Lab 3: QQplots, sampling distributions, confidence intervals, hypothesis testing (10 pts. + 2 pts. Bonus)

Objectives

- **Part 1: QQplot**
- **Part 2: Sampling Distributions**
  - **2.1) effect of size**
  - **2.2) Central Limit Theorem**
- **Part 3: Confidence Intervals/Hypothesis tests**
  - **3.1) Calculation of t critical values**
  - **3.2) Interpretation of Confidence Intervals**
  - **3.3) 1 – sample t-test**
  - **3.4) 2 – sample t-test (independent)**

Remember:
a) Please put your name, STAT 503 and the lab # on the front of the lab
b) Label each part and put them in logical order.
c) ALWAYS include your SAS code for each problem.

1. QQplots:

A Normal QQPlot can be used to visually assess the Normality of a data set. The Normal QQplot plots the observed data against the corresponding z-score which is calculated using the percentiles of the standard Normal. If the data is Normal, the points fall along a line. The QQplot code is a simple command line within proc univariate.

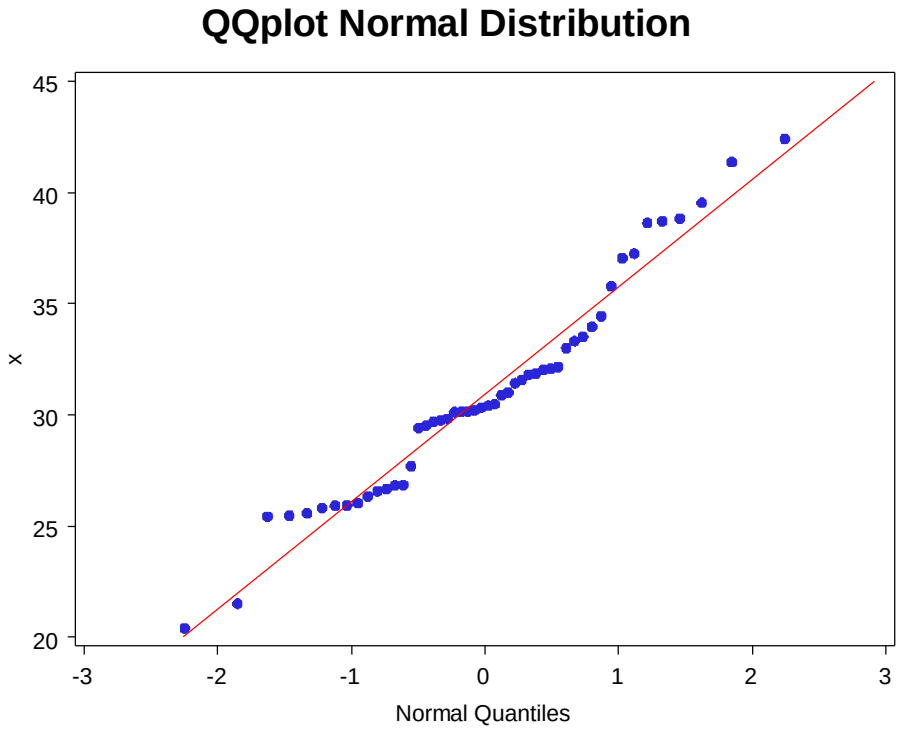
SAS Learning code: (c1.sas)

```
data normal;
infile 'H:\c1.dat';
input x;
proc print data=normal;
run;

title1 'QQplot Normal Distribution';
symbol1 v=dot height=1;
proc univariate data=normal normal;
qqplot x/height = 3 normal (mu=est sigma=est color=red) ;
run;

quit;
```

SAS Learning Output



Problem 1 (2 pts.)

The data file Ex.3.1.dat contains the amount of precipitation in March over a 30-year period in Minneapolis-St. Paul. Construct a QQPlot for the amount of precipitation. Is the data normally distributed, right skewed or left skewed?

Your submission should consist of your code, the QQPlot generated from SAS and the answer to the shape of the distribution..

2.Sampling Distributions:

In this module, we will use the same random number generation tool but look at the sampling distribution of the sample mean and sample proportion. This allows a visual inspection of how well the Normal distribution fits the randomly generated data. The larger the sample size, the better the overall fit.

The example that we are using will use the uniform distribution. The theoretical values for the population mean is ½ and for the population standard deviation is $\sqrt{\frac{1}{12}}$.

SAS Learning code: (c2.sas)

```
*Code modified from from www.stanford.edu/~kcobb/hrp259/lab1_EG.doc
6/20/2012: 4 pm;
%LET repeats=1000; *we are not changing this in this lab;
%LET n=6; *this is the only value that is changed for this lab;
data uniform;
do j=1 to &repeats by 1;
            avg=0;
            do i=1 to &n by 1;
                  avg=avg+rand('Uniform');
           end;
        avg=avg/&n;
      output;
      drop i;
end;
run;

title1 'Uniform Distribution repeats = ' &n;
proc means data=uniform; *prints out means and standard deviations;
   var avg;
run;

proc univariate data=uniform noprint;
      var avg;
      histogram / endpoints = 0 to 1 by .05;
run;

quit;
```

Problem 2 (2 pts.)

Run the preceding learning code for n = 1, n = 2 and n = 6.

Your submission should consist of the three generated histograms, the three outputs from proc means, a table consisting of the three means from the output and the theoretical value and the three standard deviations from the output and the theoretical value. The theoretical value for the standard deviation for each run is the standard error given the appropriate value of n. Does the calculated values match the theoretical values? In addition, please indicate which of the graphs look like they are a normal distribution (this can be written directly on the output). No code is required because you are making minor modifications to the learning code.

3. Confidence Intervals/Hypothesis tests:

3.1. Calculation of t critical values

As mentioned in class, the command in SAS to calculate TINV(p,df) where p = 1 - $\frac{\alpha}{2}$ and df = the degrees of freedom. The command is called TINV because this is calculating the percentile (or inverse) of the t distribution (function PROBT).

SAS Learning code: (c3.sas)

```
data tcritval;
*p = 1 - alpha/2, df = degrees of freedom;
alpha = 0.05;
p = 1 - alpha/2;
df = 20;
CritVal = TINV(p,df);

proc print data=tcritval; run;
quit;
```

SAS Learning output:

Obs	alpha	p	df	CritVal
1	0.05	0.975	20	2.08596

$t(df)_{\alpha/2} = t(20)_{0.025} = 2.08596$

Problem 3 (1 pt.)

Calculate the following:

a) t critical value when $\alpha = 0.05$, n = 22.

b) t critical value when $\alpha = 0.01$, n = 37

Your submission should consist of the answers to parts a and b, your code (clearly indicating which part is coming from which line) and the appropriate parts of the output file.

3.2. Interpretation

In this part, you will calculate 20 different 80% confidence intervals and then determine how many of them actually contain the true value of the mean. (I know that 80% is very low, but with the low number of runs, I wanted a good chance that the mean would be outside the calculated interval.) You will do this by running the learning code (c4.sas) 20 times. This code creates a sample of 10 from the normal distribution with $\mu = 5.5$, $\sigma = 1.4$. Then, you will manually count how many of the intervals contain the true value of the mean.

The confidence interval is indicated in the bolded part of the learning output as shown in the interval listed below the learning output. The procedure for calculating confidence intervals will be explained in more detail in 3.3.

SAS Learning code: (c4.sas)

```
data normal;
do i=1 to 10;
   y = rand('Normal',5.5, 1.4); output;
end;

proc ttest data=normal alpha=0.20;
  var y;
run;

quit;
```

SAS Learning output:

The TTEST Procedure

Variable: y

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	4.7456	1.4324	0.4530	2.6608	6.9958

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.7456	3.7209	5.7702	1.4324	0.9852	2.6150

DF	t Value	Pr > |t|
9	10.48	<.0001

The confidence interval is (3.7209, 5.7702)

Problem 4 (1 pts.)

Run the preceding learning code 20 times to determine the number of confidence intervals that include the theoretical mean.

You submission should consist of a sample output with the confidence interval indicated and the number of times that the confidence interval contains the true (theoretical) mean. In addition, please comment on the number of confidence intervals that you would expect to include the mean assuming a 80% confidence level.

3.3. Inference for 1 - sample for the mean:

To have SAS calculate the confidence intervals and hypothesis testing, we use the proc ttest. As I alluded to in class, we specify the α not the confidence level which is 1 - α. Therefore, the confidence level is the same for a 95% confidence interval and a hypothesis test with α = 0.5. For the hypothesis test, the value of μ_0 is given by H0 = value in the procedure line.

This example in the learning code is taken from the textbook (Example 6.7.3) concerning the thorax weight from male and female Monarch butterflies. I will be using the same example for the next section (3.4) when we will be looking a 2 – sample inference for the mean. In this example, we are a) determining the 95% confidence interval and b) performing the hypothesis test to see if the weight of the male thorax is the same as the average weight of the female thorax (63.4)

SAS Learning code: (c5.sas)

```
data thorax;
infile 'H:\c5.dat';
input male;
run;

proc print data=thorax; run;

Title 'One Sample Inference';
proc ttest data=thorax alpha=0.05 H0 = 63.4;
   *alpha: the default value is 0.05, it is required for
         both CI and hypothesis tests.;
   *H0: The default value is 0, it is only required for hypothesis tests.;
var male;
run;

quit;
```

SAS Learning output:

The TTEST Procedure
Variable: male

N	Mean	Std Dev	Std Err	Minimum	Maximum
7	75.7143	8.4007	3.1752	63.0000	85.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
75.7143	67.9450	83.4836	8.4007	5.4133	18.4989

DF	t Value	Pr > |t|
6	3.88	0.0082

The 95% Confidence interval is (67.9450, 83.4836).
For the hypothesis test with α = 0.05 (95% confidence interval),
df = 6, t value = 3.88 and the p-value = 0.0082

Problem 5 (2 pts.)
An experiment is performed to see if dietary chromium has an effect on the activity level of the liver enzyme GITH in rats. Fourteen rats were fed special low-chromium diets, and the activity level of this enzyme was measured for each rat (Inference1.dat). It is believed that this level follows a Normal distribution. The mean activity using a normal diet is 53.2.
a) Calculate the 95% confidence interval for this example (α = 0.05). Is the mean activity of the normal diet in the confidence interval?
b) Perform a hypothesis test (significance level = 0.05) to determine if the activity if a low-chromium diet is different from the activity of this enzyme with a normal diet.
Your submission should consist of one code for both parts and the relevant parts of the output in addition to the confidence interval, the mean value, whether the mean value is in the confidence interval, t-value, degrees of freedom, the p-value and the conclusion of the hypothesis test in the context of the example.

Bonus B1: (1 pt.)
Test whether the hypothesis that this data set is normal is correct.
Your submission should consist of the code and the relevant output.

3.4. Inference for 2-sample for two means:

For two means, we will again use “proc ttest”. However, like in the boxplot, we need to indicate which value is for which of the two cases. In example 6.7.3:

weight	gender
67	Male
73	Male
…	
54	Female
61	Female
…	

Be aware that you must sort the data by the group variable first, in order for the proc ttest to work correctly. Because you are sorting the data, be careful to determine what is ‘1’ and what is ‘2’. SAS always does the inference as ‘1’ – ‘2’. H0 is NOT required in the procedure statement.

For pooled variance, use the ‘Pooled’ rows, for unpooled variance, use the ‘Satterthwaite’ rows. The section titled ‘Equality of Variances’ is a hypothesis test whether the variances are the same.

SAS Learning code: (c6.sas)

```
data thorax;
infile 'H:\c6.dat';
input weight gender$;
run;

proc print data=thorax; run;

proc sort data=thorax;
by gender;
run;

Title 'Two Sample Inference';
proc ttest data=thorax alpha=0.05 ;
class gender;
var weight;
run;

quit;
```

SAS Learning output:

The TTEST Procedure
Variable: weight

gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	8	63.3750	7.5392	2.6655	54.0000	75.0000
Male	7	75.7143	8.4007	3.1752	63.0000	85.0000
Diff (1-2)		-12.3393	7.9484	4.1137		

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		63.3750	57.0721	69.6779	7.5392	4.9847	15.3443
Male		75.7143	67.9450	83.4836	8.4007	5.4133	18.4989
Diff (1-2)	Pooled	-12.3393	-21.2264	-3.4522	7.9484	5.7622	12.8052
Diff (1-2)	Satterthwaite	-12.3393	-21.3531	-3.3255			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	13	-3.00	0.0102
Satterthwaite	Unequal	12.23	-2.98	0.0114

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	6	7	1.24	0.7751

The 95% Confidence interval (pooled) is (-21.2264, -3.4522), df = 13
The 95% Confidence interval (unpooled) is (-21.3531, -3.3255), df = 12.23

For the hypothesis test with $\alpha = 0.05$ (again from the 95% CI),
pooled: df = 13, t value = -3.00 and the p-value = 0.0102
unpooled: df = 12.23, t-value = -2.98, p-value = 0.0114.

Note: the numbers are opposite those in the example in the book because the book performs the test male – female, and SAS performs the test female – male because female is first alphabetically.

Notice also that the p-value is different in the parts 3.3 and 3.4. The latter is better because it includes more information, the variability of the data for the female butterflies in addition to the average value.

Problem 6 (2 pts.)

Now suppose that we do not know the activity level of the enzyme in rats fed a normal diet. Ten additional rats are fed a normal diet, and their levels are measured and recorded (Inference2.dat). Let μ_1 refer to the mean level for the low-Cr rats, and μ_2 refer to the mean level for the normal rats. We will be assuming that the variances are NOT the same.

a) Calculate the 95% confidence interval for this example. Are the two mean levels the same?

b) Perform the hypothesis test (significance level = 0.05) where the null and alternative hypotheses are H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$.

Your submission should consist of one code for both parts and the relevant parts of the output in addition to the confidence interval, whether the two means the same (using the confidence interval), t-value, degrees of freedom, the p-value and the conclusion of the hypothesis test in the context of the example. Be sure to use the appropriate method to calculate the variance (pooled or unpooled)

Bonus B2: (1 pt.)

Test whether the hypothesis that the data set for normal chromium levels has a normal distribution.

Your submission should consist of the code and the relevant output. Note: If you have to change the infile, the modified infile needs to be included also.