Statistical Methods and Computing, STAT:2100/4200
Lab 5

1 SAS for confidence interval of population mean: proc means

We can use *proc means* to get various summary statistics in a more compact format than *proc univariate* provides. The default statistics provided are

- n = number of observations
- mean
- std dev = standard deviation
- minimum
- maximum

Exercise 14.26 from the textbook for example (homework 6 problem). Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral. Researchers measured the percent change in mineral content of the spines of 47 mothers during three months of breast-feeding. The data are on the course web page as boneloss.dat.

```
data boneloss;
input change;
datalines ;
-4.7
-2.5
...
< copy data in here>
;
run ;

proc means;
var change; * numeric variable for calculating statistics ;
run ;
```

The MEANS Procedure

Analysis Variable : change				
N	Mean	Std Dev	Minimum	Maximum
47	-3.5872340	2.5056126	-8.3000000	2.2000000

Further **SAS proc means computes confidence intervals for population means** without making the unrealistic assumption that we know the the population standard deviation σ.

```
proc means n mean clm ;
var change; * numeric variable for calculating statistics ;
run ;
```

The MEANS Procedure

Analysis Variable : change			
N	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
47	-3.5872340	-4.3229095	-2.8515586

2 SAS for one-sample t-tests

- SAS automatically does a two-sided test

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

Example: Body temperature. It is widely believed that the average body temperature for healthy humans is 98.6 F. We think that might not be true, so we decide to do a two-sided significance test at significance level $\alpha = .05$:

$$H_0 : \mu_{\text{temp}} = 98.6$$
$$H_a : \mu_{\text{temp}} \neq 98.6$$

A random sample of 130 healthy people is taken and each person's temperature is recorded in "normtemp.dat".

```
data normtemp ;
input temp gender heart ;
datalines;
* note: copy and paste data in here ;
;
run ;
```

The next two pages show two approaches to carry out a two-sided hypotheses test, they always lead to the same decision (ether reject $H_0 : \mu = \mu_0$, or fail to reject H_0).

Approach #1 is to perform a t test and obtain the p-value. Approach #2 is to build a confidence interval (using t table) and see if it contains μ_0.

In theory, the assumptions for a t-test and a confidence interval (using t table) to be valid are SRS, and normality of the population distribution. In practice, use the rules of thumb regarding one-sample t procedures (see chap18-19-extra notes, page 3). Are the rules satisfied here?

In SAS, use *proc univariate* to help check the rule of thumb:

```
proc univariate plot data = normtemp ;
var temp;
run ;
```

Approach #1. Perform a t test

In SAS, One-sample t-tests are performed using **proc univariate**

```
proc univariate mu0 = 98.6 data = normtemp ;
var temp;
run ;
```

```
                    The UNIVARIATE Procedure
                         Variable:  temp


                  Tests for Location: Mu0=98.6

          Test           -Statistic-    -----p Value------

          Student's t    t  -5.45482    Pr > |t|  <.0001
          Sign           M       -21    Pr >= |M| 0.0002
          Signed Rank    S     -1963    Pr >= |S| <.0001
```

(OPTIONAL: By the way, SAS proc univariate allows you to do two tests at the same time, for example, one for body temperature ($H_0 : \mu_{\text{temp}} = 98.6$ vs $H_a : \mu_{\text{temp}} \neq 98.6$), and one for heart rate (say, $H_0 : \mu_{\text{heart}} = 73$ vs $H_a : \mu_{\text{heart}} \neq 73$).

```
proc univariate mu0 = 98.6 73 data = normtemp ;
var temp heart;
run ;
```

```
                    The UNIVARIATE Procedure
                             Variable:  temp


                      Tests for Location: Mu0=98.6

              Test           -Statistic-    -----p Value------

              Student's t    t  -5.45482    Pr > |t|  <.0001
              Sign           M       -21    Pr >= |M| 0.0002
              Signed Rank    S     -1963    Pr >= |S| <.0001


                             Variable:  heart


                      Tests for Location: Mu0=73

              Test           -Statistic-    -----p Value------

              Student's t    t  1.229507    Pr > |t|  .2211
              Sign           M       6    Pr >= |M| 0.3153
              Signed Rank    S     522.5    Pr >= |S| .1718
```

Approach #2. Construct a 95% confidence interval, and check if the value of μ specified in the null hypothesis falls in it. If it does, then it implies that the p-value is greater than .05, so we cannot reject H_0.

```
proc means n mean stddev clm alpha = .05 ;
var temp ;
run ;
```

Variable	N	Mean	Std Dev	Lower 95.0% CLM	Upper 95.0% CLM
temp	130	98.249	0.733	98.122	98.376

Note that:

- The 95% confidence interval for μ does not contain 98.6.
- The p-value is less than .05, so we reject the null hypothesis.

3 Paired-sample problems (one-sample t test)

To carry out the hypothesis test of interest, we apply one-sample procedures to the *differences* between values measured on members of each pair.

Example: To study the effect of cigarette smoking on platelet aggregation, researchers drew blood samples from 11 individuals before and after they smoked a cigarette and measured the percentage of blood platelet aggregation. This study can be found in Levine, P. H. (1973). An acute effect of cigarette smoking on platelet function, Circulation, 48, 619-623. The data is available in ICON/modules/lab worksheets/smoking.dat.

We test the null hypothesis that the means before and after are the same. Use $\alpha = 0.05$.

We will do a two-sided test, because we are not sure in advance whether to expect μ_1 (mean percentage of blood platelet aggregation before smoking) to be higher or lower than μ_2 (mean percentage of blood platelet aggregation after smoking).

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

or equivalently:

$$H_0 : \mu_2 - \mu_1 = 0$$
$$H_a : \mu_2 - \mu_1 \neq 0$$

or equivalently:

$$H_0 : \delta = 0$$
$$H_a : \delta \neq 0$$

where δ denotes $\mu_2 - \mu_1$.

We will use the *observed differences* between the percentage after and before smoking observed on each subject as our data to carry out the hypothesis test regarding δ at the .05 significance level.

We will carry out a t-test. In theory, the assumptions for a t-test to be valid are SRS, and approximate normality of the population distribution (here the distribution of differences). In practice, use the rules of thumb regarding one-sample t procedures (see chap18-19-extra notes, page 3). Are the rules satisfied here?

Note that by default, `proc univariate` tests the null hypothesis that $\mu = 0$ ($\delta = 0$ if using the symbols above), so in this case we don't have to specify a value for `mu0` in the command line.

```
data smoking ;
input before after ;
diff = after - before ;
datalines ;
* note: copy and paste data in here ;
;
run ;

proc univariate plot data = smoking ;
var diff ;
run ;

proc means data = smoking n mean stddev stderr clm alpha = .05 ;
var diff ;
run ;
```

The UNIVARIATE Procedure
Variable: diff

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	4.271609	**Pr > |t|**	0.0016
Sign	**M**	4.5	**Pr >= |M|**	0.0117
Signed Rank	**S**	32	**Pr >= |S|**	0.0020

The SAS System

The MEANS Procedure

Analysis Variable : diff					
N	**Mean**	**Std Dev**	**Std Error**	**Lower 95% CL for Mean**	**Upper 95% CL for Mean**
11	10.2727273	7.9761007	2.4048848	4.9143099	15.6311446

An alternative way to do paired t test in SAS is to use proc ttest:

```
proc ttest data = smoking ;
paired after*before ;
run ;
```

More stuff, if time permits

4 Using `proc tabulate` to summarize the distributions of quantitative variables in different groups

Gulanick (*Heart and Lung*, 1991) studied patients who were recovering from heart surgery. She was interested in whether different combinations of supervised exercise or teaching would affect patients' self-efficacy (or confidence) to perform physical activity.

Patients were randomly assigned to one of three groups. Group 1 received teaching, treadmill exercise testing, and exercise training three times per week. Group 2 received only teaching and exercise testing. Group 3 received only routine care without supervised exercise or teaching. After 4 weeks, each patient was scored on self-efficacy.

Self-efficacy was measured on a continuous scale and scores were assumed to be distributed normally in each of the populations of interest. Her results are in the dataset "gulanick.dat." We wish to produce a table that shows the number of observations and the mean and standard deviation of scores within each of the three groups.

```
proc format ;
value grpfmt 1 = 'Teaching and Training' 2 = 'Teaching' 3 = 'Neither' ;
run ;

data gulan ;
input score group ;
format group grpfmt. ;
datalines ;
< copy data in here>
;
run ;

proc univariate plot data=gulan;
var score;
by group;
run;

proc tabulate data = gulan ;
class group ;  * class statement identifies qualitative variables ;
var score ;    * var statement identifies quantitative variables ;
tables group , score * (n mean std) ;
run ;
```

5 Another example on using `proc means` to obtain confidence intervals

In the Gulanick example, what three populations are we interested in?

In each of the populations, what variable are we interested in?

From our proc tabulate output, we already have point estimates for the three population means. What are they?

Now we can use proc means to get confidence intervals for the three population means. Note that for each of the three groups, the samples are small, so we must examine the observed data and make sure that
(1) there are no outliers, and
(2) the sample is not highly skewed.
Otherwise, a z or t confidence interval is not valid.

We already have the boxplots from proc univariate to help check the above conditions. Are the conditions satisfied?
If so, it is safe to obtain confidence intervals. (SAS always assume that σ is unknown, and hence produce a t confidence interval.)

```
proc means n mean clm ;
var score ; * numeric variable for calculating statistics ;
class group ; * qualitative variable for defining groups ;
run ;
```

```
The MEANS Procedure

Analysis Variable : score
group                  N Obs  N  Mean      Lower 95%       Upper 95%
                                           CL for Mean     CL for Mean

Teaching and Training  11  11  126.8181818  110.5254093  143.1109543
Teaching               12  12  128.4166667  112.5048195  144.3285138
Neither                13  13  103.9230769   93.2221772  114.6239767
```

What quantity are we 95% confident lies in the interval (110.53, 143.11)?

Be default, proc means gives 95% confidence intervals. We can request different confidence levels by including the alpha option. For example, the following code will produce 90% confidence intervals. Do you expect 90% confidence intervals to be wider or narrower than 95% intervals? (Answer before you run the code.)

```
proc means n mean clm alpha = 0.10 ;
var score ; * numeric variable for calculating statistics ;
class group ; * qualitative variable for defining groups ;
run ;
```