



Academic

## Simple Hypothesis tests in PROC UNIVARIATE

by Katrien Declercq

Assume we have a set of observations on the weight of cereal boxes (in pounds) from a sample of a particular brand and we would like to know whether this brand does a correct filling of its boxes or not (they claim to have 15 pounds in the boxes). This can be translated in testing the null-hypothesis that the boxes contain on the mean 15 pounds ( $H_0: \text{mean} = 0$ ) against the alternative that the mean weight of the boxes is different from 15 pounds ( $H_1: \text{mean} \neq 0$ ). Based on the data we have from the sample, we would like to know whether we could reject the null-hypothesis or not.

The following code provides the default output from PROC UNIVARIATE on these data.

```
TITLE 'Hypotheses tests with PROC UNIVARIATE';
DATA cereal;
  INPUT weight @@;
  CARDS;
  15.02 15.30 14.95 14.65 14.86 15.00 14.78 14.35 14.89 15.19
  14.97 15.10 14.75 14.62 14.89 14.79 15.10 14.68 14.95 14.88
  ;
RUN;

PROC UNIVARIATE DATA=cereal;
  VAR weight;
RUN;
```

After running this code in V6, the line of the output reading 'T: Mean=0 306.3074 Pr>|T| 0.0001' (in the section on 'Moments'), specifies the null-hypothesis it tests ( $H_0: \text{mean}=0$ ). This is not exactly what we need, since we want to test whether the mean is 15 pounds. In order to have SAS do what we want (in V6), we first need to subtract 15 from all observations and then we can use the default test of PROC UNIVARIATE on these differences.

This means we only need to add the following data step to the code and then run PROC UNIVARIATE on the variable DIFF.

```
DATA cereal;
  SET cereal;
  diff=weight-15;
RUN;

PROC UNIVARIATE DATA=cereal;
  VAR diff;
RUN;
```

Then, the line in the output with 'T: Mean=0 -2.34576 Pr>|T| 0.0300' shows the information we need to be able to test the null-hypothesis we specified. The crucial information is the P-value, which is the number after 'Pr>|T|' and indicates how plausible it is to accept the null-hypothesis with the observations of our sample. Since this value is only 0.0300, the probability that we see the observed data under the null-hypothesis that the true mean weight is 15 pounds is quite low, so we have evidence to reject the null-hypothesis and conclude that the mean weight of the cereal boxes is different from 15 pounds. In practice, the P-value is often compared to 0.05 (i.e. 5% significance level) to be able to decide whether one rejects the null-hypothesis or not: if the p-value is smaller, the null-hypothesis is rejected in favor of the alternative, otherwise, there is not enough evidence to reject H0.

In the discussion above, we focused on the results from the Student's T-test to explain the concept. SAS also provides two other tests (Sign test and Signed rank test which are non-parametric tests for the same hypothesis). The use of the P-values is similar as discussed above. Since many statistical methods assume a normal underlying distribution, it may be useful to check whether this assumption is fulfilled before proceeding with the analysis. Using the NORMAL option in the PROC UNIVARIATE statement tells SAS to do a formal hypothesis test of the null-hypothesis of normality versus the alternative of no normality and puts a line with the corresponding information in the last line of the 'Moments' section.

In V8, the output of PROC UNIVARIATE is organized in a somewhat different order. The tests for H0: mean=0 are found under the header 'Tests for Location: Mu0=0' and the NORMAL option provides more than one test for normality which are put under the header 'Tests for Normality'. A convenient new option in V8 is MU0=value(s), which allows to specify the value of the mean used in H0 and so avoids the need of the extra dataset to subtract the null-value from all observations and allows to immediately test the correct null-hypothesis (then the header of the section specifies the value that is chosen instead of 0).