Categorical Variables

Variables which record a response as a set of categories are termed categorical. Such variables fall into three classifications: Nominal, Ordinal, and Interval. Nominal variables have categories that have no natural order to them. Examples could be different crops: wheat, barley, and peas or different irrigation methods: flood, furrow, and dry land. Ordinal variables, on the other hand, do have a natural order. Examples of these could be pesticide levels: high, medium, and low or an injury scale: 0, 1, 2, 3, 4, and 5. Caution should be used with some tests designed for ordinal variables because they may assume equal 'distance' between the levels. Such distances may be hard to actually quantify. The last type is the interval variable and it is, as the name implies, created from intervals on a continuos scale. This could be categories based on weights, e.g.: 0-2 gm, 2-10 gm, and >10 gm.

Categorical Responses

The actual response in any category type is binary (i.e. it records one of two possible conditions or outcomes). In the examples above, this could be the presence or absence of a weed or insect species for the different crops, or a YES/NO answer to irrigation method, etc. For SAS, it does not matter whether the variables are numeric or alphabetic. The yes/no type variable could be entered as "YES":"NO" or 1:2 or "Y":"N", etc. However, it is important to maintain the same case with alphabetic variables. The values "N" and "n" are not the same in SAS! A sample data file could look like the following:

OBS	VAR1	VAR2	VAR3 ...
1	YES	1	Y
2	YES	2	N
3	NO	1	Y
.	.	.	.
.	.	.	.
.	.	.	.

Although SAS categorical procedures will automatically tabulate the response categories before analysis, this data type can be easily condensed to a more compact form. That is, the variables could record the number of "YES" responses, "NO" responses, etc. The summary may also record the number for combinations of categorical conditions such as the number of observations with "YES", 1, and "N". This form will be refered to as count data. The following data file example illustrates this:

VAR1	VAR2	VAR3	COUNT
YES	1	Y	34
NO	1	Y	23
YES	2	Y	12
NO	2	Y	16
.	.	.	.
.	.	.	.
.	.	.	.

It is possible to get SAS to do this summarization for you. Given a binary data form, several SAS procedures can create a summarized SAS data set (counts). As an illustration, the following SAS code summarizes a binary data file using the MEANS Procedure. Other Procedures that could be used are SUMMARY, FREQ and UNIVARIATE.

Example.

```
DATA BINARY;
     INPUT SUBJECT RESPONSE $ @@;
     CARDS;
     1 Y  2 N  3 N  4 Y  5 N  6 N  7 Y  8 Y
     9 Y  10 Y 11 Y 12 N 13 N 14 Y 15 Y 16 N;

PROC SORT DATA=BINARY;
     BY RESPONSE;

PROC MEANS NOPRINT DATA=BINARY;
     VAR SUBJECT;
     OUTPUT OUT=COUNTS N=COUNT;
          BY RESPONSE;

PROC PRINT DATA=COUNTS;
     VAR RESPONSE COUNT;
```

Output.

OBS	RESPONSE	COUNT
1	N	7
2	Y	9

Categorical Structures

A basic structure for categorical data is the **one-way frequency**, i.e.: only one variable is examined at a time. Again using an example from above, a one-way frequency could be counts from the different crops. If two categorical variables have been recorded, the cross classification is called a **two-way contingency table**. This could be a 3 x 3 table of crop types by pesticide levels. More complex structures involving three or more categorical variables are refered to as **multi-dimensional distributions**. A cross classification of crop type by pesticide level by irrigation method would be an example.