1. **The algae dataset contains data on 90 independent river water samples.**

a) **Is river size associated with season? Carry out an appropriate test. Include the appropriate hypotheses, test statistic value, p-value, and conclusion.**

Hypothesis:

H_0: Variables river size and season are independent
H_A: Variables river size and season are not independent

Test statistics:

$$\chi^2 = 3.4653$$

P-value:

$$P(\chi^2 > 3.4653) = 0.7486$$

Conclusion:

The P-value = 0.7486 is greater than 0.05, hence we do not reject the null hypothesis. There is enough evidence to conclude that river size and seasons are independent at 5% significant level.

SAS Output

Association between riversize and season

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of riversize by season

riversize	season 1	2	3	4	Total
1	4 4.44 19.05 16.67	4 4.44 19.05 23.53	7 7.78 33.33 28.00	6 6.67 28.57 25.00	21 23.33
2	13 14.44 25.00 54.17	11 12.22 21.15 64.71	15 16.67 28.85 60.00	13 14.44 25.00 54.17	52 57.78
3	7 7.78 41.18 29.17	2 2.22 11.76 11.76	3 3.33 17.65 12.00	5 5.56 29.41 20.83	17 18.89
Total	24 26.67	17 18.89	25 27.78	24 26.67	90 100.00

Statistics for Table of riversize by season

Statistic	DF	Value	Prob
Chi-Square	6	3.4653	0.7486
Likelihood Ratio Chi-Square	6	3.4695	0.7480
Mantel-Haenszel Chi-Square	1	0.9256	0.3360
Phi Coefficient		0.1962	
Contingency Coefficient		0.1926	
Cramer's V		0.1388	

WARNING: 42% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Table Probability (P)	<.0001
Pr <= P	0.7861

Sample Size = 90

b) **Create a new variable by combining the small and medium size rivers in one category. So, this new variable will have two categories -small/medium and large. Is there a significant difference in mean chem3 value for rivers of small/medium and large sizes? Carry out an appropriate test. Include the appropriate hypotheses, test statistic, p-value, and conclusion.**

We use two sample t-test. First, we use F test to determine if variances are equal. The F test statistic for testing $H_0 : \sigma_1 = \sigma_2$ is 2.23 with p-value = 0.0225. So, we reject H_0 at 5% level and conclude that the variances are unequal.

Hypothesis:

$\boldsymbol{H_0}$: There is no significant difference between mean chem3 values for rivers of small/medium and large sizes.
$\boldsymbol{H_A}$: There is significant difference between mean chem3 values for rivers of small/medium and large sizes.

Test statistics:

$$T = 0.03$$

P-value:

$$2P(T > |0.03|) = 0.9778$$

Conclusion:

The P-value = 0.9778 is greater than 0.05, hence we do not reject the null hypothesis and conclude that there is **no enough evidence** to support the claim that there is significant difference between mean chem3 values for rivers of small/medium and large sizes at 5% significant level.

SAS Output:

Updated algae dataset: combining small and medium size rivers in one category

Obs	season	riversize	fluidvelocity	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	abundance	group
1	1	1	2	8.00	9.8	60.80	6.238	578.00	105.00	170.00	50.0	0.9191	0
2	4	1	2	8.06	9.0	55.35	10.420	233.70	58.22	97.58	10.5	0.6128	0
3	1	1	3	8.25	13.1	65.75	9.248	430.00	18.25	56.67	28.4	1.1000	0
4	3	1	3	8.15	10.3	73.25	1.535	110.00	61.25	111.80	3.2	0.8325	0
5	4	1	3	8.05	10.6	59.07	4.990	205.70	44.67	77.43	6.9	0.9395	0
6	2	1	3	7.61	9.8	7.00	1.443	31.33	20.00	57.83	0.4	0.1461	0
7	3	1	3	7.35	10.4	7.00	1.718	49.00	41.50	61.50	0.8	0.9138	0
8	4	1	3	7.75	10.3	32.92	2.942	42.00	16.00	40.00	7.6	1.0450	0
9	2	1	3	7.84	9.4	10.98	1.510	12.50	3.00	11.50	1.5	0.2041	0
10	3	1	3	7.77	10.7	12.54	3.976	58.50	9.00	44.14	3.0	1.0040	0

T test for significant difference in mean chem3 value for rivers of small/medium and large sizes

The TTEST Procedure

Variable: chem3

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		73	50.0749	38.4354	4.4985	1.5490	194.8
3		17	49.6629	57.3400	13.9070	5.3260	208.4
Diff (1-2)	Pooled		0.4120	42.5027	11.4459		
Diff (1-2)	Satterthwaite		0.4120		14.6165		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		50.0749	41.1072	59.0425	38.4354	33.0539	45.9264
3		49.6629	20.1814	79.1444	57.3400	42.7051	87.2673
Diff (1-2)	Pooled	0.4120	-22.3344	23.1584	42.5027	37.0447	49.8618
Diff (1-2)	Satterthwaite	0.4120	-30.1301	30.9540			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	88	0.04	0.9714
Satterthwaite	Unequal	19.476	0.03	0.9778

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	16	72	2.23	0.0225

(c) Report the skewness statistic for chem3. Estimate its p-values by Monte Carlo method. The hypotheses of interest are $H_0: \gamma_1 = 0$ vs $H_A: \gamma_1 \neq 0$, and the test statistic is k_3. Assume the null distribution to be normal. Note: PROC UNIVARIATE provides the value of k_3

Skewness Statistic for chem 3 = 1.67952

Hypothesis:

$H_0 : \gamma_1 = 0$
$H_A : \gamma_1 \neq 0$

Conclusion:

For the variable chem3, From the Monte Carlo simulation, the P-value = 0 is less than 0.05, hence we reject the null hypothesis and conclude that there is enough evidence to support the claim that the variable chem3 is not symmetrically distributed at 5% significant level.

SAS Output

Data set B. Observed statistic

Obs	n	sobs	Nruns
1	90	$\boxed{1.67952}$	10000

Data set MC. Simulated samples

Obs	n	sobs	Nruns	SEED	MCrun	j	X
1	90	1.67952	10000	1311542125	1	1	-1.13060
2	90	1.67952	10000	1210284520	1	2	-0.48639
3	90	1.67952	10000	1243875875	1	3	-2.27440
4	90	1.67952	10000	17513725	1	4	1.38988
5	90	1.67952	10000	2129266123	1	5	1.03708

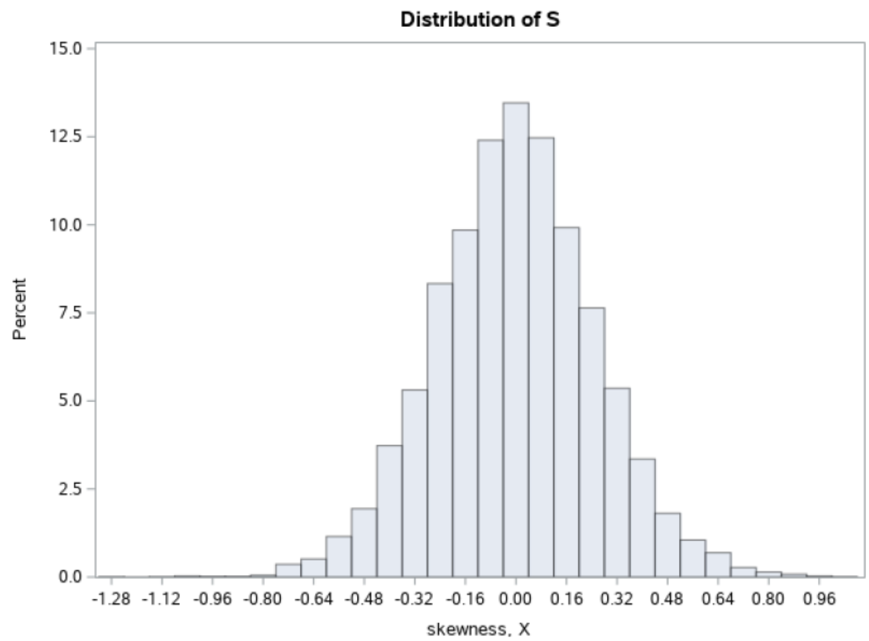
The UNIVARIATE Procedure
Variable: S (skewness, X)

Moments			
N	10000	Sum Weights	10000
Mean	-0.0028127	Sum Observations	-28.126969
Std Deviation	0.25393728	Variance	0.06448414
Skewness	0.02295643	Kurtosis	0.3761697
Uncorrected SS	644.856071	Corrected SS	644.776958
Coeff Variation	-9028.2493	Std Error Mean	0.00253937

Basic Statistical Measures			
Location		Variability	
Mean	-0.00281	Std Deviation	0.25394
Median	-0.00374	Variance	0.06448
Mode	.	Range	2.35183
		Interquartile Range	0.33226

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	-1.10763	Pr > |t|	0.2680
Sign	M	-70	Pr >= |M|	0.1645
Signed Rank	S	-379657	Pr >= |S|	0.1885

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-1.28720	7195	0.917708	1913
-1.10137	9013	0.937847	7627
-1.07455	2182	0.941663	2285
-1.06296	5984	0.976391	7573
$\boxed{-1.02225}$	1614	$\boxed{1.064629}$	4907



Data set D. Results of simulations

Obs	n	sobs	Nruns	i	MCrun	S	indicator
1	90	1.67952	10000	1	1	-0.17104	0
2	90	1.67952	10000	2	2	-0.18773	0
3	90	1.67952	10000	3	3	0.06263	0
4	90	1.67952	10000	4	4	0.13855	0
5	90	1.67952	10000	5	5	-0.28931	0

Estimated p-value

Obs	Pvalue
1	0

---

2) Implement the Monte Carlo simulation study discussed in class for estimating the coverage probability of the standard 95% confidence interval for proportion for n = 25, 50, and 100. The standard error of the estimated coverage probability should not exceed 0.005. Use p = 0:1. You can use call ranbin(seed,n,p,x). State your conclusions including the effect of increasing n on the coverage probability.

Below is a summary of coverage probabilities for the combinations of p =0.1 and n:

n	25	50	100
Coverage Probability	0.9150	0.8737	0.9320

There is no obvious pattern in the coverage probabilities when n is varied while p is constant. Hence, we can only conclude that the coverage probabilities depend on both p and n.

SAS Output:

Output for Question 2
Part of the dataset generated for Monte Carlo

Obs	sample	n	p	x	phat	lb	ub	indicator
1	1	25	0.1	4	0.16	0.016293	0.30371	1
2	2	25	0.1	2	0.08	-0.026345	0.18634	1
3	3	25	0.1	5	0.20	0.043203	0.35680	1
4	4	25	0.1	4	0.16	0.016293	0.30371	1
5	5	25	0.1	3	0.12	-0.007383	0.24738	1

Coverage probability for n = 25 and p = 0.1

Obs	coverage
1	$\boxed{0.9265}$

Coverage probability for n = 50 and p = 0.1

Obs	coverage
1	$\boxed{0.8865}$

Coverage probability for n = 100 and p = 0.1

Obs	coverage
1	$\boxed{0.94075}$

R Codes

Question 1

```
FILENAME algae '/folders/myfolders/Project1/algae.csv'; /*create a pointer to data file*/

DATA algae; /*Assign name algae to data*/
INFILE algae DSD FIRSTOBS = 2;
INPUT season riversize fluidvelocity chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 abundance;
RUN;

PROC PRINT DATA=algae (OBS=10); /* Print 10 observations from the original dataset */
TITLE 'Algae dataset';
run;

/*Part a) Finding associaton between riversize and season*/
PROC FREQ DATA=algae;
TABLES riversize*season / CHISQ FISHER; /* contigency table of riversize by season and chisquare test */
TITLE 'Association between riversize and season';
RUN;

/*Part b) Creating a new variable by combining the small and medium size rivers in one category */
DATA algae1; SET algae;
IF riversize= 1 OR riversize= 2 THEN group = 0;
ELSE group = riversize;
RUN;

PROC PRINT DATA=algae1 (OBS=10); /* Print 10 observations from the new dataset */
TITLE 'Updated algae dataset: combining small and medium size rivers in one category';
RUN;

PROC TTEST DATA=algae1; /* Testing differences between means using T-test */
CLASS group;
VAR chem3;
TITLE 'T test for significant difference in mean chem3 value for rivers of small/medium and large sizes';
RUN;
```

```
/*Part c) Monte Carlo simulation for find the skewness*/

PROC UNIVARIATE DATA=algae NOPRINT;      /* supresses the output        */
VAR chem3;
OUTPUT OUT=B SKEW=sobs N=n;        /* save skewness values and sample size in dataset newalgae */
RUN;

DATA B; SET B; Nruns = 10000;         /* Adds Nruns = number of MC runs to dataset B */
RUN;

PROC PRINT DATA=B; TITLE 'Data set B. Observed statistic';
RUN;

DATA MC; SET B; /* creates dataset MC using dataset B */
RETAIN SEED 98638;
DO MCrun=1 TO Nruns;
  DO j=1 TO N;          /* Generate Nruns samples of size N of normal variables  */
    CALL RANNOR(SEED, X); /* Generates N(0,1) variate and saves in X. Returns a new seed.*/
    OUTPUT; /* ensures no overwriting of the perviously saved X */
  END; /* at this point for a given MCrun value, a sample of size N has been generated*/
END; /* Nruns replicates generated; each replicate of size N */
RUN;

PROC PRINT DATA=MC (obs=5);
TITLE 'Data set MC. Simulated samples';
RUN;


PROC UNIVARIATE DATA=MC NOPRINT;
VAR X;
CLASS MCrun;                        /* Compute skewness for each sample */

OUTPUT OUT=C SKEW=S;
RUN;

PROC PRINT DATA=C (obs=5);
TITLE 'Data set C. Skewness values for each samples';
RUN;

PROC UNIVARIATE DATA=C;
VAR S;
HISTOGRAM; /* Null distribution of skewness value */
RUN;

DATA B; SET B;                      /* Extending dataset B        */
DO i=1 TO Nruns; OUTPUT; END;       /* to the dimension as C: Repeating content of data B Nruns times.      */
RUN;

DATA D; MERGE B C;                  /* The indicator is 1 if      */
indicator =(S<=-(sobs))+(S>=(sobs));             /* S <= Sobs and 0 if S > Sobs */
RUN;

PROC PRINT DATA=D (obs=5); TITLE 'Data set D. Results of simulations';
RUN;

PROC MEANS DATA=D NOPRINT;          /* Finding the probability  */
VAR indicator;
OUTPUT OUT=E MEAN=Pvalue;

PROC PRINT DATA=E; TITLE 'Estimated p-value';
VAR Pvalue;                      /* Report the p-value*/
RUN;
```

Question 2

```
DATA values;
 p=0.1; n=100; alpha=0.05; z=QUANTILE('normal',1-alpha/2);
 nruns=4000; /* Because SE of the estimated coverage probability should not exceed 0.005 */  seed=98638;
RUN;

DATA montecarlo;
 SET values;
 CALL streaminit(seed);
 DO sample=1 TO nruns;
    x=RAND('binomial',p,n); /* sample */
    phat=x/n;  /* Estimate for p */
    lb=phat - z*sqrt((phat*(1-phat)/n));
    ub=phat + z*sqrt((phat*(1-phat)/n));
    indicator=(lb<=p<=ub); /* Indicator is 1 if p lies with the confidence interval, otherwise 0 */
    OUTPUT;
 END;
RUN;

PROC PRINT DATA=montecarlo (OBS=5);
 VAR sample n p x phat lb ub indicator;
 TITLE1 'Output for Question 2';
 TITLE2 'Part of the dataset generated for Monte Carlo';
RUN;

PROC MEANS DATA=montecarlo NOPRINT;
 VAR indicator;
 OUTPUT OUT=results MEAN=coverage; /* Estimating coverage probability using the proportion of Indicators */
RUN;

PROC PRINT DATA=results;
 VAR coverage;
 TITLE 'Coverage probability for n = 100 and p = 0.1';
RUN;
```