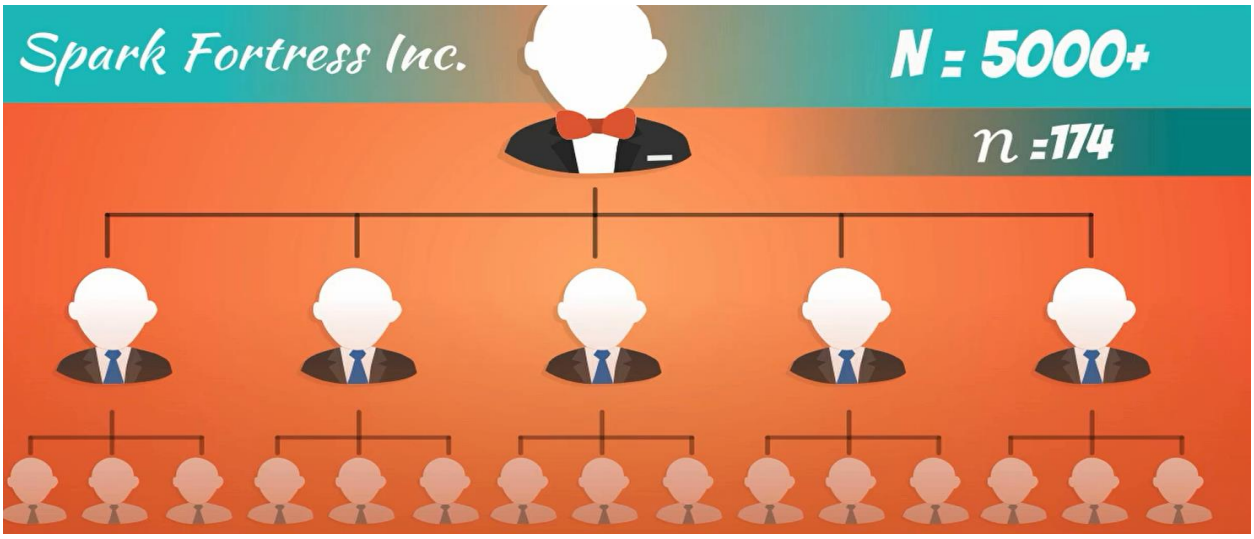
Practical Example. Hypothesis Testing

We explore the topic of gender pay gap.

we will test whether a particular company is discriminating against some of its employees on a gender basis. Our Fictitious company is called Sparke fortress incorporated it is a big company with more than 5000 employees and here will work with a sample of 174 of them.



we have tata showing us their detail.

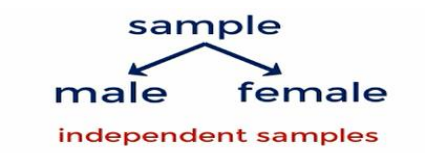
Practical example. Hypothesis testing.
Spark Fortress Inc. HR data

Surname	Name	Age	Gender	Country	Ethnicity	Start_date	Department	Position	Salary
Warfield	Sarah	39	Female	United States	Asian	3/30/2015	IT/IS	Sr. Network Engineer	$114,910.00
South	Joe	52	Male	United States	White	11/10/2014	IT/IS	Sr. Network Engineer	$110,240.00
Boutwell	Bonalyn	30	Female	United States	Asian	2/16/2015	Admin Offices	Sr. Accountant	$72,696.00
Foster-Baker	Amy	38	Female	United States	White	1/5/2009	Admin Offices	Sr. Accountant	$72,696.00
Sweetwater	Alex	51	Male	United States	White	8/15/2011	Software Engineering	Software Engineering Mana	$56,160.00
Del Bosque	Keyla	38	Female	United States	Black or African American	1/9/2012	Software Engineering	Software Engineer	$118,809.60
Carabbio	Judith	30	Female	United States	White	11/11/2013	Software Engineering	Software Engineer	$116,480.00
Martin	Sandra	30	Female	United States	Asian	11/11/2013	Software Engineering	Software Engineer	$115,460.80
Saada	Adell	31	Female	United States	White	11/5/2012	Software Engineering	Software Engineer	$102,440.00
Szabo	Andrew	34	Male	United States	White	7/7/2014	Software Engineering	Software Engineer	$99,840.00
Andreola	Colby	38	Female	United States	White	11/10/2014	Software Engineering	Software Engineer	$99,008.00
LeBlanc	Brandon R	33	Male	United States	White	1/5/2016	Admin Offices	Shared Services Manager	$114,400.00
Smith	John	33	Male	United States	Black or African American	5/18/2014	Sales	Sales Manager	$116,480.00
Daneault	Lynn	27	Female	United States	White	5/5/2014	Sales	Sales Manager	$112,320.00
Moumanil	Maliki	43	Male	United States	Black or African American	5/13/2013	Production	Production Technician II	$60,320.00
Sahoo	Adil	31	Male	United States	White	8/30/2010	Production	Production Technician II	$60,320.00
Blount	Dianna	27	Female	United States	White	4/4/2011	Production	Production Technician II	$56,160.00
Faller	Megan	39	Female	United States	Black or African American	7/7/2014	Production	Production Technician II	$56,160.00
Monkfish	Erasumus	25	Male	United States	White	11/7/2011	Production	Production Technician II	$56,160.00
Nowlan	Kristie	32	Female	United States	White	11/10/2014	Production	Production Technician II	$54,891.20
Lunquist	Lisa	35	Female	United States	Black or African American	8/19/2013	Production	Production Technician II	$54,288.00
Burkett	Benjamin	40	Male	United States	White	4/4/2011	Production	Production Technician II	$54,080.00
McCarthy	Brigit	30	Female	United States	White	3/30/2015	Production	Production Technician II	$54,080.00
Petingill	Shana	38	Female	Australia	Asian	4/2/2012	Production	Production Technician II	$54,080.00
Burke	Joelle	37	Female	United States	Black or African American	3/5/2012	Production	Production Technician II	$52,000.00
Davis	Daniel	38	Male	Australia	Two or more races	11/7/2011	Production	Production Technician II	$52,000.00
Johnston	Yen	48	Female	United States	White	7/7/2014	Production	Production Technician II	$52,000.00
Monterro	Luisa	47	Female	United States	Black or African American	5/13/2013	Production	Production Technician II	$52,000.00
Woodson	Jason	32	Male	United States	White	7/7/2014	Production	Production Technician II	$50,440.00

We are going to test if there is a significant difference in the salaries, based on gender

we are going to test if there is a significant difference in the salaries of employees are paid based on their gender.

Our 174-employee sample could be divided into two sub samples one that is exclusively male and one female. So, we have two samples drawn from the same population that are independent.



although so far, we have worked with different populations only. if the values in one sample reveal no information about the other sample, then they are considered independent.

If the values in one sample reveal no information about the other, they are independent

There are different methodologies to conduct a study and while regression analysis is my preferred one, we will have to wait until next section for that here.

We will use a hypothesis test for mean salary

let's State that two hypotheses

H_0: The average male salary is equal to the average female salary.

H_0: $\mu_m = \mu_f$

or H_0: $\mu_m - \mu_f = 0$

H_1: The average male salary differs from the average female salary.

H_0: $\mu_m \neq \mu_f$

Or H_0: $\mu_m - \mu_f \neq 0$

The test should use is t-test for the independent samples. salary population variance, it is surely unknown and we can assume it is equal.

The test we should use is the t-test for independent samples, var unknown but assumed equal

let's construct a frequency distribution table.

we have 98 females and 76 males. These are our sample sizes. assume that the population variances are equal. We should also compute the Pooled variance.

Practical example. Hypothesis testing.
Is there a gender bias in Spark Fortress

Overall	n	Mean	Sample variance	Pooled variance
Female	98	$ 65,736.91	$² 932,705,380.02	$² 1,025,188,119.29
Male	76	$ 72,300.53	$² 1,144,799,128.75	

$$H_0: \mu_m - \mu_f = 0$$

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

Pooled variance formula

Practical example. Hypothesis testing.
Is there a gender bias in Spark Fortress

Overall	n	Mean	Sample variance	Pooled variance	T-score
Female	98	$ 65,736.91	$² 932,705,380.02	$² 1,025,188,119.29	1.34
Male	76	$ 72,300.53	$² 1,144,799,128.75		

$$H_0: \mu_m - \mu_f = 0$$

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_m - \mu_f)}{\sqrt{\frac{s_p^2}{n_m} + \frac{s_p^2}{n_f}}}$$

t-score formula

Here, t-score =1.34

Degree of freedom = 98 +76-2 =172

Once we have surpassed 50 degrees of freedom. student's t distribution almost completely overlaps with the normal distribution.

Once we have surpassed 50 degrees of freedom, the Student's T ~ Normal distribution

Thus, the p-values for t-score of 1.34 and Z -score of 1.34 it will be Virtually the same. So, I will just give you the P-value is 0.182. The p-value is much greater than all common levels of significance.

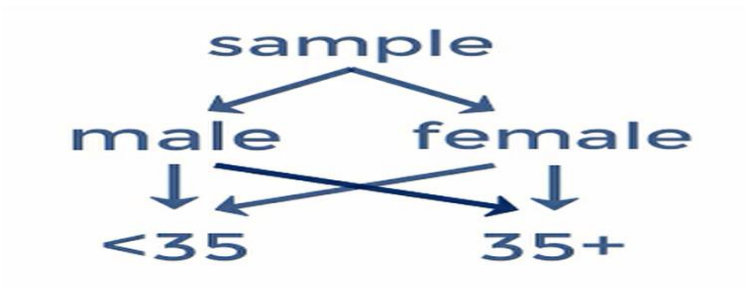
0.182 >> all common levels of significance

we conclude that we cannot reject the null hypothesis.

We cannot reject the null hypothesis =>
there isn't enough statistical evidence that there is a gender wage gap in this firm

let's order the salaries from largest to the smallest. We can see that the highest paid employee is actually the president and CEO of the company. Caroline bolds Who is female.

I would normally for the segment the data. let's divided the employees into two more groups below 35 and above 35. This will give us valuable information about the wage equality of younger and older staff.



I have created two more data sets that are based on the original one. Let's run the same tests as before, but this time we will do it in our segmented data.

The hypothesis is the same.

We get for these true tests is a test score of 0.43 for employees below 35 and 2.0 for employees over 35. the corresponding p-values are 0.668 and 0.048.

Practical example. Hypothesis testing.
Is there a gender bias in Spark Fortress

Overall	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	98	$ 65,736.91	$² 932,705,380.02	$² 1,025,188,119.29	1.34	0.182
Male	76	$ 72,300.53	$² 1,144,799,128.75			

Below35	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	46	$ 66,775.23	$² 1,063,144,850.86	$² 1,048,675,919.03	0.43	0.668
Male	37	$ 69,859.89	$² 1,030,589,754.25			

Over 35	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	52	$ 63,448.22	$² 1,129,668,678.18	$² 1,210,601,529.12	2.00	0.048
Male	40	$ 78,049.60	$² 1,316,436,795.73			

$$H_0: \mu_m - \mu_f = 0$$
$$H_1: \mu_m - \mu_f \neq 0$$

what these numbers mean that is the group below 35 there is Virtually no wage gap on a gender basis. In the older group P value is 0.048 this is very close to 0.05 but still below it. This implies that at 95% significance we reject the null hypothesis. therefore, a wage gap does exist for older employees.

Practical example. Hypothesis testing.
Is there a gender bias in Spark Fortress

$$H_0: \mu_m - \mu_f = 0$$

Overall	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	98	$ 65,736.91	$² 932,705,380.02	$² 1,025,188,119.29	1.34	0.182
Male	76	$ 72,300.53	$² 1,144,799,128.75			

Below35	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	46	$ 66,775.23	$² 1,063,144,850.86	$² 1,048,675,919.03	0.43	0.668
Male	37	$ 69,859.89	$² 1,030,589,754.25			

there is virtually no wage gap on a gender basis

Over 35	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	52	$ 63,448.22	$² 1,129,668,678.18	$² 1,210,601,529.12	2.00	0.048
Male	40	$ 78,049.60	$² 1,316,436,795.73			

at 5% significance, we reject the null hypothesis
a wage gap does exist for older employees

this is a two-sided test so we are not sure who gets more money. Well do you remember the nifty track the t score of 2 is positive. therefore, the difference in pay is positive in favor of males.

The T-score (2.00) > 0 => the difference in pay is positive (males earn more)

The limitation of this analysis.

Limitations (we omitted important factors):
1. Position
2. Ethnicity
... etc.

so, we are not completely sure what's going on in the frame but we can say that overall, there is no wage kept in Sparke Fortress and this is driven by wage equality among young employees.

Overall, there is no wage gap in Spark Fortress (driven by younger employees)