CHAPTER 4
STATISTICAL INFERENCE ON SAS
YI-TING HWANG
DEPARTMENT OF STATISTICS
NATIONAL TAIPEI UNIVERSITY
TYPE OF VARIABLES

• Variables of interest: dependent variable, response variable, outcome variable.
  – Continuous variables
  – Discrete variables
CONTINUOUS VARIABLES

- One sample $T$ Test
- Two independent sample $T$ Test
- Paired $T$ Test
- ANOVA
EXAMPLE 1

Suppose that a sample of 50 golf balls is selected and that the sample mean is 297.6 yards. This sample mean provides support for the conclusions that the population mean is larger than 295 yards. Is this statistical significant?
ONE SAMPLE PROBLEM

Consider a sample $x_1, \ldots, x_n$ drawing from an approximately normal population with true mean $\mu_0$.

Suppose we want to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu \neq \mu_0$$
Z TEST

If $\sigma_x$ is known

$$Z = \frac{\bar{X} - \mu_0}{\sigma_x} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

where $\sigma_x$ is the true population s.d., then the test follows a normal distribution with mean 0 and variance 1.
If $\sigma_x$ is unknown,

$$T = \frac{\bar{X} - \mu_0}{s_x} = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}.$$

where $s$ is the sample standard deviation, the test follows a $t$ distribution with df $n-1$. 
SAS PROCEDURES

PROC TTEST < options > ;
   CLASS variable ;
   PAIRED variables ;
   BY variables ;
   VAR variables ;
   FREQ variable ;
   WEIGHT variable ;
## COMBINATIONS FOR PAIRS

<table>
<thead>
<tr>
<th>PAIRED statements</th>
<th>Yield comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*B;</td>
<td>A-B</td>
</tr>
<tr>
<td>A<em>B C</em>D;</td>
<td>A-B and C-D</td>
</tr>
<tr>
<td>(A B)*(C D);</td>
<td>A-C, A-D, B-C, and B-D</td>
</tr>
<tr>
<td>(A B)*(C B);</td>
<td>A-C, A-B, and B-C</td>
</tr>
</tbody>
</table>
## OPTIONS

<table>
<thead>
<tr>
<th>Options</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha=p$</td>
<td>Specifies that confidence intervals are to be $100(1-p)%$ confidence intervals, where $0&lt;p&lt;1$</td>
</tr>
<tr>
<td>$H_0=m$</td>
<td>Requests tests against $m$ instead of 0 in all three situations (one-sample, two-sample, and paired observation $t$ tests).</td>
</tr>
</tbody>
</table>
EXAMPLE 2

Par, Inc., is a major manufacturer of golf equipment. Management believes that Par’s market share could be increased with the introduction of a cut-resistant, longer-lasting golf ball. The research group at Par has been investigating a new golf ball coating designed to resist cuts and provide a more durable ball. Is the new ball better?
TWO SAMPLE PROBLEM

Consider two sample

\[ x_1, \ldots, x_{n_1} \quad y_1, \ldots, y_{n_2} \]

drawing from two approximately \textbf{normal} populations with mean \( \mu_x \) and \( \mu_y \)

Suppose we want to test

\[ H_0 : \mu_x = \mu_y \quad \text{vs} \quad H_a : \mu_x \neq \mu_y \]
TEST FOR EQUAL VARIANCE

- To perform $T$ test, we need to verify whether the variances in two populations are the same. That is,

$$H_0 : \sigma_x = \sigma_y \text{ vs } H_a : \sigma_x \neq \sigma_y$$

- The test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$
T TEST WITH UNEQUAL VARIANCE

Let

\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

where \( s_1^2 \) and \( s_2^2 \) are the sample standard deviations, the test follows a \( t \) distribution with df \( n_1 + n_2 - 2 \) or

\[
\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2 \bigg/ \left( \frac{s_1^2}{n_1} \right)^2 \left( \frac{n_1}{n_1 - 1} \right) + \left( \frac{s_2^2}{n_2} \right)^2 \left( \frac{n_2}{n_2 - 1} \right)
\]
**T TEST WITH EQUAL VARIANCE**

Let

\[
T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} 
\]

where

\[
s^2_p = \frac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2}{n_1 + n_2 - 2} 
\]

is the pooled sample standard deviation, the test follows a \( t \) distribution with \( df = n_1 + n_2 - 2 \).
EXAMPLE 4

A survey was made of Book-of-the-Month-Club members to ascertain whether members spend more time watching television than they do reading. Assume a sample of 15 respondents provided the following data on weekly hours of television watching and weekly hours of reading. Can you conclude that the members spend more hours per week watching television than reading?
PAIRED SAMPLE

Consider a paired sample

\[(x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})\]

The distribution of the difference \(d_i = X_i - Y_i\) is approximately **normal** distributed with mean difference \(\mu_D\). Let

Suppose we want to test

\[H_0 : \mu_D = D_0 \quad \text{vs} \quad H_a : \mu_D \neq D_0\]
Let \( T = \frac{X_1 - X_2}{s_D / \sqrt{n}} \)

where \( s_D^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} \)

is the sample standard deviation of the differences, the test follows a \( t \) distribution with df \( n-1 \).
EXAMPLE 5

• Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients.

• A cross-sectional study is designed to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects (N = 315) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous.
VARIABLE NAMES

• Variable Names in order from left to right:
  – AGE: Age (years)
  – SEX: Sex (1=Male, 2=Female).
  – SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
  – QUETELET: Quetelet (weight/(height^2))
  – VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
  – CALORIES: Number of calories consumed per day.
  – FAT: Grams of fat consumed per day.
  – FIBER: Grams of fiber consumed per day.
  – ALCOHOL: Number of alcoholic drinks consumed per week.
  – CHOLESTEROL: Cholesterol consumed (mg per day).
  – BETADIET: Dietary beta-carotene consumed (mcg per day).
  – RETDIET: Dietary retinol consumed (mcg per day)
  – BETAPLAS: Plasma beta-carotene (ng/ml)
  – RETPLAS: Plasma Retinol (ng/ml)
DISCRETE VARIABLES

• Test of binomial situation
• Test of multinomial situation
• Test of independence (homogeneous)
TEST OF BINOMIAL – Z TEST

Consider a sample \( x_1, \ldots, x_n \) where \( x_i = 0 \) or 1 with \( P[X_i=1]=p \).

- Suppose we want to test hypothesis
  
  \[ H_0 : p = p_0 \quad \text{vs} \quad H_a : p \neq p_0 \]

- Test statistic I

  \[
  Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1)
  \]
NOTATIONS

• Let $p_1$ and $p_2$ be the proportion of successes and failures in the population.
• Let $O_1$ and $O_2$ be the numbers of successes and failures in the sample.
• Let $E_1$ and $E_2$ be the numbers of successes and failures under $H_0$. 
CHI-SQUARE TEST

\[ H_0 : p_1 = p_0 ; p_2 = 1 - p_0 \]

vs

\[ H_a : \text{at least one not equal} \]

\[ \chi^2 = \sum_{j=1}^{2} \frac{(O_j - E_j)^2}{E_j} \approx \chi^2_{df} \]

where \( df=1 \) and \( E_j = np_j \)
TEST OF MULTINOMIAL

\( H_0 : p_1 = p_{10} ; p_2 = p_{20} ; \cdots ; p_k = p_{k0} \)

vs

\( H_a : \) at least one not equal

Test statistic:

\[ \chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \approx \chi^2_{k-1}, \]

where \( E_j = np_j \).0
TEST OF INDEPENDENCE

Consider two variables of interest, $X$ and $Y$, with $I$ and $J$ categories, respectively.

Objective:

$H_0$ : the classifications are independent
$H_a$ : the classifications are dependent
**CONTINGENCY TABLE**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>.....</th>
<th>I</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td></td>
<td>$n_{1I}$</td>
<td>$n_{1+}$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td></td>
<td>$n_{2I}$</td>
<td>$n_{2+}$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
<td>.....</td>
<td>$n_{+I}$</td>
<td>$n_{++}$</td>
</tr>
</tbody>
</table>
CHI-SQUARE TEST

Test statistic:

\[ \chi^2 = \sum_{ij} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \approx \chi^2_{df} \]

where

\( \hat{E}_{ij} = \frac{(\text{row total for cell } i) \times (\text{column total for cell } j)}{n} \)

\( df = (I - 1) \times (J - 1) \)
SAS PROCEDURES

PROC FREQ < options > ;
   BY variables ;
   EXACT statistic-options < / computation-options > ;
   OUTPUT < OUT=SAS-data-set > options ;
   TABLES requests < / options > ;
   TEST options ;
   WEIGHT variable ;
## GROUPING SYNTAX

<table>
<thead>
<tr>
<th>Request</th>
<th>Equivalent to</th>
</tr>
</thead>
<tbody>
<tr>
<td>tables A*(B C);</td>
<td>tables A<em>B A</em>C;</td>
</tr>
<tr>
<td>tables (A B)*(C D);</td>
<td>tables A<em>C B</em>C A<em>D B</em>D;</td>
</tr>
<tr>
<td>tables (A B C)*D;</td>
<td>tables A<em>D B</em>D C*D;</td>
</tr>
<tr>
<td>tables A -C;</td>
<td>tables A B C;</td>
</tr>
<tr>
<td>tables (A -C)*D;</td>
<td>tables A<em>D B</em>D C*D;</td>
</tr>
</tbody>
</table>
# OPTIONS

<table>
<thead>
<tr>
<th>Options</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPHA</td>
<td>Sets the confidence level for confidence limits</td>
</tr>
<tr>
<td>BINOMIAL</td>
<td>Requests binomial proportion, confidence limits and test for one-way tables</td>
</tr>
<tr>
<td>CHISQ</td>
<td>Requests chi-square tests and measures of association based on chi-square</td>
</tr>
<tr>
<td>CMH</td>
<td>Requests all Cochran-Mantel-Haenszel statistics</td>
</tr>
<tr>
<td>TESTP</td>
<td>Specifies expected proportions for a one-way table chi-square test</td>
</tr>
<tr>
<td>EXPECTED</td>
<td>Display the expected cell frequency for each cell</td>
</tr>
</tbody>
</table>
EXAMPLE 5

This example examines whether the children's hair color has a specified multinomial distribution for the two regions. The hypothesized distribution for hair color is 30% fair, 12% red, 30% medium, 25% dark, and 3% black.
EXAMPLE 6
(CONTS EXAMPLE 5)

The binomial proportion is computed as the proportion of observations for the first level of the variable that you are studying. The following statements compute the proportion of children with brown eyes and test this value against the hypothesis that the proportion is 50%. Also, these statements test whether the proportion of children with fair hair is 28%.
EXAMPLE 7

A survey of 100 customers from stories A, B and C recorded the form of payment for each customer. The results are as follows:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cash</td>
<td>40</td>
<td>35</td>
<td>20</td>
</tr>
<tr>
<td>Credit card</td>
<td>40</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Check</td>
<td>20</td>
<td>35</td>
<td>50</td>
</tr>
</tbody>
</table>