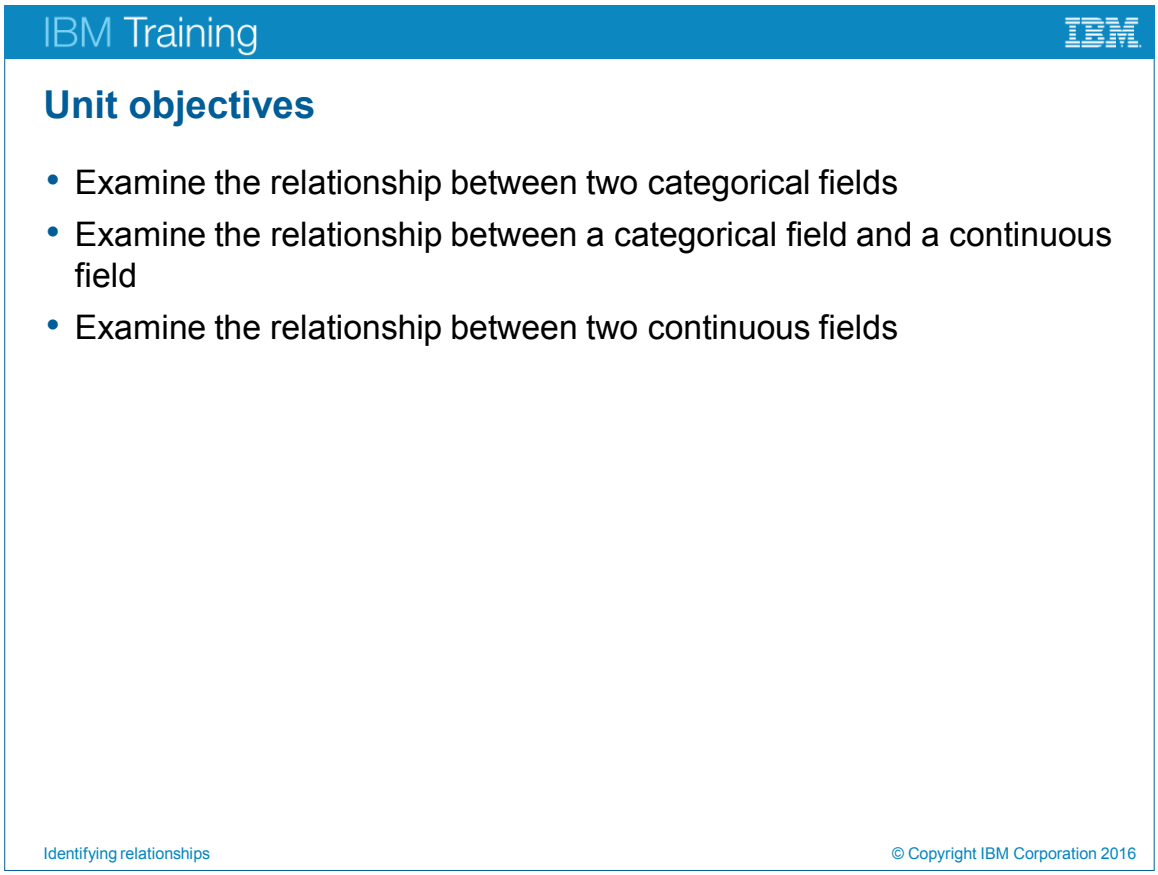
IBM Training

Identifying relationships

IBM SPSS Modeler (v18)



Unit objectives

Although building powerful models is key in data mining projects, investigating the relationships between the target field (churn, fraud, credit risk, response, and so on) and the predictors can still be helpful in answering the questions that motivated the project. You may find that revenue is directly related to length of time as a customer, or that customers with a certain mobile phone plan are more likely to switch providers. Although these patterns are not substitutes for a full model, they can often be used along with a model.

This unit presents methods to examine the relationship between two fields. Before reviewing this unit you should be familiar with the following topics:

- CRISP-DM
- IBM SPSS Modeler streams, nodes and palettes
- methods to collect initial data
- methods to explore the data

IBM Training

Examine the relationship between two fields

Relationship between	Tabular output	Graphical output
Two categorical fields	Matrix	Distribution
One categorical, one continuous field	Means	Histogram
Two continuous fields	Statistics	Plot

Identifying relationships

Examine the relationship between two fields

What to do to examine the relationship between two fields depends on the measurement level of the fields involved. This goes for exploring the relationship in both tabular and graphical format. This slide outlines what node to use.

When you use the dialog boxes, you will notice that IBM SPSS Modeler restricts the fields in the field lists to fields of a particular measurement level. For example, you can only select categorical fields in a Distribution node. If a field such as HAS_CHURNED is typed as continuous because its storage is numeric, you will not be able to run a Distribution node on the field. Therefore, it is important that the fields' measurement levels are set correctly.

IBM Training

Explore Matrix output

Matrix | Appearance | Annotations

GENDER

RISK		female	male	Total
bad	Count	457	449	906
	Column %	22.003	22.010	22.006
good	Count	1620	1591	3211
	Column %	77.997	77.990	77.994
Total	Count	2077	2040	4117
	Column %	100	100	100

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 0, df = 1, probability = 0.996

Identifying relationships

Explore Matrix output

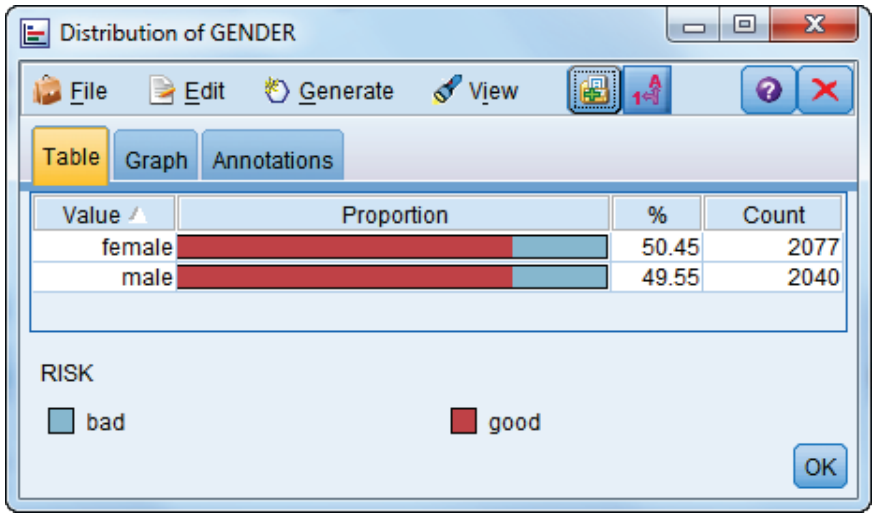
To examine the relationship between two categorical fields, cross tabulate the two fields by using the Matrix node (located in the Output palette).The output on this slide shows the relationship between GENDER and RISK, two categorical fields. Percentages in the table are based on the total column count. For example, 457 out of 2077 women are classified as bad risk, a percentage of (457/2077) * 100 = 22.003%. This percentage is 22.010% for men.

Examining the percentages, there appears to be no difference between men and women in percentage bad risk. This is confirmed by a statistical test, the Chi-square test. Roughly put, this test computes the probability that the difference between the percentages of bad risk is caused by the sampling process. (Note: probabilities range between 0 and 1).

The probability that the difference between the percentages of bad risk can be attributed to the sampling process in this dataset equals .996, almost 1. Based on this probability you may conclude that the difference that you have observed is only a sample difference, and that men and women in the population, from which this sample was drawn, do not differ with respect to bad risk.

IBM Training

Explore Distribution output



Identifying relationships

Explore Distribution output

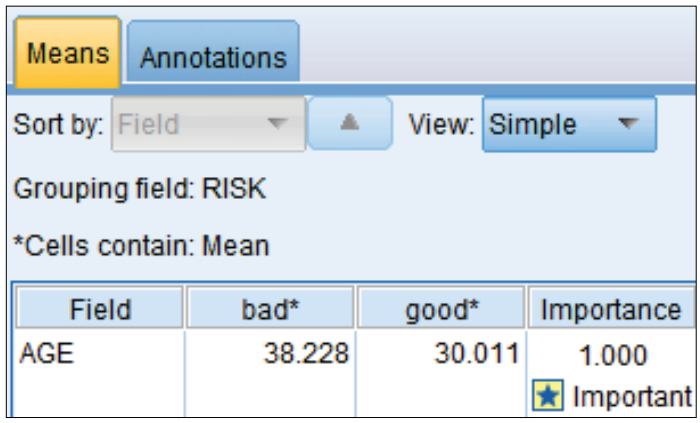
Rather than presenting a table with counts and percentages, you can visualize the relationship by using the Distribution node (located in the Graphs palette).

In the example on this slide, the categories of GENDER make up the bars, and the bars are overlaid with the values of RISK. The bars have the same length, which is the effect of using normalizing the bars. Normalization makes it easier to compare the groups for their bad risk rates.

The figure confirms the findings for the cross tabulation: men and women appear to behave the same with respect to credit risk.

IBM Training

Explore Means output



Identifying relationships

Explore Means output

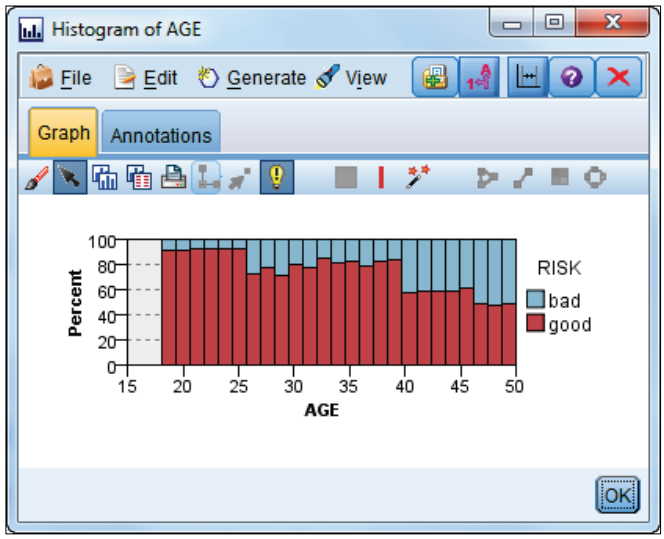
To investigate the relationship between a categorical field and a continuous field, compare the group means by using the Means node (located in the Output palette).On this slide, the mean age for the bad credit risk group is 38.228, against 30.011 for the good credit risk group. So, on average, bad risk customers are older than good risk customers.

The difference (38.228 versus 30.011) is labeled important, as shown in the Importance column. Importance equals 1 – probability, and as in the Chi-square test this is the probability that the sample difference between the means is caused by the sampling process (although the actual test performed is different from the Chi-square test, because of the measurement levels of the fields involved). An importance of 1 means that the probability must have been 0, and you may conclude that the sample difference cannot be attributed to the sampling process. Thus, there must be another reason why you have observed the difference: the sample reflects a difference in the population.

All in all, you may conclude that there are differences in mean age between the two groups.

IBM Training

Explore Histogram output



Identifying relationships

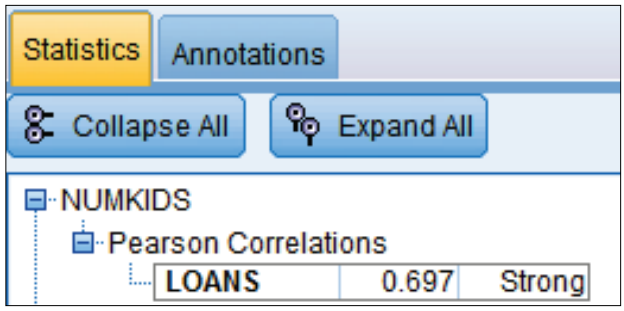
Explore Histogram output

Instead of, or in conjunction with, a table of means, you can present the relationship graphically. The Histogram node (located in the Graphs palette) is used rather than the Distribution node when one field is categorical, and the other continuous.

The figure on this slide shows a normalized histogram (the bars all have the same length). Apparently, older people have higher rates of bad risk than youngsters.

IBM Training

Explore Statistics output



Identifying relationships

Explore Statistics output

To examine the relationship between two continuous fields, request the Pearson correlation using the Statistics node (located in the Output palette).

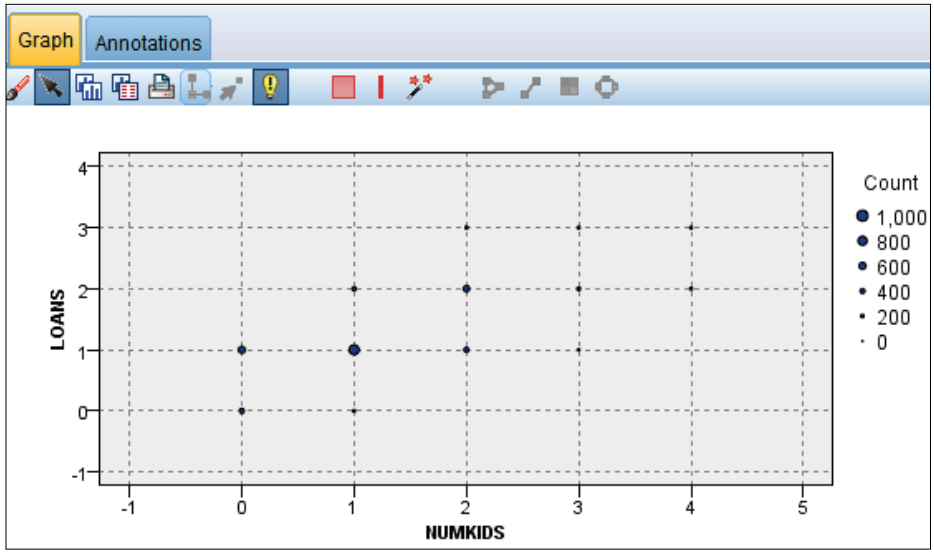
The Pearson correlation measures the extent to which two continuous fields are linearly associated, that is the degree to which the relationship between two fields can be described by a straight line.

The correlation coefficient ranges from –1 to +1, where: +1 represents a perfect positive linear relationship (as one field increases the other field increases at a constant rate), –1 represents a perfect negative relationship (as one field increases the other decreases at a constant rate), and 0 represents no linear relationship between the two fields.

In this example, the correlation between NUMKIDS and LOANS equals of .697 and is labeled Strong.

IBM Training

Explore Plot output



Identifying relationships

Explore Plot output

To visualize the relationship between two continuous fields, plot one field against the other with the Plot node (located in the Graphs palette).

In general, it is hard to deduce the value of the correlation from a plot. On the previous slide, we noticed that the correlation between NUMKIDS and LOANS was 0.697, and labeled Strong, which you probably cannot tell from this plot.

Having said this, the plot does give an impression of the form of the relationship. Non-linear relationships can be detected in this way, which can explain why the correlation between the two fields is small.