Chapter 9. Evaluating Significance

With Emily H. Griffith, Ph.D.
Department of Statistics
NC State University

> Over the last four chapters you have been looking at ways of seeing patterns in verbal data. In particular, you have been asking how the distribution of data into coding categories varied across contrast and across dimension. In some cases, you may have found small variations; in some cases, large variations. In this chapter, we turn to considering the issue of evaluating the significance of those variations.

Significance and Surprise

Generally speaking, we call something "significant" if it is *important*, if it has a bearing on what we will do. But statistical significance is better thought of as *surprising* rather than *important*. If something surprises us that means it seems outside of our expectations. It's unusual. As this definition suggests, evaluating statistical significance involves making a comparison between what we have observed and what we would usually expect to observe if nothing much was going on.

The comparison of observations and expectations guides our evaluation of significance in all kinds of everyday activities. We judge, for example, the significance of Jenna's not returning our morning greeting against our expectations for what Jenna would do if nothing much were going on. If our model of expectations is that she always returns our greeting, her failing to do so this morning can seem highly significant. If, however, she is often lost in thought on days when she has a lot of work to do, her failure to return our greeting will be seen as far less significant.

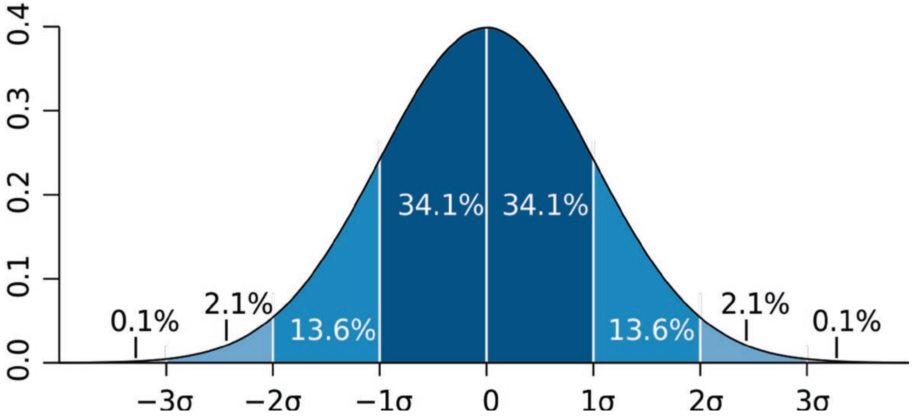

Figure 9.1: A normal distribution.[7]

For many of us, the most familiar tests of statistical significance involve comparing what was actually observed to expectations represented in a normal curve like that shown in Figure 9.1. With a normal curve, just by knowing the values of two parameters, the mean (or the average) and the standard deviation, you can draw the curve. The standard deviation is a measure of the degree to which the data is spread out around the mean. It is calculated by subtracting the mean from each data point, squaring the results (to make sure that none of them are negative numbers), taking their average, and then taking the square root of the result (to reverse the effects of the earlier squaring).

7 Graphic by M. W. Toews and used under the Creative Commons Attribution license 2.5 Generic (retrieved from https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg)

Statistics that use a normal curve shape our expectations about our observations using this mean and standard deviation. Graphically, one standard deviation is located at the graph's inflection points, where the slope changes from curving down to curving up. With a normal curve, we expect most of our observations to cluster symmetrically around the center or average, with fully 34.1% of the data lying evenly on either side of this mean. Within the further boundaries of two standard deviations from the mean we expect to find 95% of our observations. And we expect almost all of our observations (99.7%) to lie within the further boundaries of three standard deviations.

In many statistical methods, we imagine drawing a number of random samples from the expected model—100 samples, 1,000 samples, even 10,000 samples—and ask, how often would one of them look like what I've observed? In terms of Jenna's greetings—and assuming we had perfect memory—we might examine her morning behavior over the last three years when we know everything was OK between us and ask, how many times did she not return our greeting? If it just one time in 1,000, we might say the chances that this morning's behavior was expected was one in 1,000 or, written in statistical language, $p < .01$.

The assumption that an appropriate underlying model of expected distribution follows a symmetrical normal curve works very well for some phenomena. The number of times a coin turns up heads in a set of tosses, for example, follows a normal distribution, assuming the coin is not dinged up in a way that favors one side or the other. Many phenomena are not normally distributed however. If we ask about the distribution of household income in a country, for example, we often find that a few individuals have a net worth far greater than the average household. These high-income households, when averaged in with everyone else positively skew the mean household income as shown in Figure 9.2, not a normal curve.

Many researchers evaluate significance without understanding that they are implicitly making a choice about how to model their underlying expectations for the data. If the assumption of normalcy is inappropriate, such tests will tell you little about how you should evaluate the outcome of your analysis. It would be as if you had taken Jenna's behavior and inappropriately compared it to your model of expectations for Ralph: Ralph always returns my greeting, we might think, so Jenna's silence must be highly significant. As this example is meant to indicate, using the wrong model of underlying expectations can warp your evaluation of significance.

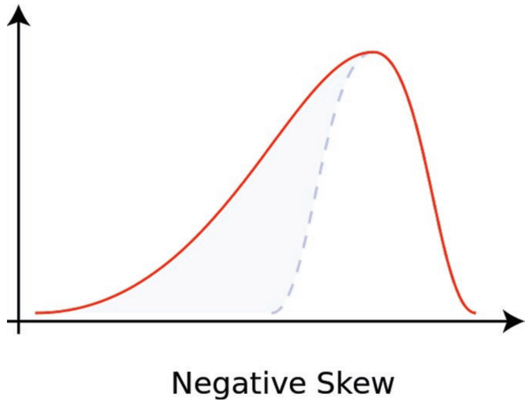

Figure 9.2: A skewed distribution.[8]

Exercise 9.1 Test Your Understanding

Decide whether you would expect the following distributions to be normal if nothing much were going on. Think about where the average might be, and then consider whether you would expect values above and below that average to be evenly distributed and increasingly less common the further from the average.

- The number of times heads arise in 150 coin tosses.
- The length of essays written in a timed writing assessment.
- The number of students who pass and who fail as the result of a writing assessment.
- The number of times a computer user checks email in an average day.
- The number of personal and work-related emails a computer user receives.

For Discussion: What aspects of the data seem to be important to consider in choosing a significance test?

8 Graphic by Rodolfo Hermans and used under the Creative Commons Attribution-Share Alike 3.0 Unported license. Retrieved from https://commons.wikimedia.org/wiki/File:Negative_and_positive_skew_diagrams_(English).svg.

Significance Tests for Coded Verbal Data

For the kind of analyses we have been discussing in this book—the analysis of verbal data gathered from a number of different cases (different speakers, different classrooms, different disciplines)—the normal curve is inappropriate as a model of underlying expectations. Verbal data coded into categories—categorical data—can never expected to approach a normal distribution because a normal distribution is continuous while categorical data is, well, categorical.

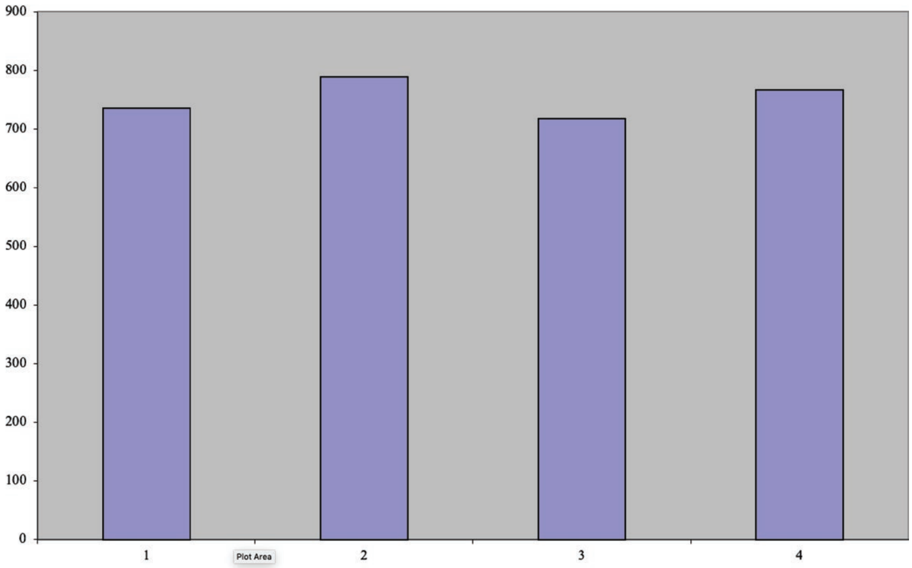

Figure 9.3: Expected distribution of categorical data with a 4-code coding scheme.

To see the difference, imagine a data set that has values up to 4. If this data were continuous, the values might include .01, 3.3, 1.2, 1.27, and so on. But if the data were categorical with only four categories, the values would always be 1, 2, 3, or 4. To see the difference, imagine we put a set of four buckets on the floor and randomly toss coins into them. Each toss is going to go into the 1 bucket, the 2 bucket, the 3 bucket, or the 4 bucket. And if our toss was truly random, after a while the buckets would have about the same number of coins in them with the kind of flat distribution shown in Figure 9.3. Indeed, if one bucket had more or fewer coins than expected, we might suspect that our tosses had not been truly random.

When evaluating the significance of a pattern of coded verbal data, the underlying model of distribution is usually the kind of flat distribution shown in Figure 9.3 where the probability of each category is equal to every other category. The most frequently used significance test for coded verbal data is the χ^2 test. χ^2 is pronounced chi-square. A second and less commonly used test is the multinomial logistic regression with a case effect. Both tests are designed to work with categorical data and both can tell us something about the extent to which the patterns in coded data are surprising. In the rest of this section, we explain how each of these tests work. Then, in the second half of this chapter, we introduce procedures for using them.

How the χ^2 test works

The χ^2 test measures the level of association among the categories of a frequency table like the one shown in Figure 9.4. An association occurs when the values along one of the dimensions generally co-occur with certain values along the other dimension. In Figure 9.4, an association would mean that one or more of the categories in the Frame dimension (*Identity*, *Object*, *Practice*) would co-occur with one or more of the categories of the Alignment dimension (*Professional*, *Social*, *Technical*). That is, they would occur more or less than we would expect. They would be surprising.

	Identity	object	practice	
Professional	9	0	4	13
Social	23	12	14	49
Technical	5	63	80	148
	37	75	98	210

Figure 9.4: Sample frequency table showing the distribution of Frame (Identity, Object, Practice) by Alignment (Professional, Social, Technical).

Whether the distribution of the data in Figure 9.4 is surprising is what a χ^2 test is designed to tell us. It does so by comparing the actual distribution of coded data like what we see in Figure 9.4 with a model of the expected distribution if nothing much was going on, like the one shown in Figure 9.5.

Let's examine the model in Figure 9.5 in more detail. First, you may have noticed that its marginal sums are the same as we saw in the actual data. This is no coincidence. The χ^2 model works by saying, "if we keep the totals in each row and column the same, what would we expect the distribution in the cells to be by random chance?" For instance, overall, *Identity* occurs about 18% of the time in the actual data, *Object* about 36%, and *Practice* about 47%, all adding up to 210 or 100%. For any other row in the table in Figure 9.5, these percentages remain true throughout. That is, in every row, about 18% of the data are *Identity*, 36% *Object*, and 47% *Practice*.

	identity	object	practice	Total
Professional	2	5	6	13
Social	9	18	23	49
Technical	26	53	69	148
Total	37	75	98	210
	18%	36%	47%	100%

Figure 9.5: Expected distribution of the data shown in Figure 9.4.

Another way to understand expected model is to see it in terms of a visualization like the block chart on the bottom in Figure 9.6, where all three planes of the chart have the same shape. Moving from the front with *Professional* to the back with *Technical*, values increase from the left. Everything is proportional.

Compare this with a block chart for the actual data, shown at the top of Figure 9.6. The front plane, the data with *Professional* alignment, shows a shallow U curve. The second plane, the data with *Social* alignment, is also shaped like a U, although a little less shallow. And the back plane, the data with *Technical* alignment, the curve slopes sharply to the right. None of these shapes looks particularly similar to each other.

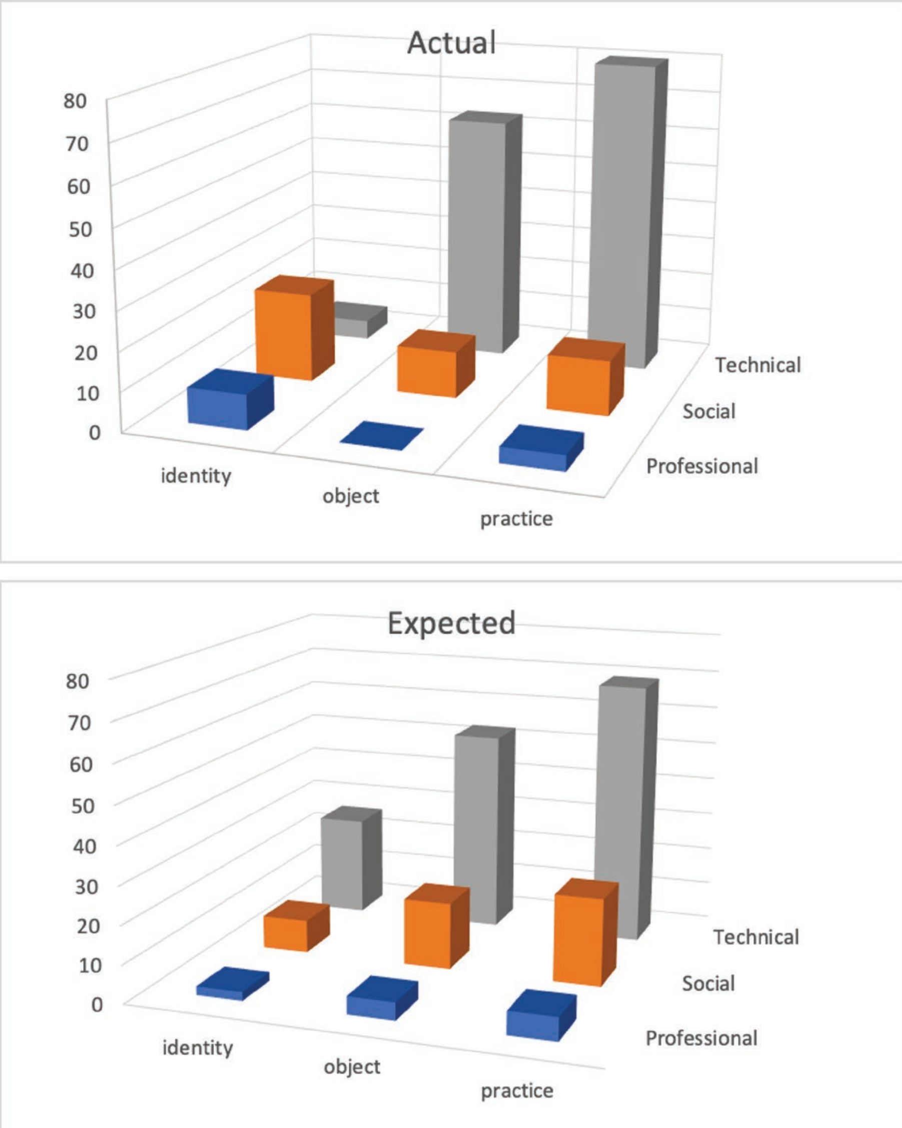

Figure 9.6: Visual representations of the actual (top) and expected (bottom) distributions from Figures 9.3 and 9.4.

The χ^2 test works by comparing these two distributions, one for the actual data and the other for the expected data using the following formula:

$$\chi^2 = \text{sum of} \left[\frac{(O-E)^2}{E} \right]$$

Translated into English, this formula means that χ^2 equals the sum of the squares of the differences between the observed and expected values for each cell in your frequency table, each difference having been divided by the expected value for that cell. The greater the sum of differences between two, the more surprising or statistically significant the result is. This decision-making rule parallels our example with Jenna's morning greeting: the more that her behavior on a given day doesn't fit with our understanding of her usual behavior, the more we find her behavior surprising or significant.

How Multinomial Logistic Regression Works

Multinomial logistic regression works in nearly the opposite way from the χ^2 test. Whereas surprise and significance for the χ^2 test lies in the lack of fit between the actual distribution and the model, for multinomial logistic regression, as we shall see, surprise and significance lies in an increasing fit.

Furthermore, unlike the model used in a χ^2 test, which uses the categorical sums in a frequency table, a multinomial logistic regression uses all of the data points, not just their sums. As shown in Figure 9.7, for example, the model created by a multinomial logistic regression tries to predict what the coding for Alignment would be, given a coding for Frame.

In this way, rather than looking for an association among dimensions as the χ^2 test does, multinomial logistic regression seeks to determine the predictive power of one factor—such as the dimension of Frame—for a dimension—such as Alignment. The first of these is often called the predictor variable and the second the outcome variable.

Unit	Year	User	Content	Frame	Alignment
1	Year1	irunepan	I'm a student of a Biomedical Engineering Master of Barcelona,	Identity	Professional
2	Year1	irunepan	and I'm doing my Master Thesis about virtual endoscopy.	Identity	Professional
3	Year1	irunepan	Specifically I'm trying the endoscopy module of the 3D slicer in the Abdominalatlas2011 data set.	Practice	Technical
4	Year1	irunepan	First I'm testing the navigation mode,	Practice	Technical
5	Year1	irunepan	I create a Fiduacil list and Fly through,	Practice	Technical
6	Year1	irunepan	but I wanted to know if it possible to record	Practice	Technical
7	Year1	irunepan	and make a video of the navigation.	Practice	Technical

Figure 9.7: Data points as modeled by a multinomial logistic regression.

In the verbal data coding dealt with in this book, predictor variables are usually one of two types. The first, as we illustrate with Figure 9.7, is a value on a first coding dimension and would answer the question: given this value, what do we predict would be the value on a second dimension?

The second possibility for a predictor variable is the contrast we have built into our data collection. In the data shown in Figure 9.7, for example, the data have been labeled by the year in which the content was produced. With this data, we could seek to answer the question: given that a piece of data was produced in *Year1*, what do we predict would be its alignment? With a well-fitted regression model, we should be able to predict with better than chance accuracy the answers to questions like these. The ability to make such a prediction would be surprising—and significant.

Like all regressions, multinomial logistic regression works by fitting lines to actual data. In a simple linear regression like that shown in Figure 9.8, a straight line is drawn to minimize the distance between the actual data points shown in blue and the line shown in red. A multinomial logistic regression fits a more complicated line like that shown in Figure 9.9. In this logistic curve, the values are limited to a range between 0 and 1, making it a good model for categorical data.

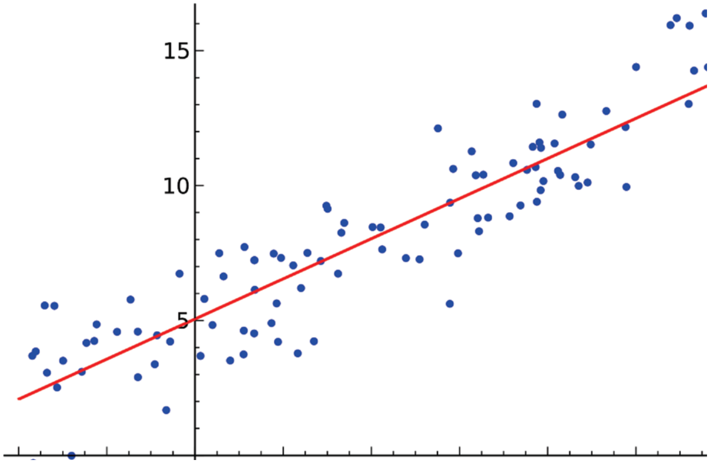

Figure 9.8: Fitting a line in simple linear regression.[9]

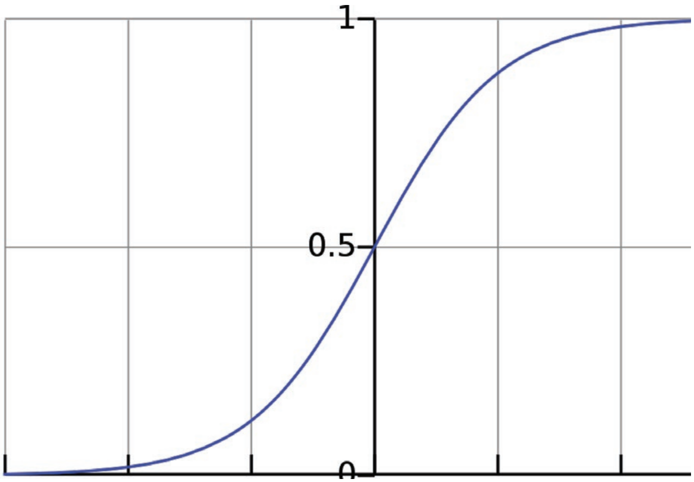

Figure 9.9: A logistic curve used by a multinomial logistic regression.[10]

9 Graphic by Sewaqu and released into the public domain (retrieved from https://commons.wikimedia.org/wiki/File:Linear_regression.svg).

10 Graphic by Qef and released into the public domain Retrieved from https://commons.wikimedia.org/wiki/File:Logistic-curve.svg

As we saw earlier, a χ^2 test works by using probabilities. Probabilities in verbal data analysis can be defined as the frequency of segments in a given category divided by the total number of segments in all categories. So, as illustrated on the right in Figure 9.10, if we have a three-category scheme with equal probabilities applied to nine pieces of data, the probability of the category *Social* is 3 divided by 9 or .33. Probabilities like these are key in a χ^2 test where they are used to model the expected values.

With multinomial logistic regression, the key is the slightly different concept of odds. Odds compare the probability of a category being used to the probability of it not being used. In verbal data, odds can be defined as the frequency of segments in a given category divided by the frequency of segments in all other categories. As illustrated on the left in Figure 9.10, the odds of the category *Social* are 3 divided by 6 or .5. In gambling contexts, this can be expressed as an odds of 2 to 1 against being coded as *Social*.

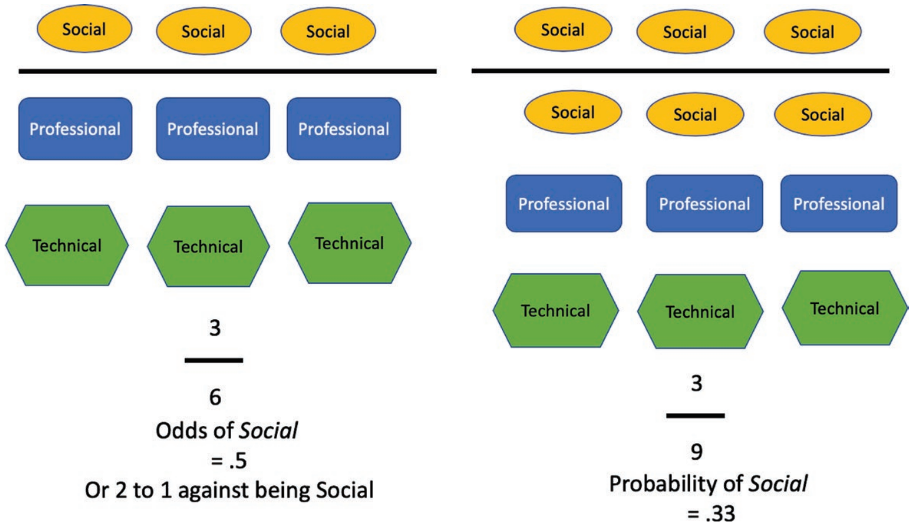

Figure 9.10: The difference between probability and odds. Modeled after Doliner (2014).

A multinomial logistic regression makes a comparison between two models, a baseline model without the value of interest and a model with the value of interest added. Thus, it begins by choosing a baseline category from a categorization scheme. It doesn't matter which category is chosen as the baseline, but the app developed for this book generally chooses the first category it encounters in the worksheet. So for coding the data shown in Figure 9.7, the baseline would be the category *Professional* as it is the first coding category used for Alignment.

Next, the multinomial logistic regression makes a series of comparisons between the odds of each of the other categories in the categorization scheme and the odds of this baseline. To do so, it computes the log odds as the comparison.[11] So with our data, for example, it computes the log odds of being coded as *Social* compared to being coded as *Professional*. Then it will compute the odds of being coded as *Technical* compared to being coded as *Professional*. If the additional information provided by the model with the added category provides a better fit than the baseline model, then we find the category to be significant predictor.

Note here that, as we said earlier, significant doesn't mean important. A variable might be significant in improving a fit between the line and the data by making a relatively small but stable improvement. In other words, our chance of being correct might be better than chance with this additional information, but it might still be relatively poor. So with this and all significance testing judging whether a significant result is an important result requires assessing the patterns you discovered using the techniques outlined in chapters 6 and 7.

Assessing Your Data

As we have just seen, all significance testing builds one or more models against which to evaluate the distribution of our actual data. To better understand which significance test (if any) to use with a given data set, we need to review the structure of our data set and then check to see whether and how it is appropriate for the significance testing we describe in this chapter.

11 Log odds are the natural logarithm of the odds ratio between the category of interest and the baseline category. If you remember from high school what a logarithm is, that's fine, but otherwise, don't worry about it.

Data Points

We begin by counting the total number of data points in our data set. We can count up the number of data points or, if we have built a frequency table, we can find the total in the bottom right-hand of the marginals. In the frequency table in Figure 9.4, for example, the total number of data points is 210. Most statistical tests are more accurate with more data points. If you have just a few data points, you may not be able to evaluate significance.

Independent Cases

Next, we count the number of independent cases in our data set. If you have followed earlier chapters, you may have put each independent case—each interview, each text, etc.—in a separate worksheet or a separate document, even though you may now have combined them to do statistical analysis. For our combined data, as shown in Figure 9.7, the cases come from different users like irunepan. In fact, our data set includes verbal data from 11 such users, or 11 cases. Some statistical tests are designed to take into account the way that a data set is structured by cases.

Keep in mind that cases should be more or less independent from one another. In our example, independence means that what irunepan says is not influenced in any direct way by what another user says. If speakers are in the same conversation, their contributions are likely to be influenced by one another and probably should not be considered separate cases. But if their contributions are from separate interviews, then they could be considered independent. In your data, you may find that you have multiple independent cases, or, if you are studying one focus group, for example, you may only have one case.

Built-In Contrast

We may have one or more built-in contrasts in the design of our data set. A built-in contrast in a data set means that we have deliberately sampled our data from different areas in the universe of our phenomenon. Perhaps we gathered transcriptions of both Design and Managerial meetings. Perhaps we have essays from students who did above and below average in their composition course. Perhaps we have scraped web texts from political discussions and from discussions about gardening. In any of these cases, we have a built-in contrast that needs to be taken into account in choosing a statistical test. In our sample data, we have data from four years so we could use Year as a possible built-in contrast.

Coding Dimensions

Much of the data we analyze has only one coding dimension. That is, we have used only one coding scheme with our data set. But, as we see in Figure 9.7, it is not uncommon to use two different dimensions such as Frame and Alignment. Knowing how many dimensions we have is important to deciding which kind of significance testing to use.

Choosing Your Significance Test(s)

Some Preliminaries

The analytic techniques introduced in this book are primarily focused on producing a descriptive analysis of verbal data. That is, they are designed to describe the data set you have collected. Some researches want to take their analysis an additional step to produce an analysis that is inferential. An inferential analysis uses a description of a data set to make inferences about the larger population from which the data set was taken. In our description of Design and Managerial meetings, we have focused largely on trying to describe what was going on in those meetings in terms of speaker participation; this is a descriptive purpose.

If we wanted to draw inferences about other meetings, we would need to consider the kind of sample we had drawn from the larger population of possible meetings. In general, inferences are only valid if the sample is drawn using random sampling, a sampling method we reviewed in Chapter 2. So if we wanted to make inferences about other Managerial and Design meetings, we would have had to collect and analyze a random sample of such meetings. In many cases with verbal data, such random sampling is neither possible nor desirable.

All statistical methods do require, however, that you have enough data. If your frequency table is sparse, the statistics will yield results that are not to be trusted. In general, a sparse frequency table is one where:

- One or more of the cells is empty.
- More than about 20% of the cells have values less than 5.

A sparse frequency table indicates that you have one or more coding categories that were not often used in coding your data. If this is the case, you may be able to combine infrequent categories into some more general category—combining some less interesting categories into a larger Miscellaneous category for example—but make sure that you maintain the categories that motivated your study in the first place.

If combining coding categories will not be possible, then you simply may not have enough data to use significance testing. Going back to our analogy with morning greetings, you may not have encountered Jenna on enough mornings to enable you to say whether her failure to greet you this morning is surprising. This doesn't mean that you cannot describe what you have seen her do, only that you cannot say if it is surprising.

Choosing Your Test

The decision about which test you use to evaluate significance depends on the structure of your data, as shown in Figure 9.11. Usually, we recommend that you always run some kind of χ^2 test first. As we shall see, such a test tells you a great deal more about your data than just its significance.

In many cases, we also recommend that you go on to run a multinomial logistic regression and compare the results. As we discussed earlier, a multinomial logistic regression is an inferential test that will tell you something about your chances of predicting a value on a second, or outcome variable, given a value on a first, or predictor variable. This is not something that a χ^2 test can do.

But there is a further and perhaps more important reason to run a multinomial logistic regression in addition to a χ^2 test. The χ^2 test assumes that each data point in your frequency table is independent. This is an assumption that is often violated with verbal data. If your data segments combine to make up continuous discourse, they are not going to be independent from one another. Furthermore, if two segments come from texts that are written by the same author, they may well not be independent. Even when segments come from essays written by students enrolled in the same writing course, they might not be independent.

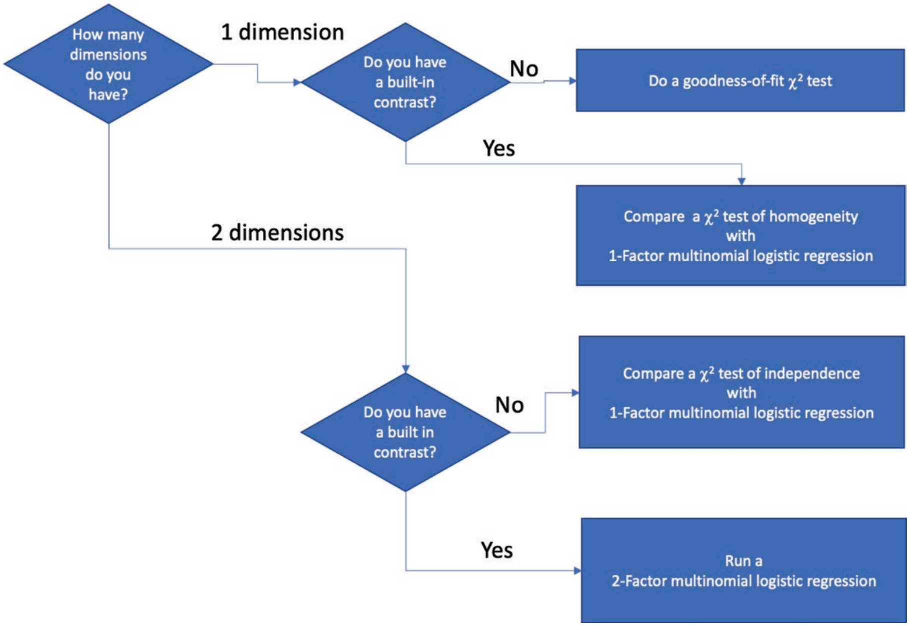

Figure 9.11: A decision tree for choosing significance test(s) for coded verbal data.

As these example are meant to indicate, it can be difficult to tell in advance whether a data set violates the requirement of independence. Sometimes the only way to tell is to run the χ^2 test. If the results suggest you need to get a more valid measure of significance, go on to multinomial logistic regression. Multinomial logistic regression can take into account the interdependency among the data points within each case, factor it out, and establish whether the remaining variation is still surprising, still significant.

Multinomial logistic regression works best when there are a lot of data points within each case. If there are few data points or something else inappropriate about the model, the regression may produce unstable results. To make sure that your multinomial logistic regression is producing stable results, we suggest you run a multinomial logistic regression at least twice. If the significance results are the same, then you can feel confident in them. If the results remain unstable, you may have too few data points for significance testing. If nevertheless you believe that you have enough data, we recommend that you consult with a statistician.

Exercise 9.2 Test Your Understanding

Use the decision tree in Figure 9.11 to make a plan for evaluating the significance of the following data sets.

1. You have gathered and coded essays from six classes, three of which were taught using the usual curriculum and three of which were taught with a new curriculum. Your coding scheme was for Engagement.
2. You have examined published journal articles from biology, physics, and medicine all dealing with the same phenomenon. You have coded the citations for Function and for Source.
3. You have coded a set of published articles for Genre.
4. You are trying to understand the patterns of interaction among students and their teachers in your program. You have coded classroom transcripts for Speaker and Contribution.

For Discussion: What is the impact on evaluating significance of adding a second dimension to your coding? What is the impact of including a built-in contrast?

Additional Notes on Procedures for Significance Testing

Some additional notes on our procedures for the statistical analyses. First, all of the procedures for χ^2 analysis start with the assumption that you have created a frequency table for your data using methods introduced in Chapter 6. If you don't yet have a frequency table, you may want to turn to this chapter.

Second, we have provided procedures for doing all of the statistical analysis using both online apps and, for the χ^2 analysis, using Excel. We have not provided any procedures using MAXQDA because the standard package does not support significance testing.

Third, multinomial logistic regression with a case effect is a relatively recent development in statistical methods and its application can be tricky. The online app for conducting a multinomial logistic regression we direct you to has been developed for us by Dr. Emily Griffith of North Carolina State University. We are grateful to Dr. Griffith for this contribution to the analysis of coded verbal data.

Memo 9.1: Plan for Significance Testing

Record your assessment of your dataset, its size, contrast, cases, and coding dimensions. What significance test(s) do you plan to use and why?

The χ^2 Test of Goodness of Fit

Although we have emphasized the value of using a built-in contrast to code your data, you may find that you want to look at a set of data that has been coded along one dimension without contrast to ask the question:

How likely is it that my segments have been coded randomly?

Answering such a question can assure you—and your readers—that the coders were coding by something more than chance.

Calculating the χ^2 Test of Goodness of Fit

The six steps shown in Figure 9.12 and discussed in Excel Procedure 9.1 will take you through the χ^2 test for goodness of fit. You can download a template for your calculations at https://wac.colostate.edu/books/practice/codingstreams/. Directions for an app to do this calculation are provided in Procedure 9.1.

Step 1. Observed	Identity	Object	Practice	Total
Observed	37	75	98	210
Total	37	75	98	210
Step 2. Expected	identity	object	practice	Total
Expected	70	70	70	210
	70	70	70	210
Step 3. O-E	identity	object	practice	Total
0	-33	5	28	0
	-33	5	28	0
Step 4. $(O\text{-}E)^2/E$	identity	object	practice	Total
0	15.56	0.36	11.20	27.11
	15.56	0.36	11.20	**27.11**
Step 5. DF				
Number of categories	3			
Number of categories - 1	2			
df	2			
Step 6. Table Lookup	http://www.z-table.com/chi-square-table.html			
The sume of Chi Square equals	27.11	df & probability	df = 2 p<.005	

Figure 9.12: The six-step calculation of the goodness-of-fit χ^2 test.

Interpreting the Results of the χ^2 Test of Goodness of Fit

The final step in the computation of the Goodness of Fit χ^2 test—looking up the values on the table—tells you what the chances are that the distribution of your data over categories is surprising. Generally speaking, we think of any probability of less than .01 as significant, less than .001 as highly significant, and less than .05 as somewhat significant. See Excel Procedure 9.1 and Procedure 9.1.

Such numbers do not tell you how your observed data is departing from the expected model in such a fashion as to lead to a significant outcome for the χ^2 test. For this, we need to compare the observed values with the expected distribution. Then we will be able to see that some observed values lay closer to their expected counterparts and some are more distant. The greater the difference between the pairs, the more they contribute to a large sum of χ^2 value.

Thus, interpreting a significant χ^2 result involves pinpointing the greatest differences in the values making up the χ^2 value. To see these, you must return to examine the table in Step 4 where you computed (O-E)2/E for each cell. Since these are the numbers that you added up to get the final sum of χ^2, extremely high values tell you what is so unexpected in the distribution of your data. In Step 4 in Figure 9.12, for example, we see that almost all of the value for the significant sum of χ^2 comes from the values for *Identity* (15.56) and *Practice* (11.20). The value for *Object* is nearly zero (.36).

Having pinpointed the cells that make the greatest contribution to your significant χ^2 value, you next try to understand what makes the observed values in these cells so different from the expected values. You can do this by looking at differences between the observed and expected values, comparing Steps 1 and 2. For example, looking at the tables in Figure 9.12, we see that the observed value for *Identity* is much lower than expected and the observed value for *Practice* is much higher than expected. This means that our coders have been using the code *Identity* much less than we would have expected had they been coding by chance, and they are using the code *Practice* much more than we would have expected by chance.

Excel Procedure 9.1: Calculating a Goodness of Fit χ^2 Test in Excel

https://goo.gl/Hx5Ay7

1. Create a frequency table holding the categories of your coding scheme, as shown in Step 1 of Figure 9.12. Make sure to include the marginal sums.
2. Create 3 more tables in the same way. Label them as shown in Figure 9.12. You may also use the Excel template at https://wac.colostate.edu/books/practice/codingstreams/ that will automatically do the calculations for steps 3-5.
3. For Step 2, Expected, divide the total number of segments by the number of categories and put the result in each cell of this table.
4. For Step 3, O-E, subtract the expected frequencies from the observed values and put the result in each of the cells.

In Figure 9.12, we used a formula like the following to accomplish this calculation in each cell:

```
=B18-B22
```

5. In Step 4, $(O-E)^2/E$, for each cell, square the value from Step 3 and divide the result by the expected value from Step 2.

In Figure 9.12, we used a formula like the following to accomplish this calculation in each cell:

```
=(B26*B26)/B22
```

The sum of χ^2 will be the grand total in the table.

6. Calculate the degrees of freedom by subtracting 1 from the number of categories in your coding scheme.
7. Use a chi-square calculator like the one at https://www.socscistatistics.com/pvalues/chidistribution.aspx to calculate the p-value for your sum of chi-squares with your degrees of freedom.

Procedure 9.1: Calculating a Goodness of Fit χ^2 Test with an Online App

https://goo.gl/Hx5Ay7

1. Create a frequency table holding the categories of your coding scheme. Make sure to include the marginal sums.
2. Create a table in the same way to hold your expected values. To calculate your expected values, divide the total number of segments by the number of categories.
3. Go to the online app at http://vassarstats.net/csfit.HTML and enter the data for the observed and expected values as shown in Figure 9.13.
4. Click **Calculate**.
5. The app will return the degrees of freedom, the sum of χ^2 and the probability value as shown in Figure 9.14.

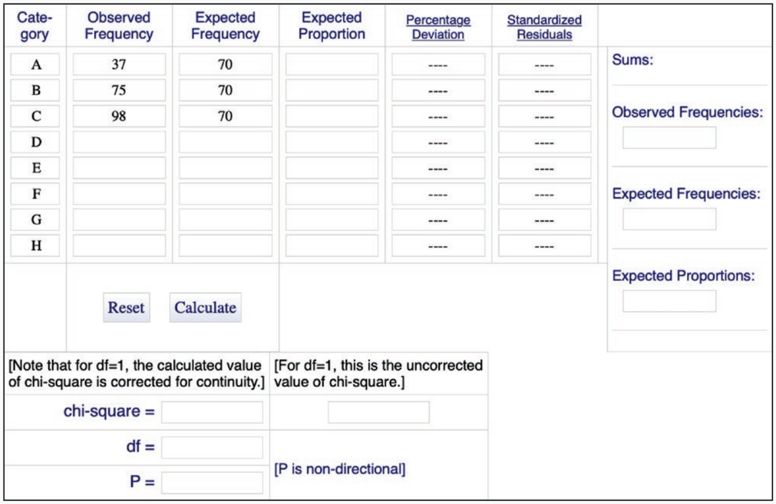

Figure 9.13: Calculating the goodness-of-fit χ^2 test online.

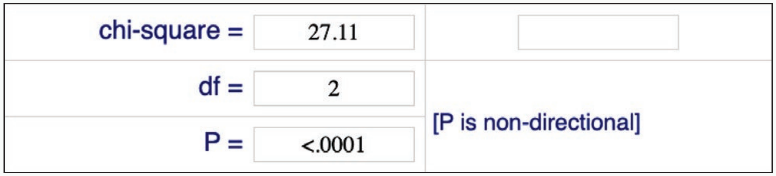

Figure 9.14: Results of the online calculation for the goodness-of-fit χ^2 test (http://vassarstats.net/csfit.HTML).

Overall then, the Goodness of Fit χ^2 test can give us a way to see the coding preferences that our coders used in coding the data. Unfortunately, the results of our example χ^2 test cannot take us much further than this because the sum of χ^2 appears inflated. This suggests that the test has not given us a valid measure of significance. As discussed earlier, inflated sums of χ^2 can result from a lack of independence among the data segments.

So while the χ^2 test gives us a way of seeing what is going on in our coders' use of the coding categories, if you find a lack of independence, the results cannot be relied on as a measure of significance. If the sum of χ^2 appears inflated, you should not infer anything about the coding patterns of the larger population from which your data set was drawn. In reporting an analysis that yields inflated sums of χ^2, then, you can point out to readers what the distribution of coding preferences was, but you should not report the results of the χ^2 test.

The χ^2 Test of Homogeneity

The χ^2 test of homogeneity works with data coded along one dimension which has a built-in contrast. It is a way of answering the question:

> How likely is it that two or more groups in my study share the same distribution across the categories in my coding scheme?

Answering such a question can help you to evaluate the significance of differences across your built-in contrast. Such a test is often called a test of homogeneity because we are asking whether the distribution in one sample of data is similar to—or homogeneous with—the distribution in another sample.

Computing a χ^2 test of homogeneity

The six steps shown in Figure 9.16 and discussed in Excel Procedure 9.2 will take you through the χ^2 test. You can download a template for your calculations at https://wac.colostate.edu/books/practice/codingstreams/. Directions for an app to do this calculation are provided in Procedure 9.2.

Excel Procedure 9.2: Calculating a χ^2 Test of Homogeneity in Excel

https://goo.gl/Hx5Ay7

1. Create a frequency table holding the categories of your coding scheme and the values of your contrast as shown in Step 1 of Figure 9.15. Make sure to include the marginal sums.
2. Create 3 more tables in the same way. Label them as shown in Figure 9.16. You may also use the Excel template available at https://wac.colostate.edu/books/practice/codingstreams/ that will automatically do the calculations for steps 3-5.

In Figure 9.15, we used the following formula to accomplish this calculation in each cell:

=($E15*B$19)/$E$19

3. For Step 3, O-E, subtract the expected frequencies from the observed values and put the result in each of the cells.

Step 1. Observed	Identity	Object	Practice	Total
Year1	25	45	53	123
Year2	8	25	23	56
Year3	0	0	2	2
Year4	4	5	20	29
Total	37	75	98	210
Step 2. Expected	Identity	Object	Practice	Total
Year1	22	44	57	123
Year2	10	20	26	56
Year3	0	1	1	2
Year4	5	10	14	29
Total	37	75	98	210
Step 3. O-E	Identity	Object	Practice	Total
Year1	3.33	1.07	-4.40	0.00
Year2	-1.87	5.00	-3.13	0.00
Year3	-0.35	-0.71	1.07	0.00
Year4	-1.11	-5.36	6.47	0.00
Total	0.00	0.00	0.00	0.00
Step 4. (O-E)^2/E	Identity	Object	Practice	Total
Year1	0.51	0.03	0.34	0.87
Year2	0.35	1.25	0.38	1.98
Year3	0.35	0.71	1.22	2.29
Year4	0.24	2.77	3.09	6.10
Total	1.46	4.76	5.02	11.24
Step 5. DF				
Number of rows -1	4			
Number of categories	3			
Number of rows -1	3			
Number of categories - 1	2			
df	6			
Step 6. Table Lookup	http://www.z-table.com/chi-square-table.html			
The sume of Chi Square equals	11.24	df & probability	df=6. p<.10	

Figure 9.15: Calculating χ^2 of homogeneity.

Continued . . .

Excel Procedure 9.2: Calculating a χ^2 Test of Homogeneity in Excel (continued)

https://goo.gl/Hx5Ay7

In Figure 9.15, we used a formula like the following to accomplish this calculation in each cell:

=($E15*B$19)/$E$19

4. For Step 3, O-E, subtract the expected frequencies from the observed values and put the result in each of the cells.

In Figure 9.15, we used a formula like the following to accomplish this calculation in each cell:

=B15-B22

5. In Step 4, (O-E)2/E, for each cell, square the value from Step 3 and divide the result by the expected value from Step 2.

In Figure 9.15, we used a formula like the following to accomplish this calculation in each cell:

=(B29*B29)/B22

The sum of χ^2 will be the grand total for the table.

6. Calculate the degrees of freedom by subtracting 1 from the number of rows in your contrast and 1 from the number categories in your coding scheme. Multiple these 2 numbers together

For Figure 9.15, we multipled together (4-1) and (3-1) to get degrees of freedom equal to 6.

7. Use a chi-square calculator like the one at https://www.socscistatistics.com/pvalues/chidistribution.aspx to calculate the p-value for your sum of chi-squares with your degrees of freedom.

Procedure 9.2: Calculating a χ^2 Test of Homogeneity with an Online App

https://goo.gl/Hx5Ay7

1. Go to the online app at http://turner.faculty.swau.edu/mathematics/math241/materials/contablecalc/ and enter the number of rows and columns in the opening page of the app. Press **Continue**.
2. Enter a title, labels and data for your frequency table as shown in Figure 9.16. Leave the option to display individual χ^2 values checked and press **Compute**.

The app will return a frequency table in which each cell holds the observed value, followed by the expected value (in italics), and the individual ² values as shown in Figure 9.17. The sum of χ^2, the degrees of freedom, and the probability value can be found below the table.

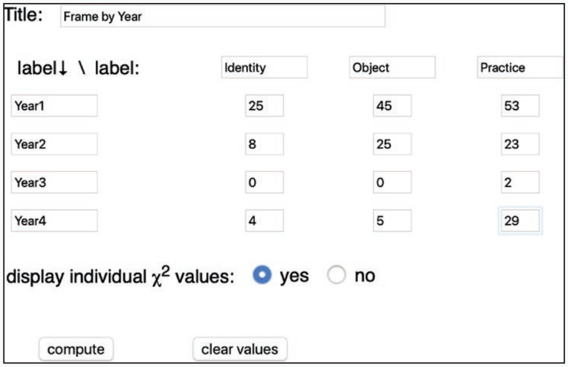

Figure 9.16: Entering data for the online app for the χ^2 test of homogeneity.

	Frame by Year			
	Identity	**Object**	**Practice**	
Year1	25 *20.78* (0.86)	45 *42.12* (0.20)	53 *60.10* (0.84)	**123**
Year2	8 *9.46* (0.23)	25 *19.18* (1.77)	23 *27.36* (0.70)	**56**
Year3	0 *0.34* (0.34)	0 *0.68* (0.68)	2 *0.98* (1.07)	**2**
Year4	4 *6.42* (0.91)	5 *13.01* (4.93)	29 *18.57* (5.86)	**38**
	37	**75**	**107**	**219**

χ^2 = 18.383, df = 6, χ^2/df = 3.06 , P(χ^2 > 18.383) = 0.0053

warning: some observed or expected frequencies are less than 5; thus the Central Limit Theorem may not apply and the resultant χ^2 may be invalid

expected values are displayed in *italics*
individual χ^2values are displayed in (parentheses)

Figure 9.17: Results of the calculations for the online app for the χ^2 test of homogeneity (http://turner.faculty.swau.edu/mathematics/math241/materials/contablecalc/).

Exercise 9.3 Try It Out

Perform a χ^2 test of homogeneity for the data in Figure 9.18 (and available at https://wac.colostate.edu/books/practice/codingstreams/).

	Ed	Cheryl	John	
Meeting 1	70	128	115	313
Meeting 2	104	106	104	314
	174	234	219	627

Figure 9.18: Observed frequency distribution of speakers in meetings 1 and 2.

For Discussion: What do the results tell you about the likelihood that Meetings 1 and 2 share the same distribution of speakers?

Interpreting the Results of the χ^2 test of Homogeneity

The final step in the computation of the χ^2 test of homogeneity tells you what the chances are that the distribution of your data over categories and across contrast is surprisingly different or not homogeneous. Interpreting a significant χ^2 result of $p < .05$ or $p < .01$ involves pinpointing the greatest differences in the values making up the χ^2 value which can be found in the table in Step 4 of Figure 9.15 or the third row in the cells of Figure 9.17 High values can tell you what is so unexpected in the distribution of your data; low values tell you what is not surprising.

Our calculations shown in Figure 9.15 suggest that there is nothing surprising about the way the distribution of data into our coding categories changes by year. The total, 11.24 with df equal to 6 show a probably of less that 1 in 10 ($p < .10$), a result that well can occur by chance. So the answer to the question with which we opened, How likely is it that two or more groups in my study share the same distribution across the categories in my coding scheme? appears to be "pretty likely."

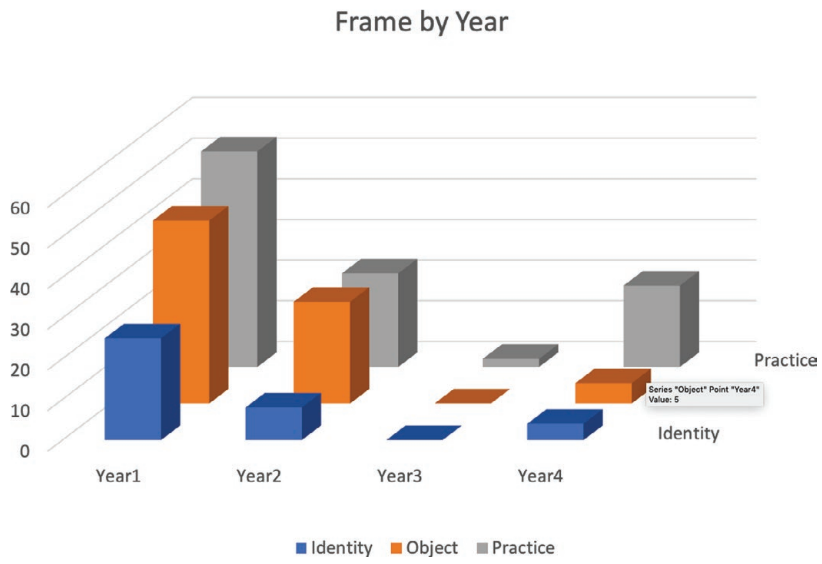

Figure 9.19: A block chart of the actual data from Figure 9.16.

A look at the individual χ^2 values shown in Step 4 of Figure 9.15 confirms that none of the values look surprisingly large. And the block chart of the same data shown in Figure 9.19 also shows homogeneity with each year's data being lower for *Identity*, medium sized for *Object* and, for the most part, highest for *Practice*. While it is the case that *Practice* shows up proportionately less in *Year2*, the difference is not large enough to reach significance.

Part of the problem with the data in our example χ^2 is that almost all of the observed values for *Year3* are very small; two are 0 and one is just 2. As mentioned earlier, scarce data can compromise the validity of a χ^2 analysis. In this case, it might be worthwhile to combine the *Year3* and *Year4* into a category like *After Year2* which we have done in the analysis shown in Figure 9.20.

The results for the analysis with combined categories yields a sum of χ^2 of 10.50. This value crosses the threshold for significance of the .05 level with df equal to 4. A look at the values in Step 4 pinpoints the After *Year2* values for *Object* and *Practice* make the largest contributions. And a comparison of the observed and expected values in Steps 1 and 2 suggests that surprise is coming from an unexpectedly low number of *Object* codes and an unexpectedly high number of *Practice* codes.

Step 1. Observed	Identity	Object	Practice	Total
Year1	25	45	53	123
Year2	8	25	23	56
After Year 2	4	5	22	31
Total	37	75	98	210

Step 2. Expected	Identity	Object	Practice	Total
Year1	22	44	57	123
Year2	10	20	26	56
After Year 2	5	11	14	31
Total	37	75	98	210

Step 3. O-E	Identity	Object	Practice	Total
Year1	3.33	1.07	-4.40	0.00
Year2	-1.87	5.00	-3.13	0.00
After Year 2	-1.46	-6.07	7.53	0.00
Total	0.00	0.00	0.00	0.00

Step 4. $(O\text{-}E)^2/E$	Identity	Object	Practice	Total
Year1	0.51	0.03	0.34	0.87
Year2	0.35	1.25	0.38	1.98
After Year 2	0.39	3.33	3.92	7.64
Total	1.26	4.61	4.64	10.50

Step 5. DF	
Number of rows -1	3
Number of categories	3
Number of rows -1	2
Number of categories - 1	2
df	4

Step 6. Table Lookup	http://www.z-table.com/chi-square-table.html		
The sume of Chi Square equals	10.50	df & probability	df=4. p<.05

Figure 9.20: A χ^2 analysis with data combined over scarce categories.

The fact is, however, that the surprise arises only in the *After Year2* category for which we have relatively little data. As a consequence, we would be somewhat conservative in making claims about the way that the data changes after *Year2*. At best, these results suggest that we should go on to do a One-Factor Multinomial Logistic Regression.

So while the χ^2 test of homogeneity gives us a way of seeing what is going on across our built-in contrast, scarce data may mean the results cannot be relied on as a measure of significance. If you have any cell values of 0 or many cell values of less than 5, you should consider combining categories.

In addition, inflated sum of χ^2 may affect a χ^2 test of homogeneity just as it did with the χ^2 test for goodness of fit. For this reason, we always recommend that you go on to do a One-Factor Multinomial Logistic Regression to confirm any significant results from a χ^2 analysis.

The χ^2 Test of Independence

The χ^2 test of independence works with data coded along two dimension without a built-in contrast. It is a way of answering the question, *"How likely is it that two dimensions in my study are independent of one another?"* Answering this question can help you to see whether there is a relationship between the way your data is coded along one dimension with the way it is coded along a second dimension.

Computing a χ^2 test of Independence

The six steps shown in Figure 9.21 and discussed in Excel Procedure 9.3 will take you through the χ^2 test of independence. You can download a template for your calculations at https://wac.colostate.edu/books/practice/codingstreams/. Directions for an app to do this calculation are provided in Procedure 9.3.

Excel Procedure 9.3: Calculating a χ^2 Test of Independence in Excel

https://goo.gl/Hx5Ay7

1. Create a frequency table holding the categories of your first and second coding schemes shown in Step 1 of Figure 9.21. Make sure to include the marginal sums.
2. Create three more tables in the same way. Label them as shown in Figure 9.21. You may also use the Excel template available at https://wac.colostate.edu/books/practice/codingstreams/ that will automatically do the calculations for steps 3-5.
3. For Step 2, Expected, for each cell, multiple the row total by its column total and then divide the result by the table's grand total.

In Figure 9.21, we used the following formula to accomplish this calculation in each cell:

=($E19*B$22)/$E$22

4. For Step 3, O-E, subtract the expected frequencies from the observed values and put the result in each of the cells.

Step 1. Observed	Identity	Object	Practice	Total
Professional	9	0	4	13
Social	23	12	14	49
Technical	5	63	80	148
Total	37	75	98	210
Step 2. Expected	identity	object	practice	Total
Professional	2	5	6	13
Social	9	18	23	49
Technical	26	53	69	148
	37	75	98	210
Step 3. O-E	identity	object	practice	Total
Professional	6.71	-4.64	-2.07	0.00
Social	14.37	-5.50	-8.87	0.00
Technical	-21.08	10.14	10.93	0.00
	0	0	0	0
Step 4. $(O-E)^2/E$	identity	object	practice	Total
Professional	19.65	4.64	0.70	25.00
Social	23.91	1.73	3.44	29.07
Technical	17.03	1.95	1.73	20.71
	60.60	8.32	5.87	**74.79**
Step 5. DF				
Number of rows	3			
Number of columns	3			
Number of rows -1	2			
Number of columns - 1	2			
df	4			
Step 6. Table Lookup	http://www.z-table.com/chi-square-table.html			
The sume of Chi Square equals	74.79	df & probability	df=4. p<.005	

Figure 9.21: Results of a χ^2 test of independence.

Continued . . .

Excel Procedure 9.3: Calculating a χ^2 Test of Independence in Excel (continued)

https://goo.gl/Hx5Ay7

In Figure 9.21, we used the following formula to accomplish this calculation in each cell:

=B19-B25

5. In Step 4, $(O-E)^2/E$, for each cell, square the value from Step 3 and divide the result by the expected value from Step 2.

In Figure 9.21, we used the following formula to accomplish this calculation in each cell:

=(B31*B31)/B25

The sum of χ^2 will be the grand total for the table.

6. Calculate the degrees of freedom by subtracting 1 from the number of rows in your contrast and 1 from the number categories in your coding scheme. Multiple these 2 numbers together

For Figure 9.21, we multipled together (3-1) and (3-1) to get degrees of freedom equal to 4.

7. Use a chi-square calculator like the one at https://www.socscistatistics.com/pvalues/chidistribution.aspx to calculate the p-value for your sum of chi-squares with your degrees of freedom.

Procedure 9.3: Calculating a χ^2 Test of Independence with an Online App

https://goo.gl/Hx5Ay7

1. Go to the online app at http://turner.faculty.swau.edu/mathematics/math241/materials/contablecalc/ and enter the number of rows and columns in the opening page of the app. Press **Continue**.
2. Enter a title, labels and data for your frequency table. Leave the option to display individual χ^2 values checked and press **Compute**.

The app will return a frequency table like that shown in Figure 9.22 in which each cell holds the observed value, followed by the expected value (in italics), and the individual χ^2 values. The sum of χ^2, the degrees of freedom, and the probability value can be found below the table.

	Frame x Alignment			
	Identity	**Object**	**Practice**	
Professional	9 *2.29* (19.65)	0 *4.64* (4.64)	4 *6.07* (0.70)	**13**
Social	23 *8.63* (23.91)	12 *17.50* (1.73)	14 *22.87* (3.44)	**49**
Technical	5 *26.08* (17.03)	63 *52.86* (1.95)	80 *69.07* (1.73)	**148**
	37	**75**	**98**	**210**

χ^2 = 74.787, df = 4, χ^2/df = 18.70 , P(χ^2 > 74.787) = 0.0000

warning: some observed or expected frequencies are less than 5; thus the Central Limit Theorem may not apply and the resultant χ^2 may be invalid

Figure 9.22: Results of the calculations for the online app for the χ^2 test of independence (http://turner.faculty.swau.edu/mathematics/math241/materials/contablecalc/).

Interpreting the Results of a χ^2 test of Independence

The final step in the computation of the χ^2 test of independence tells you the chances that the two dimensions of coding are associated with one another. That is, to what extent will values on the first dimension co-occur with values on a second dimension.

In the example shown in Figures 9.21 and 9.22, we see a very high sum of χ^2 (74.79 with 4 degrees of freedom) which could suggest that there is a very strong relationship between *Frame* and *Alignment*. A χ^2 calculator (https://www.socscistatistics.com/pvalues/chidistribution.aspx) shows that this is highly surprising.

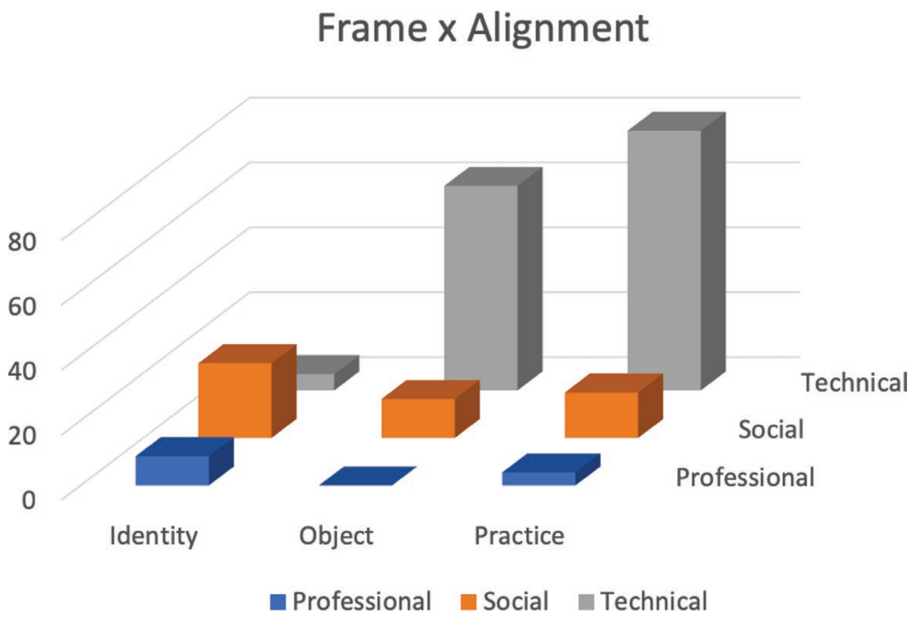

Figure 9.23: Block chart of the observed data from Figure 9.20.

A look at the block chart for the observed data in Figure 9.23 provides more detail about these surprises. In general, the observed values of *Technical* varies considerably depending on the *Frame* used. With *Object* and *Practice*, it is the most common category, but with *Identity*, it is the least common category. In fact, as we see in Step 4 of Figure 9.21, the individual χ^2 value for *Technical* with *Identity* is a high 17.03. The values of the other categories for *Identity* are also matters of surprise with high individual χ^2 values (19.65 for *Professional* and 23.91 for *Social*).

When we use Steps 1 and 2 to compare the observed and expected values in these three cells we can see that *Professional* and *Social* are both higher than expected (9 vs. 2 for *Professional* and 23 vs. 9 for *Social*) while *Technical* is lower than expected (5 vs. 26). Clearly, there is a surprising relationship between these two dimensions.

Unfortunately, because of the inflated sum of χ^2, we cannot draw any conclusions about the significance of this relationship. As discussed earlier, a high sum of χ^2 is often the result of a lack of independence among the data points. To get a clearer picture of what is going on here, we would turn to a One-Factor Multinomial Logistic Regression.

The χ^2 test of independence gives us a way of seeing a relationship between two coding dimensions. But just as with our earlier χ^2 tests, scarce data can compromise it as a measure of significance. If you have any cell values of 0 or many cell values of less than 5, you should consider combining categories.

In addition, as we have just seen, inflated sum of χ^2 may affect a χ^2 test of independence. Thus, we always recommend that you go on to do a One-Factor Multinomial Logistic Regression to confirm any significance results from a χ^2 analysis.

Memo 9.2: Your χ^2 Analysis

Record the frequency table that you used as input to the χ^2 test. Record the results: your degrees of freedom, the sum of χ^2 and the probability level. Check for an inflated χ^2 value that would limit your ability to draw conclusions about significance.

If significant, use the individual χ^2 values to determine which values are making the largest contribution. For these cells, compare the observed frequencies to the expected frequencies. Put into plain language what these comparisons mean in terms of what is surprising in your data.

One-Factor Multinomial Logistic Regression

One-factor multinomial logistic regression is an analytic tool designed to examine the impact of a predictor variable on an outcome variable. With coded verbal data, the outcome variable is always the coding along a given dimension.

The predictor variable may be the values on the contrast built into the design of the data. In this case, it is designed to answer the question:

> What is the likelihood of a given code given a value on the built-in contrast?

In this first form, we recommend you use this test as a follow-up analysis to the χ^2 test of homogeneity.

The predictor variable may also be the values on a second coding dimension. In this case, one-factor multinomial logistic regression is designed to answer the question:

> What is the likelihood of a given code along a second coding dimension given a value on a first coding dimension?

In this second form, we recommend you use this test as a follow-up analysis to the χ^2 test of independence.

Running a One-Factor Multinomial Logistic Regression

The one-factor multinomial logistic regression works with the individual data points in your data set. Prepare your data for the app as detailed in Excel Procedure 9.4 paying particular attention to the labels of your columns. Then run the app using Procedure 9.5.

Procedure 9.4: Preparing the Data for a One-Factor Multinomial Regression

https://goo.gl/Hx5Ay7

1. Combine the coded data from individual data worksheets into a single worksheet, keeping track of which data comes from which worksheet.
2. Creating a new column in the combined worksheet. Label it as **Case**.
3. In this column, next to each segment, enter the name of the data worksheet from which the segment was copied.
4. Determine which column is to be used as the predictive factor and change its heading to **Factor1**.

	A	B	C	D	E
1	**Unit**	**Year**	**Case**	**Factor1**	**Dimension**
2	1	2011	irunepan	Identity	Professional
3	2	2011	irunepan	Identity	Professional
4	3	2011	irunepan	Practice	Technical
5	4	2011	irunepan	Practice	Technical
6	5	2011	irunepan	Practice	Technical
7	6	2011	irunepan	Practice	Technical
8	7	2011	irunepan	Practice	Technical
9	8	2011	irunepan	Identity	Technical
10	9	2011	irunepan	Practice	Technical
11	10	2011	pieper	Practice	Technical
12	11	2011	pieper	Object	Technical
13	12	2011	pieper	Practice	Technical
14	13	2011	pieper	Object	Social
15	14	2011	haehn	Object	Technical

Figure 9.24: Worksheet arrangement for a one-factor multinomial logistic regression.

The Factor1 column may be the one holding the values of your built-in contrast or it may be the one holding the values of your first coding dimension, depending on which variety of the one-factor multinomial logistic regression you are preparing for.

5. Determine which column is to be used as the outcome dimension and label it **Dimension**.
6. Delete the column holding the actual verbal data.

The online app will not work properly if the verbal data is left in the worksheet. Your worksheet should look something like the one shown in Figure 9.24.

Save this worksheet in a CSV (comma separated values) format using the **File > Save As** command.

Procedure 9.5: Running a One-Factor Multinomial Regression

https://goo.gl/Hx5Ay7

1. Navigate to the online app at https://wac.colostate.edu/books/practice/codingstreams/.
2. The interface should look like that shown in Figure 9.25.

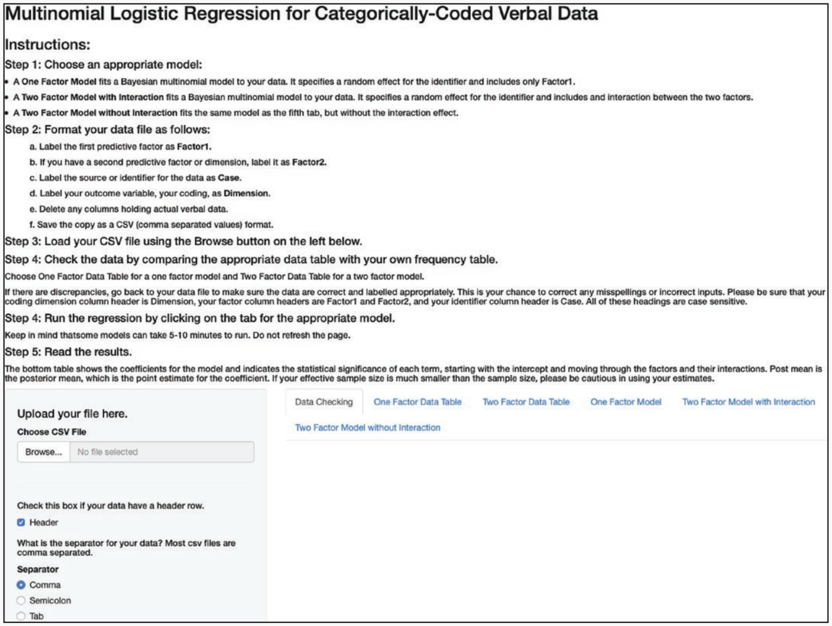

Figure 9.25: Interface for the online app for multinomial logistic regression.

3. Click on the **Browse button** on the left. Navigate to and choose the CSV file holding your data.

The data should load.

4. Click on the tab labeled **One Factor Data Table**.

If the app returns an error, check your data setup following Procedure 9.4.

5. Compare the frequency table on the tab with the frequency table you created for your earlier χ^2 analysis.

Continued . . .

Procedure 9.5: Running a One-Factor Multinomial Regression (continued)

https://goo.gl/Hx5Ay7

If the frequency table does not match a frequency table you generated earlier, check that you are using the correct data file and that the columns are labeled appropriately.

6. To get the results of the regression, click on the tab labeled **One Factor Model**.
7. Wait until the calculation is completed.

The output will look like that shown in Figure 9.26.

```
 Iterations = 10001:99901
 Thinning interval  = 100
 Sample size  = 900

 DIC: 272.2726

 G-structure:  ~idh(trait):Case

                              post.mean l-95% CI u-95% CI eff.samp
traitDimension.Social.Case       1.3330  0.04435   3.1674   142.31
traitDimension.Technical.Case    0.2383  0.01739   0.7477    53.71

 R-structure:  ~us(trait):units

                                                   post.mean l-95% CI
traitDimension.Social:traitDimension.Social.units        0.31406  0.02240
traitDimension.Technical:traitDimension.Social.units     0.01242 -0.42167
traitDimension.Social:traitDimension.Technical.units     0.01242 -0.42167
traitDimension.Technical:traitDimension.Technical.units  0.21119  0.02234
                                                   u-95% CI eff.samp
traitDimension.Social:traitDimension.Social.units        0.9841    83.77
traitDimension.Technical:traitDimension.Social.units     0.3651   116.27
traitDimension.Social:traitDimension.Technical.units     0.3651   116.27
traitDimension.Technical:traitDimension.Technical.units  0.6290   124.27

 Location effects: Dimension ~ -1 + trait + Factor1

                         post.mean l-95% CI u-95% CI eff.samp   pMCMC
traitDimension.Social      0.09732 -0.82864  1.01817   334.82 0.78444
traitDimension.Technical   1.23417  0.47790  1.97048   209.19 < 0.001 **
Factor1Object              2.17303  0.81538  3.49232    22.53 < 0.001 **
Factor1Practice            1.39528  0.38909  2.49851    92.42 0.00889 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9.26: Output of a one-factor multinomial logistic regression.

8. Make a copy of the output and place it in a worksheet along with your data worksheet. Label it Run 1.
9. Reload the app page in your browser to clear the data.
10. Click on the **Browse button** and load the same CSV file.
11. Click on the tab labeled **One Factor Model**.
12. Copy the output into a second worksheet labeled Run 2.

Interpreting a One-Factor Multinomial Logistic Regression

The first step in interpreting the results of a one-factor multinomial logistic regression is ensuring your results are stable. This involves comparing the results of the two runs you have made. For the sample data the results for our two runs are shown in Figure 9.27.

```
                          post.mean l-95% CI u-95% CI eff.samp   pMCMC
traitDimension.Social       0.09732 -0.82864  1.01817   334.82 0.78444
traitDimension.Technical    1.23417  0.47790  1.97048   209.19 < 0.001 **
Factor1Object               2.17303  0.81538  3.49232    22.53 < 0.001 **
Factor1Practice             1.39528  0.38909  2.49851    92.42 0.00889 **
---
```

Run 1

```
                          post.mean l-95% CI u-95% CI eff.samp   pMCMC
traitDimension.Social        0.1394  -0.8573   1.0766   180.34 0.75333
traitDimension.Technical     1.2388   0.5671   2.0205   198.42 0.00222 **
Factor1Object                1.9628   0.6942   3.1814    55.05 < 0.001 **
Factor1Practice              1.4305   0.4198   2.1984    74.85 < 0.001 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Run 2

Figure 9.27: Comparing the results of two runs of the app for one-factor multinomial logistic regression.

In our analysis, the code of *Professional* serves as the baseline for the outcome dimension of Alignment. The first two lines in the two outputs shown in Figure 9.25 give the results for the other two Alignment codes, *Social* and *Technical*, compared with this baseline.

The next two lines in the output give the results for the Frame dimension. Here the category of *Identity* serves as the baseline, and these two lines give the results for the other two Frame codes, *Object* and *Practice*, compared with this baseline.

On each line, the first and last columns are your main focus. In the first column, labeled post.mean (posterior mean), you find the critical log odds values computed by the multinomial logistic regression. In the last column, labeled pMCMC, you find an estimate of the probability that the log odds could have occurred by chance. Asterisks mark those that may be considered significant. P values of less than 1 in a thousand ($p < .001$) are labeled with a triple asterisk

(***). Those with p values of less than 1 in a hundred ($p < .01$) are labeled with a double asterisk (**). Those with p values of less than 5 in a hundred ($P < .05$) are labeled with a single asterisk (*).

When you compare the two runs for stability, you want to see that they have the same results in terms of significance levels and somewhat similar log odds values. When we compare the significance values of the two runs shown in Figure 9.27, for example, we see that they both show significance values at $p < .001$ for *Technical*, for *Object*, and for *Practice* and we see that the log odds are somewhat similar. This assures us that the results are stable.

To understand these results, we take a closer look at the log odds. Looking at Run 1, Figure 9.26, we see that the log odds listed for traitDimension.Social is .09732. In other words, log-odds for being coded as *Social* relative to being coded as *Professional* is .09732. The fact that this number is positive shows that any segment is more likely to be coded as *Social* than *Professional* regardless of how it was coded for Frame. If we refer back to the frequency table of observed values repeated in Figure 9.28, we can confirm that the frequencies for the *Social* row are all larger than the *Professional* row. But the small log odds from the multinomial logistic regression tells us that this difference is not large enough to be surprising.

We look next at the results for traitDimension.Technical. Here the results are suggesting that the log odds for being coded as *Technical* relative to being coded as *Professional* are 1.23417, which is significant at $p < .001$. That is, any segment is significantly more likely to be coded as *Technical* than *Professional* regardless of how it was coded for Frame. Again, a look at the frequency table in Figure 9.28 confirms that the values for *Technical* are generally much higher than those for *Professional*. That is, there is less than one chance in a thousand that this would have occurred by chance if there were no true difference.

The next two lines show the impact of the Frame dimension, our predictor variable, on Alignment, our outcome variable. For Factor1Object, the log odds for being coded as *Object* relative to being coded as *Identity* are 2.17303, which is significant at $p < .001$. This means that if a segment were to change coding from *Identity* to *Object* along the Frame dimension, the multinomial log-odds for being coded as something other than *Professional* along the Alignment dimension would be expected to increase by 2.17303 units while holding all other variables in the model constant. In other words, the impact of coding a segment as *Object* along the Frame dimension increases its chances of being coded as something other than *Professional* along the Alignment dimension. That is, when people wrote about objects, they did not often talk about the professional contexts for those objects.

Step 1. Observed	Identity	Object	Practice	Total
Professional	9	0	4	13
Social	23	12	14	49
Technical	5	63	80	148
Total	37	75	98	210
Step 2. Expected	identity	object	practice	Total
Professional	2	5	6	13
Social	9	18	23	49
Technical	26	53	69	148
	37	75	98	210
Step 3. O-E	identity	object	practice	Total
Professional	6.71	-4.64	-2.07	0.00
Social	14.37	-5.50	-8.87	0.00
Technical	-21.08	10.14	10.93	0.00
	0	0	0	0
Step 4. $(O\text{-}E)^2/E$	identity	object	practice	Total
Professional	19.65	4.64	0.70	25.00
Social	23.91	1.73	3.44	29.07
Technical	17.03	1.95	1.73	20.71
	60.60	8.32	5.87	**74.79**

Figure 9.28: Frequency tables from the χ^2 test of independence for the same data used to produce the output shown in Figure 9.24.

For Factor1Practice, the log odds for being coded as *Practice* relative to being coded as *Identity* are 1.39528, which is also significant at $p < .001$. This means that if a segment were to change coding from *Identity* to *Practice* along the dimensions of Frame, the multinomial log-odds for being coded as something other than *Professional* along the dimension of Alignment would be expected to increase by 1.39528 units while holding all other variables in the model constant. In other words, the impact of coding a segment as *Practice* along the Frame dimension also increases its chances of being coded as something other than *Professional* along the Alignment dimension. That is, when people wrote about practices they did not often talk about the professional contexts for those practices.

These results are generally consistent with those of our earlier χ^2 test of independence, where we found a very high sum of χ^2 (74.79) but were unsure of how to interpret this inflated result. With the significant results of the multinomial logistic regression, we can have greater confidence in this earlier finding and see some further patterns that are indicated by color in Figure 9.28. In gold, we see the three frequencies identified in the χ^2 analysis as being unexpectedly different from their expected values. And in shades of orange we have marked those values that have been identified by the multinomial logistic regression as significant.

As we noted in our earlier discussion, the χ^2 analysis suggested that something surprising is going on with the predictor variable of *Identity*. And, as we just noted, the significant log odds for Factor1Object and Factor1Practice also suggest something going on with this category: as coding on the Frame dimension moves out of the *Identity* category into the other two categories, the chances of being coded as *Professional* decrease significantly. In Figure 9.28, this is indicated by the light orange and medium orange cells in the *Object* and *Practice* columns compared to the uncolored cells in the *Professional* row. That is, when people wrote about practices and objects they did not often talk about them in their professional contexts.

The results of the multinomial logistic regression also tell us something else: the value for *Technical* compared to *Professional* is surprisingly high on the Alignment dimension. Though not examined by the earlier χ^2 test for independence, this result is consistent with our frequency table of observed values. The dark orange cell in Figure 9.28 pinpoints an overall frequency of 148 for *Technical* compared to the overall frequency of 13 for *Professional*. The log odds for traitDimension.Technical tell us that this is significant. And it might appear that the frequency of 49 for *Social* compared to the frequency of 13 for *Professional* would also be significant. But the log odds for traitDimension.Social tell us this is not the case.

Two-Factor Multinomial Logistic Regression

Two-factor multinomial logistic regression is an analytic tool designed to examine the impact of two predictor variables on an outcome variable. With coded verbal data, the outcome variable is always the coding along a given dimension.

The predictor variables will be the values on a build-in contrast and the values on a second dimension. It is designed to answer the question, *"Given a value on a built-in contrast, what is the likelihood of a given code along a second coding dimension given a value on a first coding dimension?"* This is the test to use when you have data coded along two dimensions as well as a built-in contrast, a complex analysis that cannot be handled by a χ^2 test.

In addition to looking for the main effects of the two predictor variables, a two-factor multinomial logistic regression can also look for a significant interaction between them. An interaction between two variables means the effect of one of those variables on a third variable is not constant—the effect differs at different values of the other. For the sample data we will be using show in Figure 9.29, an interaction would mean that the effect of Factor2 (the Frame dimension) on Dimension (the Alignment dimension) would be different depending on the built-in contrast of Year. As we noted earlier in Chapter 7, a pattern of association between two dimensions may not hold true on both sides of a contrast; this is an interaction. A two-factor multinomial logistic regression will tell us if this interaction is significant.

Adding an interaction to a two-factor model may improve the fit of the model, but it is also possible that it does not improve it. For this reason, in the following procedures, we suggest that you run a two-factor multinomial logistic regression both with and without an interaction and then determine which is the better fit.

Running a two-factor multinomial logistic regression

The two-factor multinomial logistic regression, like its one-factor counterpart, works with the individual data points in your data set. Prepare your data for the app as detailed in Procedure 9.6 and then run a two-factor multinomial logistic regression both with and without interaction using Procedures 9.7 and 9.8.

Procedure 9.6: Preparing the Data for a Two-Factor Multinomial Regression

https://goo.gl/Hx5Ay7

1. Combine the coded data from individual data worksheets into a single worksheet, keeping track of which data comes from which worksheet.
2. Create a new column in the worksheet. Label it as **Case**.
3. In this column, next to each segment, enter the name of the data worksheet from which the segment was copied.
4. Change the name of the column holding your built-in contrast to **Factor1**.
5. Change the name of the column with your first dimension to **Factor2**
6. Change the name of the column with your second dimension to **Dimension**.
7. Delete the column holding the actual verbal data.

The online app will not work properly if the verbal data is left in the worksheet. Your worksheet should look something like the one shown in Figure 9.29.

Obs	Case	Factor1	Factor2	Dimension
1	irunepan	Year1	Identity	Professional
2	irunepan	Year1	Identity	Professional
3	irunepan	Year1	Practice	Technical
4	irunepan	Year1	Practice	Technical
5	irunepan	Year1	Practice	Technical
6	irunepan	Year1	Practice	Technical
7	irunepan	Year1	Practice	Technical
8	irunepan	Year1	Identity	Technical
9	irunepan	Year1	Practice	Technical
10	pieper	Year1	Practice	Technical

Figure 9.29: Worksheet arrangement for a two-factor multinomial logistic regression.

Save this worksheet in a CSV (comma separated values) format using the **File > Save As** command.

Procedure 9.7: Running a Two-Factor Multinomial Regression with Interaction

https://goo.gl/Hx5Ay7

1. Navigate to the online app at https://wac.colostate.edu/books/practice/codingstreams/. The interface should look like that shown in Figure 9.25.
2. Click on the **Browse button** on the left. Navigate to and choose the CSV file holding your data.

The data should load.

3. Click on the tab labeled **Two Factor Data Table**.

If the app returns an error, check your data setup following Procedure 9.6.

4. Check the frequency table on the tab to make sure that the values look right.

If the frequency table does not look right, check that you are using the correct data file and that the columns are labeled appropriately.

5. To get the results of the regression with interaction, click on the tab labeled **Two Factor Model with Interaction**.
6. Wait until the calculation is completed.

The output will look like that shown in Figure 9.30.

```
 Iterations = 5001:79971
 Thinning interval  = 90
 Sample size  = 834

 DIC: 271.2435

 G-structure:  ~idh(trait):Case

                              post.mean l-95% CI u-95% CI eff.samp
traitDimension.Social.Case       1.5550  0.03869   4.2242    97.84
traitDimension.Technical.Case    0.2024  0.01506   0.6315   130.28

 R-structure:  ~us(trait):units

                                                       post.mean l-95% CI
traitDimension.Social:traitDimension.Social.units        0.36178  0.02382
traitDimension.Technical:traitDimension.Social.units     0.03881 -0.40194
traitDimension.Social:traitDimension.Technical.units     0.03881 -0.40194
traitDimension.Technical:traitDimension.Technical.units  0.22481  0.02148
                                                       u-95% CI eff.samp
traitDimension.Social:traitDimension.Social.units        1.2215    62.45
traitDimension.Technical:traitDimension.Social.units     0.6138    52.70
traitDimension.Social:traitDimension.Technical.units     0.6138    52.70
traitDimension.Technical:traitDimension.Technical.units  0.6334    76.31

 Location effects: Dimension ~ -1 + trait + Factor1 + Factor2 + Factor1 *
Factor2

                             post.mean l-95% CI u-95% CI eff.samp  pMCMC

traitDimension.Social          0.02296 -0.97040  1.04654   260.56 0.9616

traitDimension.Technical       1.15934  0.39163  1.96202   156.16 0.0024 *
*
Factor1Year2                   0.47757 -0.60851  1.70344   196.56 0.3717

Factor2Object                  2.14273  0.85537  3.36889    71.84 <0.001 *
*
Factor2Practice                1.19465  0.07297  2.32849   105.64 0.0384 *

Factor1Year2:Factor2Object     0.52472 -1.08910  2.25891    94.08 0.5755

Factor1Year2:Factor2Practice   0.24464 -0.97286  1.77297   118.38 0.7434

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9.30: Output from a two-factor multinomial logistic regression with interaction.

Procedure 9.8: Running a Two-Factor Multinomial Regression without Interaction

https://goo.gl/Hx5Ay7

1. Navigate to the online app at https://wac.colostate.edu/books/practice/codingstreams/.

The interface should look like that shown in Figure 9.25.

1. Click on the **Browse button** on the left. Navigate to and choose the CSV file holding your data.
2. The data should load.
3. To get the results of the regression without interaction, click on the tab labeled **Two Factor Model without Interaction**.
4. Wait until the calculation is completed.

The output will look like that shown in Figure 9.31.

```
 Iterations = 7001:89901
 Thinning interval  = 100
 Sample size  = 830

 DIC: 272.0983

 G-structure:  ~idh(trait):Case

                              post.mean l-95% CI u-95% CI eff.samp
traitDimension.Social.Case       1.6330  0.04846   4.6775    44.73
traitDimension.Technical.Case    0.2491  0.01519   0.8326    57.51

 R-structure:  ~us(trait):units

                                                         post.mean l-95% CI
traitDimension.Social:traitDimension.Social.units          0.34472  0.02647
traitDimension.Technical:traitDimension.Social.units       0.03253 -0.43202
traitDimension.Social:traitDimension.Technical.units       0.03253 -0.43202
traitDimension.Technical:traitDimension.Technical.units    0.24841  0.02532
                                                         u-95% CI eff.samp
traitDimension.Social:traitDimension.Social.units          1.1345   125.12
traitDimension.Technical:traitDimension.Social.units       0.5842    66.05
traitDimension.Social:traitDimension.Technical.units       0.5842    66.05
traitDimension.Technical:traitDimension.Technical.units    0.7774    72.76

 Location effects: Dimension ~ -1 + trait + Factor1 + Factor2

                         post.mean l-95% CI u-95% CI eff.samp   pMCMC
traitDimension.Social     -0.02805 -1.05551  0.91618   323.53 0.98072
traitDimension.Technical   1.07304  0.34776  1.84195   238.52 0.00482 **
Factor1Year2               0.67048 -0.36824  1.77548    60.91 0.20723
Factor2Object              2.04976  0.99334  3.25079    59.85 < 0.001 **
Factor2Practice            1.43713  0.44051  2.62174    62.19 0.01205 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9.31: Output from a two-factor multinomial

Interpreting a Two-factor multinomial logistic regression

The first step in interpreting the results of a two-factor multinomial logistic regression is choosing between the two tests you have run, one with interaction and one without. In most cases, you will choose to use the one with the interaction as it will give you more information. But occasionally, the model with interaction will be a poorer fit for the data. To check for this, compare the DIC numbers at the top of the two outputs.

With our sample data, we see that the DIC for the run with interaction, shown in Figure 9.30, is 271.2435. Without interaction, the DIC, shown in Figure 9.31, is 272.0983. In general, the run with the smaller DIC is a better fit. These two DICs are pretty close to one another, so it may not make much difference, so we choose to work with the results with interaction.

To understand the results, we take a closer look at the log odds. Looking at Figure 9.30, we see that the log odds listed for trait Dimension.Social are .02296. In other words, log-odds for being coded as *Social* relative to being coded as *Professional* are .02296. The fact that this number is positive shows that any segment is more likely to be coded as *Social* than *Professional* regardless of how it was coded for Frame or Year, although this difference is not big enough to be significant.

We look next at the results for traitDimension.Technical. Here the results are suggesting that the log odds for being coded as *Technical* relative to being coded as *Professional* are 1.15934, which is significant at $p < .01$. That is, any segment is significantly more likely to be coded as *Technical* than *Professional* regardless of how it was coded for Frame or Year.

The next line shows the impact of Year, the first of our predictor variables, on Alignment, our outcome variable. For Factor1Year2, the log odds for being *Year2* relative to being coded as *Year1* are .47757, which is not significant. This means that if a segment were to change coding from *Identity* to *Object* along the built-in contrast of Year, the multinomial log-odds for being coded as something other than *Professional* along the Alignment dimension would be expected to increase by .47757 unit while holding all other variables in the model constant. In other words, the impact of being *Year2* rather than *Year1* along the built-in contrast of Year does not have much effect on how it is coded along the Alignment dimension.

The next two lines show the impact of the Frame dimension, our second predictor variable, on Alignment. For Factor2Object, the log odds for being coded as *Object* relative to being coded as *Identity* are 2.14273, which is significant at $p < .01$. This means that if a segment were to change coding from *Identity* to *Object* along the Frame dimension, the multinomial log-odds for being coded as something other than *Professional* along the Alignment dimension would be expected to increase by 2.14273 units while holding all other variables in the model constant. In other words, the impact of coding a segment as *Object* along the Frame dimension increases its chances of being coded as something other than *Professional* along the Alignment dimension.

For Factor2Practice, the log odds for being coded as *Practice* relative to being coded as *Identity* are 1.19465, which is also significant at $p < .01$. This means that if a segment were to change coding from *Identity* to *Practice* along the dimensions of Frame, the multinomial log-odds for being coded as something other than *Professional* along the dimension of Alignment would be expected to increase by 1.19465 units while holding all other variables in the model constant. In other words, the impact of coding a segment as *Practice* along the Frame dimension also increase its chances of being coded as something other than *Professional* along the Alignment dimension.

Our final results concern the interactions between Year and Frame. For Factor1Year2:Factor2Object, the log odds for being coded as *Object* relative to being coded as *Identity* under a coding of *Year1* or *Year2* are .52472, which is not significant. For Factor1Year2:Factor2Practice, the log odds for being coded as *Practice* relative to being coded as *Identity* under a coding of *Year1* or *Year2* are .24464, which is also not significant. This means that neither being coded as *Object* nor *Practice* are significantly affected by Year.

We note that the data used here are slightly different than that used for the earlier χ^2 test of homogeneity because here we only have two values for the built-in contrast of Year.

Memo 9.3: Interpreting Your Multinomial Logistic Regression

Record the results of a multinomial logistic regression on your data. What are the log odds of each category? Which ones are significant at which level? Also record a frequency table for the data set.

For each result, write a sentence describing what each result means. Refer to the frequency tables for details. Overall, which predictor variables seem to have an impact on the way your data was coded?

For Further Reading

Dolinar, S. (2014). *Statistics—Probability vs. odds*. Retrieved from https://stats.seandolinar.com/statistics-probability-vs-odds/

Grace-Martin, K. (2019a). *Chi-square test vs. logistic regression: Is a fancier test better?* The Analysis Factor. Retrieved from https://www.theanalysisfactor.com/chi-square-test-vs-logistic-regression-is-a-fancier-test-better/

Grace-Martin, K. (2019b). *The difference between interaction and association?* The Analysis Factor. Retrieved from https://www.theanalysisfactor.com/interaction-association/

Kiernan, K. (2018). Insights into using the GLIMMIX procedure to model categorical outcomes with random effects. *SAS Global Forum Proceedings*. Retrieved from https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2179-2018.PDF

Kleinbaum, D. G. & Klein, M. (2010). *Logistic regression: A self-learning text*. New York: Springer.

Miller, S. V. (2014). *Reading a regression table: A guide for students*. Retrieved from http://svmiller.com/blog/2014/08/reading-a-regression-table-a-guide-for-students/

Papageorgiou, G., & Hinde, J. (2012). Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Stat Comput* 22, 79-92.

UCLA: Statistical Consulting Group.Multinomial logistic regression: Stata annotated output. Retrieved September 8, 2019, from https://stats.idre.ucla.edu/stata/output/multinomial-logistic-regression-2/