proportions of the other categories are calculated and used when attributing those individuals with rare categories.

- Elimination of individuals with rare categories. This solution should be avoided wherever possible. It should only be used if all of the rare categories are due to a very small number of individuals (situation which sometimes occurs when questions remain unanswered).

3.7.2 Description of a Categorical Variable or a Subpopulation

Multidimensional analysis is often supplemented by univariate analyses which are used to characterise a number of specific variables. We shall here focus on describing a specific categorical variable as well as groups of individuals defined by the categories of this variable. To do so, we can use quantitative variables, categorical variables, or the categories of categorical variables.

For example, we shall here describe the variable *type* in detail (cheapest, luxury, supermarket, etc.); one interesting feature of this variable is that it has more than two categories. The results of the **catdes** function applied to the variable *type* are detailed as follows:

```
> catdes(tea,num.var=18)
```

3.7.2.1 Description of a Categorical Variable by a Categorical Variable

To evaluate the relationship between the categorical variable we are interested in (*type*), and another categorical variable, we can conduct a χ^2 test. The smaller the p-value associated with the χ^2 test, the more questionable the independence hypothesis, and the more the categorical variable characterises the variable *type*. The categorical variables can therefore be sorted in ascending order of p-value. In the example (see Table 3.5), the variable *place of purchase* is the most closely related to the variable *type*.

TABLE 3.5
Tea Data: Description of the Variable *Type* by the Categorical Variables (Object `$test.chi2`)

	P-value	Df
Place of purchase	1.1096e-18	10
Format	8.4420e-11	10
Tearoom	1.6729e-03	5
Friends	4.2716e-02	5
Slimming	4.3292e-02	5
Variety	4.9635e-02	10

3.7.2.2 Description of a Subpopulation (or a Category) by a Quantitative Variable

For each category of the categorical variable *type* and for each quantitative variable (denoted X), the v-test (a test-value) is calculated as follows:

$$\text{v-test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{I_q}\left(\frac{I-I_q}{I-1}\right)}},$$

where $\bar{x}_q$ is the average of variable X for the individuals of category q, $\bar{x}$ is the average of X for all of the individuals, and I_q is the number of individuals carrying the category q. This value is used to test the following null hypothesis: *the values of X for the individuals who chose the category q are selected at random from all of the possible values of X.* We therefore consider the random variable $\bar{X}_q$, average of the individuals for category q. Its expected value and variance are:

$$\mathbb{E}(\bar{X}_q) = \bar{x} \quad \text{and} \quad \mathbb{V}(\bar{X}_q) = \frac{s^2}{I_q} \times \frac{I - I_q}{I-1}.$$

The v-test can therefore be considered a "standardised" deviation between the mean of those individuals with the category q and the general average. Among other things, we can attribute a probability to the v-test. If, among the participants, X is normally distributed according to the null hypothesis, the $\bar{X}_q$ distribution is as follows:

$$\bar{X}_q = \mathcal{N}\left(\bar{x}, \frac{s}{\sqrt{I_q}}\sqrt{\frac{I-I_q}{I-1}}\right).$$

If X is not normally distributed, we can still use normal distribution as an approximate distribution for $\bar{X}_q$. We consider the v-test as a statistic of the test for H_0 ("the average of X for category q is equal to the general average", or in other words, "variable X does not characterise category q") and can therefore calculate a p-value.

Remark
When categories stem from a clustering: this test can only be applied satisfactorily to supplementary variables (i.e., which were not used to determine the categories), but they are also calculated for the active variables for information.

As the p-value provides an indication of the "significance" of a given deviation, it makes sense to organise the quantitative variables in descending order of v-test by limiting oneself to p-values less than 5%.

In the example (see below), the only category to be characterised by a quantitative variable is *t_luxury*. This category is characterised by individuals of above-average age as the v-test is positive. The average age of those who buy in this class is 43.4 years whereas the average overall age is 37.1 years.

The standard deviations are provided for both the class (16.9) and the overall population (16.8).

```
> catdes(tea,num.var=18)
$quanti$cheapest
NULL

$quanti$known.brand
NULL

$quanti$luxury
    v.test Mean in category  Overall mean  sd in category  Overall sd  p.value
age   3.02             43.4          37.1            16.9        16.8  0.00256

$quanti$shop.brand
NULL

$quanti$unknown
NULL

$quanti$varies
NULL
```

3.7.2.3 Description of a Subpopulation (or a Category) by the Categories of a Categorical Variable

The description of a categorical variable can be refined by studying the relationships between categories. We thus characterise each of the categories of the variable we are interested in (variable *type*) by using the categories of the categorical variables.

These calculations are illustrated using first the variable *place of purchase* and second the contingency table for the variables *type* and *place of purchase* (see Table 3.6).

TABLE 3.6
Tea Data: Contingency Table for the Variables *Type* and *Place of Purchase*

	Supermarket	Supermarket and Specialist	Specialist	Total
Cheapest	6	1	0	7
Luxury	12	20	21	53
Unknown	10	1	1	12
Famous brand	82	11	2	95
Shop brand	20	1	0	21
Varies	62	44	6	112
Total	192	78	30	300

Let us examine the category *luxury* and consider the variable *place of purchase* which has three categories: *supermarket*, *supermarket+specialist* and *specialist shop*. We shall look more closely at *specialist shop* (see Table 3.7.2.3). The following question is raised: "Is the category *luxury* characterised by the category *specialist shop*?" The objective is to calculate the proportion of individuals who buy their tea in a *specialist shop* out of those who buy luxury tea I_{qt}/I_q from the overall percentage of individuals who buy their tea in specialist shops I_t/I.

	Specialist shop	Other	Total
Luxury	$I_{qt} = 21$	32	$I_q = 53$
Other	9	238	247
Total	$I_t = 30$	270	$I = 300$

These two proportions are equal under the null hypothesis of independence:

$$\frac{I_{qt}}{I_q} = \frac{I_t}{I}.$$

I_q individuals are randomly selected (those with the category we are interested in *luxury*) among I (the total population). We shall focus on the random variable X equal to the number I_{qt} of occurrences of individuals which have the characteristic that is being studied (purchased in a specialist shop), while it must be remembered that their sample size within the population is I_t. Under the null hypothesis, the random variable X follows the hypergeometric distribution $\mathcal{H}(I, I_t, I_q)$. The probability of having a more extreme value than the observed value can therefore be calculated. For each category of the variable being studied, each of the categories of the characterising categorical variables can be sorted in ascending order of p-value.

The first row of Table 3.7 indicates that 70% (21/30; see Table 3.6 or the extract) of the individuals who buy their tea in specialist shops also belong to the class *luxury*; 39.6% (21/53; see Table 3.6) of the individuals from the class *luxury* purchase their tea in specialist shops; 10% (30/300; see Table 3.6) of the participants purchase their tea in specialist shops. The p-value of the test (1.58e-11) is provided along with the associated v-test (6.64). The v-test here corresponds to the quantile of the normal distribution which is associated with p-value; the sign indicates an over- or underrepresentation (Lebart et al., 2006).

The categories of all the categorical variables are organised from most to least characteristic when the category is overrepresented in the given class (i.e., the category in question) compared to the other categories (the v-test is therefore positive), and from least characteristic to most when the category is underrepresented in the class (and the v-test is therefore negative). The individuals who buy luxury tea are most significantly characterised by the fact that they do not buy tea in supermarkets (the v-test for supermarkets is negative, and has the highest absolute value).

3.7.3 The Burt Table

A Burt table is a square table of $K \times K$ dimensions, where each row and each column correspond to one of the categories K of the set of variables. In the cell (k, k') we observe the number of individuals who carry both categories k

TABLE 3.7
Tea Data: Description of the Category *Luxury* of the Variable *Type* by the Categories of the Categorical Variables (Object `$category$luxury`)

	Cla/Mod	Mod/Cla	Global	P-value	V-test
Place.of.purchase=specialist.shop	70.00	39.6	10.0	3.16e-11	6.64
Format=loose	55.60	37.7	12.0	5.59e-08	5.43
Variety=black	28.40	39.6	24.7	1.15e-02	2.53
Age_Q=60 and +	31.60	22.6	12.7	3.76e-02	2.08
No.effect.health=no.effect.health	27.30	34.0	22.0	3.81e-02	2.07
No.effect.health=not.without.effect	15.00	66.0	78.0	3.81e-02	-2.07
Variety=flavoured	12.40	45.3	64.3	2.86e-03	-2.98
Age_Q=15-24	7.61	13.2	30.7	2.48e-03	-3.03
Format=sachet	8.24	26.4	56.7	1.90e-06	-4.76
Place.of.purchase=supermarket	6.25	22.6	64.0	2.62e-11	-6.67

and k'. This table is an extension of the contingency table where there are more than two categorical variables: it juxtaposes all of the information from the contingency table of variables taken as pairs (in rows and columns).

A correspondence analysis of this table is used to represent the categories. As this table is symmetrical, the representation of the cloud of row profiles is identical to that of the cloud of column profiles (only one of the two representations is therefore retained). This representation is very similar to the representation of the categories as provided by MCA and demonstrates the collinearity of the principal components of the same rank. However, the inertias associated with each component differ by a coefficient of λ_s. When λ_s is the inertia of s for the MCA, the inertia of component s for a CA of the Burt table will be λ_s^2. It can be observed that the percentages of inertia associated with the first components of the CA of the Burt table are higher than the percentages of inertia associated with the first components of the MCA alone. In the example, the percentages of inertia associated with the first two components of the MCA are worth 9.88% and 8.10% respectively, compared with 20.73% and 14.11% for those of the CA.

The Burt table is therefore useful in terms of data storage. Rather than conserving the complete table of individuals × variables, it is in fact sufficient to construct a Burt table containing the same information in terms of associations between categories, which are considered in pairs with a view to conducting the principal component method. When dealing with a very large number of individuals, the individual responses are often ignored in favour of the associations between categories.

3.7.4 Missing Values

It is very common for some data to be missing, for a survey conducted by questionnaire, for example. The easiest way to manage missing values in datasets with categorical variables is to create a new category for each variable which contains one or more missing values. A variable j with K_j categories will therefore have K_j+1 categories if at least one individual possesses missing