MATH 4044 – Statistics for Data Science

Practical Week 6 Solutions

Question 1

Is there a statistically significant difference between pulse rates for males and females? The data for this question is stored in a SAS data file called `pulse_rates.sas7bdat` located in `mydata` library on the SAS OnDemand server. Variables in that file are as follows:

Variable	Units
Height	Cm
Weight	Kg
Age	Years
Gender	1 = 'Male', 2 = 'Female'
Smokes	1 = 'Yes', 2 = 'No'
Drinks alcohol	1 = 'Yes', 2 = 'No'
Exercise Frequency	1 = 'High', 2 = 'Moderate', 3 = 'Low'
Pulse	Pulse rate, beats per minute

Check the assumptions and perform an appropriate hypothesis test. Interpret the results.

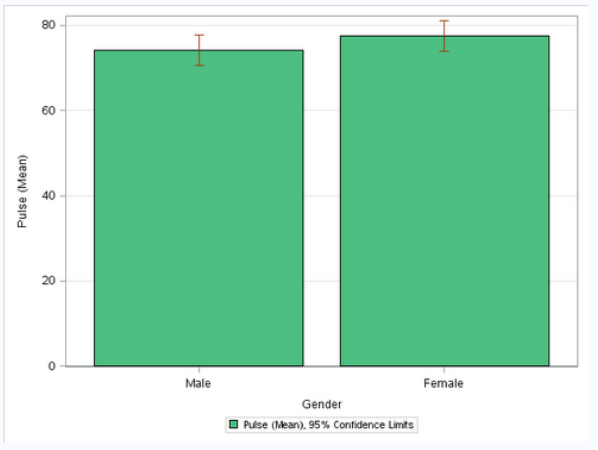

Figure 1. Bar chart of mean pulse rates by gender

Analysis Variable : Pulse						
Gender	N Obs	N	Mean	Std Dev	Minimum	Maximum
Male	59	59	74.153	13.759	49.000	145.000
Female	51	50	77.500	12.629	47.000	119.000

Figure 2. Descriptive Statistics for pulse rates by gender

The bar chart in Figure 1 shows some difference in sample pulse rates for male and female subjects, however this difference is unlikely to be statistically significant as there is a high degree of overlap in the 95% confidence limits.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.832535	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.148978	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.183609	Pr > W-Sq	0.0084
Anderson-Darling	A-Sq	1.35634	Pr > A-Sq	<0.0050

Figure 3. Normality test results for males

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.957941	Pr < W	0.0729
Kolmogorov-Smirnov	D	0.082533	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.078319	Pr > W-Sq	0.2200
Anderson-Darling	A-Sq	0.581929	Pr > A-Sq	0.1284

Figure 4. Normality test results for females

The two samples are independent and we shall assume that subjects were randomly selected. The remaining condition to check is Normality.

Results of Normality tests for males shown in Figure 3 suggest that the data cannot be assumed to have come from a Normally distributed population. *P*-values for all tests are less than 0.01 which indicates statistically significant departures from Normality. However, the sample size is large (from Descriptive Statistics in Figure 2, $n = 59 > 30$). In contrast, *P*-values for the same Normality tests for females, shown in Figure 4, are all greater than 0.05, suggesting that the distribution of pulse rates for females can be assumed to be Normal. In addition, the sample size is large (from Descriptive Statistics in Figure 2, $n = 50 > 30$).

Therefore, while only one of the populations can be assumed Normal, sample sizes are large so we will proceed with an independent *t*-test.

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Male	59	74.1525	13.7588	1.7912	49.0000	145.0
Female	50	77.5000	12.6285	1.7859	47.0000	119.0
Diff (1-2)		-3.3475	13.2532	2.5475		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Male		74.1525	70.5670	77.7381	13.7588	11.6473	16.8126
Female		77.5000	73.9110	81.0890	12.6285	10.5490	15.7368
Diff (1-2)	Pooled	-3.3475	-8.3977	1.7027	13.2532	11.6905	15.3018
Diff (1-2)	Satterthwaite	-3.3475	-8.3622	1.6673			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	107	-1.31	0.1917
Satterthwaite	Unequal	106.3	-1.32	0.1885

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	58	49	1.19	0.5404

Figure 5. Results of an independent *t*-test for pulse rates by gender

Let μ_1 denote the population mean pulse rate for males and μ_2 denote the population mean pulse rate for females. The hypotheses to be tested are then as follows:

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$

The Equality of Variance table in Figure 5 shows a *P*-value of 0.5404, which means that the hypothesis of equal variances cannot be rejected. We will therefore proceed with a pooled *t*-test.

The test statistic is t = -1.31 (df = 107). The corresponding *P*-value is 0.1917 > 0.05. Therefore, H_0 can't be rejected.

At 5% significance level, there is not enough statistical evidence to conclude that the population mean pulse rates for males and females are different.

Question 2

A random sample of 10 female university students was asked for their own height and their mother's height. The researchers wanted to know whether female university students are taller on average than their mothers. The results (in cm) were as follows:

Pair	1	2	3	4	5	6	7	8	9	10
Daughter	168	163	163	175	168	165	166	174	175	170
Mother	167	157	165	168	160	167	160	170	179	167

Find and interpret a 95% confidence interval for the parameter of interest in this situation using the *t*-distribution. Are the necessary conditions satisfied to justify using this confidence interval?

In this case we have two random and dependent samples. The population parameter of interest in this case is the population mean difference in height between daughters and mothers. We need to check whether differences in height come from a Normal distribution. The P-values in Figure 6 are greater than 0.05 for all Normality tests, so the hypothesis of Normality cannot be rejected.

Therefore, it is justifiable to use the confidence interval based on using the *t*-distribution.

From the TTEST procedure output shown in Figure 7, the 95% confidence interval for the difference in height between daughters and mothers is from -0.33cm to 5.74cm.

As this confidence contains zero, we would not be able to reject the null hypothesis of no difference in height on average between female university students and their mothers in the population at large.

[From Figure 7, the sample mean difference was 2.7cm, indicating that in our sample daughters were on average 2.7cm taller than their mothers. However, this difference is not significant at the 5% level, t(9) = 2.01, P-value = 0.0751 > 0.05.]

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.920328	Pr < W	0.3597
Kolmogorov-Smirnov	D	0.181591	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.054638	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.357349	Pr > A-Sq	>0.2500

Figure 6. Results of Normality tests for differences in height

The TTEST Procedure

Difference: daughter_height - mother_height

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	2.7000	4.2439	1.3421	-4.0000	8.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2.7000	-0.3359	5.7359	4.2439	2.9191	7.7478

DF	t Value	Pr > |t|
9	2.01	0.0751

Figure 7. Results of a dependent *t*-test for differences in heights

Question 3

Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel economy (in miles per gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Also obtain and interpret an appropriate confidence interval. Assume that conditions for inference are satisfied.

	City MPG	
	Automatic	**Manual**
Mean	16.12	19.85
Standard deviation	3.58	4.51
Sample size	26	26

Let μ_1 denote the population mean city MPG for cars with automatic transmissions, and μ_2 denote the population mean city MPG for cars with manual transmission. The hypotheses to be tested are then as follows:

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$

Since the samples are of the same size, a pooled *t*-test procedure with $n_1 + n_2 - 2$ degrees of freedom is acceptable.

Relevant statistics, P-value and confidence limits are shown in Figure 8.

x1	x2	n1	n2	sd1	sd2	df	sd_p	t	P_value	CL_Left	CL_Right
16.12	19.85	26	26	3.58	4.51	50	4.07164	-3.30302	.001771798	-5.99820	-1.46180

Figure 8. Calculation results for a pooled t-test and 95% confidence interval

The test statistic is t = -3.30 (df = 50). The corresponding P-value is 0.0018 < 0.05. Therefore, H_0 should be rejected.

At 5% significance level, there is enough statistical evidence to conclude that the population mean city MPG for cars with automatic and manual transmissions is different.

From the 95% confidence interval, the population mean city MPG for cars with automatic transmission is in fact between 1.46 and 6 miles per gallon lower than for cars with manual transmission.

APPENDIX. SAS code

```
ods graphics on;

/* Question 1 */

proc format ;
    value Gender 1='Male' 2='Female';
run;

/*--Extract first item from list--*/
/*--Set Graph Size (in inches)--*/
/*--Put statistic into macro variable--*/
 %let stat=Mean;

/*--Get variable names or labels--*/
data _null_;
    array x(1) Pulse;
    set MYDATA.PULSE_RATES;
    call symputx ("Label", vlabel(x(1)));
run;

/*--Put variabel name/label or custom label into macro variable--*/
data _null_;
    call symputx ("respLabel", "&Label");
run;

/*--Combine label and stat into statRespLabel--*/
 %let statRespLabel=&respLabel (&stat);

/*--%put statRespLabel="&respLabel";--*/
/*--Set output size--*/
ods graphics / reset width=6.4in height=4.8in imagemap;

/*--SGPLOT proc statement--*/
proc sgplot data=MYDATA.PULSE_RATES noautolegend;
format gender Gender.;
    /*--TITLE and FOOTNOTE--*/
    /*--Bar chart settings--*/
    vbar gender / response=Pulse fillattrs=(color=big) limits=Both
        limitstat=CLM numstd=1 transparency=0.00 stat=Mean dataskin=None
        name='Bar';

    /*--Category Axis--*/
    xaxis;

    /*--Response Axis--*/
    yaxis grid label="&statRespLabel";

    /*--Legend Settings--*/
    keylegend 'Bar' / location=Outside;
run;

proc means data=mydata.pulse_rates maxdec=3;
    var pulse;
    format gender Gender.;
    class gender;
    run;
```

```
proc univariate data=mydata.pulse_rates normal;
    var pulse;
    format gender Gender.;
    class gender;
run;

proc ttest data=mydata.pulse_rates;
    format gender Gender.;
    var pulse;
    class gender;
run;

/* Question 2 */

data work.paired_heights;
    input pair daughter_height mother_height difference;
    datalines;
1 168 167 1
2 163 157 6
3 163 165 -2
4 175 168 7
5 168 160 8
6 165 167 -2
7 166 160 6
8 174 170 4
9 175 179 -4
10 170 167 3
;
run;

proc print data=paired_heights;
run;

proc univariate data=work.paired_heights normal;
    var difference;
    histogram;
    probplot / normal(mu=est sigma=est);
run;

proc ttest data=work.paired_heights;
    paired daughter_height*mother_height;
run;

/* Question 3 */

data work.mpg_ttest;

    /* Defining variables to be stored in that file */
    x1=16.12;
    x2=19.85;
    n1=26;
    n2=26;
    sd1=3.58;
    sd2=4.51;

    /* Calculating degrees of freedom */
    df=n1 + n2 - 2;

    /* Calculating pooled variance*/
    sd_p=sqrt(((n1-1)*sd1**2 + (n2-1)*sd2**2)/(n1+n2-2));
```

```
    /* Calculating the t-test statistic */
    t=(x1-x2)/sqrt(sd_p**2/n1+sd_p**2/n2);

    /* Calculating P-value for a two-tailed test */
    P_value=2*(1-probt(abs(t), df));

    /* Calculating 95% confidence limits */
    CL_Left=(x1-x2) - TINV(.975, df)*sqrt(sd_p**2/n1+sd_p**2/n2);
    CL_Right=(x1-x2) + TINV(.975, df)*sqrt(sd_p**2/n1+sd_p**2/n2);
run;

proc print data=work.mpg_ttest noobs;
run;

ods graphics off;
```