Kankanamge Harsha **PROJECT 1**

(One-factor Completely Randomized Design)

Smoking habits and absenteeism rates (hours per month) in the workplace was studied based on a survey of blue- and white-collar employees. They were categorized into four groups in terms of their smoking behavior: continuous smokers, recent ex-smokers, long-term ex-smokers, and nonsmokers. The data represent absenteeism rates of randomly selected employees from each group and are recorded in **PROJ1-641.sav**.

1)Identify the response variable, factor and levels of the factor. What are the values of 'n' and 'k'?

Response variable is: absenteeism rates(absent)

Factor is: Smoking habits (group)

Levels of the factor:

continuous smokers	1
recent ex-smokers	2
long-term ex-smokers	3
nonsmokers	4

n = 10 k=4

2) Test if 'absenteeism rates' is normally distributed in all four populations.

H0: Absenteeism rates among all four populations (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) are normal.
H1: Absenteeism rates among all four populations (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) are not normal.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	groups	Statistic	df	Sig.	Statistic	df	Sig.
absenteeism	continuous smokers	.203	10	.200*	.905	10	.249
	recent ex-smokers	.178	10	.200*	.933	10	.476
	long-term ex-smokers	.219	10	.191	.810	10	.019
	nonsmokers	.196	10	.200*	.846	10	.052

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

For continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers

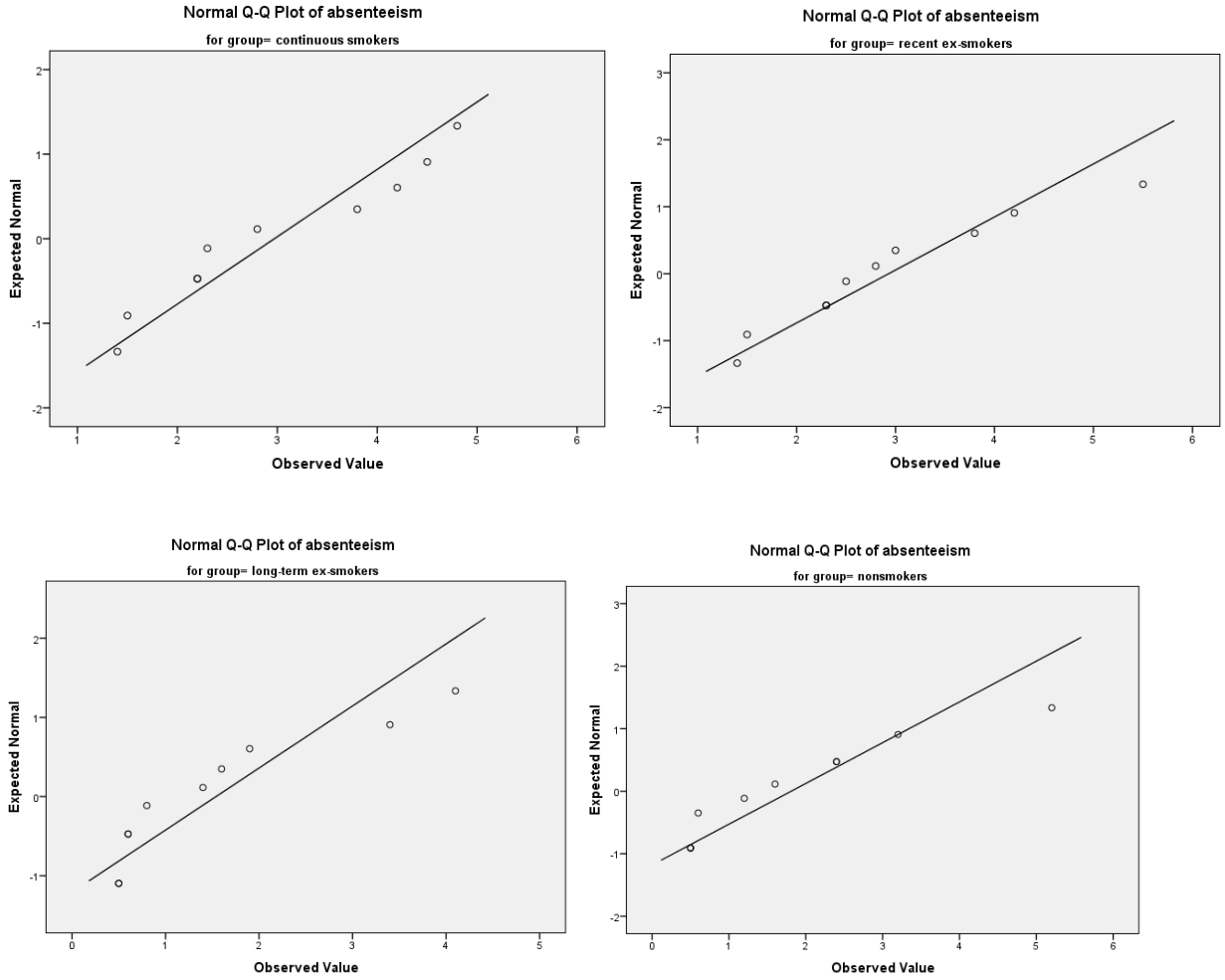
KS test statistic are: 0.203, 0.178, 0.219, 0.196 and
p-values are:0.200,0.200,0.191,0.200 respectively.

α=0.10
All p-values (0.200,0.200,0.191,0.200) > α (0.10)

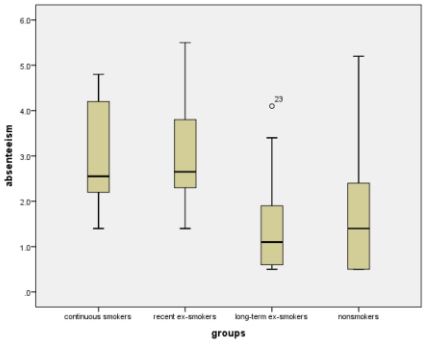
Do not reject H0
'Absenteeism rates' is normally distributed among four (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) populations.

Descriptives

	groups			Statistic	Std. Error
absenteeism	continuous smokers	Mean		2.970	.3969
		90% Confidence Interval for Mean	Lower Bound	2.242	
			Upper Bound	3.698	
		5% Trimmed Mean		2.956	
		Median		2.550	
		Variance		1.576	
		Std. Deviation		1.2553	
		Minimum		1.4	
		Maximum		4.8	
		Range		3.4	
		Interquartile Range		2.3	
		Skewness		.272	.687
		Kurtosis		-1.605	1.334
	recent ex-smokers	Mean		2.930	.3995
		90% Confidence Interval for Mean	Lower Bound	2.198	
			Upper Bound	3.662	
		5% Trimmed Mean		2.872	
		Median		2.650	
		Variance		1.596	
		Std. Deviation		1.2632	
		Minimum		1.4	
		Maximum		5.5	
		Range		4.1	
		Interquartile Range		1.8	
		Skewness		.868	.687
		Kurtosis		.530	1.334
	long-term ex-smokers	Mean		1.540	.4034
		90% Confidence Interval for Mean	Lower Bound	.801	
			Upper Bound	2.279	
		5% Trimmed Mean		1.456	
		Median		1.100	
		Variance		1.627	
		Std. Deviation		1.2756	
		Minimum		.5	
		Maximum		4.1	
		Range		3.6	
		Interquartile Range		1.7	
		Skewness		1.267	.687
		Kurtosis		.536	1.334
	nonsmokers	Mean		1.810	.4848
		90% Confidence Interval for Mean	Lower Bound	.921	
			Upper Bound	2.699	
		5% Trimmed Mean		1.694	
		Median		1.400	
		Variance		2.350	
		Std. Deviation		1.5329	
		Minimum		.5	
		Maximum		5.2	
		Range		4.7	
		Interquartile Range		2.1	
		Skewness		1.284	.687
		Kurtosis		1.477	1.334



Normal Q-Q plots of four populations (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) shows that most points are fitting onto the line, which means the all four populations are normally distributed.



There is one outlier shows in group long-term ex-smokers. Which is 3rd data point in that group.

3)Test for homogeneity of variance of 'absenteeism rates' in all four populations.

H0: σ1=σ2=σ3=σ4
H1: At least one of the variances of 'absenteeism rates' among the four populations (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) are not equal.

Test of Homogeneity of Variances

absenteeism

Levene Statistic	df1	df2	Sig.
.223	3	36	.880

Levene Statistic F (0.1,3,36) =0.223 and p-value= 0.880

P-value (0.880)>0.10

Do not reject H0

Four groups (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) have same variance for the 'absenteeism rates'.

4)Do the data provide sufficient evidence to indicate a difference among the mean absenteeism rates for the four groups? **Also show calculations for all components of the ANOVA table and make appropriate conclusions**.

H0: μ1=μ2=μ3=μ4
H1: At least one of the means of 'absenteeism rates' among the four populations (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) are not equal.

ANOVA

absenteeism

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16.629	3	5.543	3.102	.039
Within Groups	64.335	36	1.787		
Total	80.964	39			

ANOVA F test statistic: 3.102 and p-value is 0.039

α=0.1
p-value (0.039) < α (0.1)

Reject H0

Four groups (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers) do not have same means for the 'absenteeism rates'. At least one of the means of 'absenteeism rates' among four group is different. Which means among four groups there are significant differences of means of 'absenteeism rates'.

Descriptives

absenteeism

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
continuous smokers	10	2.970	1.2553	.3969	2.072	3.868	1.4	4.8
recent ex-smokers	10	2.930	1.2632	.3995	2.026	3.834	1.4	5.5
long-term ex-smokers	10	1.540	1.2756	.4034	.628	2.452	.5	4.1
nonsmokers	10	1.810	1.5329	.4848	.713	2.907	.5	5.2
Total	40	2.313	1.4408	.2278	1.852	2.773	.5	5.5

Descriptive Statistics

	N	Mean	Variance
absenteeism	40	2.312	2.076
Valid N (listwise)	40		

SST = (n-1)*S^2

SST=39*2.076 = 80.964

SSA = $\sum_{i=1}^{4} 4(y\tilde{i} - \tilde{y}..)$^2

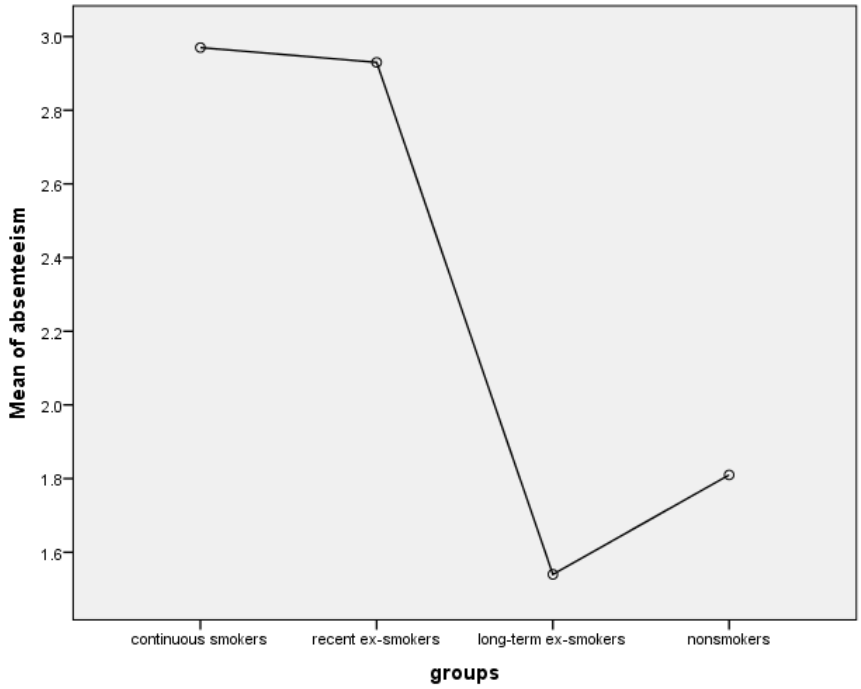
SSA=10*((2.97-2.313)^2+(2.93-2.313)^2+(1.54-2.313)^2+(1.81-2.313)^2)= 16.6287

SSE=SST-SSA=80.964-16.6287=64.3353

Source of variation	Sum of square	D.F	Mean Square	F
'absenteeism rates'(SSA)	16.6287	K-1=3	16.6287/3=5.5429	5.5429/1.787 =3.102
Error (SSE)	64.3353	K(n-1)=36	64.3353/36=1.787	
Total (SST)	80.964	n*k-1=39		

F (0.1, 3, 36) = 2.226091<3.102 (F calculated)
Reject H0

In here also it is showing that there are significant differences in mean values of 'absenteeism rates' among all four populations (continuous smokers, recent ex-smokers, long-term ex-smokers, nonsmokers).



In this graph it is clearly show that there are differences in mean values of absenteeism in four groups. We can see that "contiguous smokers" and "recent ex-smokers" have very closer mean values of absenteeism. As well as "long-term ex-smokers" and "nonsmokers" have closer mean values too but this value is significantly lower than mean of previously mentioned two categories. Simply "contiguous smokers" and "recent ex-smokers" have higher absenteeism by mean values. And "long-term ex-smokers" and "nonsmokers" have lower absenteeism by mean values.

5)Conduct Tukey and Bonferroni tests for multiple comparisons.

Multiple Comparisons

Dependent Variable: absenteeism

	(I) groups	(J) groups	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	continuous smokers	recent ex-smokers	.0400	.5978	1.000	-1.381	1.461
		long-term ex-smokers	1.4300*	.5978	.097	.009	2.851
		nonsmokers	1.1600	.5978	.230	-.261	2.581
	recent ex-smokers	continuous smokers	-.0400	.5978	1.000	-1.461	1.381
		long-term ex-smokers	1.3900	.5978	.111	-.031	2.811
		nonsmokers	1.1200	.5978	.257	-.301	2.541
	long-term ex-smokers	continuous smokers	-1.4300*	.5978	.097	-2.851	-.009
		recent ex-smokers	-1.3900	.5978	.111	-2.811	.031
		nonsmokers	-.2700	.5978	.969	-1.691	1.151
	nonsmokers	continuous smokers	-1.1600	.5978	.230	-2.581	.261
		recent ex-smokers	-1.1200	.5978	.257	-2.541	.301
		long-term ex-smokers	.2700	.5978	.969	-1.151	1.691
Bonferroni	continuous smokers	recent ex-smokers	.0400	.5978	1.000	-1.461	1.541
		long-term ex-smokers	1.4300	.5978	.133	-.071	2.931
		nonsmokers	1.1600	.5978	.361	-.341	2.661
	recent ex-smokers	continuous smokers	-.0400	.5978	1.000	-1.541	1.461
		long-term ex-smokers	1.3900	.5978	.155	-.111	2.891
		nonsmokers	1.1200	.5978	.415	-.381	2.621
	long-term ex-smokers	continuous smokers	-1.4300	.5978	.133	-2.931	.071
		recent ex-smokers	-1.3900	.5978	.155	-2.891	.111
		nonsmokers	-.2700	.5978	1.000	-1.771	1.231
	nonsmokers	continuous smokers	-1.1600	.5978	.361	-2.661	.341
		recent ex-smokers	-1.1200	.5978	.415	-2.621	.381
		long-term ex-smokers	.2700	.5978	1.000	-1.231	1.771

*. The mean difference is significant at the 0.1 level.

H0: $\mu_i = \mu_j$ i, j=1,2,3,4
H1: $\mu_i \neq \mu_j$ i≠j

α=0.10
Tukey HSD: means differences of “absenteeism “between “continuous smokers”, “long-term ex-smokers” (p-value= 0.097 <0.10 and difference= 1.4300, Reject H0) are significant.
And means differences of “absenteeism “between” recent ex-smokers”, “long-term ex-smokers” (p-value= 0.111 and difference= 1.39) are very closer to border line significant.

Bonferroni: All p-values are greater than 0.10 which means those differences are not significant.
Bonferroni: All of mean differences of “absenteeism rates” in between each group is not significant.

Mean differences very week between “continuous smokers” and “recent ex-smokers” (p-value= 1.0 and difference= 0.04) in Tukey and Bonferroni test.

Mean difference between “continuous smokers” and “long-term ex-smokers” (p-value= 0.133 and difference= 1.4300) and “recent ex-smokers” and “long-term ex-smokers” (p-value= 0.155 and difference= 1.39) considerably higher than other differences in Bonferroni test.

6) Conduct Dunnett's test to compare whether the mean absenteeism rates are higher for continuous smokers, recent ex-smokers, and long-term ex-smokers as compared to nonsmokers.

H0: $\mu_i = \mu_ß$ i=1,2,3 $\mu_ß$= mean absenteeism rates for nonsmokers
H1: $\mu_i > \mu_ß$

Multiple Comparisons

Dependent Variable: absenteeism

Dunnett t (>control)[a]

(I) groups	(J) groups	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval
					Lower Bound
continuous smokers	nonsmokers	1.1600*	.5978	.074	.098
recent ex-smokers	nonsmokers	1.1200*	.5978	.084	.058
long-term ex-smokers	nonsmokers	-.2700	.5978	.883	-1.332

*. The mean difference is significant at the 0.1 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Mean differences of absenteeism in all group compared with nonsmokers group:

Continuous smokers - nonsmokers =1.16 and p-value is 0.074<0.1
Reject H0
which means, mean value of absenteeism rates in Continuous smokers group is higher than nonsmokers group and it is statically significant.

recent ex-smokers - nonsmokers =1.12 and p-value is 0.084<0.1
Reject H0
which means, mean value of absenteeism rates in recent ex-smokers group is higher than nonsmokers group and it is statically significant.

long-term ex-smokers - nonsmokers = -0.2700 and p-value is 0.883>0.1
Do not reject H0
Which means, mean differences in this two group is not statically significant.
(which means, mean value of absenteeism rates in long-term ex-smokers group is not higher than nonsmokers group and it is not statically significant.)

Which means the mean absenteeism rates are significantly higher for continuous smokers, recent ex-smokers as compared to nonsmoker. But not for long-term ex-smokers.