Assignment 1 Submission by Krishna Damarla

Note: Please find the attached Assignment1_KD_DataAnalysis.xlsx workbook (has 6 worksheets supporting the solutions to the assignment problems in this PDF document). Please consider that I am using a German version of Excel for doing the data analysis. Hence, 10.22 (US metric) in the Excel would look like 10,22 (German metric). Calculations/results are all the same. Just that the notation is a bit different. Thanks for your understanding.

1a) Interval Estimate for Physical and E-Books

Answer all questions at the 95% confidence level.

1. A book publishing company is interested in keeping track of whether their customers are more inclined to read physical or electronic copies of books. They survey 150 customers, asking them how many physical books and ebooks they have read in the past twelve months. Use the dataset book_type.xlsx to answer the following questions:

    (a) What is the estimate for the mean value of the number of physical books and the number of ebooks their customer's read in the past year? Give your answer both as an interval estimate (value ± margin of error) and as a confidence interval (LCL, UCL).

To estimate the mean values for the number of physical books and ebooks that the customers have read in the past year, we calculate the interval estimate, confidence interval of the mean.

Method 1 : Manual calculation using the below formula

Interval Estimate = $\bar{x} \pm$ t_critical (standard_error).

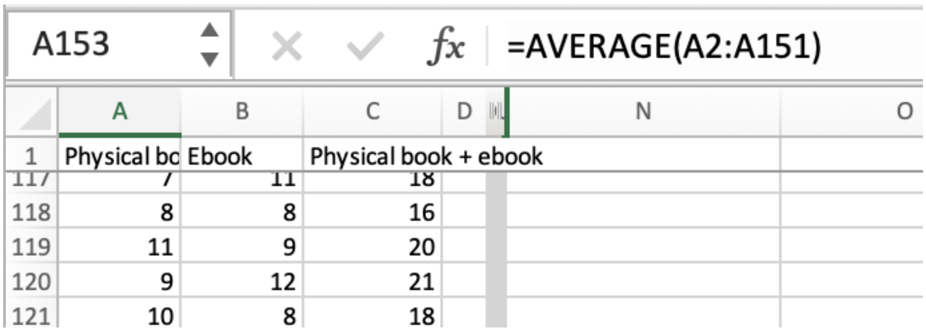
Where,

standard_error = Sample_Standard_Deviation / $\sqrt{n}$
$\bar{x}$ = Sample Mean
n = Sample size

Data = Numerical type
population parameter = μ
Population = 1

For Physical books:

$\bar{x}p$ = Mean of the data (A2: A151 from below excel) for Physical book = 10.23

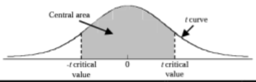

A153 =AVERAGE(A2:A151)

	A	B	C	D	N	O
1	Physical bo	Ebook	Physical book + ebook			
117	7	11	18			
118	8	8	16			
119	11	9	20			
120	9	12	21			
121	10	8	18			

Standard_error = sample_standard_deviation / sqrt (n)
= 1.963241624 / sqrt (150)
= 0.1603

T_critical value at 95% confidence level for 149 degrees of freedom (df = n-1) is 1.975 from below t_critical values table

t critical values



Confidence area captured:		0.90	0.95	0.98	0.99
Confidence level:		90%	95%	98%	99%
Degrees of Freedom	1	6.31	12.71	31.82	63.66
	2	2.92	4.30	6.97	9.93
	3	2.35	3.18	4.54	5.84
	4	2.13	2.78	3.75	4.60
	5	2.02	2.57	3.37	4.03
	6	1.94	2.45	3.14	3.71
	7	1.90	2.37	3.00	3.50
	8	1.86	2.31	2.90	3.36
	9	1.83	2.26	2.82	3.25
	10	1.81	2.23	2.76	3.17
	11	1.80	2.20	2.72	3.11
	12	1.78	2.18	2.68	3.06
	13	1.77	2.16	2.65	3.01
	14	1.76	2.15	2.62	2.98
	15	1.75	2.13	2.60	2.95
	16	1.75	2.12	2.58	2.92
	17	1.74	2.11	2.57	2.90
	18	1.73	2.10	2.55	2.88
	19	1.73	2.09	2.54	2.86
	20	1.73	2.09	2.53	2.85
	21	1.72	2.08	2.52	2.83
	22	1.72	2.07	2.51	2.82
	23	1.71	2.07	2.50	2.81
	24	1.71	2.06	2.49	2.80
	25	1.71	2.06	2.49	2.79
	26	1.71	2.06	2.48	2.78
	27	1.70	2.05	2.47	2.77
	28	1.70	2.05	2.47	2.76
	29	1.70	2.05	2.46	2.76
	30	1.70	2.04	2.46	2.75
	40	1.68	2.02	2.42	2.70
	60	1.67	2.00	2.39	2.66
	70	1.67	1.99	2.38	2.65
	80	1.66	1.99	2.37	2.64
	90	1.66	1.99	2.37	2.63
	100	1.66	1.98	2.36	2.63
	1000	1.65	1.96	2.33	2.58
z critical values	∞	1.645	1.96	2.33	2.58
α for 2-tailed tests		0.10	0.05	0.02	0.01
α for 1-tailed tests		0.05	0.025	0.01	0.005

Margin of Error = t_critical * (standard_error) = 1.975*0.1603 = 0.31659

Interval Estimate = $\bar{x} \pm$ t_critical (standard_error) = 10.22666667 $\pm$ 1.975*0.1603 = 10.22666667 $\pm$ 0.31659

Confidence Interval **= (**LCL, UCL) = (9.910, 10.543)

For e-books:

$\bar{x}e$ = Mean of the data (B2: B151 from below excel) for e-books = 10.77

SUM =AVERAGE(B2:B151)

	A	B	C	D	N
1	Physical bo	Ebook	Physical book + ebook		
17	7	11	18		
18	8	8	16		
19	11	9	20		
20	9	12	21		
21	10	8	18		
22	11	15	26		
23	6	8	14		
24	12	14	26		
25	8	13	21		
26	9	5	14		
27	13	10	23		
28	5	9	14		
29	6	9	15		
30	12	11	23		
31	9	11	20		
32	8	16	24		
33	7	11	18		
34	10	14	24		
35	8	10	18		
36	9	12	21		
37	15	11	26		
38	9	15	24		
39	8	17	25		
40	8	17	25		
41	10	12	22		
42	9	11	20		
43	12	13	25		
44	11	11	22		
45	10	12	22		
46	9	9	18		
47	11	9	20		
48	10	12	22		
49	9	11	20		
50	12	6	18		
51	11	13	24		
52	1534				
53	10,22667	=AVERAGE(B2:B151)			
54	1,963242	2,719916			

Standard_error = sample_standard_deviation / sqrt (n)
= 2,719916/ sqrt (150)
= 0.222

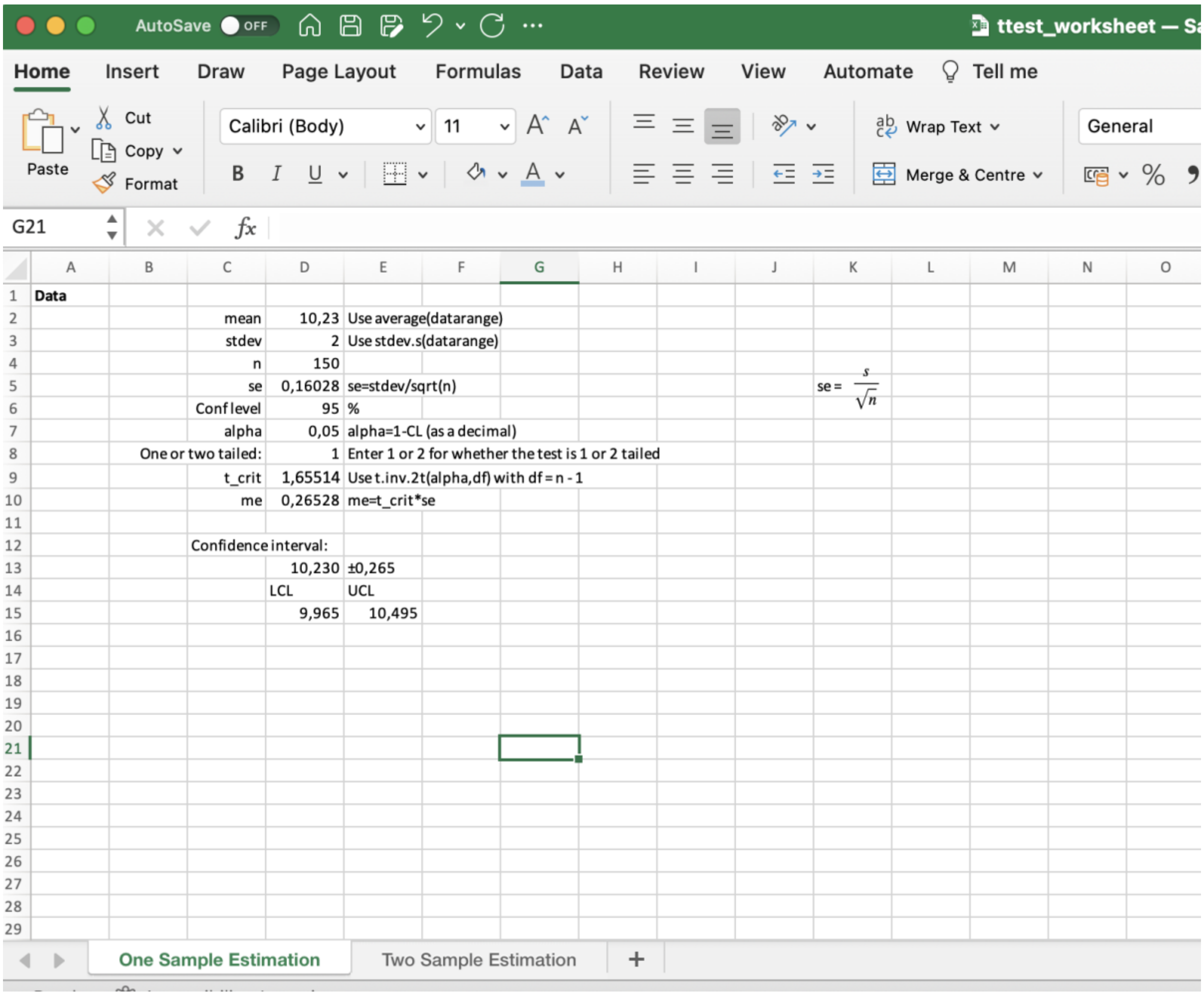
T_critical value at 95% confidence level for 149 (=150 -1) degress of freedom is 1.975 from below t_critical values table

Margin of Error = t_critical * (standard_error) = 1.975*0.222 = 0.4386

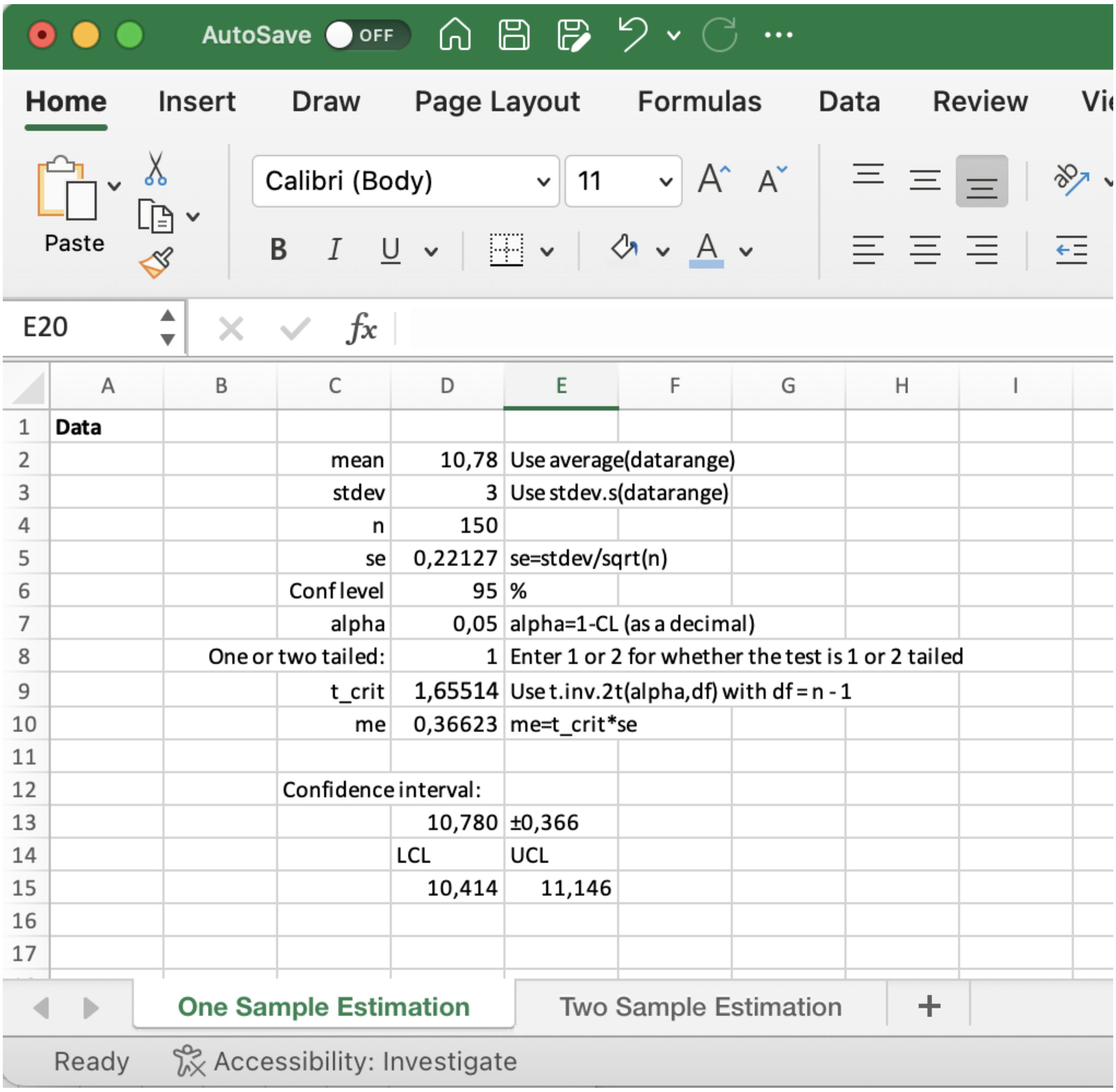
Interval Estimate = $\bar{x} \pm$ t_critical (standard_error) = 10.7733 $\pm$ 0.4386

Confidence Interval = **(**LCL, UCL) **=** (10.335, 11.212)

Method 2: Using the t_test_worksheet one-sample estimation provided in the class resources as shown below. The interval estimate & Confidence interval for physical books & e-boooks as shown in below excel screenshots is close to what we calculated in method 1.



	A	B	C	D	E	K
1	**Data**					
2			mean	10,23	Use average(datarange)	
3			stdev	2	Use stdev.s(datarange)	
4			n	150		
5			se	0,16028	se=stdev/sqrt(n)	$se = \frac{s}{\sqrt{n}}$
6			Conf level	95	%	
7			alpha	0,05	alpha=1-CL (as a decimal)	
8			One or two tailed:	1	Enter 1 or 2 for whether the test is 1 or 2 tailed	
9			t_crit	1,65514	Use t.inv.2t(alpha,df) with df = n - 1	
10			me	0,26528	me=t_crit*se	
11						
12			Confidence interval:			
13				10,230	±0,265	
14				LCL	UCL	
15				9,965	10,495	



	A	B	C	D	E	F	G	H	I
1	**Data**								
2			mean	10,78	Use average(datarange)				
3			stdev	3	Use stdev.s(datarange)				
4			n	150					
5			se	0,22127	se=stdev/sqrt(n)				
6			Conf level	95	%				
7			alpha	0,05	alpha=1-CL (as a decimal)				
8			One or two tailed:	1	Enter 1 or 2 for whether the test is 1 or 2 tailed				
9			t_crit	1,65514	Use t.inv.2t(alpha,df) with df = n - 1				
10			me	0,36623	me=t_crit*se				
11									
12			Confidence interval:						
13				10,780	±0,366				
14				LCL	UCL				
15				10,414	11,146				
16									
17									

1b) Hypothesis test to determine difference between number of Physical and E-Books

(b) Run a hypothesis test to determine if there is a difference between the number of physical books and the number of ebooks. Your work should include the following steps:

- Explain which test you need to use.
- State your hypotheses.
- Calculate the appropriate test statistic.
- Calculate the p-value that corresponds to the test statistic.
- Interpret the p-value and draw a conclusion from your results.

To determine if there is a statistically significant difference between the number of physical books and ebooks, we follow below steps:

Population = 2
Data = Numerical
Parameter = μ

1. **Test type**
   We perform a two-sample t-test for the difference in means. Because we have dependent/related data of 2 populations (physical books, ebooks) read by the same group of customers.

2. **Hypothesis**

   Null Hypothesis (H_0): There is no difference between the mean number of physical books & mean number of ebooks read in the past year. μ_physical = μ_ebook.

   Alternative Hypothesis (H_1): There is a difference between the mean number of physical books & mean number of ebooks read in the past year.
   μ_physical ≠ μ_ebook.

3. **Determining the appropriate test_statistic**
   Perform a F-test to determine a two-sample t-test with equal variance (or) without equal variance

F-Test Two-Sample for Variances		
	Physbook	*Ebook*
Mean	10,22666667	10,77333333
Variance	3,854317673	7,397941834
Observations	150	150
df	149	149
F	0,520998645	
P(F<=f) one-tail	4,10126E-05	
F Critical one-tail	0,763100731	

At 95% confidence level => $\propto = 1 - 0.95 = 0.05$

From above F-Test calculation in the excel, we understand that P-value = 4,10126E-05 = 0.00004.

P is clearly less than ∝. So, We reject the null hypothesis and conclude that variances of two populations are not equal.

Hence, we proceed with test_statistic for 2 sample t-test with unequal variances

Test_statistic for 2 sample t-test with unequal variances = ($\bar{x}p$ – $\bar{x}e$) / (Standard_error)

$$= (\bar{x}p - \bar{x}e) \; / \; \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

= - 0.5467 / 0.274 = -1.995

4. **P-value calculation for 2-sample t-test with unequal variances**

   From the below 2 sample t-test with unequal variances analysis in excel, we can identify the p-value is approximately 0.047

t-Test: Two-Sample Assuming Unequal Variances		
	Physbook	*Ebook*
Mean	10,22666667	10,77333333
Variance	3,854317673	7,397941834
Observations	150	150
Hypothesized Mean Diffe	0	
df	271	
t Stat	-1,995944004	
P(T<=t) one-tail	0,023470337	
t Critical one-tail	1,650495779	
P(T<=t) two-tail	0,046940674	
t Critical two-tail	1,968756314	

5. Conclusion

At 95% confidence level, P-value for two-tail is approximately 0.047 which is less than $\propto of$ 0.05.

Hence, we reject the null hypothesis and conclude that there is a difference between the mean number of physical books & mean number of ebooks read in the past year.

1c) Do more than 2/3 of customers read ≥ 20 books/year. Estimate for proportion of customers

(c) Do more than two-thirds of the company's customers read twenty or more books a year? What is your best estimate for the minimum value of the proportion of customers who have read twenty or more books in the last twelve months?

Q1: Do more than two-thirds of the company's customers read twenty or more books a year ?

Data = categorical
Population parameter = π
Point estimate = p

Test type: 1-sample Z-test (one-sided). As we are discussing here about proportion of customers and the data type is categorical, we are proceeding with 1-sample Z-test. For greater than or less than type comparisons, we use a one-sided test.

Sample proportion (p) = Number of customers who have read 20+ books / Total number of customers = 107 /150 = 0.7133

Physbook	Ebook	PHYSBOOK + EBOOK
9	11	20
12	6	18
11	13	24
1534		107
10,226667	10,773333	
1,9632416	2,7199158	

Null Hypothesis (H0) => two-thirds or less than two-thirds of the company's customers read twenty or more books a year <= 2/3 (100) => **<= 66.66%**

Alternative Hypothesis (H1) => More than two-thirds of the company's customers read twenty or more books a year => **> 66.66%**

Using the Z_test worksheet given in the class resources. Choosing 1 sample test with one tail (or) one-sided test. We got the p-value as 0,12.

Conclusion: As the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that more than two-thirds of the company's customers read twenty or more books a year.

	A	B	C	D	E	F	G	H
1	Enter the following values:							
2	Number of Successes:		107					
3	Sample Size:		150					
4	Confidence Level:		95			Confidence Interval:	0,713	±-0,072
5	Null Hypothesis Value:		0,67			*p*-value:	0,12027073	
6	One or Two Sided Test:		1					
7								
8								

One sample test | Two sample test | +

Ready   Accessibility: Good to go

Q2: Best estimate for minimum value of proportion of customers

Method 1: From the Z_test worksheet above, confidence interval = 0.713 $\pm$ 0.072 = (0.641, 0.785). Best estimate for minimum value or Lower confidence interval is 0.641.

Method 2:
Confidence Interval = $p \pm$ z_critical (standard_error)

Standard_error = $$se = \sqrt{\frac{p(1-p)}{n}}$$

Standard_error = sqrt(0.2045/150) = sqrt(0.00136) = 0.0368

At 95% confidence level, the critical z-score is approximately 1.96 from the below table.

Confidence level	Critical (z) value to be used in confidence interval calculation
50%	0.67449
75%	1.15035
90%	1.64485
95%	1.95996
97%	2.17009
99%	2.57583
99.9%	3.29053

Margin of Error = critical_z-score * (standard_error) = 1.96 * 0.0368 = 0.0723

Confidence Interval (CI) = $p \pm$ Margin of Error = 0.7133 $\pm$ 0.0723 = (0.641, 0.7856)

Minimum value of the proportion of customers who have read 20+ books in last 12 months = LCL (Lower CI) from above CI = 0.641

2a) Run appropriate test. Draw conclusion on mortgage payments

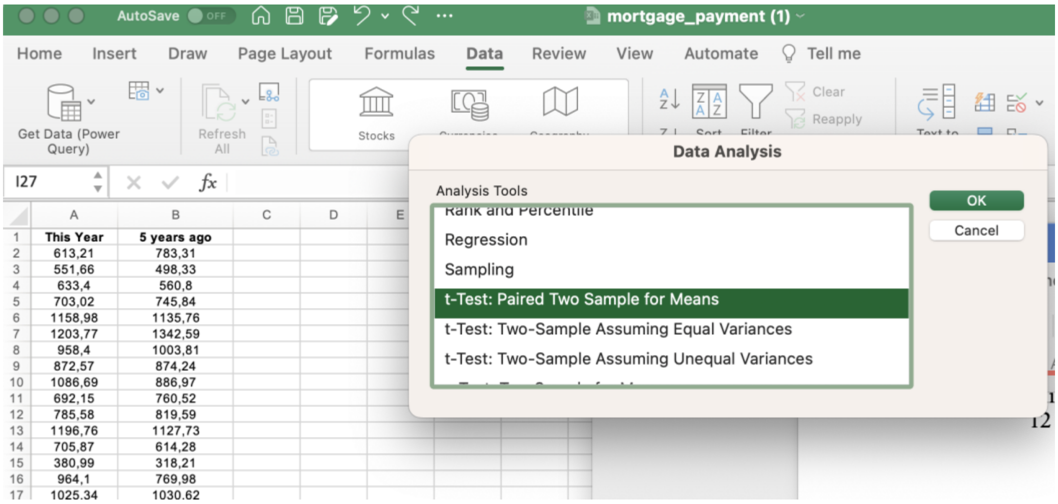
2. The file mortgage_payments.xlsx shows two ramdom samples, one of mortgage payments from this year, the other of mortgage payments from five years ago.

   (a) Assuming that both samples were collected from homeowners living in the same house as they were five years ago, what type of test would you run to see if there is a difference between mortgage payments now and five years ago? Run the appropriate test and draw a conclusion from your results.

Population = 2 (this year, 5 years ago)
Data = Numerical
Parameter = μ

Samples: Random ? Yes
independant? No. We are assuming that both samples were collected from homeowners living in the same house as they were five years ago. Hence, we choose t-test: Paired two samples for means (or) Matched pairs t-test

H0 (Null hypothesis) = There is no difference between mortgage payments now & 5 years ago
H1 (Alternative hypothesis) = There is difference between mortgage payments now & 5 years ago



t-Test: Paired Two Sample for Means		
	This Year	*5 years ago*
Mean	937,9098667	925,5089333
Variance	34193,13389	44637,76851
Observations	150	150
Pearson Correlation	0,883101999	
Hypothesized Mean Dif	0	
df	149	
t Stat	1,531958697	
P(T<=t) one-tail	0,063826543	
t Critical one-tail	1,655144534	
P(T<=t) two-tail	0,127653086	
t Critical two-tail	1,976013178	

T_statistic = 1.531 < T_critical (2-tail) = 1.97 => P(2 tail) = 0.127 > $\propto = 0.05$

Conclusion:
At 95% confidence level, P value for two-tail is 0.127 is greater than $\propto of$ 0.05

If P-value > ∝ => we donot reject the Null hypothesis and conclude that there is no difference between mortgage payments now & 5 years ago

2b) Conclusion from a different test on mortgage payments

(b) What if, instead of being from homeowners in the same house, each sample is a random sample of local homeowners, but there is no connection between homeowners included in this years sample versus the sample from five years ago. How would that change both the test you run to determine if the mortgage payments are different? Does your conclusion change as a result of running a different test?

Population = 2 (this year, 5 years ago)
Data = Numerical
Parameter = μ
Samples: Random ? Yes

independent? Yes. We are assuming that there is no connection between homeowners in this year's random samples with those from five years ago. Hence, we choose a 2-sample t-test with independent samples.

We perform f-test to check for equal variance as below:

F-Test Two-Sample for Variances		
	This Year	*5 years ago*
Mean	937,9098667	925,5089333
Variance	34193,13389	44637,76851
Observations	150	150
df	149	149
F	0,766013514	
P(F<=f) one-tail	0,052429063	
F Critical one-tail	0,763100731	

At 95% confidence level, P = 0.0524 > ∝ = 0.05.

If P-value > ∝ => $We\ do\ not$ reject Null hypothesis and conclude that variances of two samples are equal. Hence, we proceed with 2 sample t-test with equal variances

t-Test: Two-Sample Assuming Equal Variances		
	This Year	*5 years ago*
Mean	937,9098667	925,5089333
Variance	34193,13389	44637,76851
Observations	150	150
Pooled Variance	39415,4512	
Hypothesized Me	0	
df	298	
t Stat	0,540943309	
P(T<=t) one-tail	0,294475389	
t Critical one-tail	1,649982976	
P(T<=t) two-tail	0,588950778	
t Critical two-tail	1,967956506	

Conclusion

At 95% confidence level, P(two-tail)= 0.588 > ∝= 0.05
If P-value > ∝ => we donot reject null hypothesis and conclude there is no difference between mortgage payments now & 5 years ago. The conclusion didn't change as a result of running a different test (2 sample t-test with equal variances instead of t-test with paired sample).

2c) Interval Estimate difference between 2a), 2b). Why margin of error is small in 2a)

(c) What is the estimate of the difference between mortgage payments for both parts (a) and (b)? Why is the margin of error smaller in part (a)?

Q1: Difference of Interval Estimate between 2a) and 2b).

	A	B	C	D
1	This Year	5 years ago	Diff	Null
2	613,21	783,31	-170,1	0
3	551,66	498,33	53,33	0
4	633,4	560,8	72,6	
5	703,02	745,84	-42,82	
6	1158,98	1135,76	23,22	
7	1203,77	1342,59	-138,82	
8	958,4	1003,81	-45,41	
9	872,57	874,24	-1,67	
10	1086,69	886,97	199,72	
11	692,15	760,52	-68,37	
12	785,58	819,59	-34,01	
13	1196,76	1127,73	69,03	
14	705,87	614,28	91,59	
15	380,99	318,21	62,78	
16	964,1	769,98	194,12	
17	1025,34	1030,62	-5,28	
18	726,33	593,46	132,87	
19	700,16	731,64	-31,48	
20	847,21	885,72	-38,51	
21	767,43	813,06	-45,63	
22	858,47	732,88	125,59	
23	966,11	922,84	43,27	
24	501,64	428,1	73,54	
25	921,37	801,76	119,61	
26	747,42	558,12	189,3	
27	993,44	1044,33	-50,89	
28	872,8	814,49	58,31	
29	1006,41	981,71	24,7	
30	957,56	862,93	94,63	
31	927,7	981,66	-53,96	
32	791,51	829,34	-37,83	
33	926,52	937,86	-11,34	
34	916,45	1022,83	-106,38	
35	790,54	749,18	41,36	
36	1026,06	1053,49	-27,43	
37	1071,33	1046,71	24,62	
38	954,09	838,61	115,48	
39	823,69	966,36	-142,67	
40	973,28	901,78	71,5	
41	851,19	879,88	-28,69	
42	829,46	705,31	124,15	
43	845,07	597,36	247,71	
44	1150,59	817,73	332,86	
45	865,7	687,39	178,31	
46	992,31	1136,48	-144,17	
47	1105,74	1162,46	-56,72	
48	1098,17	1056,31	41,86	
49	949,96	971,49	-21,53	
50	832,38	723,45	108,93	
51	706,99	579,4	127,59	
52	776,6	712,53	64,07	
53	914,53	919,6	-5,07	

	F	G	H	I	J	K
1	t-Test: Paired Two Sample for Means					
3		This Year	5 years ago			
4	Mean	937,9098667	925,5089333	SD	Se = SD/s	Mean_diff
5	Variance	34193,13389	44637,76851	99,14092	8,094822	12,40093333
6	Observations	150	150			
7	Pearson Correlation	0,883101999		ME	15,99548	
8	Hypothesized Mean Diff	0		Estimate =	28,39641	-3,594542091
9	df	149				
10	t Stat	1,531958697				
11	P(T<=t) one-tail	0,063826543				
12	t Critical one-tail	1,655144534				
13	P(T<=t) two-tail	0,127653086				
14	t Critical two-tail	1,976013178				

	L	M	N
1	F-Test Two-Sample for Variances		
3		This Year	5 years ago
4	Mean	937,9098667	925,5089333
5	Variance	34193,13389	44637,76851
6	Observations	150	150
7	df	149	149
8	F	0,766013514	
9	P(F<=f) one-tail	0,052429063	
10	F Critical one-tail	0,763100731	

	L	M	N	O	P	Q	R	S
16	t-Test: Two-Sample Assuming Equal Variances							
18		This Year	5 years ago			Mean_diff		
19	Mean	937,9098667	925,5089333			12,4		
20	Variance	34193,13389	44637,76851	184,9139	211,2765			
21	Observations	150	150					
22	Pooled Variance	39415,4512						
23	Hypothesized Me	0		S.e = sqrt(	sqrt(0.01	sqrt(0.013	sqrt(524.2	22.89
24	df	298		SE = 22.89				
25	t Stat	0,540943309		ME = se *	22.89 * 1.9	45		
26	P(T<=t) one-tail	0,294475389		Eastimate =				
27	t Critical one-tail	1,649982976		(1/n1 + 1/n2) = 2/150 = 0.0133				
28	P(T<=t) two-tail	0,588950778						
29	t Critical two-tail	1,967956506						
30				Estimate =	45 + 12.4	45 - 12.4		
31					57.4	33.05		

CI (confidence interval) = $(\bar{x}d) \pm$ t_critical (standard_error)

= $(\bar{x}1 - \bar{x}2) \pm$ t_critical (standard_error)

= $(\bar{x}1 - \bar{x}2) \pm$ t_critical (S_D / sqrt(n)))

Where, S_D is the Standard deviation difference of 2 samples = S1 – S2
$\bar{x}d$ is the mean difference of 2 samples = $\bar{x}1 - \bar{x}2$

Margin of error for the matched pair test is approximately 16 as shown from calculations in above screenshot.

Interval Estimate for matched pairs = 12.4 $\pm$ 16 = (-3,6, 28,34)

t-Test: Paired Two Sample for Means					
	This Year	*5 years ago*			
Mean	937,9098667	925,5089333	SD	Se = SD/sc	Mean_diff
Variance	34193,13389	44637,76851	99,14092	8,094822	12,40093333
Observations	150	150			
Pearson Correlation	0,883101999		ME	15,99548	
Hypothesized Mean Diff	0		Estimate =|	28,39641	-3,594542091
df	149				
t Stat	1,531958697				
P(T<=t) one-tail	0,063826543				
t Critical one-tail	1,655144534				
P(T<=t) two-tail	0,127653086				
t Critical two-tail	1,976013178				

CI (confidence interval) = $(\bar{x}d) \pm$ t_critical (standard_error)

= $(\bar{x}1 - \bar{x}2) \pm$ t_critical (standard_error)

= $(\bar{x}1 - \bar{x}2) \pm$ t_critical (sqrt (sp^2 (1/n1 + 1/n2)))

Margin of error for the 2-sample t- test with equal variances is approximately 45 as shown in above screenshot.

Interval Estimate for two-sample t-test with equal variances = 12.4 $\pm$ 45 = (33.05, 57.4)

Estimate of the difference between mortgage payments of 2a) and 2b) = 45 – 16 = 29

t-Test: Two-Sample Assuming Equal Variances							
	This Year	*5 years ago*			Mean_diff		
Mean	937,9098667	925,5089333			12,4		
Variance	34193,13389	44637,76851	184,9139	211,2765			
Observations	150	150					
Pooled Variance	39415,4512						
Hypothesized Me	0		S.e = sqrt(!	sqrt(0.01:	sqrt(0.013	sqrt(524.2:	22.89
df	298		SE = 22.89				
t Stat	0,540943309		ME = se *	22.89 * 1.9	45		
P(T<=t) one-tail	0,294475389		Eastimate =				
t Critical one-tail	1,649982976		(1/n1 + 1/n2) = 2/150 = 0.0133				
P(T<=t) two-tail	0,588950778						
t Critical two-tail	1,967956506						
			Estimate =	45 + 12.4	45 - 12.4		
				57.4	33.05		

Q2: Why the margin of error is small in 2a)

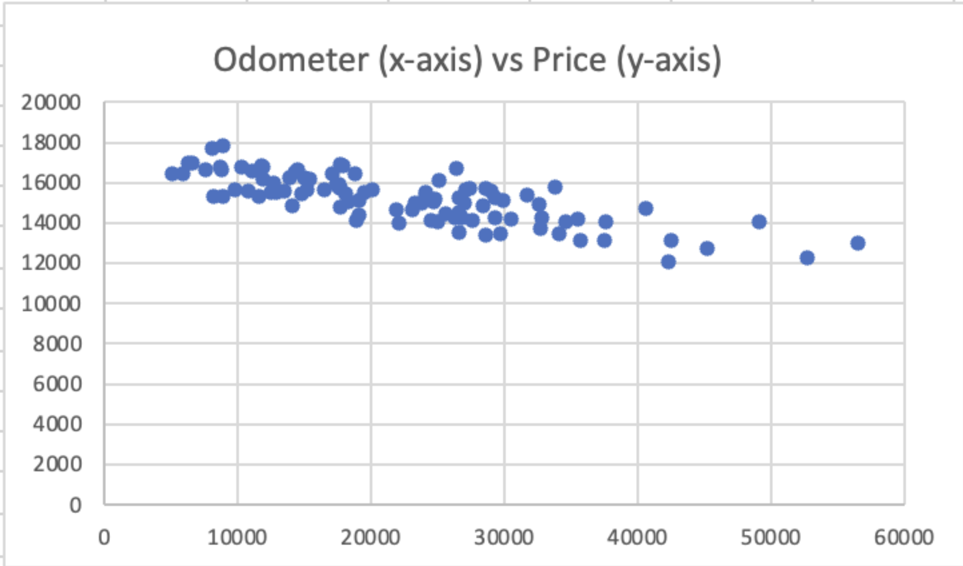
The margin of error in 2a) is 16 which is smaller than the margin of error in 2b). Because in 2a) we assumed dependant samples (or) both samples were collected from homeowners living in the same house as they were five years ago.

3a) Create a simple linear regression model. Interpret variable coefficient

3. A used car dealership wants to build a model for the price of used cars based on the miles on the odometer. Use the file used_cars.xlsx to create a simple linear regression model and answer the following questions.

   (a) Interpret the variable coefficient in terms of how the independent variable effects the dependent variable.

From the below scatterplot, we can identify that there is a strong negative linear correlation between the dependent variable (price) and the independent variable (odometer). As the price of the used_car depends on how much mileage it already consumed, the dependent variable is price. Hence, the price is taken on the y-axis. And, the odometer is on the x-axis.



The regression model built at a 95% confidence level for the given used_cars data is as shown below:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,783879							
R Square	0,614466							
Adjusted R	0,610532							
Standard E	741,8207							
Observatio	100							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	85952464	85952464	156,1926	5,31E-22			
Residual	98	53929203	550298					
Total	99	1,4E+08						
	Coefficients	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*Jpper 95,0%*
Intercept	17160,81	173,2084	99,0761	4,6E-100	16817,09	17504,54	16817,09	17504,54
Odometer	-0,086887	0,006952	-12,4977	5,31E-22	-0,100684	-0,073091	-0,100684	-0,073091

From the above regression analysis table, we can identify the correlation coefficient (Multiple R) value of 0.78 (absolute) which tells us that there is a strong correlation between Price and Odometer. A b1 coefficient of -0.087 tells us about a negative slope.

y = bo+(b1)x.

b_0 = the value you would predict for y when x equals zero (may or may not be meaningful)

b_1 = the predicted change in y for a one unit increase in $x \Rightarrow$ *if x increased by one unit, y would change by b_1 units*

For every 1 unit (mile) increase in x (odometer), there is b1(coefficient) unit ($) change in y (price).

y = 17160 + (-0.087) x

price = 17160 + (-0.087) * odometer

Variable coefficient Interpretation: For every 1-mile increase in the odometer, there is a -0.087$ change in price.

3b) Variation in the dependent variable

(b) How much of the variation in the dependent variable is explained by the model?

R square (coefficient of determination) value multiplied by 100 gives us the percentage of the variation in Y (dependent variable). A higher R-squared value indicates a better fit of the model to the data.

The current regression model explains 61.44% of the variation in the data.

Extra optional findings: The F-test determines if we have a statistically significant model. As p < alpha (at 95% confidence level) from the above model => we reject the null hypothesis. I.e., we conclude that the model is statistically significant. i.e., Both the intercept and the variable coefficient are statistically significant.

3c) Estimate price for 15,000 miles

(c) What price would you estimate for a used car that has 15,000 miles on it?

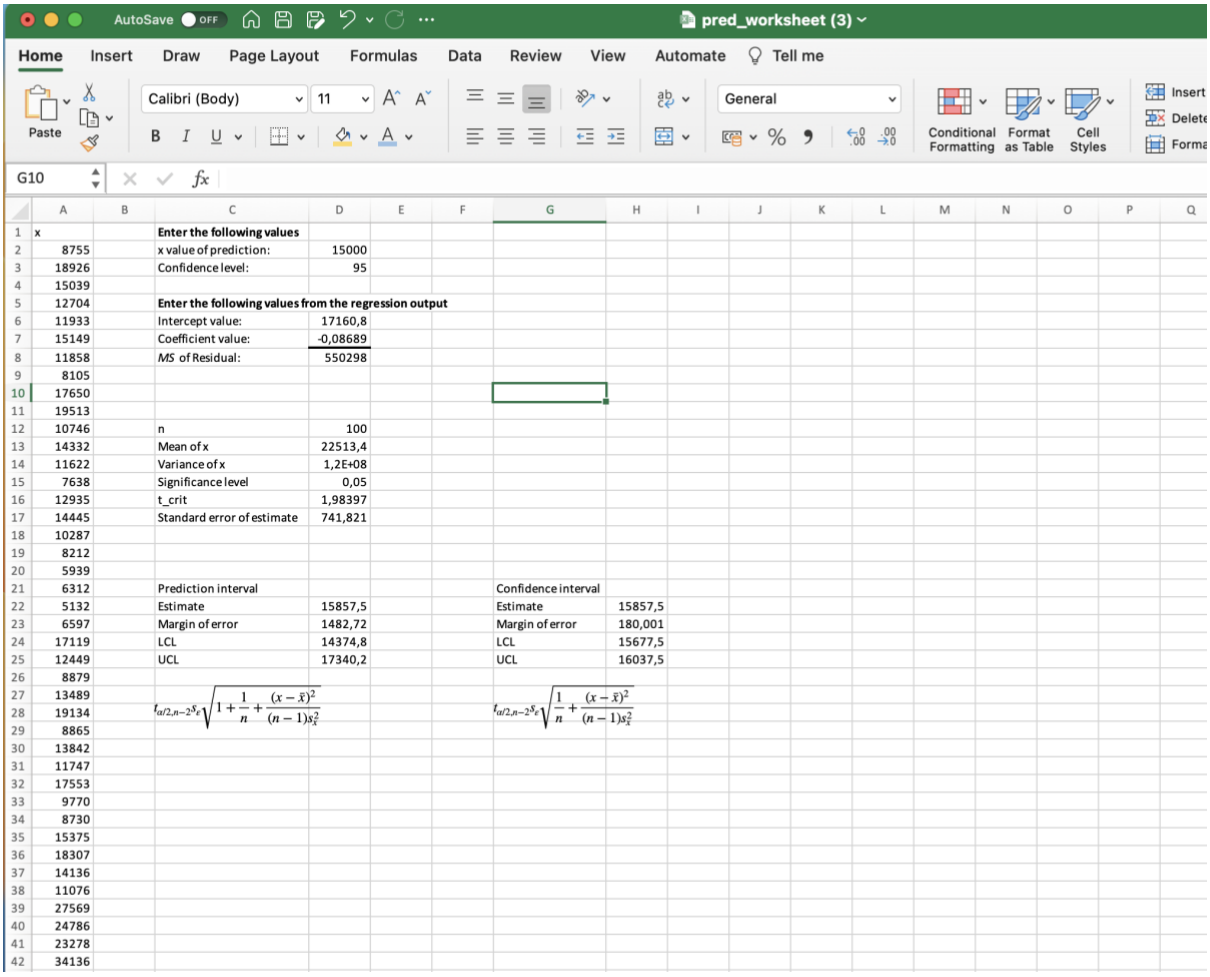
We can estimate the price in either one of the below 2 ways:

Method 1 : Using the equation we derived above

price = 17160 + (-0.087) * odometer

Estimated price = 17160 + (-0.087) * 15,000 = 17160 - 1305 = 15855$

Method 2: Using the prediction worksheet from class. Estimated price = 15857.5$ ± 1482.72$.



	A	C	D	G	H
1	x	**Enter the following values**			
2	8755	x value of prediction:	15000		
3	18926	Confidence level:	95		
4	15039				
5	12704	**Enter the following values from the regression output**			
6	11933	Intercept value:	17160,8		
7	15149	Coefficient value:	-0,08689		
8	11858	*MS* of Residual:	550298		
9	8105				
10	17650				
11	19513				
12	10746	n	100		
13	14332	Mean of x	22513,4		
14	11622	Variance of x	1,2E+08		
15	7638	Significance level	0,05		
16	12935	t_crit	1,98397		
17	14445	Standard error of estimate	741,821		
18	10287				
19	8212				
20	5939				
21	6312	Prediction interval		Confidence interval	
22	5132	Estimate	15857,5	Estimate	15857,5
23	6597	Margin of error	1482,72	Margin of error	180,001
24	17119	LCL	14374,8	LCL	15677,5
25	12449	UCL	17340,2	UCL	16037,5
26	8879				
27	13489	$t_{\alpha/2,n-2}s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{(n-1)s_x^2}}$		$t_{\alpha/2,n-2}s_\varepsilon\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{(n-1)s_x^2}}$	
28	19134				
29	8865				
30	13842				
31	11747				
32	17553				
33	9770				
34	8730				
35	15375				
36	18307				
37	14136				
38	11076				
39	27569				
40	24786				
41	23278				
42	34136				

3d) Would you use regression equation to estimate price for 30,000 miles ? Explain

(d) Would you want to use your regression equation to estimate the price of a used car that has 30,000 miles? Explain your answer.

Yes, I would use the regression equation to estimate the price of a used car that has 30,000 miles. Because 30,000 miles fits within the sample data range that is shared with us (or) within the range of data we used to build the regression model. See the below Excel snippet, especially odometer Max_value (56519) and Min value (5132).

Min_value < 30,000 < Max_value

Price	Odometer	
12972,45	56519	
12083,32	42322	
12704,13	45235	
13379,59	28580	
15405,31	31721	
14965,57	26942	
13151,81	37479	
12267,2	52658	
13154,01	42486	
16723,48	26400	
13461,77	29710	
17816,26	56519	Max
12083,32	5132	Min

Extra optional findings: The price of the used car which consumed 30,000 miles would depreciate to be sold at 14554.2$ ± 1482.69$. This matches exactly with the negative slope / negative correlation between price and odometer values. When miles increase, the price of used_car would decrease.

G14

	A	B	C	D	E	F	G	H
1	x		**Enter the following values**					
2	8755		x value of prediction:	30000				
3	18926		Confidence level:	95				
4	15039							
5	12704		**Enter the following values from the regression output**					
6	11933		Intercept value:	17160,8				
7	15149		Coefficient value:	-0,08689				
8	11858		*MS* of Residual:	550298				
9	8105							
10	17650							
11	19513							
12	10746		n	100				
13	14332		Mean of x	22513,4				
14	11622		Variance of x	1,2E+08				
15	7638		Significance level	0,05				
16	12935		t_crit	1,98397				
17	14445		Standard error of estimate	741,821				
18	10287							
19	8212							
20	5939							
21	6312		Prediction interval				Confidence interval	
22	5132		Estimate	14554,2			Estimate	14554,2
23	6597		Margin of error	1482,69			Margin of error	179,788
24	17119		LCL	13071,5			LCL	14374,4
25	12449		UCL	16036,9			UCL	14734
26	8879							
27	13489							
28	19134		$t_{\alpha/2,n-2}s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{(n-1)s_x^2}}$				$t_{\alpha/2,n-2}s_\varepsilon\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{(n-1)s_x^2}}$	
29	8865							
30	13842							
31	11747							
32	17553							
33	9770							
34	8730							
35	15375							
36	18307							
37	14136							
38	11076							
39	27569							
40	24786							
41	23278							
42	34136							

Sheet1 +