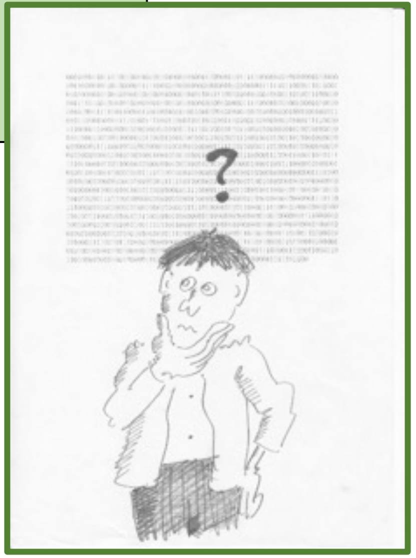
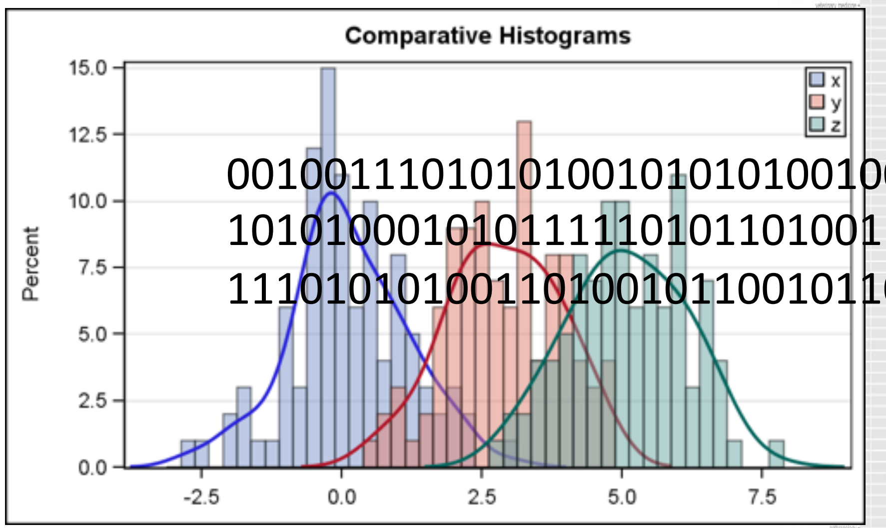
Wilcoxon Rank Sum and t-test

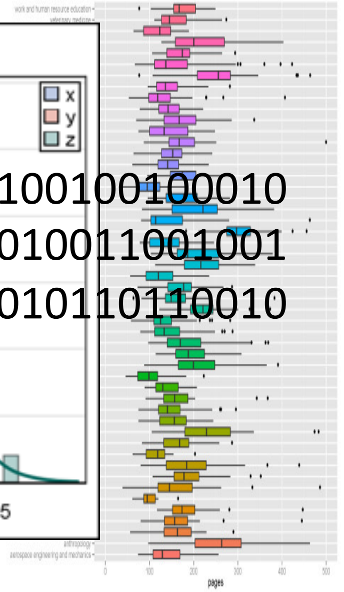
Statistics and data analysis

Zohar Yakhini, Leon Anavy, Ben Galili

IDC, Herzeliya







IDC HERZLIYA

Wilcoxon Rank Sum test

Frank Wilcoxon
American statistician

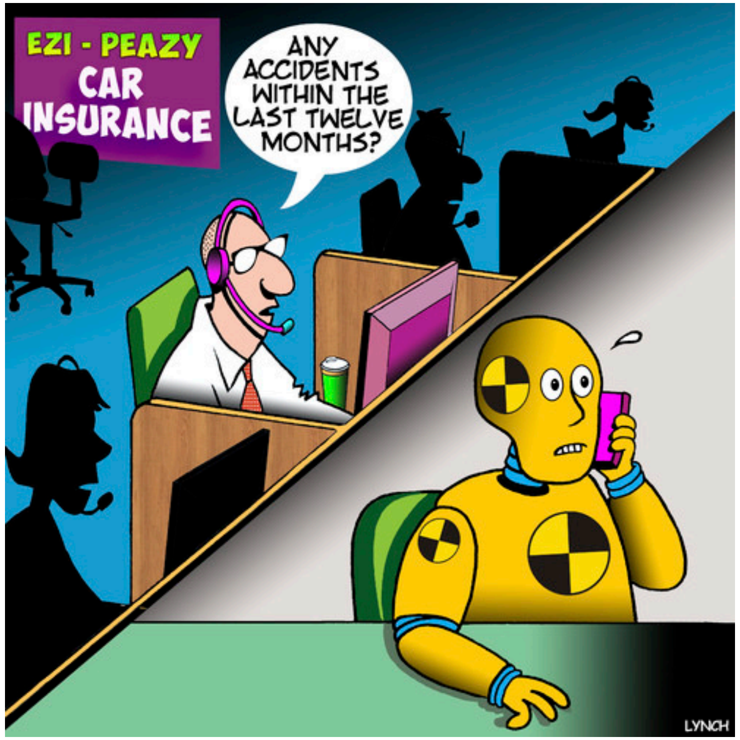
- We are comparing numbers, or measured quantities, obtained for two different populations.
- Individuals are assumed to be sampled randomly and independently.
- We have two vectors of measured values – one for each population.

Wilcoxon Rank Sum test

Consider two sample sets of independently acquired observations from two different labels/populations.

Example: safety test results for cars manufactured in the Randomistan VW factory vs those made in the German factory.

German factory	Stochastic Heights factory
3.2	3.7
4.5	8.5
8.1	6.1
9.9	9.3
4.1	9.1
7.3	4.3
5.2	7.2
6.0	
nG = 8	nR = 7



Wilcoxon Rank Sum test

Null assumption:
When considering samples from both factories then all rank configurations are equiprobable.

German factory	Stochastic Heights factory
3.2	3.7
4.5	8.5
8.1	6.1
9.9	9.3
4.1	9.1
7.3	4.3
5.2	7.2
6.0	
nG = 8	nR = 7

9.9	G
9.3	R
9.1	R
8.5	R
8.1	G
7.3	G
7.2	R
6.1	R
6.0	G
5.2	G
4.5	G
4.3	R
4.1	G
3.7	R
3.2	G

N = nG + nR = 15

Wilcoxon Rank Sum test

Null model, abstracted:
All N choose B binary configurations in
$$\{0,1\}^{(N,B)}$$
are equiprobable.
Let T = sum of the ranks of the entries labeled 1.

$\{0,1\}^{(N,B)}$: All binary vectors with N elements out of which B are 1s

9.9	G
9.3	R
9.1	R
8.5	R
8.1	G
7.3	G
7.2	R
6.1	R
6.0	G
5.2	G
4.5	G
4.3	R
4.1	G
3.7	R
3.2	G

N = nG + nR = 15

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

N = 15, B = 7, T = 50

Wilcoxon Rank Sum test

Under the null model we have

$$E(T) = 56$$

The result here points to better (lower numbers) ranks for R.

But is it <u>significant</u>?

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

$N = 15$, $B = 7$, $T = 50$

Wilcoxon Rank Sum test

We want to compute

$P(T \leq 50)$,

under the null model.

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

$N = 15$, $B = 7$, $T = 50$

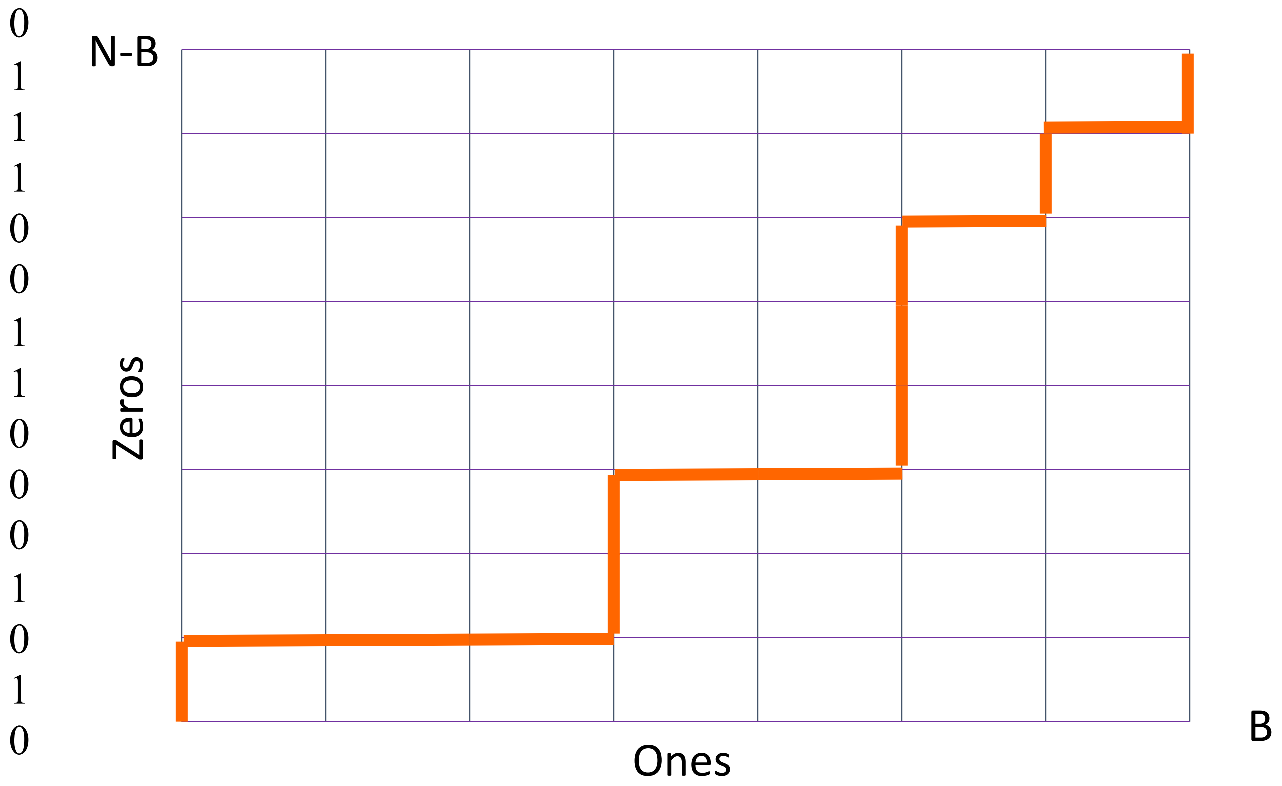
Wilcoxon Rank Sum test

Can we calculate an exact p-value?

$P_{Null}(T \leq 50) = ?$

Number of binary vectors for which $T \leq 50$ out of all possible binary vectors

Idea:
Map patterns to paths on the lattice
Use DP to count paths in which $T \leq 50$



Wilcoxon Rank Sum test

Normal approximation (Wilcoxon 1947)

$P_{Null}(T \leq 50)$?

When B and N-B are sufficiently large and depending on the desired accuracy

Let

$$\mu_T = \frac{B(N+1)}{2}$$

and

$$\sigma_T = \sqrt{\frac{B(N-B)(N+1)}{12}}$$

then

$$Z(T) = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

Wilcoxon Rank Sum test

Back to Randomistan VW factory: N = 15, B = 7, T = 50

Using the normal approximation even though the numbers are not sufficiently large:

$$\mu_T = \frac{7(15+1)}{2} = 56, \qquad \sigma_T = \sqrt{\frac{7(15-7)(15+1)}{12}} \approx 8.64, \qquad Z \approx \frac{50-56}{8.64} = -0.7$$

and therefore

$$P_{Null}(T \leq 50) \approx 0.24$$

and we **<u>do not have sufficient confidence for rejecting the hypothesis</u>** that the German factory is as good as the one in Stochastic Heights.

Wilcoxon Rank Sum test

- You want to check a new workshop for the course. Does the **new** workshop help the students improve?
  You collected the following data:

User	Before	After
Donna	88	98
Santosha	76	78
Sam	83	90
Tamika	80	99
Brian	68	74
Jorge	85	84

Wilcoxon Rank Sum test

User	Before	After
Donna	88	98
Santosha	76	78
Sam	83	90
Tamika	80	99
Brian	68	74
Jorge	85	84

68	Na = 6	1
74	Nb = 6	2
76	N = 12	3
78		4
80		5
83		6
84		7
85		8
88		9
90		10
98		11
99		12

Wilcoxon Rank Sum test

- Let T = sum of the ranks of the **BLACK** entries
- B=6, N=12
- E(T) = 39
- T = 1 + 3 + 5 + 6 + 8 + 9 = 32
- We want to compute $P(T \leq 32)$ under the null model

68	Na = 6	1
74	Nb = 6	2
76	N = 12	3
78		4
80		5
83		6
84		7
85		8
88		9
90		10
98		11
99		12

Wilcoxon Rank Sum test

Let

$$\mu_T = \frac{B(N+1)}{2} = \frac{6*13}{2} = 39$$

and

$$\sigma_T = \sqrt{\frac{B(N-B)(N+1)}{12}} = \sqrt{\frac{6*6*13}{12}} = \sqrt{39}$$

then

$$Z(T) = \frac{T-\mu_T}{\sigma_T} = \frac{32-39}{\sqrt{39}} = -1.12$$

$$Z(T) \sim N(0,1)$$

$$\downarrow$$

$$P(T \leq 32) = p - value = 0.13$$

68	1
74	2
76	3
78	4
80	5
83	6
84	7
85	8
88	9
90	10
98	11
99	12

Na = 6
Nb = 6
N = 12

Wilcoxon Signed Rank test

- **<u>But, is this really the correct test?</u>**
- Tests Probability Distributions of 2 Matched Populations
- Assumptions:
  - Sampling is independent
  - **<u>Paired (matched) samples</u>**

User	Before	After
Donna	88	98
Santosha	76	78
Sam	83	90
Tamika	80	99
Brian	68	74
Jorge	85	84

Wilcoxon Signed Rank test

- Obtain Difference Scores, $D_i = X_{1i} - X_{2i}$
- Take Absolute Value $|D_i|$ and *rank them* (Do not count $D_i = 0$)
- Assign Ranks, R_i, with Smallest = 1
- Calculate ranks and mean rank for $|D_i|$
- Sum '+' Ranks (T_+) & '-' Ranks (T_-)
- Take T as $\min\{T_+, T_-\}$

Wilcoxon Signed Rank Test – Procedure

- If n is large enough we will use normal approximation:

$$\mu_T = \frac{n(n+1)}{4}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Let

$$Z(T) = \frac{T - \mu_T}{\sigma_T}$$

$$Z(T) \sim N(0,1)$$

Wilcoxon Signed Rank test

- You want to check a new workshop for the course. Does the **new** workshop help the students improve?

Before	After	D_i	$\|D_i\|$	R_i	Sign
88	98	10	10	5	+
76	78	2	2	2	+
83	90	7	7	4	+
80	99	19	19	6	+
68	74	6	6	3	+
85	84	-1	1	1	-

- $T_+ = 20$
- $T_- = 1$
- $T = 1$

Wilcoxon Signed Rank Test – Procedure

- If n is large enough we will use normal approximation:

$$\mu_T = \frac{n(n+1)}{4} = \frac{6*7}{4} = 10.5$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{6*7*13}{24}} = 4.77$$

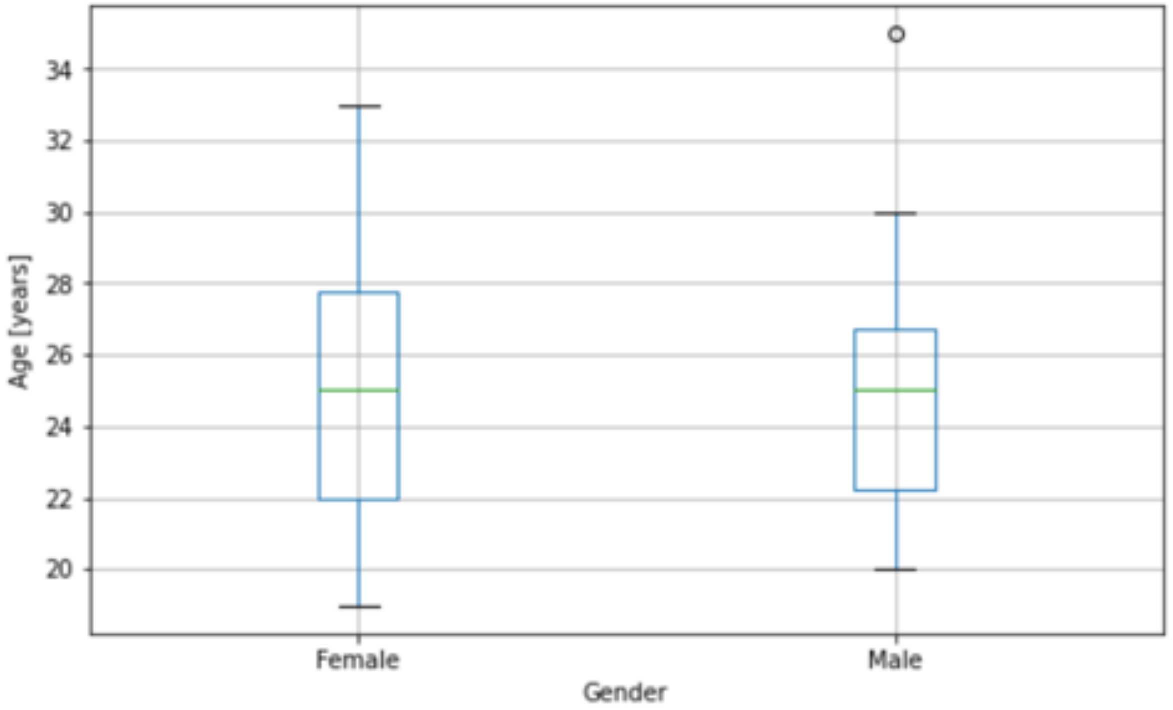
Let

$$Z(T) = \frac{T-\mu_T}{\sigma_T} = \frac{1-10.5}{4.77} = -2$$

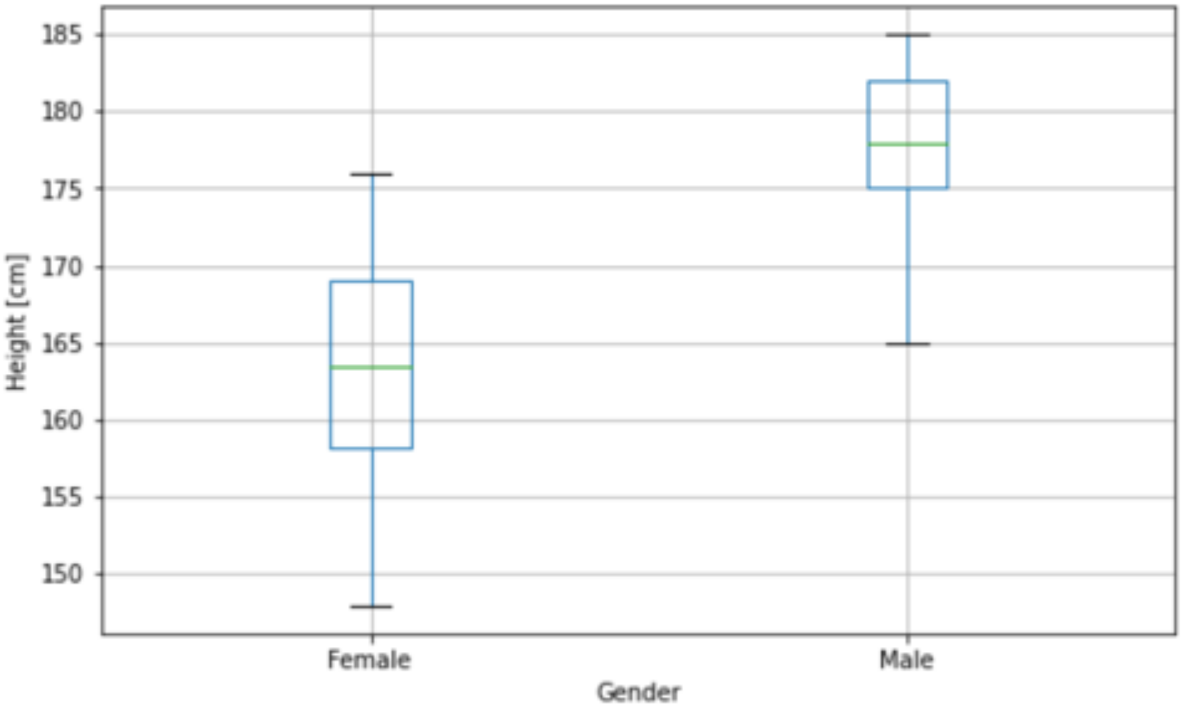
$$p-value = 0.023$$

Comparing two independent samples – t-test

Are men older than women?



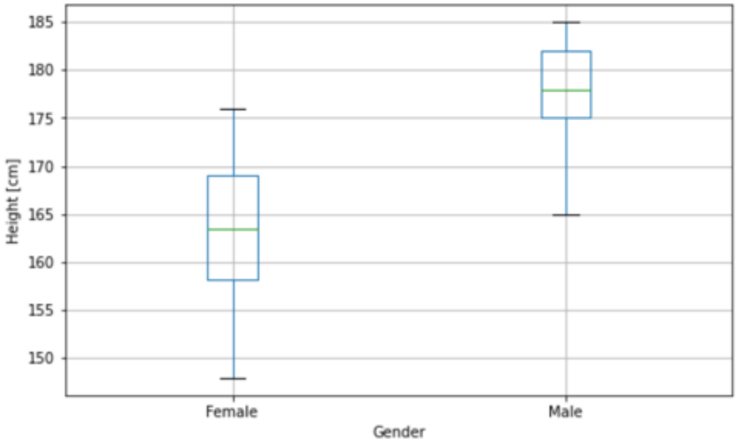
Are men taller than women?



Comparing two independent samples – t-test

- **Null hypothesis $\boldsymbol{H_0}$** - the hypothesis on which we want to defend
  $\boldsymbol{H_0}: \mu_{men} \leq \mu_{women}$
- **Alternative hypothesis $\boldsymbol{H_1}$** - a new hypothesis that we want to check
  $\boldsymbol{H_1}: \mu_{men} > \mu_{women}$

Are men taller than women?



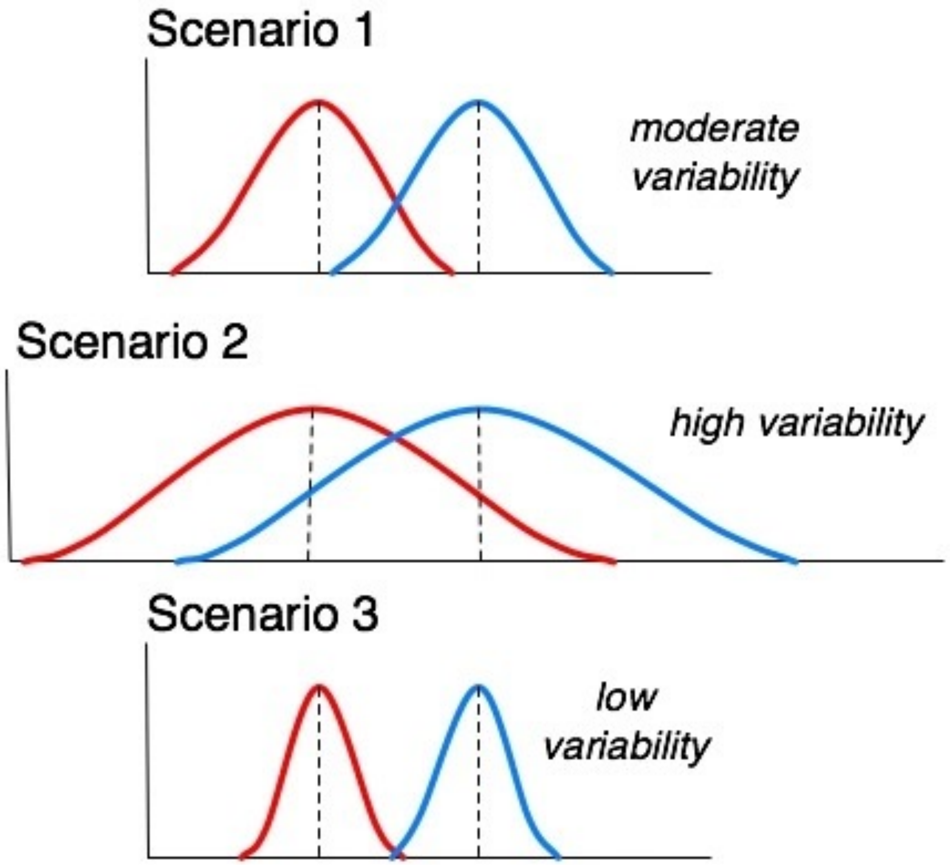
- **Test statistic**
  - Can be calculated from the sample
  - We know its distribution under H_0
- **p-value**
  - **Under $\boldsymbol{H_0}$**, What is the probability to get a test statistic which is “more extreme” than the observed.

Comparing two independent samples – t-test

- **Test statistic**
    - Can be calculated from the sample
    - We know its distribution under H_0

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S}$$

Where S is a scaling factor so that t has a **<u>Student's t-distribution</u>** with $n_1 + n_2 - 2$ degrees of freedom.



- **p-value**
    - **Under $\boldsymbol{H_0}$**, What is the probability to get a test statistic which is "more extreme" than the observed.

Student's t-distribution

Let $X_1, \dots, X_n$ be i.i.d from $N(\mu, \sigma^2)$ and let $\bar{X}$ and S^2 be the sample mean and variance.

Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

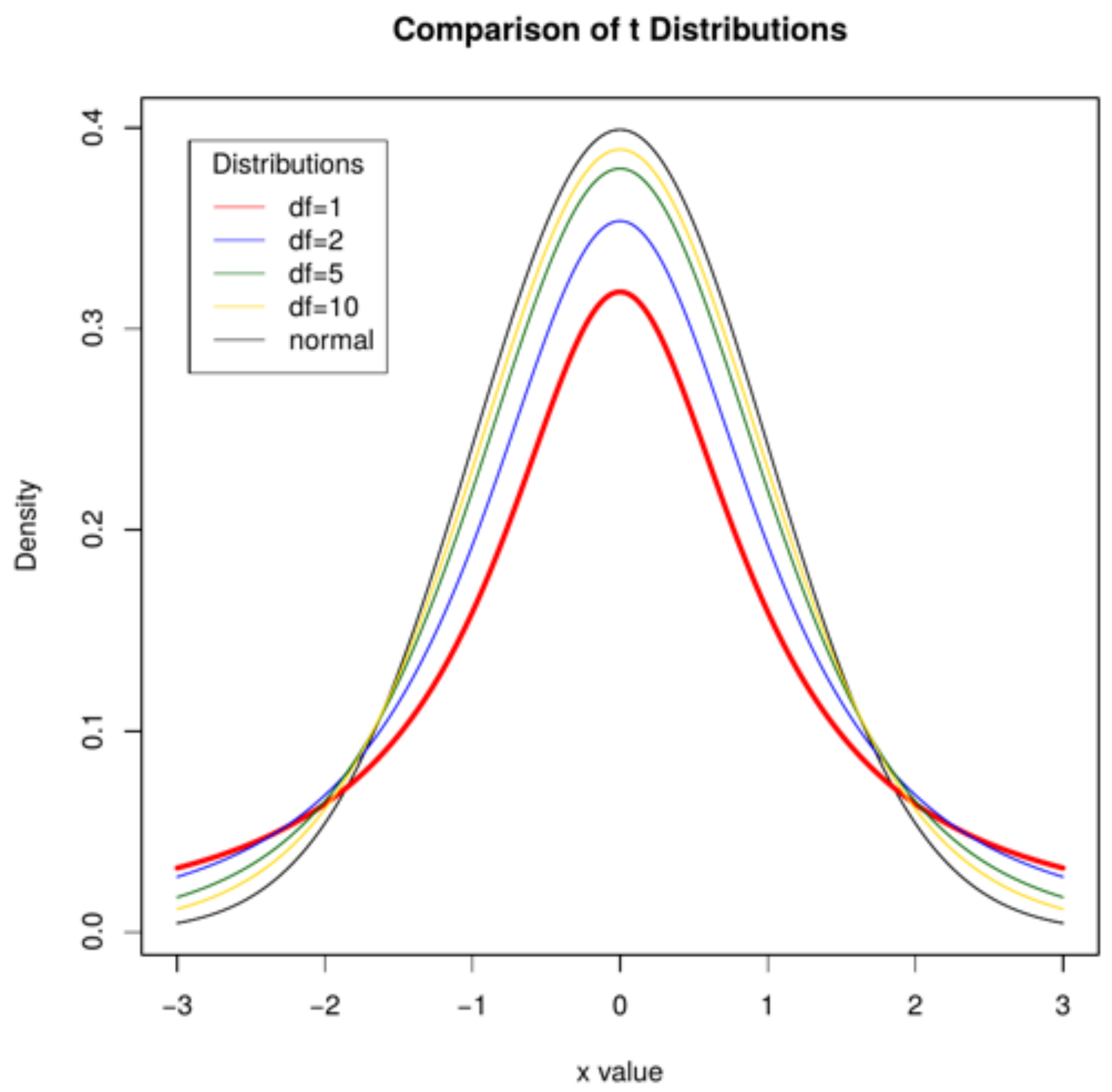
And

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Has a Student's t-distribution with $n - 1$ degree of freedom

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Student's t-distribution



Comparing two independent samples – t-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S}$$

Where S is a scaling factor so that t has a student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Assuming similar variance to both samples:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad S_p = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$$

$$p - value = P(T_{n-1} > t)$$

t-test example

Female		Male	
163	162	168	173
160	174	173	178
158	155	183	183
158	150	178	175
155	164	176	180
168	153	176	170
159	176	175	170
169	168	179	185
170	174	182	173
170	169	175	178
160	166	180	175
173	157	185	175
160	167	165	183
169	148	178	183
163	172	183	182

$$\bar{X}_{men} = 168 + 173 + \cdots + 182 = 177.3$$

$$\bar{X}_{women} = 163 + 160 + \cdots + 172 = 163.67$$

$$S_p = \sqrt{\frac{29 \cdot 26.7 + 29 \cdot 55.61}{58}} = 6.41$$

$$t = \frac{177.3 - 163.67}{6.51\sqrt{\left(\frac{1}{29} + \frac{1}{29}\right)}} = 8.23$$

$$p - value = P(T_{58} > 8.23) = 1.25\times 10^{-11}$$

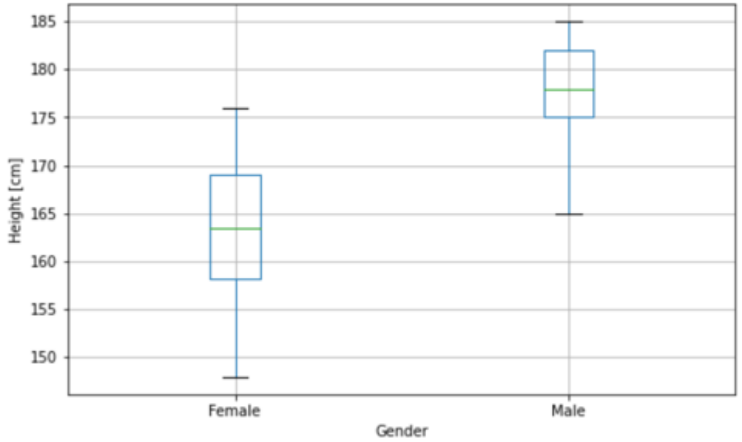
Are men taller than women?

$$\boldsymbol{H_0}: \mu_{men} \leq \mu_{women}$$

$$\boldsymbol{H_1}: \mu_{men} > \mu_{women}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p - value = P(T_{n-1} > t)$$



t-test example

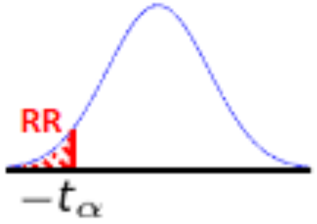
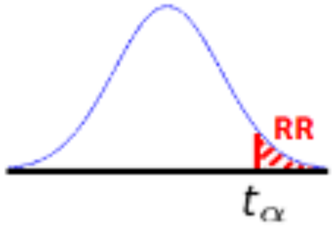
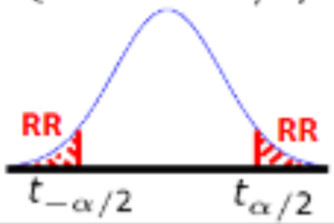
```python
f,m = [x['Height'].values for g,x in data.groupby('Gender')]\

# let's calculate
f_mean = f.mean()
f_var_unbiased = np.var(f,ddof=1)
f_n = len(f)
m_mean = m.mean()
m_var_unbiased = np.var(m,ddof=1)
m_n = len(m)
print('calculate:')
print('f:',f,len(f),f_mean,f_var_unbiased)
print('m:',m,len(m),m_mean,m_var_unbiased)
s_pooled = np.sqrt(((m_n-1)*m_var_unbiased+(f_n-1)*f_var_unbiased)/(m_n+f_n-2))
print('S_p:',s_pooled)
t_stat = (m_mean-f_mean)/(s_pooled*np.sqrt((1/m_n+1/f_n)))
print('t-test:',t_stat, t_dist.sf(t_stat,m_n+f_n-2),'\n')
```

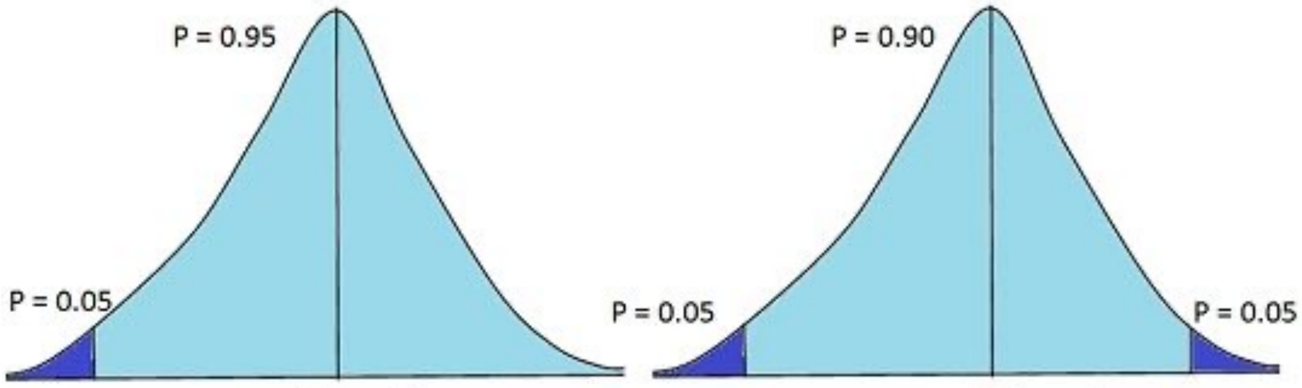
```
calculate:
f: n=30, mean=163.67, S^2=55.61
m: n=30, mean=177.30, S^2=26.70
S_p: 6.42
t-test: t=8.23, p_val=1.26e-11
```

```python
# using the built in method
print('using the built in method:')
t_stat,pval = ttest_ind(m,f)
print('t-test: t={:.2f}, p_val={:.2e}\n'.format(t_stat,pval))
```

```
using the built in method:
t-test: t=8.23, p_val=2.52e-11
```

One-tailed vs two-tailed t-test

H_o	H_a	REJECTION REGION
$\mu_1 \ge \mu_2$	$\mu_1 < \mu_2$	$RR\ \{TS \le -t_{\alpha,df}\}$ 
$\mu_1 \le \mu_2$	$\mu_1 > \mu_2$	$RR\ \{TS \ge t_{\alpha,df}\}$ 
$\mu_1 = \mu_2$	$\mu_1 \ne \mu_2$	$RR\ \{\lvert TS\rvert \ge t_{\alpha/2,df}\}$ 



One-tailed Test Vs Two-tailed Test

Comparing two independent samples – t-test

- Assumptions:
    - The means of the two samples follow normal distributions. (CLT…)
    - The variances of the two samples are equal or similar.
    - The two samples are independent.

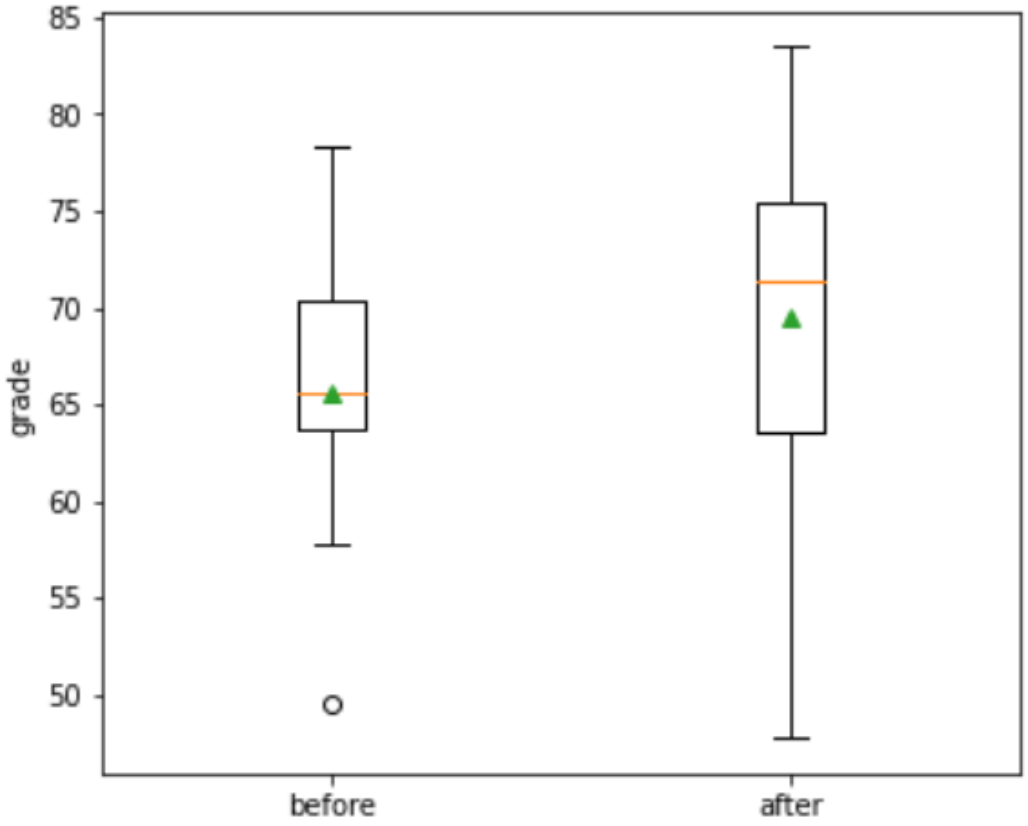
- Variations:
    - Unequal variance – Welch's t-test
    - Two sided t-test
    - Paired samples (soon)

Comparing two paired samples – t-test

Suppose we are examining the grades of 10 students. We record their understandings in statistics and ML before and after this course.

Did the course help?

ID	Before	After
1	78.3	83.5
2	72.2	81.2
3	49.5	47.7
4	64.9	73.1
5	71.2	75.4
6	57.8	63
7	67.7	65
8	66.1	69.8
9	65	75.5
10	63.3	60.9

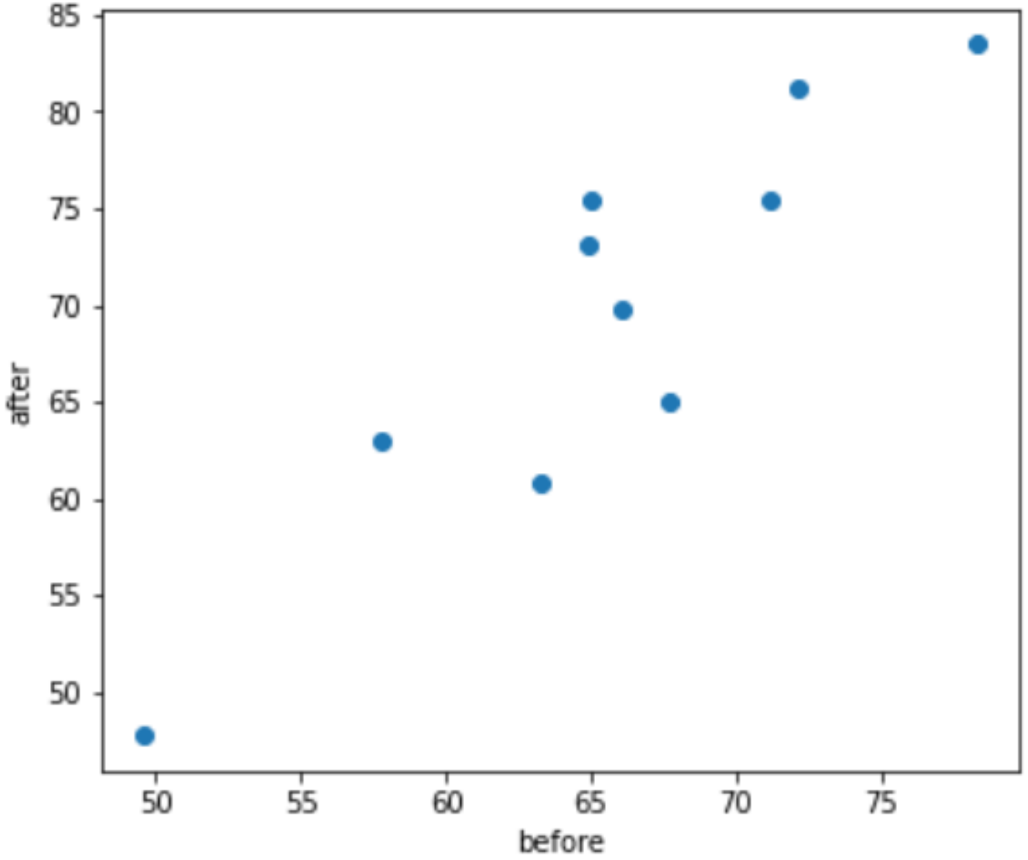


Comparing two paired samples – t-test

Those are the same 10 students measured twice. We can use this extra information.

Did the course help?

ID	Before	After
1	78.3	83.5
2	72.2	81.2
3	49.5	47.7
4	64.9	73.1
5	71.2	75.4
6	57.8	63
7	67.7	65
8	66.1	69.8
9	65	75.5
10	63.3	60.9

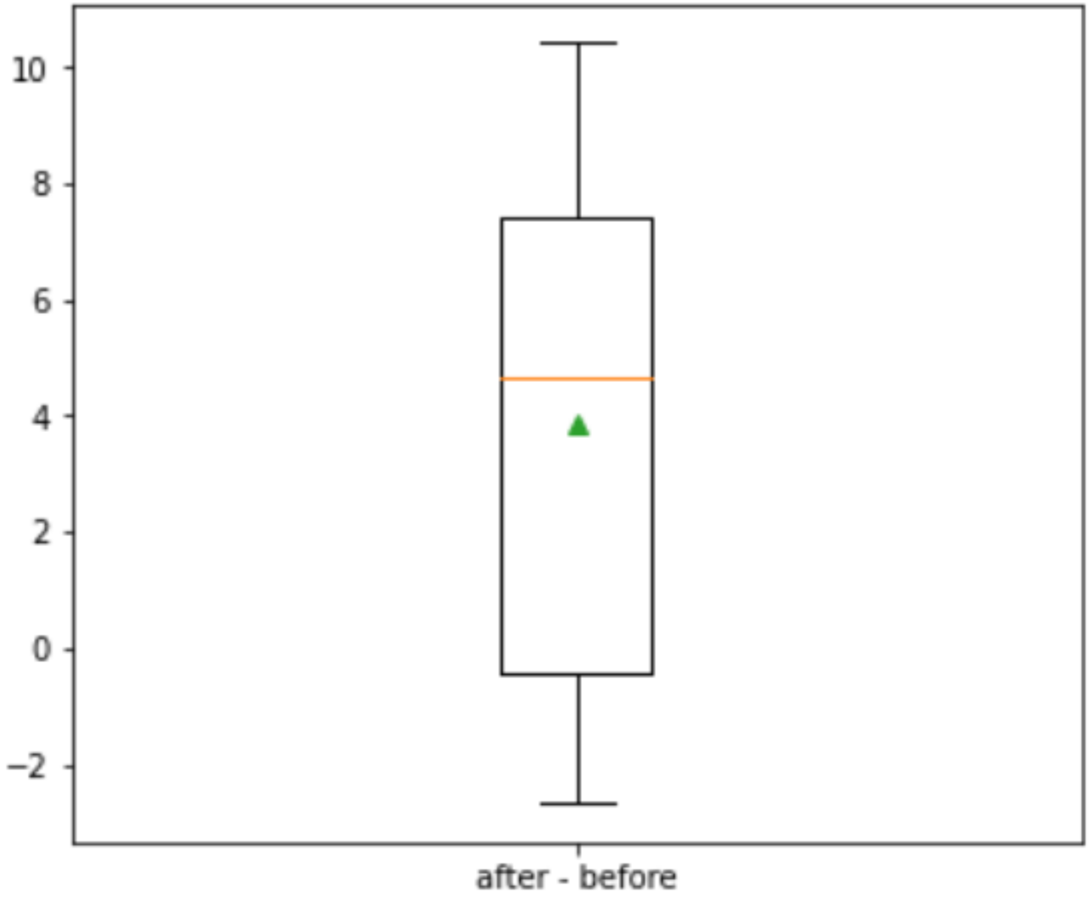


Comparing two paired samples – t-test

Let’s calculate the difference for each student.

Did the course help?

ID	Before	After	**Diff**
1	78.3	83.5	**5.2**
2	72.2	81.2	**9**
3	49.5	47.7	**-1.8**
4	64.9	73.1	**8.1**
5	71.2	75.4	**4.1**
6	57.8	63	**5.2**
7	67.7	65	**-2.7**
8	66.1	69.8	**3.7**
9	65	75.5	**10.4**
10	63.3	60.9	**-2.4**



Comparing two independent samples – t-test

X_D - The set of differences of between all pairs. $\bar{X}_D, \mathrm{S_D}$ are the mean and the standard deviation of the differences.

$H_0: the\ mean\ differnce\ is\ \mu_0$. (Usually $\mu_0 = 0$)
$H_1: the\ mean\ differnce\ is\ larger\ than\ \mu_0$. (Right tailed…)

$$t = \frac{\bar{X}_D - \mu_0}{S_D/\sqrt{n}}$$

$$p - value = P(T_{n-1} > t)$$

Paired t-test example

ID	Before	After	Diff
1	78.3	83.5	**5.2**
2	72.2	81.2	**9**
3	49.5	47.7	**-1.8**
4	64.9	73.1	**8.1**
5	71.2	75.4	**4.1**
6	57.8	63	**5.2**
7	67.7	65	**-2.7**
8	66.1	69.8	**3.7**
9	65	75.5	**10.4**
10	63.3	60.9	**-2.4**

$\bar{X}_D = 5.2 + 9 + \cdots - 2.4 = 3.89$

$S_D = 4.54$

$\mu_0 = 0$

$$t = \frac{3.89 - 0}{4.54/\sqrt{10}} = 2.57$$

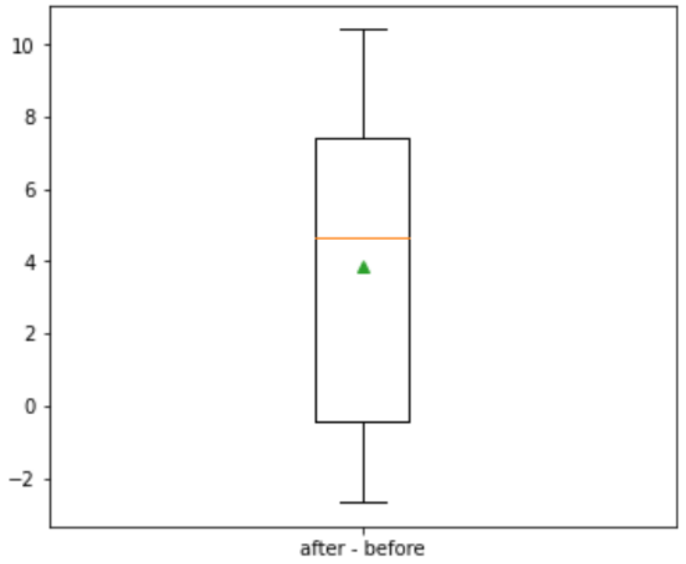
$p - value = P(T_9 > 2.57)$ =0.015

$\boldsymbol{H_0}: \mu_0 = 0$

$\boldsymbol{H_1}: \mu_0 > 0$

$$t = \frac{\bar{X}_D - \mu_0}{S_D/\sqrt{n}}$$

$$p - value = P(T_{n-1} > t)$$



Paired t-test example

$t = 2.57$

$p - value =$0.015

```
# indepedent t-test
t_stat ,pval = ttest_ind(after,before)
print('independet t-test: t={:.2f}, p_val={:.2e}'.format(t_stat,pval))
# paired t-test
t_stat ,pval = ttest_rel(after,before)
print('paired t-test: t={:.2f}, p_val={:.2e}'.format(t_stat,pval))
# one sample t-test
t_stat ,pval = ttest_1samp(diff,0)
print('one sample t-test: t={:.2f}, p_val={:.2e}'.format(t_stat,pval))

# Manual calculation
m_D = diff.mean()
s_D = diff.std(ddof=1)
t_stat = m_D/(s_D/np.sqrt(n))
print('manual calculation: t={:.2f}, p_val={:.2e}'.format(t_stat,t_dist.sf(t_stat,n-1)))
```

```
independet t-test: t=0.92, p_val=3.67e-01
paired t-test: t=2.57, p_val=3.02e-02
one sample t-test: t=2.57, p_val=3.02e-02
manual calculation: t=2.57, p_val=1.51e-02
```