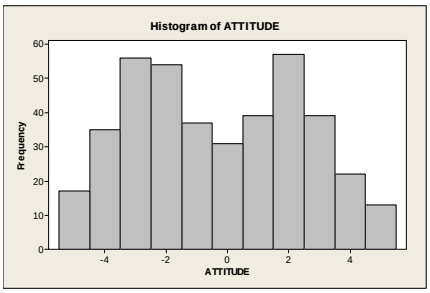
Union Card - Solution

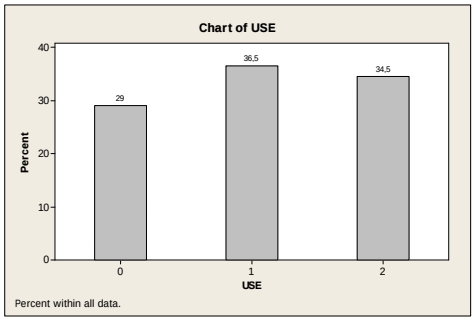
We want to explain ATTITUDE and USE by the categorical variables GENDER, GROUP and STATUS and the numerical variable AGE (categorical after grouping).
We may want to present the possible dependencies by suitable graphics or by crosstabulation, or some more refined statistical methods, and also perhaps perform some testing for independence.

ATTITUDE is an ordinal categorical variable which may be analyzed as such, or by categorizing in broader categories, say negative, indifferent and positive. It is also tempting to consider the variable as numerical of interval type, as this gives the opportunity of using more refined methods like t-tests, analysis of variance and standard regression. However, such methods assume normality (at least for significance testing), which may be questionable, as seen in the following Histogram:



The histogram shows a bimodal distribution, indicating a tendency to polarization within the membership into negative and positive members and less frequent indifference, compared to the unimodal normal distribution.

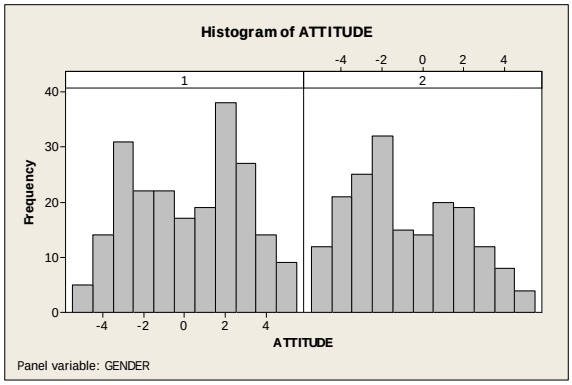
A first insight to the variable USE may be provided by a Bar Chart of percentages.



The bar chart shows that 29% of the respondents state they will not use the card (0), 36.5% undecided (1) and 34.5% state they would use it (2). The fraction of the middle category who are likely to go to one side or the other when the card is offered, may be debated.

Now let us look at ATTITUDE explained by the other variables, first GENDER.

Histograms for ATTITUDE among each gender separately indicate a similar bimodal pattern as the overall pattern, maybe not so pronounced for males as for females. If there is more to be seen, females are a bit more likely to be positive than the males.



Alternatively we may crosstabulate ATTITUDE and GENDER, and at the same time perform a chisquare test of independence between ATTITUDE and GENDER.

Tabulated statistics: GENDER; ATTITUDE

```
Rows: GENDER   Columns: ATTITUDE

       -5  -4  -3  -2  -1   0   1   2   3   4   5  All

1       5  14  31  22  22  17  19  38  27  14   9  218
2      12  21  25  32  15  14  20  19  12   8   4  182
All    17  35  56  54  37  31  39  57  39  22  13  400

Cell Contents:      Count

Pearson Chi-Square = 21.010; DF = 10; P-Value = 0.021
Likelihood Ratio Chi-Square = 21.298; DF = 10; P-Value = 0.019
```

The hypothesis of independence is rejected at the 5% level (P-value less than 5%). However, the test itself does not answer what kind of difference of attitude there is between females and males. Most likely the difference is as mentioned above.

Another possibility is to use the t-test for the hypothesis of equal expected level on the attitude scale of the two genders:

Two Sample T-Test and Confidence Interval

```
Two sample T for ATTITUDE

GENDER         N      Mean     StDev   SE Mean
1            218      0.17      2.71      0.18
2            182     -0.84      2.69      0.20

95% CI for mu (1) - mu (2): ( 0.47;  1.54)
T-Test mu (1) = mu (2) (vs not =): T = 3.71  P = 0.0002  DF = 398
Both use Pooled StDev = 2.70
```

We see that the females and the males have a mean slightly above and below zero respectively. We reject the hypothesis of equal expectations, and conclude that the expected value on the attitude scale is larger for females than males. We do this safely,

since the probability of obtaining a result at least as disparate when the expectations are equal (reported by the P-value) is very small, and the violation of the normal distributional assumption is not likely to ruin this. An alternative not depending on the normal assumption is to test equal medians by the nonparametric Mann-Whitney test

Mann-Whitney Test and CI: ATTITUDE_1; ATTITUDE_2

```
              N   Median
ATTITUDE_1  218   0.0000
ATTITUDE_2  182  -1.0000

Point estimate for ETA1-ETA2 is 1.0000
95,0 Percent CI for ETA1-ETA2 is (0.0002;2.0000)
W = 47881.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0003
The test is significant at 0.0003 (adjusted for ties)
```

We see that the sample medians for the females and the males are 0 and -1 respectively, and that the P-value is small. This means that the hypothesis of equal medians is rejected even at very small a priori chosen significance level (the risk of false rejection of equality).

Similar graphs and analyses may be performed for ATTITUDE versus STATUS. Histograms show bimodal distribution and there is significant difference between the single status (0) and married/partnership (1) indicating that the singles are more in favour than the married/partnership ones. This is confirmed by a t-test for testing equality of expected scores (P=0.001).

The categorical variable GROUP has 4 categories and differences in ATTITUDE may be illustrated by separate or stacked histogram, and is omitted here. A comparison of the expectations of the groups on the ATTITUDE scale may be performed by a one-way analysis of variance (ANOVA) as follows:

One-way Analysis of Variance

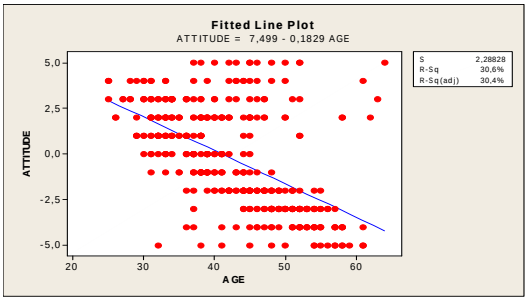
```
Analysis of Variance for ATTITUDE
Source     DF        SS        MS        F        P
GROUP       3     71.28     23.76     3.21    0.023
Error     396   2932.66      7.41
Total     399   3003.94
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N      Mean     StDev  -+---------+---------+---------+-----
1         112    -0.723     2.735   (-------*-------)
2          96    -0.333     2.586        (--------*---------)
3          94    -0.447     2.726      (---------*--------)
4          98     0.408     2.828                    (--------*--------)
                                  -+---------+---------+---------+-----
Pooled StDev =    2.721          -1.20     -0.60      0.00      0.60
```

We see that the hypothesis of equal expectations is rejected at 5% level (P-value less than 5%). Moreover group D (one of the newcomer groups) stands out as more favourable to the benefit card than the others, although the individual confidence interval still have some overlap. This may be investigated further by using the multiple comparison option (if available). With four groups we may make six pairwise comparisons, and with 95% individual confidence (80 % joint confidence) we can conclude that group D has expectation greater than all groups A, B and C. However, if we require 95% joint

confidence (which requires 99% individual) we can just conclude that group D has expectation greater than group A.

An alternative to the above testing we could have used the non-parametric Kruskal-Wallis test, not dependent on normal assumptions. However, here software do not provide multiple comparisons.

To investigate how ATTITUDE depends on AGE it is tempting to use regression analysis. However, there is no good reason that this dependence is linear. The scatterplot below is rather odd when the left hand side variable takes discrete values only.



However, we see as general tendency that the extent of favouring the introduction of the card is decreasing with age, but contrary to this, a few older members are strongly in favour. Formally the explanatory power of the regression is about 30%, and the regression coefficient is strongly significant (T=12.53, P=0.000). Here the standard assumption of constant variance is violated, besides the lack of normality (among others the residuals are skewed to the right).

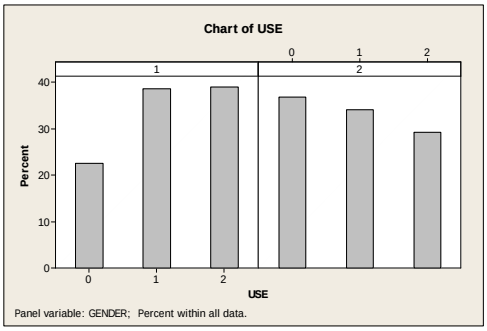
Above we have studied each of the variables separately. We may study the dependence of the categorical variables by cross-tabulate more than two variables, studying ATTITUDE for combinations of the other. More sophisticated log-linear modelling exists also. Here we limit ourselves to a multiple regression using AGE, GENDER, STATUS and GROUP as explanatory variables. Since GROUP is categorical with more than two categories, we have to introduce indicators, here taking Group A as basis. Each of the three group variables GROUPB, GROUPC, GROUPD is then 1 if the respondent belongs to the respective group and 0 otherwise. The regression output is:

Regression Analysis

```
The regression equation is
ATTITUDE = 7.41 - 0.171 AGE - 0.695 GENDER - 0.642 STATUS + 0.282 GROUPB
           + 0.166 GROUPC + 0.839 GROUPD

Predictor        Coef       StDev          T        P
Constant       7.4052      0.6315      11.73    0.000
AGE          -0.17100     0.01369     -12.49    0.000
GENDER         -0.6951      0.2267      -3.07    0.002
STATUS         -0.6425      0.2342      -2.74    0.006
GROUPB          0.2824      0.3117       0.91    0.366
GROUPC          0.1662      0.3133       0.53    0.596
GROUPD          0.8392      0.3115       2.69    0.007

S = 2.235       R-Sq = 34.7%     R-Sq(adj) = 33.7%
```

We see that all variables are statistically significant, but that we could merge Group B and Group C with Group A, leaving Group D apart from others more in favour of the card. Note that the regression coefficients now measure the marginal change in the attitude score for each variable, given that the others are kept constants. Example: Whatever gender the singles are more in favour than the married/partnership ones.

We now want to see how the variable USE relates to AGE, GENDER, STATUS and GROUP. Here follows Bar Charts for USE among each GENDER. For both gender we see a large percentage undecided (1=middle column in both panels). Among the females (left panel) there are substantial more stated users (2) than non-users (0), while the statement goes the other way among the males (right panel).



Alternatively we may tabulate GENDER vs. USE, exhibiting row percentages and a chi-square statistic to see if the difference is statistically significant.

Tabulated statistics: GENDER; USE

Rows: GENDER Columns: USE

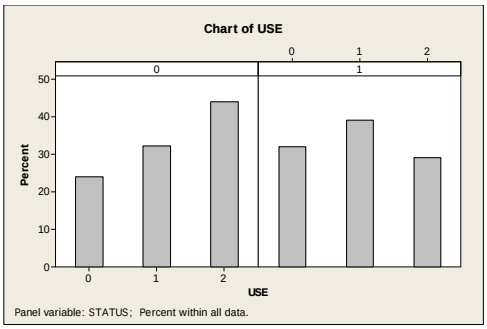
	0	1	2	All
1	22.48	38.53	38.99	100.00
2	36.81	34.07	29.12	100.00
All	29.00	36.50	34.50	100.00

Cell Contents: % of Row

Pearson Chi-Square = 10.372; DF = 2; P-Value = 0.006
Likelihood Ratio Chi-Square = 10.376; DF = 2; P-Value = 0.006

We see that the hypothesis of independence between USE and GENDER is rejected at 1% significance level (P=0.006) supporting the above comments on gender differences. However, we may wonder where the undecided ones will settle when the card is offered, if so. If we conservatively merge the undecided (2) with the no use group (0), it turns out that the hypothesis of independence is not rejected (P=0.039). The same happens if the not decided group is split in two halves, one joining the no-use group and the other joining the use group (P=0.016), both due to the fact that females (state that they) are more likely to use the card than the males.

Next follows a Bar Chart for USE among each categories of STATUS:



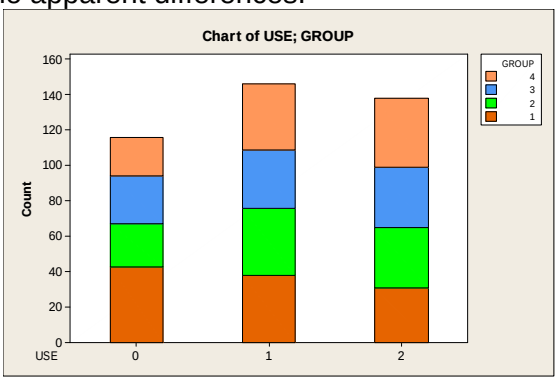
We see that there are relatively more stated users (2) and less undecided (1) among the singles (0=left panel) than among the married/partnership group (1=right panel). Consequently the percentage of stated users becomes lower in married/partnership group than the singles group.

Here follows the tabulation of STATUS vs. USE, exhibiting row percentages. The hypothesis of independence is rejected at 5% significance level, but not at 1% level. If the undecided are moved to the No use (0) category or split between the two categories, the significance is lost.

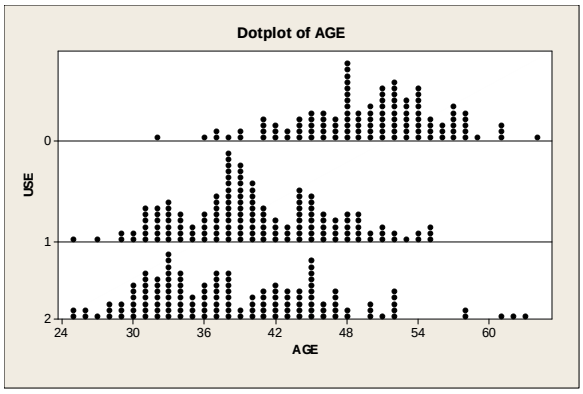
Tabulated statistics: STATUS; USE

```
Rows: STATUS   Columns: USE

           0      1      2     All
0      23.97  32.19  43.84  100.00
1      31.89  38.98  29.13  100.00
All    29.00  36.50  34.50  100.00

Cell Contents:      % of Row

Pearson Chi-Square = 8.981; DF = 2; P-Value = 0.011
Likelihood Ratio Chi-Square = 8.886; DF = 2; P-Value = 0.012
```

To see how USE depends on the GROUP variable, we present a stacked Bar Chart, showing that there are no apparent differences.



How USE depends on the last variable AGE, may be illustrated in various ways. One possibility is a Dot Plot for AGE for each category of USE:



We see that the ages of the No use group are higher than the Use group, and that in the undecided group, the younger ones are more frequent (and few old ones). The statistical analysis of a categorical variable depending on a numerical variable (or ordinal categorical variable) is typically not treated in elementary textbooks. One possibility is to categorize AGE say in two categories: Age 39 or below (1) and age 40 and above (2). Then we can perform ordinary cross-tabulation, and at the same time avoid some extra assumptions. This gives the following, where we clearly see that independence between AGEGROUP and USE is rejected, and that the row percentages agree with the statements above.

Tabulated statistics: AGEGROUP; USE

```
Rows: AGEGROUP   Columns: USE

           0      1      2     All
1       4.38  46.25  49.38  100.00
2      45.42  30.00  24.58  100.00
All    29.00  36.50  34.50  100.00

Cell Contents:      % of Row

Pearson Chi-Square = 79.808; DF = 2; P-Value = 0.000
Likelihood Ratio Chi-Square = 94.762; DF = 2; P-Value = 0.000
```

We now want to explain USE jointly by the variables AGE, GENDER, STATUS and GROUP. One possibility is binary logistic regression. For this we have to merge the USE data to two categories. Here we merge the undecided to the No-use group by recoding USE=0 if USE= 0 or 1 and otherwise use the same variables as we did for the regression analysis explaining ATTITUDE previously. We then get

Binary Logistic Regression

```
Link Function:  Logit

Response Information

Variable  Value       Count
USE       2             138  (Event)
          0             262
          Total         400

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef      StDev        Z     P    Ratio    Lower    Upper
Constant      3.2366     0.6697     4.83 0.000
AGE         -0.08895    0.01515    -5.87 0.000     0.91     0.89     0.94
GENDER       -0.3539     0.2313    -1.53 0.126     0.70     0.45     1.10
STATUS       -0.5466     0.2321    -2.35 0.019     0.58     0.37     0.91
```

```
GROUPB        0.3736     0.3229     1.16 0.247     1.45     0.77     2.74
GROUPC        0.4085     0.3240     1.26 0.207     1.50     0.80     2.84
GROUPD        0.4945     0.3154     1.57 0.117     1.64     0.88     3.04

Log-Likelihood = -229.411
Test that all slopes are zero: G = 56.614; DF = 6; P-Value = 0.000
```

We see that AGE is highly significant and STATUS is significant at the 5% level, but not on 1% level. However, GENDER and GROUP now turn out not significant. For GROUP this was no surprise, but may come as a surprise for GENDER. However, this may be explained by covariation between GENDER and AGE. Covariation between GENDER and STATUS is less likely (since most marriages/partnerships are one of each).

Some statistical software may provide a logistic regression opportunity for analyzing USE as an ordinal categorical variable (here three categories). This gives a slightly different result concerning GENDER, STATUS and GROUP, in that STATUS looses some predictive power which is picked up by GENDER and GROUP (relevant P-values in the region 3-7%). This shows that we may have lost some useful information by the recoding above. We will not try to interpret how here.