingly

Carefully plan the number of experiments and choose a suitable statistical approach.

Assignment (continuous evaluation – RT2)

❑ Write a report composed of:

- Hypotheses made (null hypothesis and alternative hypothesis)
- Description and motivation of the experimental setup (types of experiments, number of repetitions)
- Results
- Discussion of the results with statistical analysis
- Conclusion (is the hypothesis proven?)