Chapter 4

Some statistical inferences using SAS

We have already used PROC MEANS as a method of obtaining summary statistics for a dataset. In this chapter we shall look at other summaries for data, both numerical and graphical. We shall also look at some simple regression models. We begin with a motivating example.

Normal body temperatures

The dataset `normtemp` lists the body temperature (in degrees Fahrenheit), gender (1 = male, 2 = female) and heart rate (in beats per minute) for a group of 130 individuals. The dataset was presented in Mackowiak, P., Wasserman, S., and Levine, M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich.", Journal of the American Medical Association, 268, 1578-1580. The dataset was obtained from the article "What’s Normal? – Temperature, Gender, and Heart Rate" in the "Journal of Statistics Education" (Shoemaker, 1996).

Using the Windows Explorer create a new folder `normtemp` in the folder My Computer → Drive C: → My SAS files. Now download the file `normtemp.txt` and store it in this folder. Here is the program to read in the dataset and store it in a permanent dataset.

LIBNAME normtemp "C:\My SAS Files\normtemp";

/* Load the 'normtemp' dataset */
DATA normtemp.normtemp;
   INFILE "C:\My SAS Files\normtemp\normtemp.txt";
   INPUT temp gender heart_rate;

PROC PRINT data=normtemp.normtemp NOOBS;
RUN;

There are three pages of output.
4.1 Summarizing the univariate distribution

We have already used the PROC MEANS to summarize variables. A complete univariate summary is given by PROC UNIVARIATE. The command is

```
PROC UNIVARIATE DATA=dataset;
   VAR variable_names;
```


Here is a program to summarize the univariate distribution of the body temperatures

```
OPTIONS NODATE;
PROC UNIVARIATE DATA=normtemp.normtemp;
   VAR temp;
RUN;
```

In the Results double click on Univariate: The SAS system and then double click on temp. There are five sections of output to view: moments, basic measures of location and variability, tests of location, quantiles and extreme observations.

Look through the output for yourself. Try to understand as much as you can. PROC UNIVARIATE can also produce plots. Add the PLOT command after the DATA= statement. PLOT produces a stem and leaf plot, box plot and normal probability plot. Try this program.

```
PROC UNIVARIATE DATA=normtemp.normtemp PLOT;
   VAR temp;
RUN;
```

Submit the program. Note how the graphs are character plots (not graphics). To produce graphical plots you need to use the HISTOGRAM and/or QQPLOT commands. We start by producing a histogram of the data:

```
PROC UNIVARIATE DATA=normtemp.normtemp;
   VAR temp;
   HISTOGRAM;
RUN;
```
Note the default histogram has percent for a y axis. To change to counts:

```sas
PROC UNIVARIATE DATA=normtemp.normtemp;
   VAR temp;
   HISTOGRAM / VSCALE=count;
RUN;
```

We could also use VSCALE=proportion to display the y axis as a proportion. Now we specify the midpoints of the histogram bins, label the y axis, and give the plot a title.

```sas
PROC UNIVARIATE DATA=normtemp.normtemp NOPRINT;
   VAR temp;
   HISTOGRAM / VSCALE=count MIDPOINTS=95 to 101 by 1
            VAXISLABEL="the counts";
   TITLE "Histogram of the body temperature data";
RUN;
```

We use the NOPRINT option to only produce the plot (and not the summary statistics).

To label the x axis you need to create a LABEL for the variable temp. You can think of a LABEL as being a title for a variable. The easiest way to do this is to add a line to our original program to read in the data, and then draw the histogram.

```sas
LIBNAME normtemp "C:\My SAS Files\normtemp";

/* Load the 'normtemp' dataset */
DATA normtemp.normtemp;
   INFILE "C:\My SAS Files\normtemp\normtemp.txt";
   INPUT temp gender heart_rate;
   LABEL temp="normal body temperatures (degrees F)"
PROC UNIVARIATE DATA=normtemp.normtemp NOPRINT;
   VAR temp;
   HISTOGRAM / VSCALE=count MIDPOINTS=95 to 101 by 1
            VAXISLABEL="the counts";
   TITLE "Histogram of the body temperature data";
RUN;
```
Finally here is the graphical Q-Q plot of the temperature data

```sas
PROC UNIVARIATE DATA=normtemp.normtemp NOPRINT;
   VAR temp;
   QQPLOT;
   TITLE "Normal Q-Q plot of the body temperature data";
RUN;
```

To add a line which can aid in identifying normality we use the following program.

```sas
PROC UNIVARIATE DATA=normtemp.normtemp PLOT NOPRINT;
   VAR temp;
   QQPLOT / NORMAL (MU=EST SIGMA=EST);
RUN;
```

Here `MU=EST` means 'estimate the mu parameter', and `SIGMA=EST` means 'estimate the sigma parameter.

### 4.2 Multivariable summaries

We can use the `BY` command in the `UNIVARIATE` procedure to produce a summary by another variable in the dataset. As always we need to `SORT` the dataset in increasing order of the `BY` variables. Here is a program to summarize the body temperature by gender.

```sas
PROC SORT DATA=normtemp.normtemp OUT=normtemp_by_gender;
   BY gender;
PROC UNIVARIATE DATA=normtemp_by_gender PLOT;
   VAR temp;
   BY gender;
RUN;
```

Note how the output uses the label for the `temp` variable. Can you see a difference in the body temperature by gender? Are there any outliers in the observations (denoted by O in the characters plots).
4.3 Producing a boxplot

We can produce a graphical boxplot of the body temperature data using the following program.

```
DATA temp_normtemp;
   SET normtemp.normtemp;
   ones = 1;

PROC BOXPLOT DATA=temp_normtemp;
   PLOT temp * ones / BOXSTYLE=schematic CFRAME=vligb CBOXES=dagr
       CBOXFILL=ywh NOHLABEL;
   TITLE "Boxplot of body temperature";
RUN;
```

Note how we create a dummy variable called 'ones' which has value one for every observation in the dataset. This is because the `BOXPLOT` procedure always expects an x variable.


**Exercise**: Produce a side by side boxplot of the body temperature by gender.

4.4 Performing t-tests

As well as getting the t statistic from `MEANS` procedure you can use the `TTEST` procedure. Here is the syntax for the command:

```
PROC TTEST DATA=<dataset> <options>;
   VAR variables;
   CLASS variables;
   BY variables;
   PAIRED variables;
```

Some options include `ALPHA=` which sets the size of the test, and `HO=` which states which value to test (instead of the default value of zero) in the one sample t-test.

Let us test whether there evidence in the data that the population mean body temperatures is equal to 98.6°F. Here is the program.
PROC TTEST DATA=normtemp.normtemp ALPHA=.05 H0=98.6;
  VAR temp;
RUN;

Here is the output

The TTEST Procedure
Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>Lower CL</th>
<th>Upper CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>98.122</td>
<td>98.249</td>
<td>98.376</td>
<td>0.6536</td>
</tr>
<tr>
<td></td>
<td>0.7332</td>
<td>0.835</td>
<td>0.0643</td>
<td></td>
</tr>
</tbody>
</table>

T-Tests

| Variable | DF | t Value | Pr > |t| |
|----------|----|---------|------|---|
| temp     | 129 | -5.45   | <.0001 |

We reject the null hypothesis. There is strong evidence that (with P-value < .0001) the mean body temperature is not equal to 98.6 Fahrenheit. Let us now test whether the mean body temperature is same for males and females.

We use the CLASS command in the TTEST procedure to do a two sample t-test.

PROC TTEST DATA=normtemp.normtemp ALPHA=.05;
  VAR temp;
  CLASS gender;
RUN;

We do not need to sort to use the CLASS variable. Here is the output

The TTEST Procedure
Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>Lower CL</th>
<th>Upper CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>97.931</td>
<td>98.105</td>
<td>98.278</td>
<td>0.5959</td>
</tr>
<tr>
<td></td>
<td>0.6988</td>
<td>0.8449</td>
<td>0.0867</td>
<td></td>
</tr>
<tr>
<td>temp</td>
<td>98.21</td>
<td>98.394</td>
<td>98.578</td>
<td>0.634</td>
</tr>
<tr>
<td></td>
<td>0.7435</td>
<td>0.899</td>
<td>0.0922</td>
<td></td>
</tr>
<tr>
<td>temp</td>
<td>-0.54</td>
<td>-0.289</td>
<td>-0.039</td>
<td>0.6429</td>
</tr>
<tr>
<td></td>
<td>0.7215</td>
<td>0.8221</td>
<td>0.1266</td>
<td></td>
</tr>
</tbody>
</table>
According to both methods of performing the t-test, Pooled (assumes equal variance), and Satterthwaite (assumes unequal variance), there is evidence that the mean body temperature is different for the two genders. The body temperature is lower for the males (gender=1). The TTEST also carries out a test that the variances are equal. We conclude there is no evidence that the population variances are different for the two genders.


### 4.5 Scatterplots

We use the GPLOT procedure to produce scatterplots (documentation on the command is at [http://support.sas.com/onlinedoc/913/getDoc/en/graphref.hlp/gplotchap.htm](http://support.sas.com/onlinedoc/913/getDoc/en/graphref.hlp/gplotchap.htm)).

```sas
PROC GPLOT DATA=normtemp.normtemp;
   PLOT temp * heart_rate;
   TITLE "Plot of heart rate versus body temperature";
RUN;
QUIT;
```

Note the addition of the QUIT command. This tells SAS to stop working with the GPLOT procedure. This is because you can keep using GPLOT to add features to your plot until you QUIT.

### 4.6 Calculating the correlation between two variables

We calculate the sample correlations between variable using the CORR procedure. Here is a program to calculate the correlations between the body temperature and the heart rate.

```sas
PROC CORR DATA=normtemp.normtemp;
   VAR temp heart_rate;
RUN;
```
Here is the output from the program

The CORR Procedure

2 Variables:  temp  heart_rate

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>130</td>
<td>98.24923</td>
<td>0.73318</td>
<td>12772</td>
<td>96.30000</td>
<td>100.80000</td>
</tr>
<tr>
<td>heart_rate</td>
<td>130</td>
<td>73.76154</td>
<td>7.06208</td>
<td>9589</td>
<td>57.00000</td>
<td>89.00000</td>
</tr>
</tbody>
</table>

Pearson Correlation Coefficients, N = 130
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>temp</th>
<th>heart_rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>1.00000</td>
<td>0.25366</td>
</tr>
<tr>
<td>heart_rate</td>
<td>0.25366</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

From the output we conclude that correlation between body temperature and heart rate is 0.254. According to the P-value this correlation is significantly larger than zero.

We can calculate the correlations, broken down by gender, using the following program:

PROC SORT DATA=normtemp.normtemp OUT=normtemp_by_gender;
   BY gender;
PROC CORR DATA=normtemp_by_gender;
   VAR temp heart_rate;
   BY gender;
RUN;

After we submit the program, we find that the correlation for males is 0.196 (which is not significantly different from zero, with P-value 0.118) and for females is 0.287 (which is significantly different from zero with P-value 0.0205).

Exercise: Try to produce different scatterplots of the heart rate versus body temperature for the two genders.
4.7 Fitting regression models using SAS

We use the REG procedure to fit regression models to data. Here is a program to fit the model

\[
\text{body temperature}_i = \alpha + \beta \text{ heart rate}_i + \epsilon_i,
\]

for \( i = 1, \ldots, n \), where \( \{\epsilon_i\} \) are a set of IID \( N(0, \sigma^2) \) random variables.

PROC REG DATA=normtemp.normtemp;
  MODEL temp = heart_rate;
  PLOT temp * heart_rate;
  TITLE "Results of regression analysis";
RUN;
QUIT;

Documentation for the REG procedure is at

Note the use of the QUIT command again. Here are the results in the output window.

Results of regression analysis

The REG Procedure
Model: MODEL1
Dependent Variable: temp

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>4.46176</td>
<td>4.46176</td>
<td>8.80</td>
<td>0.0036</td>
</tr>
<tr>
<td>Error</td>
<td>128</td>
<td>64.88316</td>
<td>0.50690</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>129</td>
<td>69.34492</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.71197
R-Square 0.0643
Dependent Mean 98.24923
Adj R-Sq 0.0570
Coeff Var 0.72466

Parameter Estimates

| Variable       | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------------|----|--------------------|----------------|---------|-------|
| Intercept      | 1  | 96.30675           | 0.65770        | 146.43  | < .0001 |
| heart_rate     | 1  | 0.02633            | 0.00888        | 2.97    | 0.0036 |
What is the estimated model? Are the parameters in the model significant? How well does the model fit?

We add the / R command after the MODEL command to provide information on the residuals.

```
PROC REG DATA=normtemp.normtemp;
   MODEL temp = heart_rate / R;
   PLOT temp * heart_rate;
   TITLE "Results of regression analysis";
RUN;
```

There is no QUIT here. Now we can write a new program which adds output to that created previously. To plot the residuals versus the predicted values we use

```
PLOT r.*p. / CFRAME=ligr;
RUN;
QUIT;
```

Now we are finished with the REG procedure and we QUIT.

**Exercise:** How would write a program to check the model assumptions in SAS?

To fit a different model for each gender use the following program.

```
PROC REG DATA=normtemp_by_gender;
   MODEL temp = heart_rate / R;
   PLOT temp * heart_rate;
   BY gender;
   TITLE "Results of regression analysis";
RUN;
QUIT;
```