HW 1 - Solution

Problem 1

Prompt: A consumer group is collecting data on the mean cost (in dollars) of a shoulder MRI across different state imaging facilities. They use the following SAS program and output to analyze their data.

Problem 1.a.

Prompt: Test whether the mean cost is equal to $1000. What is the p-value?

Solution: You should have most, if not all, of the following (either mathematically or in words):

State your null:

$$H_o : \mu_{cost} = 1000$$

State your alternative:

$$H_1 : \mu_{cost} \neq 1000$$

Identify what key info you have (given in SAS output):

$$\bar{X} = 1709 \quad SE = 482.3$$

Identify your test statistic (also given by SAS):

$$t = \frac{\bar{X} - \mu_o}{SE} = \frac{1709 - 1000}{482.3} = 1.47$$

Identify the rejection region, your decision:

At α=0.05 and df=N-1=9: $t_{0.05/2;9} = 2.262$. Reject the null if $|t| > t_{0.05/2;9}$. Since $t < t_{0.05/2;9}$ (1.47 vs 2.262), we fail to reject the null hypothesis. *If you're using the p-value from SAS: at α=0.05, reject the null if $p < 0.05$. Since p=0.1757 > 0.05, we fail to reject the null hypothesis.*

Your interpretation: *(Brief 1-3 sentences here) e.g. The difference we observed could be due to random chance alone–we don't have enough evidence to suggest that the mean MRI cost is not $1000*

The TTEST Procedure

Variable: cost

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	1709.0	1525.3	482.3	440.0	4500.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1709.0	617.9	2800.1	1525.3	1049.2	2784.6

DF	t Value	Pr > |t|
9	1.47	0.1757

Problem 1.b.

Prompt: Construct the 95% Confidence Interval for the mean cost. Then calculate the 99% (Yes 99%) Confidence Interval for the mean cost.

Solution:

The 95% confidence interval is already given in the SAS output: (617.9, 2800.1).

To construct a 99% confidence interval, you may use the confidence interval formula which is

$$\bar{X} \pm t * SE = 1709 \pm t_{0.01/2;9}(482.3) = 1709 \pm 3.250(482.3) = (141.525, 3276.475)$$

Alternatively, create a modified version of the sample code given in the homework but with $\alpha = 0.01$:

```
data problem1; input cost @@; datalines;
1200 800 700 440 1100 4500 3000 600 4000 750
;

proc ttest data=problem1 h0=1000 sides=2 alpha=0.01;
run;
```

The TTEST Procedure

Variable: cost

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	1709.0	1525.3	482.3	440.0	4500.0

Mean	99% CL Mean		Std Dev	99% CL Std Dev	
1709.0	141.5	3276.5	1525.3	942.2	3474.1

Problem 1.c.

Prompt: Now test the null hypothesis that the mean cost is less than or equal to 1000 versus the alternative hypothesis that the mean cost is greater than 1000. What is the p-value?

Solution: You should have most, if not all, of the following:

State your null:

$$H_o : \mu_{cost} \leq 1000$$

State your alternative:

$$H_1 : \mu_{cost} > 1000$$

Identify what key info you have (given in SAS output):

$$\bar{X} = 1709 \quad SE = 482.3$$

Identify your test statistic (also given by SAS):

$$t = \frac{\bar{X} - \mu_o}{SE} = \frac{1709 - 1000}{482.3} = 1.47$$

Identify your rejection region, your decision:

At $\alpha = 0.05$ and df=N-1=9: $t_{0.05;9} = 1.833$. Reject the null if $t > t_{0.05;9}$. Since $t < t_{0.05;9}$ (1.47 vs 1.833), fail to reject the null hypothesis. *If you're using the p-value from SAS: at* α*=0.05, reject the null if* $p < 0.05$*. Since p=0.0878 > 0.05, we fail to reject the null hypothesis.*

Your interpretation: *(Brief 1-3 sentences here)*

The TTEST Procedure

Variable: cost

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	1709.0	1525.3	482.3	440.0	4500.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1709.0	824.8	Infty	1525.3	1049.2	2784.6

DF	t Value	Pr > t
9	1.47	0.0878

Problem 2

Prompt: The admission committee was wondering whether the average GRE scores had increased over time. To examine this question they looked at data from students applying in two different years, 2000 and 2019.

2000	2019
500.00	560.00
450.00	460.00
600.00	620.00
700.00	720.00
550.00	540.00
551.00	600.00
552.00	750.00

Problem 2.a.

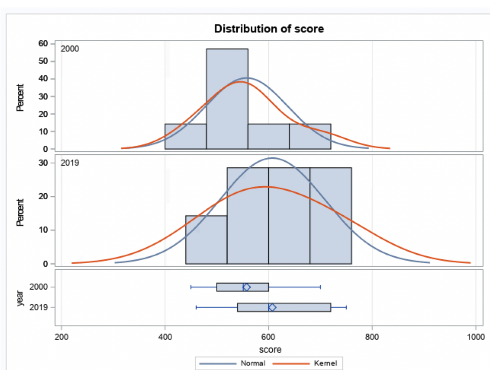
Prompt: Does the assumption that the samples come from two normal populations seem reasonable here? Why? Is the assumption that the two population variances are equal here correct? Why?

Solution: Many ways to solve this but always good to visualize. Below is a way to do the data input with proc ttest.

```
data problem2; input year score; datalines;
2000 500
2000 450
2000 600
2000 700
2000 550
2000 551
2000 552
2019 560
2019 460
2019 620
2019 720
2019 540
2019 600
2019 750
;

proc ttest data=problem2; class year; var score;
run;
```

For normality: *Approximately normal* for both years. You may use the subsequent plot (or any other graph/plot from other SAS procedures) to justify your answer. *Advanced: t-test is what we could call 'robust' against non-normality so we are not looking for a perfect bell curve. The next page will demostrate a more formal way to test for normality.*



A more formal way to test for normality: Adding `normal` to `proc univariate` will give you formal statistics on normality. In general, *Shapiro-Wilk* tests the null hypothesis that the sample in question came from a normally distributed population. The following demonstrates the code you'll need with two partial outputs for 2000 and 2019, respectively:

```
proc univariate data=problem2 normal;
      class year;
run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.961984	Pr < W	0.8356
Kolmogorov-Smirnov	D	0.163857	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.030801	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.204519	Pr > A-Sq	>0.2500

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.936035	Pr < W	0.6033
Kolmogorov-Smirnov	D	0.242527	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.061526	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.336416	Pr > A-Sq	>0.2500

For variance: You may use this output from `proc ttest` which tests the following null hypothesis: $\sigma^2_{year=2000} = \sigma^2_{year=2019}$. Based on the p-value of 0.5516, we fail to reject the null hypothesis–equality of variance is indeed a reasonable assumption.

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6	6	1.66	0.5516

Problem 2.b.

Prompt: What is the mean score (with 95% Confidence) for students in 2000? What about for 2019?

Solution:

Using the subsequent `proc ttest` output, we have:

- $\bar{X}_{2000} = 557.6$ with 95% CI of (484.8,630.3).
- $\bar{X}_{2019} = 607.1$ with 95% CI of (513.3,701.0).

year	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2000		557.6	484.8	630.3	78.6381	50.6739	173.2
2019		607.1	513.3	701.0	101.4	65.3685	223.4
Diff (1-2)	Pooled	-49.5714	-155.3	56.1288	90.7591	65.0821	149.8
Diff (1-2)	Satterthwaite	-49.5714	-156.0	56.8619			

Problem 2.c.

Prompt: Formally test whether the mean score differs for the two years at alpha=0.05.

Solution:

Please follow the framework presented in Problem 1.a.

We need to test this null hypothesis: $\mu_{2000} - \mu_{2019} = 0$.

Testing at $\alpha = 0.05$, reject our null if p < 0.05. Since $p = 0.3270 > 0.05$, we fail to reject our null. Any difference we may have observed could be due to random chance alone.

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	12	-1.02	0.3270
Satterthwaite	Unequal	11.298	-1.02	0.3282

Problem 3

Prompt: A Yale student wishes to invest some money in a new cosmetics company. The CEO of the cosmetics company says that they have a new product that makes users look years younger after using it. The Yale student is intrigued but asks for some data to support this claim. The CEO provides data for 5 persons before and after using the product. The data gives the age that the user felt they looked before and after the product use. The Yale student took this data and ran the following SAS program. The output for this program is also listed.

Problem 3.a.

Prompt: Do the data suggest that the cosmetic makes any difference in how old the user felt s/he looked? Use alpha = 0.05.

Solution: On average, users reported feeling 37.80 years younger after using the product. Since the 95% confidence interval does not contain 0, the result is indeed significant (or words to that effect). If you want to be more formal, you may use the t statistic and p-value for `agelook` to perform a formal test–the t statistic corresponds to the following null hypothesis: $\mu_{agelook} = 0$.

```
data problem3; input before after @@;
agelook=before-after;
cards;
76 23 75 44 88 56 56 29 66 20
run;

proc means mean std stderr t prt clm;
run;
```

The MEANS Procedure

Variable	Mean	Std Dev	Std Error	t Value	Pr > |t|	Lower 95% CL for Mean	Upper 95% CL for Mean
before	72.2000000	11.9666202	5.3516353	13.49	0.0002	57.3414785	87.0585215
after	34.4000000	15.2085502	6.8014704	5.06	0.0072	15.5160907	53.2839093
agelook	37.8000000	11.1220502	4.9739320	7.60	0.0016	23.9901507	51.6098493

Problem 3.b.

Prompt: Now test the CEO's specific claim that the cosmetic is associated with a 10 year reduction in perceived age.

Solution: *Follow the framework presented in Problem 1.a. since* `test` *is mentioned in the prompt.* You may hand calculate this or allow SAS to do the work for you using the code below which gives you a p-value of 0.0002. Using $\alpha = 0.05$, reject the null of $\mu_{agelook} = 10$ if p < 0.05. Since $p = 0.005 < 0.05$, we reject the null hypothesis. *(Your brief 1-2 sentences here for conclusion)*

```
proc ttest data=problem3 h0=10 sides=2 alpha=0.05;
      var agelook;
run;
```

Variable: agelook

N	Mean	Std Dev	Std Err	Minimum	Maximum
5	37.8000	11.1221	4.9739	27.0000	53.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
37.8000	23.9902	51.6098	11.1221	6.6636	31.9598

DF	t Value	Pr > |t|
4	5.59	0.0050

Problem 4

Prompt: A group of 12 friends tried three diets (1=Keto, 2=Weight Watchers, 3=South Beach) in an effort to get ready for the summer. Each friend tried each diet for one month and lost weight (in pounds) as below.

Problem 4.a.

Prompt: What is the overall variance in weight loss? How much of this variance is explained by the full model?

Solution: The key here is to fill out the partial ANOVA output given in the homework. Or to run the ANOVA code given in the homework. Once you have the relevant numbers, we can calculate the total variance using the following equation: $Var[Y] = \frac{SS_{total}}{DF_{total}} = \frac{182.75}{35} = 5.22$. According to the SAS output ($R^2 = 0.5294$), 52.94% of the variation is explained by the full model.

Problem 4.b.

Prompt: Fill in the statistical formulas for the table below.

Solution: Let `f` be the number of unique friends and `d` be the number of diets, then the table with the relevant formulas is simply the following:

Source	DF	SS
Model	(f-1)+(d-1)	$SS_{friends} + SS_{diet}$
Error	(f-1)(d-1)	$SS_{total} - SS_{friends} - SS_{diet}$
Total	fd-1	SS_{total}

where:

$$SS_{diet} = f\sum_{i=1}^{d}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2;$$

$$SS_{friends} = d\sum_{j=1}^{f}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2;$$

$$SS_{total} = \sum_{i,j}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

Note:

$$SS_{total} - SS_{friends} - SS_{diet} = \sum_{i,j}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$$

Problem 4.c.

Prompt: Is there any difference among the three diets in terms of weight loss?

Solution:

Here are some relevant output you may use to answer this:

Level of diet	N	weight	
		Mean	Std Dev
1	12	8.75000000	2.22076973
2	12	6.41666667	1.78164037
3	12	5.58333333	1.62135372

Source	DF	Anova SS	Mean Square	F Value	Pr > F
diet	2	64.66666667	32.33333333	8.27	0.0021
friend	11	32.08333333	2.91666667	0.75	0.6857

For diet: using $\alpha = 0.05$, reject the null of $\mu_K = \mu_{WW} = \mu_{SB}$ if p < 0.05. Since $p = 0.0021 < 0.05$, we reject the null hypothesis–the mean for at least one of the diets is significantly different.

Problem 4.d.

Prompt: If the variable FRIEND was not included in the analysis would your answer to part c be the same?

Solution:

Use the output below to comment on any difference. But do observe that removing `friend` from the model caused R^2 to drop from 0.53 to 0.35. For the `diet` only model, the F-statistic did increase from 8.27 to 9.04, which caused the p-value to drop from 0.0021 to 0.0007.

Partial Output With Diet and Friend:

R-Square	Coeff Var	Root MSE	weight Mean
0.529412	28.58519	1.977142	6.916667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
diet	2	64.66666667	32.33333333	8.27	0.0021
friend	11	32.08333333	2.91666667	0.75	0.6857

Partial Output for Diet Only:

R-Square	Coeff Var	Root MSE	weight Mean
0.353853	27.34894	1.891635	6.916667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
diet	2	64.66666667	32.33333333	9.04	0.0007

Problem 4.e.

Prompt: Which diet is the best in terms of weight loss?

Solution: The mean is highest for diet 1 (Keto). However, you need to make sure that the difference is significant. Here are three different ways to detect significance when multiple comparisons are involved:

```
data problem4; input friend diet weight @@; datalines;
1 1 6 1 2 5 1 3 5 2 1 7 2 2 10 2 3 4 3 1 10 3 2 7 3 3 7 4 1 10 4 2 8 4 3 4
5 1 6 5 2 8 5 3 8 6 1 12 6 2 8 6 3 4 7 1 8 7 2 4 7 3 4 8 1 6 8 2 6 8 3 8
9 1 12 9 2 6 9 3 6 10 1 8 10 2 5 10 3 4 11 1 10 11 2 5 11 3 6 12 1 10 12 2 5 12 3
7
;

proc anova data=problem4;
      class friend diet;
      model weight=diet friend;
      means diet / T CLDIFF;
      means diet / T CLM;
      means diet / T;
run;
```

Comparisons significant at the 0.05 level are indicated by ***.				
diet Comparison	Difference Between Means	95% Confidence Limits		
1 - 2	2.3333	0.6594	4.0073	***
1 - 3	3.1667	1.4927	4.8406	***
2 - 1	-2.3333	-4.0073	-0.6594	***
2 - 3	0.8333	-0.8406	2.5073	
3 - 1	-3.1667	-4.8406	-1.4927	***
3 - 2	-0.8333	-2.5073	0.8406	

diet	N	Mean	95% Confidence Limits	
1	12	8.7500	7.5663	9.9337
2	12	6.4167	5.2330	7.6003
3	12	5.5833	4.3997	6.7670