17.8. How to report factor analysis ①

When reporting factor analysis we should provide our readers with form an informed opinion about what we've done. We should be for extracting factors and the method of rotation used. We shoul of the rotated factor loadings of all items and flag (in bold) values (I would personally choose .40, but see Section 17.4.6.2). We sho centage of variance that each factor explains and possibly the eige shows an example of such a table for the SAQ data (oblique rotatio reported the sample size in the title.

In my opinion, a table of factor loadings and a description of minimum. You could consider (if it's not too large) including the tal which someone could reproduce your analysis (should they want tion on sample size adequacy. For this example we might write som

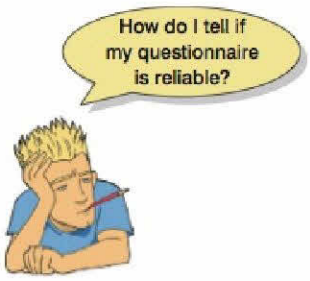
✓ A principal axis factor analysis was conducted on the 23 item (direct oblimin). The Kaiser–Meyer–Olkin measure verified for the analysis, KMO = .93 ('marvellous' according to Hu 1999), and all KMO values for individual items were greater above the acceptable limit of .5 (Field, 2013). An initial ana eigenvalues for each factor in the data. Four factors had eig criterion of 1 and in combination explained 50.32% of the v was ambiguous and showed inflexions that would justify reta tors. We retained 4 factors because of the large sample size a the scree plot and Kaiser's criterion on this value. Table 17.1 ings after rotation. The items that cluster on the same facto represents a fear of statistics, factor 2 represents peer evaluati fear of computers and factor 4 a fear of maths.

17.9. Reliability analysis ②

17.9.1. Measures of reliability ③

If you're using factor analysis to validate a questionnaire, it is useful of your scale.

SELF-TEST Thinking back to Chapter 1, what and test–retest reliability?



Reliability means that a measure (or in this case que
sistently reflect the construct that it is measuring. One
that, other things being equal, a person should get the
tionnaire if they complete it at two different points in
discovered that this is called test–retest reliability). So,
fied of SPSS and who scores highly on our SAQ shoul
if we tested them a month later (assuming they hadn't

TABLE 17.1 Summary of exploratory factor analysis results for the SPSS anxiety q ($N = 2571$)

	Rotated Factor Loadin		
Item	Fear of Statistics	Peer Evaluation	Fear of Compute
I wake up under my duvet thinking that I am trapped under a normal distribution	**.54**	−.04	.17
I can't sleep for thoughts of eigenvectors	**.47**	−.14	−.08
I weep openly at the mention of central tendency	**.45**	−.05	.17
I dream that Pearson is attacking me with correlation coefficients	**.44**	.08	.18
Standard deviations excite me	**−.44**	.32	−.05
Statistics makes me cry	**.43**	.10	.11
People try to tell you that SPSS makes statistics easier to understand but it doesn't	**.41**	−.04	.36
I don't understand statistics	**.36**	.05	.20
My friends are better at statistics than me	−.09	**.56**	−.02
My friends are better at SPSS than I am	.07	**.47**	−.11
My friends will think I'm stupid for not being able to cope with SPSS	−.18	**.45**	.04
If I'm good at statistics my friends will think I'm a nerd	.10	**.35**	.00
Everybody looks at me when I use SPSS	−.22	**.34**	−.08
I have little experience of computers	−.22	−.01	**.86**
SPSS always crashes when I try to use it	.18	−.01	**.64**
All computers hate me	.19	−.02	**.56**
I worry that I will cause irreparable damage because of my incompetence with computers	.08	−.04	**.56**
Computers have minds of their own and deliberately go wrong whenever I use them	.24	−.02	**.47**
Computers are useful only for playing games	.00	−.06	**.39**
Computers are out to get me	.11	−.13	**.32**
I have never been good at mathematics	.01	.05	−.09
I did badly at mathematics at school	−.01	−.11	.06
I slip into a coma whenever I see an equation	.08	.02	.09
Eigenvalues	7.29	1.74	1.32
% of variance	31.70	7.56	5.73
α	.82	.57	.82

Note: Factor loadings over .40 appear in bold

Note: Factor loadings over .40 appear in bold

SPSS-anxiety therapy in that month). Another way to look at reliability is to s
people who are the same in terms of the construct being measured should g
score. So, if we took two people who were equally SPSS-phobic, then they shou
or less identical scores on the SAQ. Likewise, if we took two people who lovec
should both get equally low scores. It should be apparent that the SAQ wou
accurate measure of SPSS anxiety if we took someone who loved SPSS and so

LABCOAT LENI'S REAL RESEARCH 17.1

Worldwide addiction? ②

In 2007 it was estimated that around 179 million people worldwide used the Internet. From the increasing popularity (and usefulness) of the Internet has emerged a serious and recognized problem of internet addiction. To research this construct it's helpful to be able to measure it, so Laura Nichols and Richard Nicki developed the Internet Addiction Scale (Nichols & Nicki, 2004). Nichols and Nicki's 36-item questionnaire contains items such as 'I have stayed on the Internet longer than I intended to' and 'My grades/work have suffered because of my Internet use' to which responses are made on a 5-point scale (Never, Rarely, Sometimes, Frequently, Always). (Incidentally, while researching this topic I encountered an Internet addiction recovery whole host of resources (e.g support groups, videos, podca you online for ages. It struck m a heroin addiction recovery ce of free heroin in the reception a

The data from 207 people i **Nichols & Nicki (2004).sav**. items because they had low m dropped three others because tions with other items. They pe ponent analysis on the remaini wants you to run some descrip which two items were droppe variances, then inspect a corr three items that were droppe tions. Finally, he wants you to r analysis on the data. Answers a rial on the companion websit article).

NICHOLS, L. A., & NICKI, R. (2004). *PSYCHOLOGY OF ADDICTIVE BEHAVIORS*, 18A, 381–384.

was terrified of it and they got the same score! In statistical terms, t reliability is based on the idea that individual items (or sets of items) consistent with the overall questionnaire. So, if we take someone sc overall score on the SAQ will be high; if the SAQ is reliable ther some items from it the person's score on those items should also be

The simplest way to do this in practice is to use **split-half reliabi** the scale set into two randomly selected sets of items. A score for e lated on each half of the scale. If a scale is reliable a person's score should be the same (or similar) to their score on the other half. Acr scores from the two halves of the questionnaire should correlate very between the two halves is the statistic computed in the split-half met tions being a sign of reliability. The problem with this method is tha in which a set of data can be randomly split into two and so the res of the way in which the data were split. To overcome this problem, up with a measure that is loosely equivalent to creating two sets of sible and computing the correlation coefficient for each split. The a equivalent to **Cronbach's alpha**, α, which is the most common measu

$$\alpha = \frac{N^2\overline{\text{cov}}}{\sum \quad \sum}$$

$$\sum s^2_{\text{item}} + \sum \text{cov}_{\text{item}}$$

[10] Although this is the easiest way to conceptualize Cronbach's, α, whether or not it i
of all possible split-half reliabilities depends on exactly how you calculate the split-h
for computational details). If you use the Spearman–Brown formula, which takes
deviations, then Cronbach's will be equal to the average split-half reliability only w
tions are equal; otherwise α will be smaller than the average. However, if you use
ability that does account for item standard deviations (such as Flanagan, 1937; Rul
equal the average split-half reliability (see Cortina, 1993).

This equation may look complicated, but actually isn't. For each item on our s calculate two things: the variance within the item, and the covariance between item and any other item on the scale. Put another way, we can construct a varia ance matrix of all items. In this matrix the diagonal elements will be the varia particular item, and the off-diagonal elements will be covariances between pa The top half of the equation is simply the number of items (*N*) squared multi average covariance between items (the average of the off-diagonal elements i mentioned variance–covariance matrix). The bottom half is the sum of all the ite and item covariances (i.e., the sum of everything in the variance–covariance ma

There is a standardized version of the coefficient too, which essentially us equation except that correlations are used rather than covariances, and the bo the equation uses the sum of the elements in the correlation matrix of items (i 1s that appear on the diagonal of that matrix). The normal alpha is appropriate on a scale are summed to produce a single score for that scale (the standardi not appropriate in these cases). The standardized alpha is useful, though, whe scale are standardized before being summed.

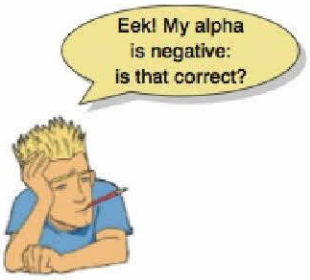
17.9.2. Interpreting Cronbach's α (some cautionary ta

You'll often see in books or journal articles, or be told by people, that a valu is an acceptable value for Cronbach's α; values substantially lower indicate a scale. Kline (1999) notes that although the generally accepted value of .8 is for cognitive tests such as intelligence tests, for ability tests a cut-off point o suitable. He goes on to say that when dealing with psychological constructs, v even .7 can, realistically, be expected because of the diversity of the constructs ured. Some even suggest that in the early stages of research, values as low as .5 (Nunnally, 1978). However, there are many reasons not to use these general gui least of which is that they distract you from thinking about what the value mean context of the research you're doing (Pedhazur & Schmelkin, 1991).

We'll now look at some issues in interpreting alpha, which have been di ticularly well by Cortina (1993) and Pedhazur and Schmelkin (1991). First, th depends on the number of items on the scale. You'll notice that the top half of t for α includes the number of items squared. Therefore, as the number of items increases, α will increase. As such, it's possible to get a large value of α becaus lot of items on the scale, and not because your scale is reliable. For example, Co reports data from two scales, both of which have $\alpha = .8$. The first scale has only and the average correlation between items was a respectable .57; however, the had 10 items with an average correlation between these items of a less resp Clearly the internal consistency of these scales differs, but according to Cronb are both equally reliable.

Second, people tend to think that alpha measures 'unidimensionality', or t which the scale measures one underlying factor or construct. This is true when

factor underlying the data (see Cortina, 1993), but Grayson (2004) demonstra
sets with the same α can have very different factor structures. He showed tha
be achieved in a scale with one underlying factor, with two moderately correl
and with two uncorrelated factors. Cortina (1993) has also shown that with m
items, and fairly high correlations between items ($r > .5$), α can reach values
above .7 (.65 to .84). These results show that α should not be used as a meas
dimensionality'. Indeed, Cronbach (1951) suggested that if several factors ex
formula should be applied separately to items relating to different factors. In c
if your questionnaire has subscales, α should be applied separately to these sub



The final warning is about items that have a revers ple, in the SAQ there is one item (question 3) that v site way around to all other items. The item was 'star me'. Compare this to any other item and you'll see it response. For example, item 1 is 'statistics make me statistics then you'll strongly agree with this statement of 5 on our scale. For item 3, if you hate statistics th are unlikely to excite you so you'll strongly disagree a the scale. These reverse-phrased items are important bias; participants will need to pay attention to the questions. F reverse phrasing doesn't matter; all that happens is you get a nega any reversed items (in fact, you'll see that item 3 has a negative fa 17.9). However, these reverse-scored items will affect alpha. To se equation for Cronbach's α. The top half incorporates the *avera* items. If an item is reverse-phrased then it will have a negative r items, hence the covariances between this item and other items average covariance is the sum of covariances divided by the numb by including a bunch of negative values we reduce the sum of cov also reduce Cronbach's α, because the top half of the equation g cases, it is even possible to get a negative value for Cronbach's α, si nitude of negative covariances is bigger than the magnitude of po Cronbach's α doesn't make much sense, but it does happen, and whether you included any reverse-phrased items.

If you have reverse-phrased items then you also have to rev they're scored before you conduct reliability analysis. This is quite data, we have one item which is currently scored as 1 = strongly 3 = neither, 4 = agree and 5 = strongly agree. This is fine for it way that agreement indicates statistics anxiety, but for item 3 (star me), disagreement indicates statistics anxiety. To reflect this nu reverse the scale such that 1 = strongly agree, 2 = agree, 3 = nei 5 = strongly disagree. In doing so, an anxious person still gets 5 they'd strongly disagree with it).

To reverse the scoring find the maximum value of your respon and add 1 to it (so you get 6 in this case). Then for each person, subtract from it the score they actually got. Therefore, someone v now scores 6−5 = 1, and someone who scored 1 originally now g in the middle of the scale with a score of 3 will still get 6−3 = 3. (a long time to do this for each person, but we can get SPSS to do it

SELF-TEST Using what you learnt in Chapter *compute* command to reverse-score item 3. (Cl that you are simply changing the variable to 6 m value.)

17.9.3. Reliability analysis in SPSS ②

Let's test the reliability of the SAQ using the data in **SAQ.sav**. Yo
scored item 3 (see above), but if you can't be bothered then loa

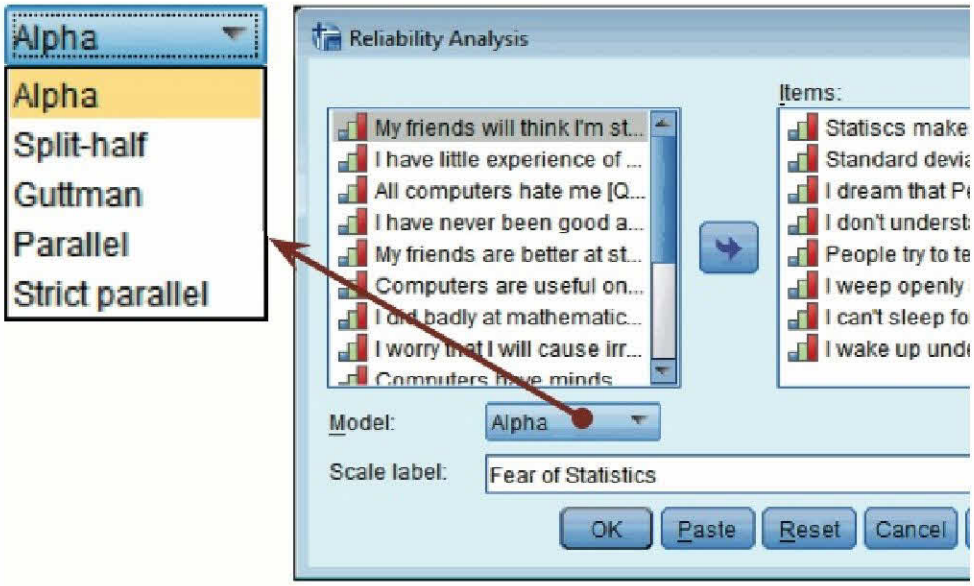


FIGURE 17.15 Main dialog box for reliability analysis.

Reversed).sav instead. Remember also that I said we should conduct reliability any subscales individually. If we use the results from our oblique rotation (O then we have four subscales:

1. Subscale 1 (*Fear of statistics*): items 1, 3, 4, 5, 12, 16, 20, 21
2. Subscale 2 (*Peer evaluation*): items 2, 9, 19, 22, 23
3. Subscale 3 (*Fear of computers*): items 6, 7, 10, 13, 14, 15, 18
4. Subscale 4 (*Fear of mathematics*): items 8, 11, 17

To conduct each reliability analysis on these data you need to s Scale ▸ Reliability Analysis... to display the dialog box in Figure 1 any items from the list that you want to analyse (to begin with, let's do the ite fear of statistics subscale: items 1, 3, 4, 5, 12, 16, 20 and 21) on the left-hanc dialog box and drag them to the box labelled Items (or click on). Rememl can select several items at the same time if you hold down the *Ctrl* (*Cmd* on while you select the variables.

There are several reliability analyses you can run, but the default option is α. You can change the method (e.g., to the split-half method) by clicking on reveal a drop-down list of possibilities, but the default method is a good one to it's a good idea to type the name of the scale (in this case 'Fear of Statistics') i labelled *Scale label* because this will add a header to the SPSS output with w

type in this box: typing a sensible name here will make your output easier to f

If you click on Statistics... you can access the dialog box in Figure 17.16. In t dialog box you can select several things, but the one most important for questi ability is: *Scale if item deleted*. This option tells us what the value of α woul item were deleted. If our questionnaire is reliable then we would not expect a to greatly affect the overall reliability. In other words, no item should cause decrease in α. If it does then you should consider dropping that item from t naire to improve reliability.

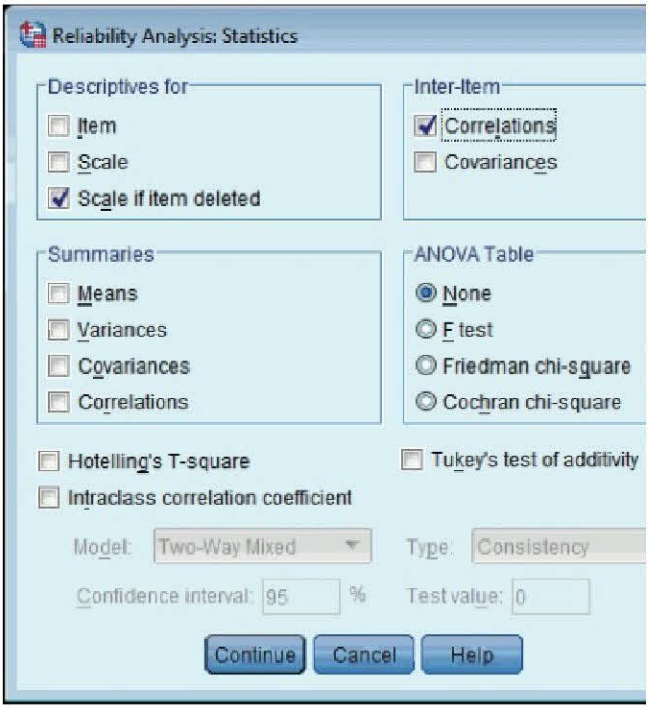


FIGURE 17.16
Statistics for reliability analysis

The inter-item correlations and covariances (and summaries) p
tion coefficients and averages for items on our scale. We should al
from our factor analysis, so there is little point in selecting these o
F test, *Friedman chi-square* (if your data are ranked), *Cochran chi-*
dichotomous), and *Hotelling's T-square* use these tests to compare
different items on the questionnaire. These tests might be useful t
similar distributional properties (i.e., the same average value), but
sizes you ought to be using for factor analysis, they will inevital
results even when only small differences exist between the questio

You can also request an **intraclass correlation coefficient (ICC)**.
cients that we encountered earlier in this book measure the relation
measure different things. For example, the correlation between
Omega and Satanism involves two classes of measures: the type
and their religious beliefs. Intraclass correlations measure the rel
variables that measure the same thing (i.e., variables within the sa
uses are in comparing paired data (such as twins) on the same me
consistency between judges' ratings of a set of objects (hence the re
the reliability statistics in SPSS). If you'd like to know more, see Se

Use the simple set of options in Figure 17.16 to run a basic relia
Continue to return to the main dialog box and then click on OK to

17.9.4. Reliability analysis output ②

Output 17.13 shows the results of this basic reliability analysis f
subscale. The value of Cronbach's α is presented in a small table a
reliability of the scale. Bearing in mind what we've already noted
number of items, and how daft it is to apply general rules, we're lo
region of about .7 to .8. In this case α is .821, which is certainly in
Kline (1999), and probably indicates good reliability.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	
Statistics makes me cry	21.76	21.442	
Standard deviations excite me	20.72	19.825	
I dream that Pearson is attacking me with correlation coefficients	21.35	20.410	
I don't understand statistics	21.41	20.942	
People try to tell you that SPSS makes statistics easier to understand but it doesn't	20.97	20.639	
I weep openly at the mention of central tendency	21.25	20.451	
I can't sleep for thoughts of eigenvectors	20.51	21.176	
I wake up under my duvet thinking that I am trapped under a normal distribution	20.96	19.939	

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.821	.823	8

OUTPUT 17.13

In the table labelled *Item-Total Statistics* the column labelled *Corrected Correlation* has the correlations between each item and the total score from the naire. In a reliable scale all items should correlate with the total. So, we're looking that don't correlate with the overall score from the scale: if any of these values are about .3 then we've got problems, because it means that a particular item does late very well with the scale overall. Items with low correlations may have to be For these data, all data have item–total correlations above .3, which is encouraging

The values in the column labelled *Cronbach's Alpha if Item Deleted* are the values overall α if that item isn't included in the calculation. As such, they reflect the Cronbach's α that would be seen if a particular item were deleted. The overall and so all values in this column should be around that same value. We're actually for values of alpha greater than the overall α. If you think about it, if the deletion item increases Cronbach's α then this means that the deletion of that item improves ability. Therefore, any items that have values of α in this column greater than α may need to be deleted from the scale to improve its reliability. None of these would increase alpha if they were deleted, which is good news. It's worth noting items do need to be removed at this stage then you should rerun your factor analysis to make sure that the deletion of the item has not affected the factor structure.

SELF-TEST Run reliability analyses on the other three subscales.

Just to illustrate the importance of reverse-scoring items before running relia
sis, Output 17.14 shows the reliability analysis for the fear of statistics subsca
on the original data (i.e., without item 3 being reverse-scored). Note that the
considerably lower (.605 rather than .821). Also, note that this item has a ne
total correlation (which is a good way to spot if you have a potential reverse-sc
the data that hasn't been reverse-scored). Finally, note that for item 3, the α if i

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation
Statistics makes me cry	20.93	12.125	.505
Standard deviations excite me	20.72	19.825	-.549
I dream that Pearson is attacking me with correlation coefficients	20.52	11.447	.526
I don't understand statistics	20.58	11.714	.466
People try to tell you that SPSS makes statistics easier to understand but it doesn't	20.14	11.739	.501
I weep openly at the mention of central tendency	20.42	11.584	.529
I can't sleep for thoughts of eigenvectors	19.68	12.107	.353
I wake up under my duvet thinking that I am trapped under a normal distribution	20.13	11.189	.541

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.605	.641	8

OUTPUT 17.14

is .8. That is, if this item were deleted then the reliability would im about .8. This, I hope, illustrates that failing to reverse-score items oppositely to other items on the scale will mess up your reliability

Let's now look at our subscale of peer evaluation. For our subs you should get the output in Output 17.15. The overall reliabilit ing to bake a cake for. The overall α is quite low, and although what Kline says we should expect for this kind of social science d statistics subscale and (as we shall see) the other two. The scale h to seven, eight and three on the other scales, so its reliability relati not going to be dramatically affected by the number of items. Th labelled *Corrected Item-Total Correlation* are all around .3, and sm results again indicate questionable internal consistency and identif problem. The values in the column labelled *Cronbach's Alpha if Ite* none of the items here would increase the reliability if they were de in this column are less than the overall reliability of .57. The item quite diverse themes of peer evaluation, and this might explain the ency; we probably need to rethink this subscale.

Moving on to the fear of computers subscale, Output 17.16 sho which is pretty good. The values in the column labelled *Corrected*

OUTPUT 17.15

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation
My friends will think I'm stupid for not being able to cope with SPSS	11.46	8.119	.339

My friends are better at statistics than me	10.24	6.395	.391
Everybody looks at me when I use SPSS	10.79	7.381	.316
My friends are better at SPSS than I am	10.20	7.282	.378
If I'm good at statistics my friends will think I'm a nerd	9.65	7.988	.239

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.570	.572	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronb Alpha Dele
I have little experience of computers	15.87	17.614	.619	.398	
All computers hate me	15.17	17.737	.619	.395	
Computers are useful only for playing games	15.81	20.736	.400	.167	
I worry that I will cause irreparable damage because of my incompetenece with computers	15.64	18.809	.607	.384	
Computers have minds of their own and deliberately go wrong whenever I use them	15.22	18.719	.577	.350	
Computers are out to get me	15.33	19.322	.491	.250	
SPSS always crashes when I try to use it	15.52	17.832	.647	.447	

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.823	.821	7

are again all above .3, which is also good. The values in the column labelled *Alpha if Item Deleted* show that none of the items would increase the reliability deleted. This indicates that all items are positively contributing to the overall

Finally, for the fear of maths subscale, Output 17.17 shows an overall reliab which indicates good reliability. The values in the column labelled *Correcte Correlation* are all above .3, which is good, and the values in the column labelle *Alpha if Item Deleted* indicate that none of the items here would increase the reli were deleted because all values in this column are less than the overall reliability

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cro Alph D
I have never been good at mathematics	4.72	2.470	.684	.470	
I did badly at mathematics at school	4.70	2.453	.682	.467	
I slip into a coma whenever I see an equation	4.49	2.504	.652	.425	

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.819	.819	3

CRAMMING SAM'S TIPS Reliability

- Reliability analysis is used to measure the consistency of a measure.

- Remember to reverse-score any items that were reverse-phrased on the original questionna analysis.
- Run separate reliability analyses for all subscales of your questionnaire.
- Cronbach's α indicates the overall reliability of a questionnaire, and values around .8 are go the like).
- The *Cronbach's Alpha if Item Deleted* column tells you whether removing an item will impro Values greater than the overall reliability indicate that removing that item will improve the ov Look for items that dramatically increase the value of α and remove them.
- If you remove items, rerun your factor analysis to check that the factor structure still holds.

17.10. How to report reliability analys

You can report the reliabilities in the text using the symbol α and ren Cronbach's α can't be larger than 1 we drop the zero before the d following APA practice):

- ✓ The fear of computers, fear of statistics and fear of maths subs high reliabilities, all Cronbach's $\alpha = .82$. However, the fear c tion subscale had relatively low reliability, Cronbach's $\alpha = .5$

However, the most common way to report reliability analysis wi analysis is to report the values of Cronbach's α as part of the ta For example, in Table 17.1 notice that in the last row of the table Cronbach's α for each subscale in turn.

17.11. Brian's attempt to woo Jane ①

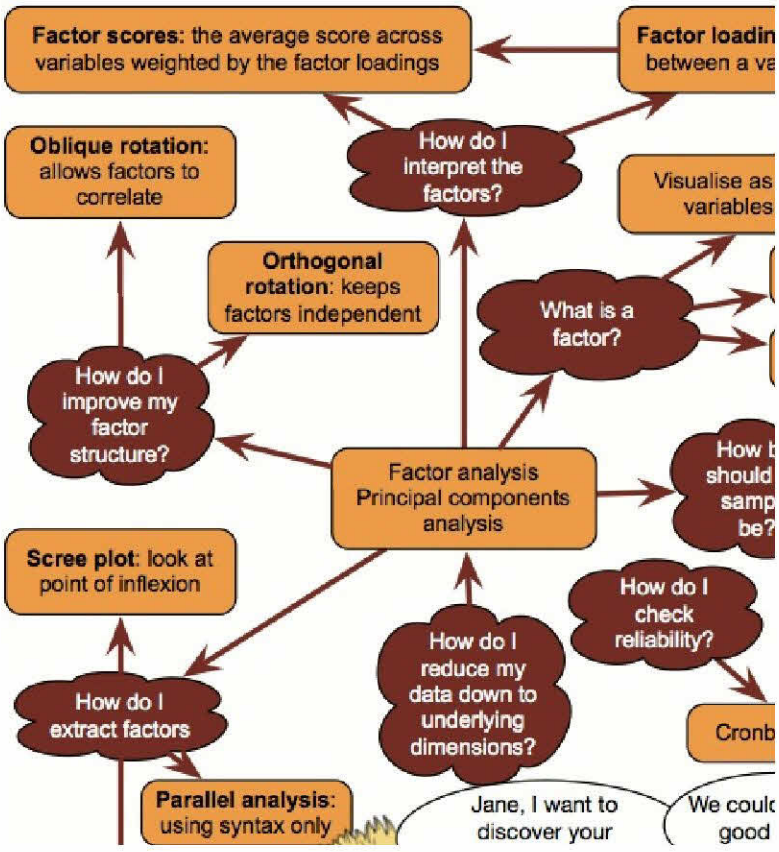


FIGURE 17.17
What Brian learnt from this chapter

17.12. What next? ②

At the age of 23 I took it upon myself to become a living homage to the diges
I furiously devoured articles and books on statistics (some of them I even un
mentally chewed over them, I broke them down with the stomach acid of m
stripped them of their goodness and nutrients, I compacted them down, and
two years I forced the smelly brown remnants of those intellectual meals out
form of a book. I was mentally exhausted at the end of it. 'It's a good job I'll n
do that again', I thought.

17.13. Key terms that I've discovered

Alpha factoring
Anderson–Rubin method
Common factor
Common variance
Communality
Component matrix
Confirmatory factor analysis
Cronbach's α
Direct oblimin
Extraction
Equamax
Factor analysis
Factor loading
Factor matrix
Factor scores
Factor transformation matrix, Λ
Intraclass correlation coefficient (ICC)
Kaiser's criterion
Latent variable
Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy
Oblique rotation
Orthogonal rotation
Pattern matrix
Principal componer analysis (PCA)
Promax
Quartimax
Random variance
Rotation
Scree plot
Singularity
Split-half reliability
Structure matrix
Unique factor
Unique variance
Varimax

17.14. Smart Alex's tasks

- **Task 1**: Rerun the analysis in this chapter using principal component analy
pare the results to those in the chapter. (Set the iterations to convergence t
- **Task 2**: The University of Sussex constantly seeks to employ the best peopl
lecturers. They wanted to revise the 'Teaching of Statistics for Scientific E
(TOSSE) questionnaire, which is based on Bland's theory that says that g
methods lecturers should have: (1) a profound love of statistics; (2) an
for experimental design; (3) a love of teaching; and (4) a complete abs
mal interpersonal skills. These characteristics should be related (i.e., corr
University revised this questionnaire to become the 'Teaching of Statistics f
Experiments – Revised' (TOSSE-R). They gave this questionnaire to 2
methods lecturers around the world to see if it supported Bland's theor

tionnaire is in Figure 17.18, and the data are in **TOSSE-R.sav**. Conduct a f
sis (with appropriate rotation) and interpret the factor structure. ②

- **Task 3:** Dr Sian Williams (University of Brighton) devised a questionnaire
organizational ability. She predicted five factors to do with organizationa
preference for organization; (2) goal achievement; (3) planning approacl
tance of delays; and (5) preference for routine. These dimensions are t
independent. Williams' questionnaire contains 28 items using a 7-point
(1 = strongly disagree, 4 = neither, 7 = strongly agree). She gave it to
Run a principal component analysis on the data in **Williams.sav**. ②

- **Task 4:** Zibarras, Port, and Woods (2008) looked at the relatic ality and creativity. They used the Hogan Development Surv sures 11 dysfunctional dispositions of employed adults: beir **cautious, detached, passive-aggressive, arrogant, manipulati perfectionist,** and **dependent.** Zibarras et al. wanted to redu based on parallel analysis, found that they could be reduced They ran a principal component analysis with varimax rotati **(Zibarras et al. (2008).sav)** to see which personality dimens (see page 210 of the original paper). ②

Answers can be found on the companion website.

FIGURE 17.18
The TOSSE-R questionnaire

Teaching of Statistics for Scientific Experiments — Revised (TOSSE-R)

		SD
1.	I once woke up in a vegetable patch hugging a turnip that I'd mistakenly dug up thinking it was Roy's largest root	○
2.	If I had a big gun I'd shoot all the students I have to teach	○
3.	I memorize probability values for the *F*-distribution	○
4.	I worship at the shrine of Pearson	○
5.	I still live with my mother and have little personal hygiene	○
6.	Teaching others makes me want to swallow a large bottle of bleach because the pain of my burning oesophagus would be light relief in comparison	○
7.	Helping others to understand sums of squares is a great feeling	○
8.	I like control conditions	○
9.	I calculate 3 ANOVAs in my head before getting out of bed	○
10.	I could spend all day explaining statistics to people	○
11.	I like it when I've helped people to understand factor rotation	○
12.	People fall asleep as soon as I open my mouth to speak	○
13.	Designing experiments is fun	○
14.	I'd rather think about appropriate dependent variables than go to the pub	○
15.	I soil my pants with excitement at the mere mention of Factor Analysis	○
16.	Thinking about whether to use repeated- or independent-measures thrills me	○
17.	I enjoy sitting in the park contemplating whether to use participant observation in my next experiment	○
18.	Standing in front of 300 people in no way makes me lose control of my bowels	○
19.	I like to help students	○
20.	Passing on knowledge is the greatest gift you can bestow on an individual	○
21.	Thinking about Bonferroni corrections gives me a tingly feeling in my groin	○
22.	I quiver with excitement when thinking about designing my next	○

	experiment	
22.	I often spend my spare time talking to the pigeons ... and even they die of boredom	○
23.	I tried to build myself a time machine so that I could go back to the 1930s and follow Fisher around on my hands and knees licking the floor on which he'd just trodden	○
25.	I love teaching	○
26.	I spend lots of time helping students	○
27.	I love teaching because students have to pretend to like me or they'll get bad marks	○
28.	My cat is my only friend	○

17.15. Further reading

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applica
of Applied Psychology, *78*, 98–104. (A very readable paper on Cronbach's α.)

Dunteman, G. E. (1989). *Principal components analysis*. Sage University Paper Series on
Applications in the Social Sciences, 07-069. Newbury Park, CA: Sage. (This monog
high level but comprehensive.)

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design and analysis*. Hillsdale, I
(Chapter 22 is an excellent introduction to the theory of factor analysis.)

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Bos
Bacon.

18 Categorical data

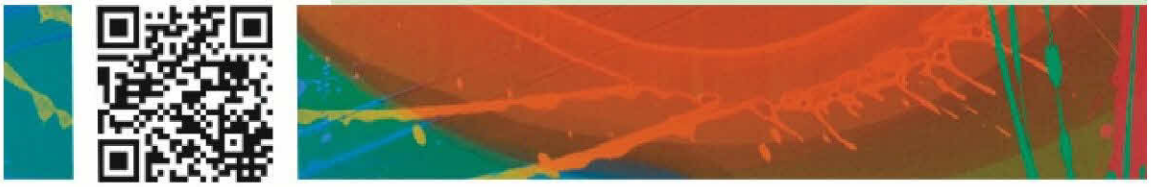


FIGURE 18.1
Midway through writing the second edition of this book, things became a little strange

18.1. What will this chapter tell me?

We discovered in the previous chapter that I wrote a book. This b
good things about writing books. The main benefit is that your
They're not *that* impressed, because they think that a good boo
as *Harry Potter* and that people should queue outside bookshops
ling instalment of *Discovering statistics* .... My parents are, cons
about how this book is seen as successful, yet I don't get invited t
Nevertheless, given that my family don't really understand what I
proof that I do *something*. The size of this book and the fact it h

added bonus because it makes me look cleverer than I actually ar
price to pay, which is immeasurable mental anguish. In England w
emotions, because we fear that if they get out into the open, civiliz
collapse, so I definitely will not mention that the writing process
was so stressful that I came within one of Fuzzy's whiskers of a t
me two years to recover, just in time to start thinking about the th
worth it because the feedback suggests that some people found th
Of course, the publishers focus less on the book's helpfulness and m
comparisons with other books. They have databases that have sal

and its competitors in different 'markets' (you are not a person, you are a 'con
you don't live in a country, you live in a 'market'), and they gibber and twitch
soles creating pink frequency distributions (with 3-D effects) of these values. T
get are frequency data (the number of books sold in a certain period of time).
if they wanted to compare sales of this book to its competitors, in different cou
would need to read this chapter because it's all about analysing data for whic
only the frequency with which events occur. Of course, they won't read this
they should ...

18.2. Analysing categorical data ①

So far we have looked at fitting models with categorical predictor variables,
predicting a continuous outcome variable. Sometimes, however, we want
categorical outcome variables. In other words, we want to predict into whi
an entity falls. For example, we might want to predict whether someone is
not, for which political party a person voted, whether a tumour is benign or
whether a sports team will win, lose or draw. In all of these cases, an entity
only one category, for example a woman can be pregnant or not; she can't be
nant'. The next two chapters deal with statistical models for categorical outc
begin with some basic models of associations between categorical variables
at predicting categorical outcomes from categorical predictors, then in the n
we'll move on to look at predicting categorical outcomes from both categoric
tinuous predictor variables.

18.3. Theory of analysing categorical data ①

We will begin by looking at the simplest situation that you could encounter; t
tifying the relationship between two categorical variables. With categorical
can't use the mean or any similar statistic because the mean of a categorica
completely meaningless: the numeric values you attach to different categories a
and the mean of those numeric values will depend on how many members ea
has. Therefore, when we've measured only categorical variables, we analyse
of things that fall into each combination of categories (i.e., the frequencies). F
a researcher was interested in whether animals could be trained to line-dan
200 cats and tried to train them to line-dance by giving them either food or
a reward for dance-like behaviour. At the end of the week he counted how m
could line-dance and how many could not. There are two categorical variables
ing (the animal was trained using either food or affection, not both) and **dance**
either learnt to line-dance or it did not). By combining categories, we end up w
ferent categories. All we then need to do is to count how many cats fall into ea

we can tabulate these frequencies as in Table 18.1 (which shows the data for th
and this is known as a **contingency table**.

18.3.1. Pearson's chi-square test ①

If we want to see whether there's a relationship between two categorical variabl
the number of cats that line-dance relate to the type of training used?) we can u
chi-square test (Fisher, 1922; Pearson, 1900). This is an extremely elegant st

TABLE 18.1 Contingency table showing how many cats will line-dance different rewards

		Training	
		Food as Reward	Affection a
Could They Dance?	Yes	28	4
	No	10	11
	Total	38	16

on the simple idea of comparing the frequencies you observe in c frequencies you might expect to get in those categories by chance (equation (2.6)) that if we want to calculate the fit (or total error) of squared differences between the observed values of the outcome, a that come from the model:

$$\text{Total error} = \sum_{i=1}^{n} \left(\text{observed}_i - \text{model}_i\right)^2$$

This equation was the basis of our sums of squares in regression a have categorical data we use essentially the same equation. There is we divide by the model scores as well, which is actually much the s the sum of squares by the degrees of freedom in ANOVA. Basically, deviation for each observation. If we add all of these standardized resulting statistic is Pearson's chi-square (χ^2) given by:

$$\chi^2 = \sum \frac{\left(\text{observed}_{ij} - \text{model}_{ij}\right)^2}{\text{model}_{ij}}$$

in which i represents the rows in the contingency table and j r The observed data are, obviously, the frequencies in Table 18.1, b what the model is. When we have categorical predictors but a co ANOVA) the model we use is group means, but, as I've mention means when we have a categorical outcome variable so we work w Therefore, we use 'expected frequencies'. One way to estimate th would be to say 'well, we've got 200 cats in total, and four cate value is simply 200/4 = 50'. This would be fine if, for example, we h cats that had affection as a reward as we did cats that had food as a 38 got food and 162 got affection as a reward. Likewise there that could and couldn't dance. To take account of these inequalit expected frequencies for each cell in the table (in this case there ar

expected frequencies for each cell in the table (in this case there ar
column and row totals for a particular cell to calculate the expecte

$$\text{model}_{ij} = E_{ij} = \frac{\text{row total}_i \times \text{column total}_j}{n}$$

In which n is the total number of observations (in this case 200).
expected frequencies for the four cells within our table (row tota
abbreviated to RT and CT, respectively):

$$\text{model}_{\text{Food, Yes}} = \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Food}}}{n} = \frac{76 \times 38}{200} = 14.44$$

$$\text{model}_{\text{Food, No}} = \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Food}}}{n} = \frac{124 \times 38}{200} = 23.56$$

$$\text{model}_{\text{Affection, Yes}} = \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Affection}}}{n} = \frac{76 \times 162}{200} = 61.56$$

$$\text{model}_{\text{Affection, No}} = \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Affection}}}{n} = \frac{124 \times 162}{200} = 100.44$$

Given that we now have these model values, all we need to do is take each v cell of our data table, subtract from it the corresponding model value, squar and then divide by the corresponding model value. Once we've done this for the table, we just add them up!

$$\chi^2 = \frac{(28-14.44)^2}{14.44} + \frac{(10-23.56)^2}{23.56} + \frac{(48-61.56)^2}{61.56} + \frac{(114-100.44)^2}{100.44}$$

$$= \frac{13.56^2}{14.44} + \frac{-13.56^2}{23.56} + \frac{-13.56^2}{61.56} + \frac{13.56^2}{100.44}$$

$$= 12.73 + 7.80 + 2.99 + 1.83$$

$$= 25.35$$

This statistic can then be checked against a distribution with known prop the **chi-square distribution**. All we need to know is the degrees of freedom, ar calculated as $(r-1)(c-1)$ in which r is the number of rows and c is the number Another way to think of it is as the number of levels of each variable minus plied together. In this case we get $df = (2 - 1)(2 - 1) = 1$. If you were doing hand, you would find a critical value for the chi-square distribution with $df =$ observed value was bigger than this critical value you would say that there was relationship between the two variables. These critical values are produced in th and for $df = 1$ they are 3.84 ($p = .05$) and 6.63 ($p = .01$). So because the ob square is bigger than these values it is significant at $p < .01$. However, SPSS v the precise probability of obtaining a chi-square statistic at least as big as (i 25.35 if there were no association between the variables in the population.

18.3.2. Fisher's exact test ①

There is one problem with the chi-square test, which is that the sampling dis
the test statistic has an *approximate* chi-square distribution. The larger the sa
better this approximation becomes, and in large samples the approximation is g
to not worry about the fact that it is an approximation. However, in small s
approximation is not good enough, making significance tests of the chi-square
inaccurate. This is why you often read that to use the chi-square test the expec
cies in each cell must be greater than 5 (see Section 18.4). When the expected
are greater than 5, the sampling distribution is probably close enough to a
square distribution for us not to worry. However, when the expected frequen

low, it probably means that the sample size is too small and that th
of the test statistic is too deviant from a chi-square distribution to b

Fisher came up with a method for computing the exact probabili
tistic that is accurate when sample sizes are small. This method is c
(Fisher, 1922) even though it's not so much a test as a way of con
ability of the chi-square statistic. This procedure is normally used
tables (i.e., two variables each with two options) and with small sa
be used on larger contingency tables and with large samples, but
tables it becomes computationally intensive and you might find S
to give you an answer. In large samples there is really no point bec
overcome the problem of small samples, so you don't need to use it

18.3.3. The likelihood ratio ②

An alternative to Pearson's chi-square is the likelihood ratio stati
maximum-likelihood theory. The general idea behind this theory i
data and create a model for which the probability of obtaining the
maximized, then you compare this model to the probability of obt
the null hypothesis. The resulting statistic is, therefore, based on c
quencies with those predicted by the model:

$$L\chi^2 = 2\sum \text{observed}_{ij} \ln\left(\frac{\text{observed}_{ij}}{\text{model}_{ij}}\right)$$

in which i and j are the rows and columns of the contingency tab
logarithm (this is the standard mathematical function that we can
and you can find it on your calculator usually labelled as ln or $\log_e$)
and observed values as in the previous section, this would give us:

$$\begin{aligned} L\chi^2 &= 2\left[28\times\ln\left(\frac{28}{14.44}\right)+10\times\ln\left(\frac{10}{23.56}\right)+48\times\ln\left(\frac{48}{61.56}\right)+1\right. \\ &= 2\left[28\times 0.662+10\times -0.857+48\times -0.249+114\times 0.127\right] \\ &= 2\left[18.54-8.57-11.94+14.44\right] \\ &= 24.94 \end{aligned}$$

As with Pearson's chi-square, this statistic has a chi-square distr
degrees of freedom (in this case 1). As such, it is tested in the sa
up the critical value of chi-square for the number of degrees of fre
before, the value we have here will be significant because it is bigge

of 3.84 ($p = .05$) and 6.63 ($p = .01$). For large samples this statistic v
as Pearson's chi-square, but is preferred when samples are small.

18.3.4. Yates's correction ②

When you have a 2 × 2 contingency table (i.e., two categorical v
categories) then Pearson's chi-square tends to produce significance v
(in other words, it tends to make a Type I error). Therefore, Yates

to the Pearson formula (usually referred to as **Yates's continuity correction**). Th is that when you calculate the deviation from the model (the observed$_{ij}$ − moc tion (18.2)) you subtract 0.5 from the absolute value of this deviation before it. In plain English this means you calculate the deviation, ignore whether it is negative, subtract 0.5 from the value and then square it. Pearson's equation the

$$\chi^2 = \sum \frac{\left(\left|\text{observed}_{ij} - \text{model}_{ij}\right| - 0.5\right)^2}{\text{model}_{ij}}$$

For the data in our example this just translates into:

$$\begin{aligned}\chi^2 &= \frac{(13.56-0.5)^2}{14.44} + \frac{(13.56-0.5)^2}{23.56} + \frac{(13.56-0.5)^2}{61.56} + \frac{(13.56-0.5)^2}{100.44} \\ &= 11.81 + 7.24 + 2.77 + 1.70 \\ &= 23.52\end{aligned}$$

The key thing to note is that it lowers the value of the chi-square statistic anc makes it less significant. Although this seems like a nice solution to the proble fair bit of evidence that this overcorrects and produces chi-square values that ar Howell (2012) provides an excellent discussion of the problem with Yates's co continuity if you're interested; all I will say is that although it's worth knowin probably best ignored.

18.3.5. Other measures of association ①

There are measures of the strength of association other than the chi-square measures modify the chi-square statistic to take account of sample size and deg dom and try to restrict the range of the test statistic from 0 to 1 (to make the the correlation coefficient described in Chapter 7). Three related measures are

- *Phi*: This statistic is accurate for 2 × 2 contingency tables. However, for more than two dimensions the value of phi may not lie between 0 and 1 chi-square value can exceed the sample size. Therefore, Pearson suggeste the contingency coefficient.
- *Contingency coefficient*: This coefficient ensures a value between 0 and 1 tunately, it seldom reaches its upper limit of 1 and for this reason Cramé alternative denoted by *V*.
- *Cramér's V*: When both variables have only two categories, phi and Cra identical. However, when variables have more than two categories Cram

can attain its maximum of 1 – unlike the other two – and so it is the mos

18.3.6. Several categorical variables: loglinear analysis

So far we've looked at situations in which there are only two categorical variable
we often want to analyse more complex contingency tables in which there are th
variables. For example, suppose we took the example we've just used but al

data from a sample of 70 dogs. We might want to compare the beha cats. We would now have three variables: **Animal** (dog or cat), **Trai** affection as reward) and **Dance** (did they dance or not?). This could Pearson chi-square and instead has to be analysed with a technique

18.3.6.1. Chi-square as regression ④

SMART ALEX ONLY

To begin with, let's have a look at how our simple chi-square exa as a regression model. Although we already know about as much a chi-square test, if we want to understand more complex situations ably easier if we consider our model as a general linear model (i.e. general linear models we've considered in this book take the gener

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

For example, when we encountered multiple regression in Chap model was written as (see equation (8.6)):

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni}) + \varepsilon_i$$

Also, when we came across one-way ANOVA, we adapted this regre tualize our Viagra example, as (see equation (11.1)):

$$\text{Libido}_i = b_0 + b_2\text{High}_i + b_1\text{Low}_i + \varepsilon_i$$

The *t*-test was conceptualized in a similar way. In all cases the same it's just the complexity of the model that changes. With categoric same model in much the same way as with regression to produce current example we have two categorical variables: training (food (yes they did dance or no they didn't dance). Both variables have we can represent each one with a single dummy variable (see Sectio category is coded as 0 and the other as 1. So for training, we coul 'affection' as 1, and we could code the dancing variable as 1 for ' Table 18.2).

This situation might be familiar if you think back to factorial AI in which we also had two variables as predictors. In that situation are two variables the general linear model became (think back to e

$$\text{Outcome}_i = (b_0 + b_1A_i + b_2B_i + b_3AB_i) + \varepsilon_i$$

TABLE 18.2 Coding scheme for dancing cats

Training	Dance	Dummy (Training)	Dummy (Dance)	Inte
Food	No	0	0	
Food	Yes	0	1	
Affection	No	1	0	
Affection	Yes	1	1	

in which *A* represents the first variable, *B* represents the second and *AB* rep interaction between the two variables. Therefore, we can construct a linear r these dummy variables that is exactly the same as the one we used for factor (above). The interaction term will be the training variable multiplied by the da (look at Section 10.3.2, and if it doesn't make sense look back to Section 13. the coding is exactly the same as this example):

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

$$\text{outcome}_{ij} = \left(b_0 + b_1\text{Training}_i + b_2\text{Dance}_j + b_3\text{Interaction}_{ij}\right) + \varepsilon_{ij}$$

However, because we're using categorical data, to make this model linear actually use log values and so the actual model becomes:[1]

$$\ln(O_i) = \ln(\text{model}) + \ln(\varepsilon_i)$$

$$\ln(O_{ij}) = \left(b_0 + b_1\text{Training}_i + b_2\text{Dance}_j + b_3\text{Interaction}_{ij}\right) + \ln(\varepsilon_{ij})$$

The training and dance variables and the interaction can take the values 0 and ing on which combination of categories we're looking at (Table 18.2). Therefo out what the *b*-values represent in this model we can do the same as we did f and ANOVA and look at what happens when we replace training and dance wi 0 and 1. To begin with, let's see what happens when we look at when trainin are both zero. This situation represents the category of cats that got food rewar line-dance. When we used this sort of model for the *t*-test and ANOVA the o used were taken from the observed data: we used the group means (e.g., see Se and 11.2.1). However, with a categorical outcome we use the observed frequer than observed means). In Table 18.1 we saw that there were 10 cats that hac reward and didn't line-dance. If we use this as the observed outcome then the be written as (if we ignore the error term for the time being):

$$\ln(O_{ij}) = b_0 + b_1\text{Training}_i + b_2\text{Dance}_j + b_3\text{Interaction}_{ij}$$

For cats that had food reward and didn't dance, the training and dance varia interaction will all be 0 and so the equation reduces to:

$$\ln(O_{\text{Food,No}}) = b_0 + (b_1 \times 0) + (b_2 \times 0) + (b_3 \times 0)$$

$$\ln(O_{\text{Food,No}}) = b_0$$

$$\ln(10) = b_0$$

$$b_0 = 2.303$$

Therefore, b_0 in the model represents the log of the observed value when all of
ries are zero. As such it's the log of the observed value of the base category (in t
that got food and didn't dance).

[1] Actually, the convention is to denote b_0 as q and the b-values as l, but I think these notational cha
to confuse people so I'm sticking with b because I want to emphasize the similarities to regression

Now, let's see what happens when we look at cats that had aff didn't dance. In this case, the training variable is 1 and the dance vari are still 0. Also, our outcome now changes to be the observed valu affection and didn't dance (from Table 18.1 we can see the value equation becomes:

$$\ln\left(O_{\text{Affection, No}}\right) = b_0 + \left(b_1 \times 1\right) + \left(b_2 \times 0\right) + \left(b_3 \times 0\right)$$

$$\ln\left(O_{\text{Affection, No}}\right) = b_0 + b_1$$

$$b_1 = \ln\left(O_{\text{Affection, No}}\right) - b_0$$

Remembering that b_0 is the expected value for cats that had foo get:

$$\begin{aligned} b_1 &= \ln\left(O_{\text{Affection, No}}\right) - \ln\left(O_{\text{Food, No}}\right) \\ &= \ln(114) - \ln(10) \\ &= 4.736 - 2.303 \\ &= 2.433 \end{aligned}$$

The important thing is that b_1 is the difference between the log of for cats that received affection and didn't dance, and the log of the that received food and didn't dance. Put another way, within the g dance it represents the difference between those trained using food affection.

Now, let's see what happens when we look at cats that had food a In this case, the training variable is 0, the dance variable is 1 and 0. Our outcome now changes to be the observed frequency for cats danced (from Table 18.1 we can see the value is 28). Therefore, th

$$\ln\left(O_{\text{Food, Yes}}\right) = b_0 + \left(b_1 \times 0\right) + \left(b_2 \times 1\right) + \left(b_3 \times 0\right)$$

$$\ln\left(O_{\text{Food, Yes}}\right) = b_0 + b_2$$

$$b_2 = \ln\left(O_{\text{Food, Yes}}\right) - b_0$$

Remembering that b_0 is the expected value for cats that had food an

$$
\begin{aligned}
b_2 &= \ln\left(O_{\text{Food, Yes}}\right) - \ln\left(O_{\text{Food, No}}\right) \\
&= \ln(28) - \ln(10) \\
&= 3.332 - 2.303 \\
&= 1.029
\end{aligned}
$$

The important thing is that b_2 is the difference between the log of the observe for cats that received food and danced, and the log of the observed frequency received food and didn't dance. Put another way, within the group of cats tl food as a reward it represents the difference between cats that didn't dance an did.

Finally, we can look at cats that had affection and danced. In this case, the dance variables are both 1 and the interaction (which is the value of training m the value of dance) is also 1. We can also replace b_0, b_1 and b_2 with what we nov represent. The outcome is the log of the observed frequency for cats that receiv and danced (this expected value is 48 – see Table 18.1). Therefore, the equati (I've used the shorthand of A for affection, F for food, Y for yes, and N for no

$$\ln(O_{A,Y}) = b_0 + (b_1 \times 1) + (b_2 \times 1) + (b_3 \times 1)$$

$$\ln(O_{A,Y}) = b_0 + b_1 + b_2 + b_3$$

$$\ln(O_{A,Y}) = \ln(O_{F,N}) + (\ln(O_{A,N}) - \ln(O_{F,N})) + (\ln(O_{F,Y}) - \ln(O_{F,N})) + b_3$$

$$\ln(O_{A,Y}) = \ln(O_{A,N}) + \ln(O_{F,Y}) - \ln(O_{F,N}) + b_3$$

$$b_3 = \ln(O_{A,Y}) - \ln(O_{F,Y}) + \ln(O_{F,N}) - \ln(O_{A,N})$$

$$= \ln(48) - \ln(28) + \ln(10) - \ln(114)$$

$$= -1.895$$

So, b_3 in the model really compares the difference between affection and foc cats didn't dance to the difference between food and affection when the cat Put another way, it compares the effect of training when cats didn't dance to training when they did dance.

The final model is therefore:

$$\ln(O_{ij}) = 2.303 + 2.433\text{Training}_i + 1.029\text{Dance}_j - 1.895\text{Interaction}_{ij} + \ln(\varepsilon$$

The important thing to note here is that everything is exactly the same ANOVA, except that we dealt with log-transformed values (compare this Section 13.2.2 to see just how similar everything is). In case you still don't this works as a general linear model, I've prepared a file called **Cat Regressior** contains the two variables **Dance** (0 = no, 1 = yes) and **Training** (0 = food tion) and the interaction (**Interaction**). There is also a variable called **Ob** contains the observed frequencies in Table 18.1 for each combination of **Training**. Finally, there is a variable called **LnObserved**, which is the natural l these observed frequencies (remember that throughout this section we've de

log observed values).

SELF-TEST Run a multiple regression analysis using **Cat Regression.sav** with **LnObserved** as the outcome, and **Training, Dance** and **Interaction** as your three predictors.

OUTPUT 18.1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	2.303	.000	
	Type of Training	2.434	.000	1.385
	Did they dance?	1.030	.000	.725
	Interaction	-1.895	.000	-1.174

a. Dependent Variable: LN (Observed Frequencies)

Output 18.1 shows the resulting coefficients table from this reg thing to note is that the constant, b_0, is 2.303 as calculated above, of training, b_1, is 2.434 and for dance, b_2, is 1.030, both of whic error of what was calculated above. Also the coefficient for the in as predicted. There is one interesting point, though: all of the star there is *no* error at all in this model (which is also why there are no is because the various combinations of coding variables completel values. This is known as a **saturated model** – I will return to this p mind. For the time being, I hope this convinces you that chi-square as a linear model.

OK, this is all very well, but the heading of this section did rat show you how the chi-square test can be conceptualized as a linear the chi-square test looks at whether two variables are independer interest in the combined effect of the two variables, only their uniq conceptualize chi-square in much the same way as the saturated don't include the interaction term. If we remove the interaction ter

$$\ln\left(\text{model}_{ij}\right) = b_0 + b_1\text{Training}_i + b_2\text{Dance}_j$$

With this new model, we cannot predict the observed values like model because we've lost some information (namely, the interactio outcome from the model changes, and therefore the beta values cha that the chi-square test is based on 'expected frequencies'. Therefor ing the chi-square test as a linear model, our outcomes will be these look back to the beginning of this chapter you'll see we already hav cies based on this model. We can recalculate the beta values based or

$$\ln\left(E_{ij}\right) = b_0 + b_1\text{Training}_i + b_2\text{Dance}_j$$

For cats that had food reward and didn't dance, the training and 0 and so the equation reduces to:

$$\ln\left(E_{\text{Food, No}}\right) = b_0 + \left(b_1 \times 0\right) + \left(b_2 \times 0\right)$$

$$\ln\left(E_{\text{Food, No}}\right) = b_0$$

$$b_0 = \ln\left(23.56\right)$$

$$= 3.16$$

Therefore, b_0 represents the log of the expected value when all of the categori

When we look at cats that had affection as a reward and didn't dance, the ti able is 1 and the dance variable is still 0. Also, our outcome now changes to be t value for cats that received affection and didn't dance:

$$\ln\left(E_{\text{Affection, No}}\right) = b_0 + (b_1 \times 1) + (b_2 \times 0)$$

$$\ln\left(E_{\text{Affection, No}}\right) = b_0 + b_1$$

$$\begin{aligned} b_1 &= \ln\left(E_{\text{Affection, No}}\right) - b_0 \\ &= \ln\left(E_{\text{Affection, No}}\right) - \ln\left(E_{\text{Food, No}}\right) \\ &= \ln(100.44) - \ln(23.56) \\ &= 1.45 \end{aligned}$$

The important thing is that b_1 is the difference between the log of the expecte for cats that received affection and didn't dance and the log of the expected va that received food and didn't dance. In fact, the value is the same as the colun that is, the difference between the total number of cats getting affection and th ber of cats getting food: $\ln(162) - \ln(38) = 1.45$. Put simply, it represents the of the type of training.

When we look at cats that had food as a reward and danced, the training and the dance variable is 1. Our outcome now changes to be the expected fr cats that received food and danced:

$$\ln\left(E_{\text{Food, Yes}}\right) = b_0 + (b_1 \times 0) + (b_2 \times 1)$$

$$\ln\left(E_{\text{Food, Yes}}\right) = b_0 + b_2$$

$$\begin{aligned} b_2 &= \ln\left(E_{\text{Food, Yes}}\right) - b_0 \\ &= \ln\left(E_{\text{Food, Yes}}\right) - \ln\left(E_{\text{Food, No}}\right) \\ &= \ln(14.44) - \ln(23.56) \\ &= -0.49 \end{aligned}$$

Therefore, b_2 is the difference between the log of the expected frequencies f received food and did or didn't dance. In fact, the value is the same as the ro that is the difference between the total number of cats that did and didn't dan $\ln(124) = -0.49$. In simpler terms, it is the main effect of whether or not the

We can double-check all of this by looking at the final cell (cats that had affection

$$\ln\left(E_{\text{Affection, Yes}}\right) = b_0 + \left(b_1 \times 1\right) + \left(b_2 \times 1\right)$$

$$\ln\left(E_{\text{Affection, Yes}}\right) = b_0 + b_1 + b_2$$

$$\ln\left(61.56\right) = 3.16 + 1.45 - 0.49$$

$$4.12 = 4.12$$

The final chi-square model is therefore:

$$\ln(O_i) = \ln(\text{model}) + \ln(\varepsilon_i)$$
$$= 3.16 + 1.45\text{Training} - 0.49\text{Dance} + \ln(\varepsilon_i)$$

We can rearrange this equation to get some residuals (the error ter

$$\ln(\varepsilon_i) = \ln(O_i) - \ln(\text{model})$$

In this case, the model is merely the expected frequencies that were square test, so the residuals are the differences between the observed ar

SELF-TEST To show that this all actually work multiple regression analysis using **Cat Regress** This time the outcome is the log of expected fre **(LnExpected)** and **Training** and **Dance** are the interaction is not included).

This section demonstrates how chi-square can work as a linear mo and ANOVA, in which the beta values tell us something about the frequencies across categories of our two variables. If nothing else to leave this section aware that chi-square (and analysis of categor be expressed as a linear model (although we have to use log value egories of a variable using dummy variables, just as we did with r and the resulting beta values can be calculated in exactly the same and ANOVA. In ANOVA, these beta values represented differences a particular category compared against a baseline category. With ca values represent the same thing, the only difference being that ra means, we're dealing with expected values. Grasping this idea (ANOVAs and categorical data analysis are basically the same) will in the next section.

EVERYBODY

18.3.6.2. Loglinear analysis ③

In the previous section, after nearly reducing my brain to even mor than it already is trying to explain how categorical data analysis regression, I ran the data through an ordinary regression using SPS talking complete gibberish. At the time I rather glibly said 'oh, by th

in the model, that's odd isn't it?' and sort of passed this off by telling rated' model and not to worry too much about it because I'd explai as I'd worked out what the hell was going on. That seemed like a g the time, but unfortunately I now have to explain what I was going

To begin with, I hope you're now happy with the idea that c expressed in the form of a linear model provided that we use log v is why the technique we're discussing is called *log*linear analysis). fully already know about ANOVA and linear models generally, yo tucked up in bed with the idea that we can extend any linear model

of predictors and any resulting interaction terms between predictors. If we ca a simple two-variable categorical analysis in terms of a linear model, then amaze you to discover that if we have more than two variables this is no probl extend the model by adding any new variables and the resulting interaction giving each one a parameter (b). This is all you really need to know. So, just as regression and ANOVA, if we think of things in terms of a linear model, then c it becomes very easy to understand how the model expands to incorporate ne So, for example, if we have three predictors (A, B and C) in ANOVA we end u two-way interactions (AB, AC, BC) and one three-way interaction (ABC). Th resulting linear model is:

$$\text{outcome}_{ijk} = \left(b_0 + b_1 A_i + b_2 B_j + b_3 C_k + b_4 AB_{ij} + b_5 AC_{ik} + b_6 BC_{jk} + b_7 ABC_{ijk}\right.$$

In exactly the same way, if we have three variables in a categorical data ana an identical model, but with an outcome in terms of logs:

$$\ln\left(O_{ijk}\right) = \left(b_0 + b_1 A_i + b_2 B_j + b_3 C_k + b_4 AB_{ij} + b_5 AC_{ik} + b_6 BC_{jk} + b_7 ABC_{ijk}\right) +$$

Obviously the calculation of beta values and expected values from the model be siderably more cumbersome and confusing, but that's why we invented comput we don't have to worry about it. Loglinear analysis works on these principles. I we've seen in the two-variable case, when our data are categorical and we inclu available terms (main effects and interactions) we get no error: our predictors c predict our outcome (the expected values). So, if we start with the most com possible, we will get no error. The job of loglinear analysis is to try to fit a sin to the data without any substantial loss of predictive power. Therefore, loglin typically works on a principle of backward elimination (yes, the same kind c elimination as in multiple regression – see Section 8.5.1.3). So we begin with t model, and then remove a predictor from the model and using this new mode our data (calculate expected frequencies, just like the chi-square test) and then s the model fits the data (i.e., are the expected frequencies close to the observ cies?). If the fit of the new model is not very different from the more complex we abandon the complex model in favour of the new one. Put another way, we term we removed was not having a significant impact on the ability of our mod the observed data.

However, we don't just remove terms randomly, we do it hierarchically. S with the saturated model and then remove the highest-order interaction, an effect that this has. If removing the interaction term has no effect on the mo obviously not having much of an effect; therefore, we get rid of it and move o any lower-order interactions. If removing these interactions has no effect th on to any main effects until we find an effect that does affect the fit of the n removed.

To put this in more concrete terms, at the beginning of the section on loglin

To put this in more concrete terms, at the beginning of the section on logline asked you to imagine we'd extended our training and line-dancing example to a sample of dogs. So, we now have three variables: **Animal** (dog or cat), **Train** affection) and **Dance** (did they dance or not?). Just as in ANOVA this results in effects:

- **Animal**
- **Training**
- **Dance**

three interactions involving two variables:

- **Animal × Training**
- **Animal × Dance**
- **Training × Dance**

and one interaction involving all three variables:

- **Animal × Training × Dance**

When I talk about backward elimination I mean that loglinear a
ing all of these effects; we then take the highest-order interaction
way interaction of **Animal × Training × Dance**) and remove it. We
without this interaction, and from the model calculate expected
the computer) then compare these expected frequencies (or mo
observed frequencies using the standard equation for the likelih
Section 18.3.3). If the new model significantly changes the likeliho
removing this interaction term has a significant effect on the fit of t
that this effect is statistically important. If this is the case then we
that we have a significant three-way interaction. We won't test an
all lower-order effects are consumed within higher-order effects. If,
three-way interaction doesn't significantly affect the fit of the mod
lower-order interactions. Therefore, we look at the **animal × traini**
training × dance interactions in turn and construct models in whi
present. For each model the computer again calculates expected val
to the observed data using a likelihood ratio statistic.[2] Again, if ar
results in a significant change in the likelihood ratio then the term is
move on to look at any main effects involved in that interaction (so
interaction is significant the computer won't look at the main effe
ing). However, if the likelihood ratio is unchanged then the analysis
interaction term and moves on to look at main effects.

I mentioned that the likelihood ratio statistic (see Section 18.3.3
model. It should be clear how this equation can be adapted to fit a
values are the same throughout, and the model frequencies are simply t
from the model being tested. For the saturated model, this statistic will
observed and model frequencies are the same so the ratio of observed t
be 1, and ln(0) = 0), but as we've seen, in other cases it will provide a
model fits the observed frequencies. To test whether a new model has
ratio, all we need do is to take the likelihood ratio for a model and su
hood ratio for the previous model (provided the models are hierarchic

$$L\chi^2_{\text{Change}} = L\chi^2_{\text{Current Model}} - L\chi^2_{\text{Previous Model}}$$

I've tried in this section to give you a flavour of how loglinear a

getting too much into the nitty-gritty of the calculations. I've trie
can conceptualize a chi-square analysis as a linear model and then re
ously told you about ANOVA to hope that you can extrapolate th
understand roughly what's going on. The curious among you migh

[2] It's worth mentioning that for every model, the computation of expected values c
more complex, the computation gets increasingly tedious and incomprehensible (a
don't need to know the calculations to get a feel for what is going on.

how everything is calculated and to these people I have two things to say: 'I c
and 'I know a really good place where you can buy a straitjacket'. If you're tha
then Tabachnick and Fidell (2007) has a wonderfully detailed and lucid cha
subject, which puts this feeble attempt to shame.

18.4. Assumptions when analysing categorical data ①

It should be obvious that the chi-square test does not rely on the assumptio
in Chapter 5 (for example, categorical data cannot have a normal sampling
because they aren't continuous). However, the chi-square test still has two
assumptions relating to (1) independence and (2) expected frequencies.

18.4.1. Independence ①

Pretty much all of the tests we have encountered in this book have made an
about the independence of residuals, and the chi-square test is no exception.
square test to be meaningful it is imperative that each person, item or entity
to only one cell of the contingency table. Therefore, you cannot use a chi-squa
repeated-measures design (e.g., if we had trained some cats with food to see if
dance and then trained the same cats with affection to see if they would dance
analyse the resulting data with Pearson's chi-square test).

18.4.2. Expected frequencies ①

With 2 × 2 contingency tables (two categorical variables both with two cat
expected values should be below 5. In larger tables, and when looking at
between three or more categorical variables (loglinear analysis, covered in Sect
the rule is that all expected counts should be greater than 1 and no more th
expected counts should be less than 5. Howell (2012) gives a nice explanation
lating this assumption creates problems. If this assumption is broken the result
reduction in test power – so dramatic, in fact, that it may not be worth bother
analysis at all.

In terms of remedies, if you're looking at associations between only two va
consider using Fisher's exact test (Section 18.3.2). With three or more variabl
linear analysis) your options are: (1) collapse the data across one of the varia
ably the one you least expect to have an effect); (2) collapse levels of one of th
(3) collect more data; or (4) accept the loss of power. If you want to collapse

(3) collect more data, or (4) accept the loss of power. If you want to collapse one of the variables then:

1. The highest-order interaction should be non-significant.
2. At least one of the lower-order interaction terms involving the variable t should be non-significant.

Let's think about our loglinear example in which we're looking at the between training (food vs. affection), whether the animal danced (yes vs. n