SAS Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery
Chapter 8: Test of Hypotheses

Note 8.1 – Testing a hypothesis about the mean assuming the variance is unknown

The SAS procedure **PROC TTEST** can be used to test a hypothesis about the mean when the population variance in unknown. As an example, we use the **DIG200** data set to test the null hypothesis that the population mean is significantly different from 122.3 mmHg. In order to do this, we must use the option H0, a capital h followed by the number zero, after the procedure **PROC TTEST** and set it equal to value under the null hypothesis, 122.3. Finally, we use the VAR statement followed by SYSBP, the variable containing systolic blood pressure information for each individual.

Under the heading Statistics, the SAS output provides the number of observations (N), the mean (Mean), the lower and upper 95% confidence interval values for the mean (Lower CL Mean and Upper CL Mean), standard deviation (Std Dev), the lower and upper 95% confidence interval values for the standard deviation (Lower CL Std Dev and Upper CL Std Dev), and the standard error (Std Err). Under the heading T-Tests, the SAS output provides the degrees of freedom (DF), the value of the t-statistic (t Value), and a p-value associated with a two-tailed test (Pr > | t |).

SAS commands:

```
PROC IMPORT FILE='C:\DIG200.XLS' OUT=DIGDATA REPLACE;
RUN;

DATA DIG200;
  SET DIGDATA;
PROC TTEST H0=122.3;
  VAR SYSBP;
RUN;
```

SAS output:

The SAS System

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
sysbp	199	123.28	125.82	128.37	16.551	18.179	20.164	1.2886

T-Tests

Variable	DF	t Value	Pr > |t|
sysbp	198	2.73	0.0068

The ALPHA option can be used to change the level of the confidence interval. For a 99% confidence interval, we can set ALPHA = 0.01. In this case, the confidence interval will become wider compare to the 95% confidence and extends from 122.47 to 129.18 as shown in the SAS output below. From the textbook, the test statistic is calculated as follows,

$$t = \frac{\bar{x} - \mu_0}{\left(s / \sqrt{n} \right)}.$$

Therefore the lower and upper limits of the (1-α)*100% confidence interval are calculated as follows,

$$\left(\bar{x} - t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right).$$

For a 99% confidence interval, we simply have to do the following calculation ($125.82 - t_{199\text{-}1,\ 1\text{-}0.01/2}*1.29$, $125.82 + t_{199\text{-}1,\ 1\text{-}0.01/2}*1.29$). Using $t_{199\text{-}1,\ 1\text{-}0.01/2} = t_{198,\ 0.995} = 2.60$, we are able to obtain the 99% confidence interval (122.47, 129.17) which is consistent with the SAS output below although there is a slight discrepancy do to rounding.

SAS commands:

```
DATA DIG200;
  SET DIGDATA;
PROC TTEST H0=122.3 ALPHA=0.01;
  VAR SYSBP;
RUN;
```

SAS output:

The SAS System

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
sysbp	199	122.47	125.82	129.18	16.081	18.179	20.851	1.2886

T-Tests

Variable	DF	t Value	Pr > |t|
sysbp	198	2.73	0.0068

Note 8.2 – Testing the hypothesis about a population proportion

Using Example 8.4 to illustrate the use of the SAS command **PROC TTEST**, we first create the data set of interest on immunization of 5 year olds using the SAS commands below. Notice that the DO loop creates two variables. The first variable ID will contain values from '1' to '140'. Then SAS creates 140 values, all ones, for the variable IMMUNIZATION. The IF – THEN statement identifies those values corresponding to the variable ID that are greater than or equal to 87 and replaces them with zeros. Therefore 86 ones correspond to children who have been immunized and the 54 zeros correspond to children who have not been immunized.

SAS commands:

```
DATA IMLEVEL;
DO ID = 1 TO 140;
IMMUNIZATION = 1;
IF ID >= 87 THEN IMMUNIZATION = 0;
OUTPUT;
END;
PROC PRINT;
RUN;
```

Now that we have a data set named **IMLEVEL** that contains 86 children who have had an immunization. First, we will consider the follow null and alternative hypotheses as shown in the textbook:

H0: $\pi = \pi_0 = 0.75$ versus
Ha: $\pi < \pi_0 = 0.75$.

The test statistic which does not contain a continuity correction and is displayed below represents a case where we can use the normal approximation to the binomial distribution,

$$z = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}}.$$

In this case, we are assuming that the above test statistic, *z*, follows a standard normal distribution. Since we are using **PROC TTEST**, we must specify H0 = 0.75 which is the value under the null hypothesis. SAS calculates its t Value or test statistic using the following equation,

$$t\ Value = \frac{\bar{X} - \mu}{s/\sqrt{n}}.$$

The t Value = (0.6143 – 0.75)/(0.4885/sqrt(140)) = -3.29, where sqrt(140) is the square root of 140. Another thing to be aware of is that the Pr > | t | or p-value provided from the SAS output below is calculated using a t distribution with 139 degrees of freedom.

SAS commands:

```
PROC TTEST DATA = IMLEVEL H0=0.75;
  VAR IMMUNIZATION;
RUN;
```

SAS output:

The SAS System

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
IMMUNIZATION	140	0.5327	0.6143	0.6959	0.4372	0.4885	0.5536	0.0413

T-Tests

Variable	DF	t Value	Pr > |t|
IMMUNIZATION	139	-3.29	0.0013

Note 8.3 – Correlation coefficients and their p-values

SAS uses the procedure **PROC CORR** to calculate the Pearson correlation coefficient as mentioned in the program notes for chapter 3. In Example 8.7, we provide the following null and alternative hypotheses for the correlation between infant mortality rates for 1988 (MRATE) and total health expenditures as a percentage of GDP in 1987 (HEALTHEXP) for 21 countries:

H0: $\rho = \rho_0 = 0$ versus
Ha: $\rho \neq \rho_0 = 0$.

The two values provided by the SAS output below under Simple Statistics are the Pearson correlation coefficient and its associated p-value indicated by Prob > | r | under H0: Rho = 0.

SAS commands:

```
DATA CORRELATE;
INPUT MRATE HEALTHEXP;
DATALINES;
 4.8  6.8
 5.8  9.0
 6.1  7.4
 6.8  8.5
 6.8  7.7
 7.2  8.6
 7.5  8.2
 7.5  6.0
 7.8  8.6
 8.1  6.0
 8.1  7.1
 8.3  7.5
 8.7  7.1
 8.9  7.4
 9.0  6.1
 9.2  7.2
 9.3  6.9
10.0 11.2
10.8  6.9
11.0  5.3
13.1  6.4
;
PROC CORR;
  VAR MRATE HEALTHEXP;
RUN;
```

SAS output:

```
                                The SAS System

                              The CORR Procedure
```

2 Variables: MRATE HEALTHEXP

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
MRATE	21	8.32381	1.90969	174.80000	4.80000	13.10000
HEALTHEXP	21	7.42381	1.30227	155.90000	5.30000	11.20000

Pearson Correlation Coefficients, N = 21
Prob > |r| under H0: Rho=0

	MRATE	HEALTHEXP
MRATE	1.00000	-0.24291 0.2887
HEALTHEXP	-0.24291 0.2887	1.00000

Note 8.4 – Testing the hypothesis of no difference in two population means assuming unknown but equal variances

In Example 8.9, we would like to test the hypothesis that there is no difference in the population mean proportions of total calories coming from fat between fifth/sixth-grade boys and seventh/eighth-grade boys. The null and alternative hypotheses where μ_1 is the population mean proportion of total calories from fat for fifth/sixth-grade boys and μ_2 is the population mean proportion of total calories from fat for seventh/eighth-grade boys can be expressed as follows:

H0: $\mu_1 - \mu_2 = 0$ versus
Ha: $\mu_1 - \mu_2 \neq 0$,

The SAS procedure **PROC TTEST** can again be used to test the hypothesis of no difference between two population means. The SAS procedure **PROC TTEST** provides a test statistic and p value for both the situation that assumes that the two population variances are equal and the situation that assumes that the two population variances are unequal. In Example 8.9, we are assuming equal variances, and we use **PROC TTEST** with the CLASS option followed by the variable GROUP to distinguish between the two groups being compared as shown in the SAS commands below. The variable GROUP is equal to '0' to indicate the boy is from the fifth/sixth grade and equal to '1' to indicate the boy is from the seventh/eighth grade. The variable named PROP_FROM_FAT follows the SAS option VAR and contains the proportion of total calories from fat for each individual. Under T-Tests, you should notice that we are using the t Value associated with Equal variances. The equation for the test statistic contains n_1 and n_2 which are equal to the size of the samples for the two groups, $\Delta_0 = 0$ under the null hypothesis, and s_p

referred to as the s-pooled. The equation for the test statistic as provided by the textbook is shown below:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

The formula for s_p, where s_1^2 and s_2^2 are the sample variances for the two groups (refer to Chapter 7 for more details), can be calculated as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

SAS commands:

```
PROC IMPORT FILE = 'C:\TABLE7-7.XLS' OUT = TABLE7_7 REPLACE;
RUN;

DATA CALORIES;
  SET TABLE7_7;
PROC TTEST;
  CLASS GROUP;
  VAR PROP_FROM_FAT;
RUN;
```

SAS output:

The SAS System

The TTEST Procedure

Statistics

Variable	group	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
prop_from_fat	0	19	0.3096	0.3527	0.3958	0.0675	0.0894	0.1321	0.0205
prop_from_fat	1	14	0.2731	0.3293	0.3855	0.0706	0.0974	0.1569	0.026
prop_from_fat	Diff (1-2)		-0.043	0.0234	0.0901	0.0744	0.0928	0.1234	0.0327

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
prop_from_fat	Pooled	Equal	31	0.72	0.4795
prop_from_fat	Satterthwaite	Unequal	26.7	0.71	0.4861

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
prop_from_fat	Folded F	13	18	1.19	0.7202

Note 8.5 – Testing the hypothesis of no difference in two population means assuming unequal variances

In Example 8.10, we are examining the mean ages of AML and ALL patients.

H0: $\mu_1 - \mu_2 = 5$ versus
Ha: $\mu_1 - \mu_2 > 5$.

In this example, we are assuming unequal variances, and we use **PROC TTEST** with the CLASS option followed by the variable DX_TYPE to distinguish between the two diagnosis groups being compared as shown in the SAS commands below. The variable DX_TYPE is equal to '0' to indicate the patients are AML and equal to '1' to indicate that they are ALL. The variable AGE following the VAR option contains the age of each patient. In the SAS output below, we select Satterthwaite under Method and choose the corresponding DF, degrees of freedom, and t Value.

SAS commands:

```
PROC IMPORT FILE = 'C:\TABLE7-8.XLS' OUT=TABLE7_8 REPLACE;
RUN;

DATA CALORIES;
  SET TABLE7_8;
PROC TTEST H0=5;
  CLASS DX_TYPE;
  VAR AGE;
RUN;
```

SAS output:

The SAS System

The TTEST Procedure

Statistics

Variable	dx_type	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
age	0	51	45.219	49.863	54.506	13.815	16.511	20.524	2.312
age	1	20	28.297	36.65	45.003	13.573	17.848	26.068	3.991
age	Diff (1-2)		4.3233	13.213	22.102	14.481	16.889	20.266	4.456

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
age	Pooled	Equal	69	1.84	0.0696
age	Satterthwaite	Unequal	32.5	1.78	0.0843

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
age	Folded F	19	50	1.17	0.6402

Note 8.6 – Paired t-test

To show how SAS can be used to conduct a paired t-test, let us suppose that a research would like to examine the effect of a diet pill on weight loss. In the SAS commands below, we display data on the weight of ten patients before and after being on the diet pill for 3 months. Here we create the new variable DIFF to calculate the difference in weight before and after the introduction of the diet pill for each patient. Finally we can use the procedure **PROC TTEST** and the option VAR followed by the variable DIFF to test the null hypothesis that the population mean difference in weight loss is equal to zero. In symbols, the null and alternative hypotheses are

H0: $\mu_d = 0$ versus
Ha: $\mu_d \neq 0$.

The test statistic is calculated using the following formula:

$$t_d = \frac{\bar{x}_d - 0}{s_d / \sqrt{n}}.$$

In the SAS output under t Value, you will find the value of the test statistic which is equal to -4.25 and follows a *t*-distribution with *n*-1 degrees of freedom.

SAS commands:

```
DATA WEIGHTLOSS;
INPUT PATIENT BEFORE AFTER;
DIFF = AFTER - BEFORE;
DATALINES;
 1 157 153
 2 173 161
 3 249 254
 4 154 147
 5 261 245
 6 135 122
 7 245 239
 8 227 220
 9 135 127
10 173 164
;
PROC TTEST;
  VAR DIFF;
RUN;
```

SAS output:

The SAS System

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
DIFF	10	-11.8	-7.7	-3.597	3.9453	5.7359	10.471	1.8138

T-Tests

Variable	DF	t Value	Pr > |t|
DIFF	9	-4.25	0.0022

Note 8.7 – Testing a hypothesis about the difference of two proportions

Here we can refer back to Program Note 8.4 to obtain the SAS commands used to obtain a test statistic as the one displayed in the SAS output under t Value in Note 8.4.