Creativity.sas: Explanation of code

Goals of code:

- Display the creativity data (histogram, boxplot, dotplot)
  - As one group or for each treatment
- Calculate summary statistics for each group
- Randomization test
- T-test with equal variances

Reading the data set: data creativity;
The data creativity; statement tells SAS that you want to create a SAS data set called creativity. All
the statements from data ; to the run; are part of the data step. The infile command names the file
in the working directory that contains the data. This file has a header line with variable names, so
the real data starts on the second line. The firstobs=2 option tells SAS to start reading data from
the second line. The input line tells SAS how many values to read on each line, what names to give
the variables, and how to read the values. This is described in the Introduction to SAS document.
Treatment is a character string, so it is followed by a ; scoreisnumericsoitisnotfollowedbya.

I generally check the line in the log that says “The data set WORK.CREATIVITY has 47 observa-
tions and 2 variables”. I check that SAS read the number of observations I expected and produced
the number of variables I expected.

Descriptive statistics and histograms: proc univariate;
PROC UNIVARIATE calculates and reports many different statistics. The VAR statement names
the variable to be summarized. You can name multiple variables if you want more than one sum-
mary. The optional HISTOGRAM statement requests a histogram in addition to the numeric
summaries. The optional title statement provides a title (in quotes).

SAS reports a lot of numbers, including many that we won’t talk about. You will probably be
interested in:
In the box of results labeled Moments:
  N: number of observations
  Mean: sample average
  Std Deviation: sample standard deviation
In the box of results labeled Basic Statistical Measures:
  Mean: a repeat of the sample average
  Median: sample median
  Std Deviation: a repeat of the sample standard deviation
  Interquartile Range: the IQR
  Coeff Variation: the coefficient of variation
You may be interested in:
In the Quantiles box: 100% Max, 75% Q3, 25% Q1 and 0% Min:
the max, 75'th percentile, 25'th percentile, and minimum.

Some of the other numbers will be used later. The rest have specialized uses.

**Descriptive statistics and histograms for groups:** `proc univariate; class treatment;`
Adding a class statement to the `proc univariate` code tells SAS to do everything twice, once for each unique value of the treatment. Same format for the numeric output, but the top of the page is labeled `treatment =` to indicate which treatment is being described. The histogram statement with a class statement produces nice side by side histograms.

Note: Remember that the grouping variable goes in the class statement and the result variable (to be described) goes in the var statement. If you reverse these, e.g. class score; var treatment;, you get mush or worse.

**Boxplots:** `proc boxplot;`
Proc boxplot produces side-by-side boxplots. One quirk is that it requires all the intrinsic observations to be together and all the extrinsic observations to be together. The easiest way to ensure this is to sort the data by treatment first. The `proc sort; by treatment; run;` does this. The `plot statement` specifies the response variable and the grouping variable. The response is first, then the group, separated by a `*`. Again, if you reverse these, the result is either mush or a lot of errors.

Note: If you expected two boxes (for two groups) and got many, the data set is probably not sorted.

**Permutation test:** `proc npar1way scores=data; exact /maxtime=10;`
If the problem is very large, the p-value can be estimated from a sample of permutations of group
labels to observations. It isn’t necessary to enumerate them all. The /mc option requests a Monte-Carlo analysis, i.e. a random sample of possible permutations. This is usually called a randomization test. The n=10000 requests 10000 samples.

The most difficult part of this analysis is finding the number you really want. That is the two-sided p-value from the Exact test (or Monte Carlo approximation). That number is the Two-sided ..., Estimate number in the box labeled Monte Carlo Estimates for the Exact Test.

**Two sample t-test:** proc ttest;
Proc ttest requests a two-sample t-test. Again, the class statement names the groups and the var statement names the response.

We will talk about many of the numbers reported here over the next week. For now, the p-value that is usually reported in scientific papers is found in the third box of results: Look for the Pooled Method, Equal variances. The p-value is found in the Pr >| t | column. Here it is 0.0060. That is a two-sided p-value.