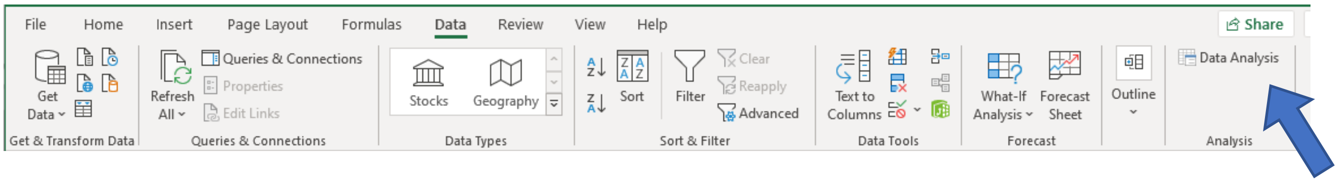
Basic Statistical Analysis in Excel

1. Enable ANALYSIS TOOLPAK

File → Options → Add-Ins → Manage → Go → Analysis ToolPak



2. Descriptive Statistics

Height of Students

Gender	Height (m)
Female	1.75
Female	1.59
Female	1.88
Female	1.65
Female	1.70
Female	1.69
Female	1.68
Female	1.80
Male	1.89
Male	1.64
Male	1.89
Male	1.94
Male	1.93
Male	1.74
Male	1.82
Male	1.91

Data -→ Data Analysis → Descriptive Statistics → Summary

Height (m)	
Mean	1.78125
Standard Error	0.02892339
Median	1.775
Mode	1.89
Standard Deviation	0.115693561
Sample Variance	0.013385
Kurtosis	-1.466948258
Skewness	-0.072961873
Range	0.35
Minimum	1.59
Maximum	1.94
Sum	28.5
Count	16

Age Range of Employees.

Data -→ Data Analysis → Descriptive Statistics → Summary

	A	B
1	*age*	
2		
3	Mean	39.1884058
4	Standard Error	1.482819393
5	Median	39
6	Mode	29
7	Standard Deviation	12.31722295
8	Sample Variance	151.7139812
9	Kurtosis	-1.124885205
10	Skewness	0.084622833
11	Range	45
12	Minimum	18
13	Maximum	63
14	Sum	2704
15	Count	69

3. Correlation

Relationship between ability in maths and sports.

Pupil No	Sports	Maths
1	15	24
2	19	29
3	12	20
4	14	27
5	9	16
6	16	28
7	20	28
8	14	23
9	15	21
10	19	29

Data -→ Data Analysis → Correlation

	A	B	C
1		*Sports*	*Maths*
2	Sports	1	
3	Maths	0.88398	1

Yes, there is a very strong relationship (r = 0.89).

4. Regression

Can we predict reading ability from IQ score? What is the relationship?

Pupil No	IQ	READING
1	98	54
2	115	74
3	124	81
4	97	59
5	135	88
6	105	70
7	117	78
8	101	73
9	110	84
10	127	86

Data -→ Data Analysis → Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.853166
R Square	0.727893
Adjusted R Square	0.693879
Standard Error	6.232712
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	831.3264139	831.3264	21.40018	0.001696799
Residual	8	310.7735861	38.8467		
Total	9	1142.1			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-8.82474	18.16262859	-0.48587	0.640085
IQ	0.739812	0.159923566	4.626033	0.001697

Yes, there is a very strong relationship (r=0.853, p = 0.0017)

Therefore: Reading = 0.74 x IQ – 8.82

Multiple Regression

ID	IQ Score	Books Avaliable	Reading Score
1	98	5	54
2	115	25	74
3	124	37	81
4	97	10	59
5	135	90	88
6	105	80	70
7	117	88	78
8	101	73	73
9	110	80	84
10	127	100	86

Data -→ Data Analysis → Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932165375
R Square	0.868932287
Adjusted R Squ	0.831484369
Standard Error	4.624352231
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	992.4075651	496.2037826	23.20375429	0.000815145
Residual	7	149.6924349	21.38463355		
Total	9	1142.1			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.478369037	14.58373096	0.444218908	0.670292701	-28.00667486	40.96341294	-28.00667486	40.96341294
IQ Score	0.531111151	0.140930393	3.768606193	0.006995669	0.197863726	0.864358575	0.197863726	0.864358575
Books Avaliabl	0.14046228	0.051178583	2.744551953	0.028728858	0.019444161	0.261480398	0.019444161	0.261480398

Yes, there is a very strong relationship (r=0.932, p = 0.0008)

Therefore: Reading = 6.47 + -531 x IQ + 0.14 x Books

We can add more. How about male or female differences (male = 1, Female = 2).

	A	B	C	D	E	F
1	ID	IQ	Books	Sex	Reading	
2	1	98	5	1	54	
3	2	115	25	2	74	
4	3	124	37	2	81	
5	4	97	10	1	59	
6	5	135	90	1	88	
7	6	105	80	1	70	
8	7	117	88	2	78	
9	8	101	73	2	73	
10	9	110	80	1	84	
11	10	127	100	2	86	
12						

Data -→ Data Analysis → Regression

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.935617027							
5	R Square	0.875379221							
6	Adjusted R Square	0.813068831							
7	Standard Error	4.87047896							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	999.7706	333.2569	14.04869	0.004030636			
13	Residual	6	142.3294	23.72157					
14	Total	9	1142.1						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	6.386250241	15.36083	0.415749	0.692053	-31.200337	43.97284	-31.2003	43.97284
18	IQ	0.507893834	0.15417	3.294369	0.016523	0.130652668	0.885135	0.130653	0.885135
19	Books	0.140456372	0.053903	2.605748	0.04035	0.008561668	0.272351	0.008562	0.272351
20	Sex	1.809134187	3.247235	0.557131	0.597591	-6.136564758	9.754833	-6.13656	9.754833
21									

Yes, there is a very strong relationship (r=0.936, p = 0.004)

Therefore: Reading = 6.39 + -51 x IQ + 0.14 x Books + 1.8 x Sex

Note that Sex is a categorical variable with two levels (Males = 1 & Females = 2).

In this case the regression weighting (1.8) is added to the predicted Reading value. So for females 1.8 (1.8 x 1) would be added and for males 3.6 (1.8 x 2) would be added.

Which variables are important?

We cannot compare these coefficients as they use different scales. We need to standardize them – put them on the same scale. We will then have the standardised regression coefficients, which we can compare.

=STANDARDIZE(B2:B11,AVERAGE(B2:B11),STDEV(B2:B11))

	A	B	C	D	E	F	G	H	I
1	ID	IQ	Books	Sex	Reading	IQ Standardised	Books Standardised	Sex Standardised	Reading Standardised
2	1	98	5	1	54	-1.146945897	-1.503912381	-0.948683298	-1.837551562
3	2	115	25	2	74	0.161650093	-0.944837146	0.948683298	-0.062139425
4	3	124	37	2	81	0.854436205	-0.609392006	0.948683298	0.559254823
5	4	97	10	1	59	-1.223922132	-1.364143572	-0.948683298	-1.393698528
6	5	135	90	1	88	1.701174787	0.872157366	-0.948683298	1.180649072
7	6	105	80	1	70	-0.608112254	0.592619748	-0.948683298	-0.417221852
8	7	117	88	2	78	0.315602562	0.816249842	0.948683298	0.292943003
9	8	101	73	2	73	-0.916017193	0.396943416	0.948683298	-0.150910032
10	9	110	80	1	84	-0.223231081	0.592619748	-0.948683298	0.825566644
11	10	127	100	2	86	1.085364909	1.151694983	0.948683298	1.003107858

Run regression function on standardised values.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.935617027							
5	R Square	0.875379221							
6	Adjusted R Square	0.813068831							
7	Standard Error	0.432355373							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	3	7.878413	2.626138	14.04869	0.004031			
13	Residual	6	1.121587	0.186931					
14	Total	9	9						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-1.38519E-17	0.136723	-1E-16	1	-0.33455	0.334549	-0.33455	0.334549
18	IQ Standardised	0.585713812	0.177792	3.294369	0.016523	0.150671	1.020756	0.150671	1.020756
19	Books Standardised	0.446036475	0.171174	2.605748	0.04035	0.027189	0.864884	0.027189	0.864884
20	Sex Standardised	0.084642546	0.151926	0.557131	0.597591	-0.28711	0.456392	-0.28711	0.456392

You now have the standardised regression coefficients, which you can now compare.

IQ: 0.58571381

Sex: 0.08464255

Books Available: 0.44603647

And you can see that the most important predictor is IQ, followed closely by Books Available. Sex is not contributing very much.

5. t-Test

Is there a difference between oral announcement of objects and memory?

No announcement:	Announcement:
3	5
8	7
5	8
6	9
3	5
5	4
9	9
3	8
3	7
4	4
7	
4	

Data -→ Data Analysis → t-Test Two Sample Assuming Equal Variances

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*No announcement*	*Announcement:*
4	Mean	5	6.6
5	Variance	4.363636364	3.822222222
6	Observations	12	10
7	Pooled Variance	4.12	
8	Hypothesized Me	0	
9	df	20	
10	t Stat	-1.840986726	
11	P(T<=t) one-tail	0.040253936	
12	t Critical one-tail	1.724718243	
13	P(T<=t) two-tail	0.080507873	
14	t Critical two-tail	2.085963447	
15			

No, there is no relationship between oral announcement of objects and memory (p = 0.08).

6. ANOVA – One-Way

Is there a difference betweem amount of alcoohol consumed and stopping distance?

NO ALCOHOL	ONE UNIT	TWO UNITS	THREE UNITS
31.2	31.2	42.1	46.7
27.4	33.7	40.6	45.1
29.8	34.3	43.4	43.2
33.5	39.2	37.9	41.7
34	36	39	40.3

Stopping distance in metres.

Data -→ Data Analysis → ANOVA Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
NO ALCOHOL	5	155.9	31.18	7.382
ONE UNIT	5	174.4	34.88	8.797
TWO UNITS	5	203	40.6	4.985
THREE UNITS	5	217	43.4	6.58

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	456.1295	3	152.0432	21.92087	6.52E-06	3.238872
Within Groups	110.976	16	6.936			
Total	567.1055	19				

Yes ($p = 6.52E\text{-}06 = 6.52 \times 10^{-6} = 0.00000652 < 0.05$). Therefore there is a significant differencein stopping distance according to amount of alcohol consumed.

7. ANOVA – Two-Way

Is there a difference between gender, amount of alcohol consumed and stopping distance?

Gender	No Alcohol	1 Unit	2 Units	3 Units
Male	31.20	31.20	42.10	46.70
Male	27.40	33.70	40.60	45.10
Male	29.80	34.30	43.40	43.20
Male	33.50	39.20	37.90	41.70
Male	34.00	36.00	39.00	40.30
Female	35.10	32.80	38.10	39.80
Female	32.30	35.90	35.90	31.70
Female	35.40	36.10	40.20	41.20
Female	38.10	40.90	39.80	39.10
Female	35.90	38.20	37.60	37.90

Stopping distance in metres.

Data -→ Data Analysis → ANOVA Two-Factor with Replication

Anova: Two-Factor With Replication						
SUMMARY	No Alcohol	1 Unit	2 Units	3 Units	Total	
Male						
Count	5	5	5	5	20	
Sum	155.9	174.4	203	217	750.3	
Average	31.18	34.88	40.6	43.4	37.515	
Variance	7.382	8.797	4.985	6.58	29.84766	
Female						
Count	5	5	5	5	20	
Sum	176.8	183.9	191.6	189.7	742	
Average	35.36	36.78	38.32	37.94	37.1	
Variance	4.308	9.017	3.037	13.593	7.707368	
Total						
Count	10	10	10	10		
Sum	332.7	358.3	394.6	406.7		
Average	33.27	35.83	39.46	40.67		
Variance	10.049	8.92011111	5.009333	17.24678		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	1.72225	1	1.72225	0.238791	0.628413	4.149097
Columns	344.24075	3	114.7469	15.90973	0.000002	2.90112
Interaction	138.50875	3	46.16958	6.40144	0.001603	2.90112
Within	230.796	32	7.212375			
Total	715.26775	39				

Sample = Row factor = Gender – not significant (p=0.63)
Columns = Amount of Alcohol Consumed – significant (p=0.000002 < 0.05
Interaction = that the influence of alcohol on stopping distance depends on gender – significant (p=0.0016)

8. Chi-Squared (χ^2)

(Test for independence}

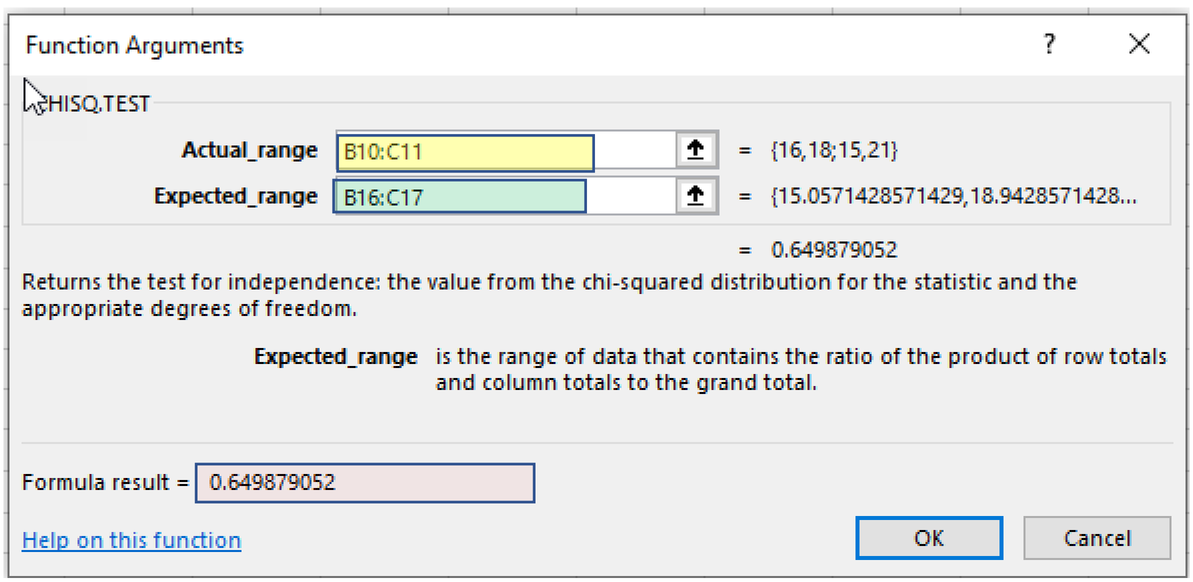
Is there a significant difference between the proportion of males and females who attended the firm's meeting last month? First we need to calculate the expected values, then use CHISQ.TEST

8	**Observed**			
9	Attended me	Male	Female	Total
10	No	16.00	18.00	34
11	Yes	15.00	21.00	36
12	Total	31	39	70
13				
14	**Expected**			
15	Attended me	Male	Female	Total
16	No	15.06	18.94	34
17	Yes	15.94	20.06	36
18	Total	31	39	70

Expected Value =

$$\frac{\text{Row Total x ColumnTotal}}{\text{Grand Total}}$$

Formulas → More Functions → Statistical → CHISQ.TEST



p = $\boxed{0.649879052,}$ which > 0.05, therefore not significant.

So, we cannot conclude that there is a significant difference in attendance at the meeting between male and female employees.

Calculating the Chi-Squared (χ^2) statistic

If we want the Ch-Squared (χ^2) statistic, we need to use the formula for each cell for Observed (O) and Expected (E) values: $(O_{ij}-E_{ij})^2/E_{ij}$

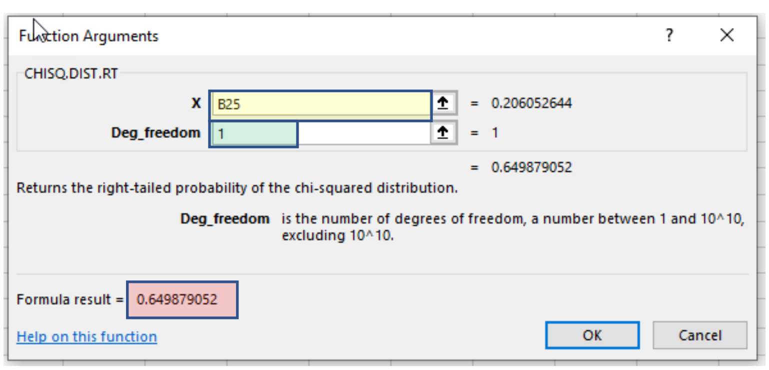
$(O-E)^2/E$		
Attended Meeting	Male	Female
No	0.05904039	0.04693
Yes	0.055760369	0.044322

Then the Chi-Squared (χ^2) statistic is the sum:
$\chi^2 = \Sigma(O_{ij}-E_{ij})^2/E_{ij} = 0.206052644$

	A	B	C	D
8	**Observed**			
9	Attended meeting	Male	Female	Total
10	No	16	18	34
11	Yes	15	21	36
12	Total	31	39	70
13				
14	**Expected**			
15	Attended meeting	Male	Female	Total
16	No	=D16*B18/D18	=D16*C18/D18	34
17	Yes	=D17*B18/D18	=D17*C18/D18	36
18	Total	31	39	70
19				
20	$(O-E)^2/E$			
21	Attended meeting	Male	Female	
22	No	=(B10-B16)^2/B16	=(C10-C16)^2/C16	
23	Yes	=(B11-B17)^2/B17	=(C11-C17)^2/C17	
24				
25	$\chi^2 = \Sigma(O_{ij}-E_{ij})^2/E_{ij}$	= SUM(B22:C23)		
26				

We can then calculate the degrees of freedom (df) = (no of rows – 1) x (no of columns – 1) = 1.

And then calculate the significance with
Formulas → More Functions → Statistical → CHISQ.DIST.RT



p = CHISQ.DIST.RT(χ^2, df) = 0.649879052, which > 0.05 – as before.