

- 2 The expected frequencies should be greater than 5. Although it is acceptable in larger contingency tables to have up to 20% of expected frequencies below 5, the result is a loss of statistical power (so, the test may fail to detect a genuine effect). Even in larger contingency tables no expected frequencies should be below 1. Howell (2006) gives a nice explanation of why violating this assumption creates problems. If you find yourself in this situation consider using Fisher's exact test (section 18.3.2).

Finally, although it's not an assumption, it seems fitting to mention in a section in which a gloomy and foreboding tone is being used that proportionately small differences in cell frequencies can result in statistically significant associations between variables if the sample is large enough (although it might need to be very large indeed). Therefore, we must look at row and column percentages to interpret any effects we get. These percentages will reflect the patterns of data far better than the frequencies themselves (because these frequencies will be dependent on the sample sizes in different categories).

18.5. Doing chi-square on SPSS ①

There are two ways in which categorical data can be entered: enter the raw scores, or enter weighted cases. We'll look at both in turn.

18.5.1. Entering data: raw scores ①



If we input the raw scores, it means that every row of the data editor represents each entity about which we have data (in this example, each row represents a cat). So, you would create two coding variables (**Training** and **Dance**) and specify appropriate numeric codes for each. The **Training** could be coded with 0 to represent a food reward and 1 to represent affection, and **Dance** could be coded with 0 to represent an animal that danced and 1 to represent one that did not. For each animal, you put the appropriate numeric code into each column. So a cat that was trained with food that did not dance would have 0 in the training column and 1 in the dance column. The data in the file *Cats.sav* are entered in this way and you should be able to identify the variables described. There were 200 cats in all and so there are 200 rows of data.

18.5.2. Entering data: weight cases ①

An alternative method of data entry is to create the same coding variables as before, but to have a third variable that represents the number of animals that fell into each combination of categories. In other words, we input the frequency data (the number of cases that fall into a particular category). We could call this variable **Frequency**. Figure 18.2 shows the data editor with this third variable added. Now, instead of having 200 rows, each one representing a different animal, we have one row representing each combination of categories and a variable telling us how many animals fell into this category combination. So, the first row represents cats that had food as a reward and then danced. The variable **Frequency** tells us that there were 28 cats that had food as a reward and then danced. This information was previously represented by 28 different rows in the file *Cats.sav* and so you can see how this method of data entry saves you a lot of time! Extending this principle, we can see that when affection was used as a reward 114 cats did not dance.

	Training	Dance	Frequency	var
1	Food as Reward	Yes	28	
2	Food as Reward	No	10	
3	Affection as Reward	Yes	48	
4	Affection as Reward	No	114	
5				

FIGURE 18.2
Data entry using
weighted cases

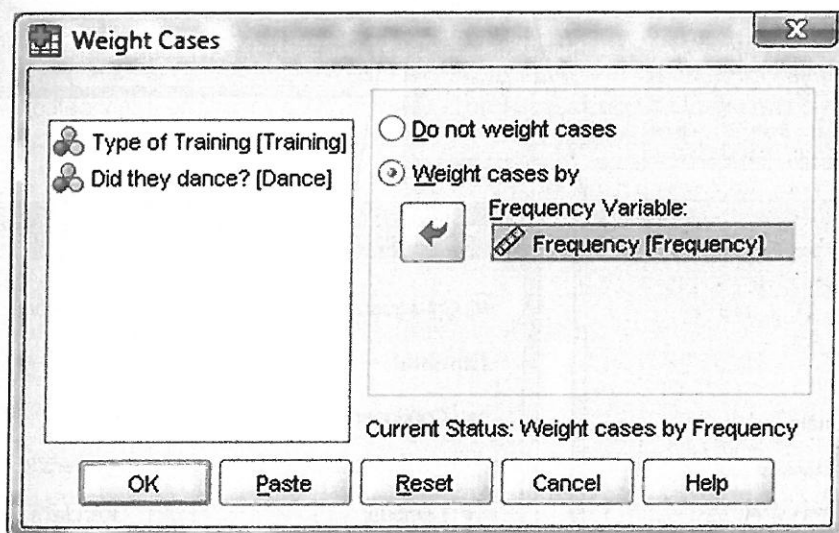


FIGURE 18.3
The dialog box
for the *weight*
cases command

Entering data using a variable representing the number of cases that fall into a combination of categories can be quite labour saving. However, to analyse data entered in this way we must tell the computer that the variable *Frequency* represents the number of cases that fell into a particular combination of categories. To do this, access the *Weight Cases* dialog box in Figure 18.3 by selecting **Data** → **Weight Cases...** Select **Weight cases by** and then select the variable in which the number of cases is specified (in this case *Frequency*) and drag it to the box labelled *Frequency variable* (or click on **→**). This process tells the computer that it should weight each category combination by the number in the column labelled *Frequency*. Therefore, the computer will pretend, for example, that there are 28 rows of data that have the category combination 0, 0 (representing cats trained with food and that danced). Data entered in this way are in the file *CatsWeight.sav* and if you use this file you must remember to weight the cases as described.



18.5.3. Running the analysis ①

Summarizing data that fall into categories is done using the *crosstabs* command (which also produces the chi-square test). *Crosstabs* is accessed by selecting **Analyze** **Descriptive Statistics** **Crosstabs...** Figure 18.4 shows the dialog boxes for the *crosstabs* command (the variable *Frequency* is in the diagram because I ran the analysis on the

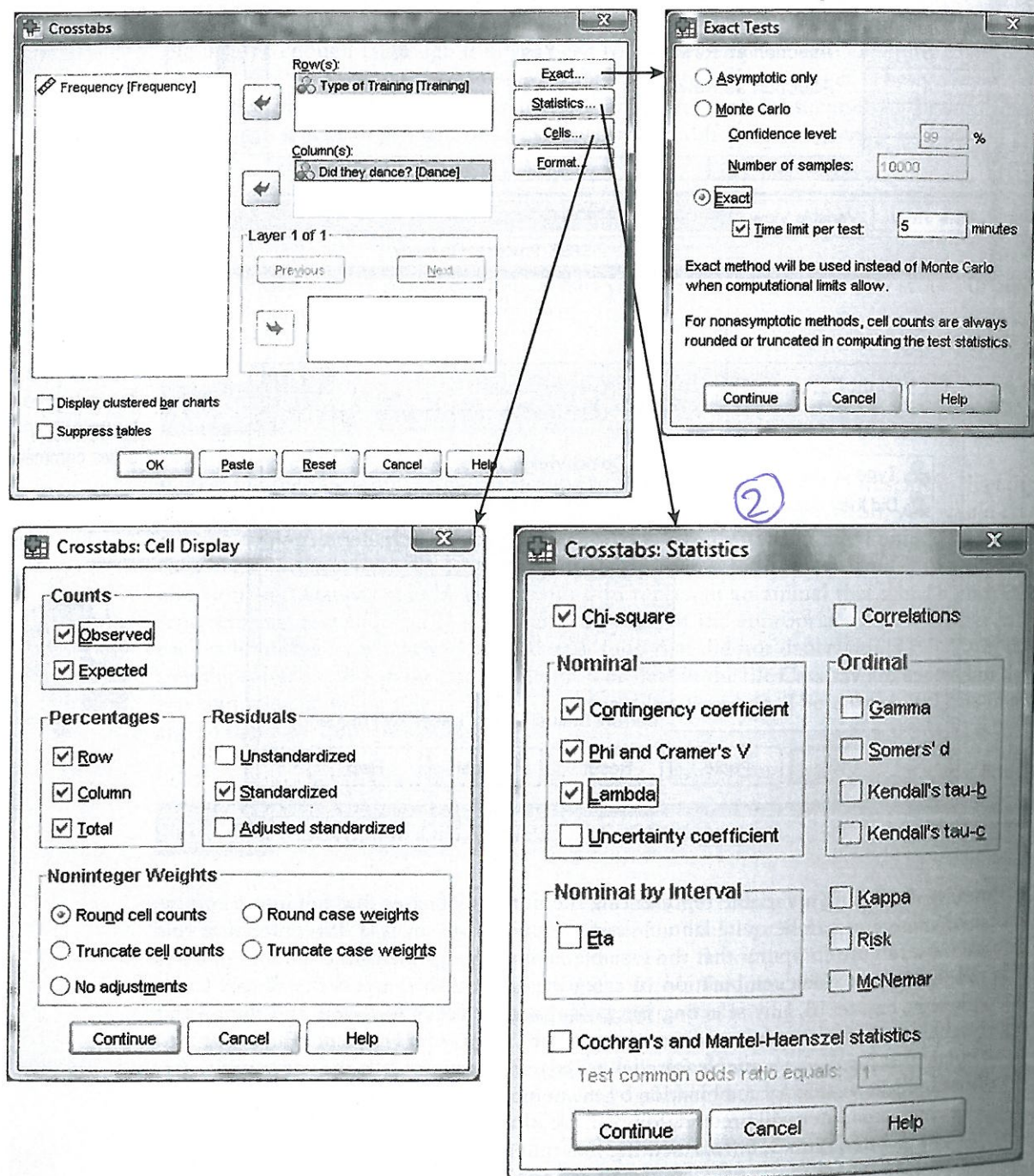


FIGURE 18.4 Dialog boxes for the *crosstabs* command

CatsWeight.sav data). First, select one of the variables of interest in the variable list and drag it into the box labelled *Row(s)* (or click on ). For this example, I selected **Training** to be the rows of the table. Next, select the other variable of interest (**Dance**) and drag it to the box labelled *Column(s)* (or click on ). In addition, it is possible to select a layer variable (i.e. you can split the rows of the table into further categories). If you had a third categorical variable (as we will later in this chapter) you could split the contingency table by this variable (so layers of the table represent different categories of this third variable).

If you click on **Statistics...** a dialog box appears in which you can specify various statistical tests. The most important options under the statistics menu for categorical data are described in SPSS Tip 18.1.



SPSS TIP 18.1

Statistical options for *crosstabs*

- **Chi-square:** This performs the basic Pearson chi-square test (section 18.3.1).
- **Phi and Cramer's V:** These are measures of the strength of association between two categorical variables. Phi is used with 2×2 contingency tables (tables in which you have two categorical variables and each variable has only two categories). Phi is calculated by taking the chi-square value and dividing it by the sample size and then taking the square root of this value. If one of the two categorical variables contains more than two categories then Cramer's V is preferred to phi because phi fails to reach its minimum value of 0 (indicating no association) in these circumstances.
- **Goodman and Kruskal's lambda (λ):** This measures the proportional reduction in error that is achieved when membership of a category of one variable is used to predict category membership of the other variable. A value of 1 means that one variable perfectly predicts the other, whereas a value of 0 indicates that one variable in no way predicts the other.
- **Kendall's statistic:** This statistic is discussed in section 6.5.4.

Select the chi-square test, the continuity correction, phi and lambda, and then click on **Continue**. If you click on **Cells...** a dialog box appears in which you can specify the type of data displayed in the crosstabulation table. It is important that you ask for expected counts because for chi-square to be accurate these expected counts must exceed certain values. The basic rule of thumb is that with 2×2 contingency tables no expected values should be below 5. In larger tables the rule is that all expected counts should be greater than 1 and no more than 20% of expected counts should be less than 5. It is also useful to have a look at the row, column and total percentages because these values are usually more easily interpreted than the actual frequencies and provide some idea of the origin of any significant effects. The second important option is to select some standardized residuals. These values are important for breaking down a significant chi-square test (should we get one). Once these options have been selected, click on **Continue** to return to the main dialog box.

From here you can click on **Exact...** (if you have exact tests installed) to compute Fisher's exact test (section 18.3.2). You can use this option if your sample is small or if your expected frequencies are too low (see 18.4). Select the **Exact** test option; we don't really need it for these data but it will be a useful way to see how it is used. Click on **Continue** to return to the main dialog box and then click on **OK** to run the analysis.

18.5.4. Output for the chi-square test ①

The crosstabulation table produced by SPSS (SPSS Output 18.1) contains the number of cases that fall into each combination of categories and is rather like our original contingency table. We can see that in total 76 cats danced (38% of the total) and of these 28 were trained using food (36.8% of the total that danced) and 48 were trained with affection (63.2% of the total that danced). Further, 124 cats didn't dance at all (62% of the total) and of those that didn't dance, 10 were trained using food as a reward (8.1% of the total that didn't dance) and a massive 114 were trained using affection (91.9% of the total that didn't dance). The numbers of cats can be read from the rows labelled *Count* and the percentages are read from the rows labelled *% within Did they dance?* We can also look at the percentages within the training categories by looking at the rows labelled *% within Type of Training*. This tells us, for example, that of those trained with food as a reward, 73.7% danced and 26.3% did not. Similarly, for those trained with affection only 29.6% danced compared to 70.4% that didn't. In summary, when food was used as a reward most cats would dance, but when affection was used most cats refused to dance.

SPSS OUTPUT 18.1

Type of Training * Did they dance? Crosstabulation

			Did they dance?		
			Yes	No	Total
Type of Training	Food as Reward	Count	28	10	38
		Expected Count	14.4	23.6	38.0
		% within Type of Training	73.7%	26.3%	100.0%
		% within Did they dance?	36.8%	8.1%	19.0%
		% of Total	14.0%	5.0%	19.0%
		Std. Residual	3.6	-2.8	
	Affection as Reward	Count	48	114	162
		Expected Count	61.6	100.4	162.0
		% within Type of Training	29.6%	70.4%	100.0%
		% within Did they dance?	63.2%	91.9%	81.0%
		% of Total	24.0%	57.0%	81.0%
		Std. Residual	-1.7	1.4	
Total	Total	Count	76	124	200
		Expected Count	76.0	124.0	200.0
		% within Type of Training	38.0%	62.0%	100.0%
		% within Did they dance?	100.0%	100.0%	100.0%
		% of Total	38.0%	62.0%	100.0%

Before moving on to look at the test statistics itself it is vital that we check that the assumption for chi-square has been met. The assumption is that in 2×2 tables (which is what we have here), all expected frequencies should be greater than 5. If you look at the expected counts in the crosstabulation table (which incidentally are the same as we calculated earlier), it should be clear that the smallest expected count is 14.4 (for cats that were trained with food and did dance). This value exceeds 5 and so the assumption has been met. If you found an expected count lower than 5 the best remedy is to collect more data to try to boost the proportion of cases falling into each category.

As we saw earlier, Pearson's chi-square test examines whether there is an association between two categorical variables (in this case the type of training and whether the animal danced or not). As part of the *crosstabs* procedure SPSS produces a table that includes the chi-square statistic and its significance value (SPSS Output 18.2). The Pearson chi-square

* may be in different order.

Chi-Square Tests

SPSS OUTPUT 18.2

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	25.356 ^a	1	.000	.000	.000	
Continuity Correction ^b	23.520	1	.000			
Likelihood Ratio	24.932	1	.000	.000	.000	
Fisher's Exact Test				.000	.000	
Linear-by-Linear Association	25.229 ^c	1	.000	.000	.000	.000
N of Valid Cases	200					

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.44.

b. Computed only for a 2x2 table

c. The standardized statistic is 5.023.

statistic tests whether the two variables are independent. If the significance value is small enough (conventionally *Sig.* must be less than .05) then we reject the hypothesis that the variables are independent and gain confidence in the hypothesis that they are in some way related. The value of the chi-square statistic is given in the table (and the degrees of freedom) as is the significance value. The value of the chi-square statistic is 25.356, which is within rounding error of what we calculated in section 18.3.1. This value is highly significant ($p < .001$), indicating that the type of training used had a significant effect on whether an animal would dance.

A series of other statistics are also included in the table (many of which have to be specifically requested using the options in the dialog box in Figure 18.4). *Continuity Correction* is Yates's continuity corrected chi-square (see section 18.3.4) and its value is the same as the value we calculated earlier (23.52). As I mentioned earlier, this test is probably best ignored anyway, but it does confirm the result from the main chi-square test. The *Likelihood Ratio* is the statistic we encountered in section 18.3.3 (and is again within rounding error of the value we calculated: 24.93). Again this confirms the main chi-square result, but this statistic would be preferred in smaller samples.

Underneath the chi-square table there are several footnotes relating to the assumption that expected counts should be greater than 5. If you forgot to check this assumption yourself, SPSS kindly gives a summary of the number of expected counts below 5. In this case, there were no expected frequencies less than 5 so we know that the chi-square statistic should be accurate.

The highly significant result indicates that there is an association between the type of training and whether the cat danced or not. What we mean by an association is that the pattern of responses (i.e. the proportion of cats that danced to the proportion that did not) in the two training conditions is significantly different. This significant finding reflects the fact that when food is used as a reward, about 74% of cats learn to dance and 26% do not, whereas when affection is used, the opposite is true (about 70% refuse to dance and 30% do dance). Therefore, we can conclude that the type of training used significantly influences the cats: they will dance for food but not for love! Having lived with a lovely cat for many years now, this supports my cynical view that they will do nothing unless there is a bowl of cat-food waiting for them at the end of it!

If requested, SPSS will produce another table of output containing some additional statistical tests. Most of these tests are measures of the strength of association. These measures are based on modifying the chi-square statistic to take account of sample size and degrees of freedom and they try to restrict the range of the test statistic from 0 to 1 (to make them similar to the correlation coefficient described in Chapter 6). These are shown in SPSS Output 18.3.

How do I interpret
chi-square?



- *Phi*: This statistic is accurate for 2×2 contingency tables. However, for tables with greater than two dimensions the value of phi may not lie between 0 and 1 because the chi-square value can exceed the sample size. Therefore, Pearson suggested the use of the coefficient of contingency.
- *Contingency Coefficient*: This coefficient ensures a value between 0 and 1 but, unfortunately, it seldom reaches its upper limit of 1 and for this reason Cramer devised Cramer's V.
- *Cramer's V*: When both variables have only two categories, phi and Cramer's V are identical. However, when variables have more than two categories Cramer's statistic can attain its maximum of one – unlike the other two – and so it is the most useful.

For these data, Cramer's statistic is 0.36 out of a possible maximum value of 1. This represents a medium association between the type of training and whether the cats danced or not (if you think of it like a correlation coefficient then this represents a medium effect size). This value is highly significant ($p < .001$) indicating that a value of the test statistic that is this big is unlikely to have happened by chance, and therefore the strength of the relationship is significant. These results confirm what the chi-square test already told us but also give us some idea of the size of effect.

SPSS OUTPUT 18.3

Symmetric Measures

		Value	Approx. Sig.	Exact Sig.
Nominal by Nominal	Phi	.356	.000	.000
	Cramer's V	.356	.000	.000
	Contingency Coefficient	.335	.000	.000
	N of Valid Cases	200		

18.5.5. Breaking down a significant chi-square test with standardized residuals ②

Although in a 2×2 contingency table like the one we have in this example the nature of the association can be quite clear from just the cell percentages or counts, in larger contingency tables it can be useful to do a finer-grained investigation of the table. In a way, you can think of a significant chi-square test in much the same way as a significant interaction in ANOVA: it is an effect that needs to be broken down further. One very easy way to break down a significant chi-square test is to use data that we already have – the standardized residual.

Just like regression, the residual is simply the error between what the model predicts (the expected frequency) and the data actually observed (the observed frequency):

$$\text{residual}_{ij} = \text{observed}_{ij} - \text{model}_{ij}$$

in which i and j represent the two variables (i.e. the rows and columns in the contingency table). This is the same as every other residual or deviation that we have encountered in this book (compare this equation to, for example, equation (2.4)). To standardize this equation, we simply divide by the square root of the expected frequency:

$$\text{standardized residual} = \frac{\text{observed}_{ij} - \text{model}_{ij}}{\sqrt{\text{model}_{ij}}}$$

Does this equation look familiar? Well, it's basically part of equation (18.1). The only difference is that rather than looking at squared deviations, we're looking at the pure deviation. Remember that the rationale for squaring deviations in the first place is simply to make them positive so that they don't cancel out when we add them. The chi-square statistic is based on adding together values, so it is important that the deviations are squared so that they don't cancel out. However, if we're not planning to add up the deviations or residuals then we can inspect them in their unsquared form. There are two important things about these standardized residuals:

- 1 Given that the chi-square statistic is the sum of these standardized residuals (sort of), then if we want to decompose what contributes to the overall association that the chi-square statistic measures, then looking at the individual standardized residuals is a good idea because they have a direct relationship with the test statistic.
- 2 These standardized residuals behave like any other (see section 7.6.1.1) in the sense that each one is a z -score. This is very useful because it means that just by looking at a standardized residual we can assess its significance (see section 1.7.4). As we have learnt many times before, if the value lies outside of ± 1.96 then it is significant at $p < .05$, if it lies outside ± 2.58 then it is significant at $p < .01$ and if it lies outside ± 3.29 then it is significant at $p < .001$.

Fortunately, when we selected ☒ Standardized in Figure 18.4, SPSS produced these standardized residuals and we can see them in SPSS Output 18.1. There are four residuals: one for each combination of the type of training and whether the cats danced. When food was used as a reward the standardized residual was significant for both those that danced ($z = 3.6$) and those that didn't dance ($z = -2.8$) because both values are bigger than 1.96 (when you ignore the minus sign). The plus or minus sign (and the counts and expected counts within the cells) tells us that when food was used as a reward significantly more cats than expected danced, and significantly less cats than expected did not dance. When affection was used as a reward the standardized residual was not significant for both those that danced ($z = -1.7$) and those that didn't dance ($z = 1.4$) because they are both smaller than 1.96 (when you ignore the minus sign). This tells us that when affection was used as a reward as many cats as expected danced and did not dance. In a nutshell, the cells for when food was used as a reward both significantly contribute to the overall chi-square statistic. Put another way, the association between the type of reward and dancing is mainly driven by when food is a reward.

18.5.6. Calculating an effect size ②

Although Cramer's V is an adequate effect size (in the sense that it is constrained to fall between 0 and 1 and is, therefore, easily interpretable), a more common and possibly more useful measure of effect size for categorical data is the odds ratio, which we encountered in Chapter 8 (it was called $Exp(B)$ in that chapter). Odds ratios are most interpretable in 2×2 contingency tables and are probably not useful for larger contingency tables. However, this isn't as restrictive as you might think because as I've said more times than I care to recall in the GLM chapters, effect sizes are only ever useful when they summarize a focused comparison. A 2×2 contingency table is the categorical data equivalent of a focused comparison!

The odds ratio is simple enough to calculate. If we look at our example, we can first calculate the odds that a cat danced given that they had food as a reward. This is simply

the number of cats that were given food and danced, divided by the number of cats given food that didn't dance:

$$\begin{aligned}\text{odds}_{\text{dancing after food}} &= \frac{\text{number that had food and danced}}{\text{number that had food but didn't dance}} \\ &= \frac{28}{10} \\ &= 2.8\end{aligned}$$

Next we calculate the odds that a cat danced given that they had affection as a reward. This is simply the number of cats that were given affection and danced, divided by the number of cats given affection that didn't dance:

$$\begin{aligned}\text{odds}_{\text{dancing after affection}} &= \frac{\text{number that had affection and danced}}{\text{number that had affection but didn't dance}} \\ &= \frac{48}{114} \\ &= 0.421\end{aligned}$$

The odds ratio is simply the odds of dancing after food divided by the odds of dancing after affection:

$$\begin{aligned}\text{odds ratio} &= \frac{\text{odds}_{\text{dancing after food}}}{\text{odds}_{\text{dancing after affection}}} \\ &= \frac{2.8}{0.421} \\ &= 6.65\end{aligned}$$

What this tells us is that if a cat was trained with food the odds of their dancing were 6.65 times higher than if they had been trained with affection. As you can see, this is an extremely elegant and easily understood metric for expressing the effect you've got!



18.5.7. Reporting the results of chi-square ①

When reporting Pearson's chi-square we simply report the value of the test statistic with its associated degrees of freedom and the significance value. The test statistic, as we've seen, is denoted by χ^2 . The SPSS output tells us that the value of χ^2 was 25.36, that the degrees of freedom on which this was based were 1, and that it was significant at $p < .001$. It's also useful to reproduce the contingency table and my vote would go to quoting the odds ratio too. As such, we could report:

- ✓ There was a significant association between the type of training and whether or not cats would dance $\chi^2 (1) = 25.36, p < .001$. This seems to represent the fact that, based on the odds ratio, the odds of cats dancing were 6.65 times higher if they were trained with food than if trained with affection.