



Comparing Two Groups with PROC TTEST

by Valérie Boquia

Academic

In the previous newsletter, we discussed how to perform simple hypothesis test using PROC UNIVARIATE; we used it to test whether boxes of cereal contain on the mean 15 pounds. Now suppose a consumer advocacy group wants to determine whether two popular cereal brands, Kellogg's and Nestle, have the same amount of cereal. Both brands advertise that they have 15 ounces of cereal per box. A random sample of both brands is selected and the number of ounces of cereal is recorded. This question can be translated in testing the null hypothesis that the two mean weights are the same whatever the brand of cereal is (H_0 : $\text{mean}_{\text{KELLOGS}} = \text{mean}_{\text{NESTLE}}$) against the alternative that they are different (H_a : $\text{mean}_{\text{KELLOGS}} \neq \text{mean}_{\text{NESTLE}}$).

There are three assumptions for this test to be valid for comparing two group means: the observations are independent, observations for each group are a random sample from a population with a normal distribution, variances for the two independent groups are equal.

The following code provides the default output from PROC TTEST on the data:

```
data cereal;
  input brand $ weight 7.4 @@;
cards;

Kellogs 14.9982 Nestle 15.0136 Kellogs 15.0100 Nestle 14.9982 Kellogs 15.0052
Nestle 14.9930 Kellogs 14.9733 Nestle 15.0812
Kellogs 15.0037 Nestle 15.0418 Kellogs 14.9957 Nestle 15.0639 Kellogs 15.0099
Nestle 15.0613 Kellogs 14.9943 Nestle 15.0255
Kellogs 14.9779 Nestle 15.0176 Kellogs 14.9862 Nestle 15.0122 Kellogs 14.9907
Nestle 15.0122 Kellogs 14.9785 Nestle 15.0322
Kellogs 15.0716 Nestle 15.0164 Kellogs 14.9787 Nestle 15.0093 Kellogs 14.9935
Nestle 15.0156 Kellogs 15.0270 Nestle 15.0393
Kellogs 14.9855 Nestle 15.0298 Kellogs 14.9982 Nestle 15.0204 Kellogs 15.0194
Nestle 15.0633 Kellogs 14.9720 Nestle 15.0464
Kellogs 14.9793 Nestle 15.0858 Kellogs 15.0304 Nestle 15.0418 Kellogs 15.0187
Nestle 15.0101 Kellogs 15.0134 Nestle 15.0580
Kellogs 14.9930 Nestle 15.0550 Kellogs 14.9690 Nestle 15.0500 Kellogs 14.9955
Nestle 15.0868 Kellogs 15.0032 Nestle 15.0196
Kellogs 14.9737 Nestle 15.0413 Kellogs 15.0254 Nestle 15.0267 Kellogs 14.9885
Nestle 15.0374 Kellogs 15.0223 Nestle 15.0437
Kellogs 15.0057 Nestle 15.0194 Kellogs 15.0039 Nestle 15.0623 Kellogs 14.9515
Nestle 15.0980 Kellogs 14.9894 Nestle 15.0234
Kellogs 15.0169 Nestle 14.9831 Kellogs 14.9803 Nestle 15.0435 Kellogs 14.9730
Nestle 15.0497 Kellogs 14.9779 Nestle 15.0096
;
run;

proc ttest data=cereal;
  class brand;
  var weight;
  title 'Testing the Equality of Means for Two Cereal Brands';
run;
```

The PROC TTEST statement requests a two-sample t-test. The CLASS statement names the variable that classifies the data set into two groups (in this case the cereal manufactured by Kellogg's or Nestle). The VAR statement names the measurement variable to be analyzed (the weight of cereal boxes). Before interpreting the output, remember the validity of this test supposes some assumptions that we first have to check: the assumption of independent observations is met because each cereal box's weight is unrelated to every other cereal box's weight. For the assumption of normality, you can test the normality for each group and find that this assumption is acceptable (this will be the subject of a next newsletter). The assumption of equal variances is given in the last line of the output labeled "**For H0: Variances are equal**". The number after **Prob>F'** gives the P-value for the test of equal variances. For the cereal data, this P-value is 0.2460, as it is quite high compared to 0.05 (a 5% confidence level) we have not enough evidence to reject the null-hypothesis the weight variances are equal. It is thus acceptable to assume equal variances between the 2 groups.

Based on whether one rejects the null-hypothesis of equality of variances or not, we can use respectively the two-sample t-test for unequal variances (in the column labeled "**Vari ances**", look at the appropriate row labeled "**Unequal**") or the two-sample t-test for equal variances (in the column labeled "**Vari ances**", it corresponds to the row labeled "**Equal**"). The column labeled "**T**" gives the value of the t-test statistic, the column labeled "**DF**" gives the degrees of freedom, and the column labeled "**Prob> |T|**" gives the P-value that has to be interpreted.

Since we can assume equal variances for the cereal data, the P-value for the exact two-sample t-test in the row labeled "**Equal**" is 0.0000, which is smaller than 0.05. The null hypothesis that the group means are equal is thus rejected at the 5% significance level. You conclude that the mean weight of the cereal boxes from Kellogg's is significantly different from the mean weight of cereal boxes from Nestle.