Table of Contents

Preface

In 1999 the Public Broadcasting System (PBS) aired a documentary entitled *Intimate strangers*. These so-called “strangers” are microbes, which can be found everywhere on earth. They play an important role in our ecological and biological systems and thus we cannot live without them, but we understand so little about them. Similarly, "Degrees of freedom" (DF) is an "intimate stranger" to statistics students. It is everywhere and quantitative researchers cannot work without it. Every quantitative-based research paper requires reporting the test statistics with the degrees of freedom, such as "*F*(DF1, DF2)," yet very few people understand what it is. Although DF is taught in almost every introductory statistics classes, many students learn the literal definition of this term rather than its deeper meaning.

Back in 1940 Walker found that very few textbooks attempted to clarify what the term “degrees of freedom” means. For mathematicians the concept is a familiar one and thus no further explanation is needed. For over six decades this polarity has not been bridged. Many elementary statistics textbooks introduce this concept in terms of the numbers that are free to vary (e.g. Howell, 1992; Jaccard & Becker, 1990; Steinberg, 2008; Weisstein, 2011). Some statistics textbooks just give the DF of various distributions (e.g. Agresti & Finlay, 1986; Moore & McCabe, 1989; Pagano, 2010). Johnson (1992) simply said that DF indicates the index number for identifying which distribution is used. The preceding explanations cannot clearly show the purpose of DF. Even some advanced statistics textbooks do not discuss the degrees of freedom in detail (e.g. Hays, 1981; Maxwell and Delany, 2003; Winner, 1985).

There are other approaches taken to present the concept of degrees of freedom, but most of them are mathematical in essence (Cramer, 1946; Galfo, 1985). While these highly mathematical explanations carry some merits, they may still be difficult and confusing to statistical students, especially in social sciences, who generally do not have a strong mathematical background. As a result, it is not uncommon that many students enrolled in advanced statistics and experienced researchers have a vague idea of the DF concept.

Failure to understand DF has several side effects. First, students and inexperienced researchers tend to misinterpret a "perfectly” fitted model or an "over-fitted" model as a good model. Second, they have a false sense of security that DF is adequate while n is large. This reflects the problem that most students and researchers fail to comprehend that DF is a function of both the number of observations and the number of parameters in one's model. Third, some students do not realize that there are exceptional cases. For example, conventional DF is no longer applicable to data analysis based on multi-stage sampling. Rather, the DF is tied to the number of strata or sampling segments.

In this book this author adopted the unpacking approach, meaning that a concept consisting of a web of other concepts will be decomposed or unpacked one by one. Some readers may wonder why some chapters have many “side trips” that are seemingly unrelated to DF. After reflecting on over a decade of teaching, research, and consulting experience, the author found that students are confused or even overwhelmed when the instructor or the writer forgot how much he has already known. Many times the professor or the writer assumes that the readers share the same information as them, and thus it is not necessary to explain rudimentary information. Although some authors have provided a meaningful definition of DF, comprehension of DF requires a web of knowledge. For example, McCollow (1992) conceptualized DF as "the dimension of the subspace containing the error vector" (p. 10). Further, "this definition displays the error vector as a bridge between the observation vector and the projection of the observation vector onto the model space" (p.10). In the initial step, one must fit the model by "constructing the orthogonal projection of Y onto the k-dimensional subspace of n-space that is dictated by the model. For simplicity, this subspace can be called the model space" (p. 8). In order to understand the preceding description, the readers must be familiar with "vector", "orthogonal", and "model space", otherwise the sentences would sound like a Martian language.   The author adopted a similar approach to McCollow's in one of the chapters, but the author will unpack these concepts step by step.

Although cognitive psychologists classify knowledge into declarative (conceptual) knowledge and procedural knowledge, this author does not see a sharp demarcation between them. Rather, it is the conviction of the author that going through the procedures could enhance our conceptual understanding (learning by doing). In retreat camps and parties, sometimes the event organizer explained the game rules over and over but the game participants still could not follow what he said. Usually I shouted, "Just play the game once or twice. And we will get the idea." By the same token, if you read a chapter several times but still cannot digest the content, go to a computer and get your hands dirty. Eventually the "hand knowledge" will be transformed into "head knowledge." To actualize this pedagogy, instructions for computing exercises are appended to the end of the book. Each exercise will take about 30 minutes or less. Although certain statistical software applications are required (e.g. SAS, SPSS, AMOS, JMP, G*Power, R…etc.), some are freeware modules that can be downloaded from the Internet, and some are standard statistical packages that are available in many university computer labs.

In summary, this book is written in lay-person terms with many "side-trips," hand-on exercises, metaphors from daily life, and concrete examples. The target audience of this book includes both novice students learning statistics and researchers who need the conceptual background information of DF, but not the mathematical and computational details. Readers who are well-versed in statistics might find some of the illustrations simplistic and even elementary. Nevertheless, these accessible examples and exercises could be helpful when a professor needs to communicate with a lay audience.

Chapter 1
DF in terms of effective sample size and necessary relationships

Toothaker (1986) explained DF as the number of independent components minus the number of parameters estimated. This approach is based upon the definition provided by Walker (1940): the number of observations minus the number of necessary relations, which is obtainable from these observations (DF = *n* - *r*). Although Good (1973) criticized that Walker's approach is not obvious in the meaning of necessary relations, the number of necessary relationships is indeed intuitive when there are just a few variables. In following illustration it is simply defined as the relationship between a dependent variable (Y) and each independent variable (X). Please keep in mind that this illustration is simplified for conceptual clarity. Although Walker regards the preceding equation as a universal rule, it is important to point out that DF = *n* - *r* cannot be applied to all situations.

No degree of freedom and effective sample size

Figure 1.1 shows that there is one relationship under investigation (*r* = 1) when there are two variables only. There is one and only one datum point in the scatterplot. The analyst cannot do any estimation of the regression line because the line can go in any direction. In other words, there isn't any useful or effective information.

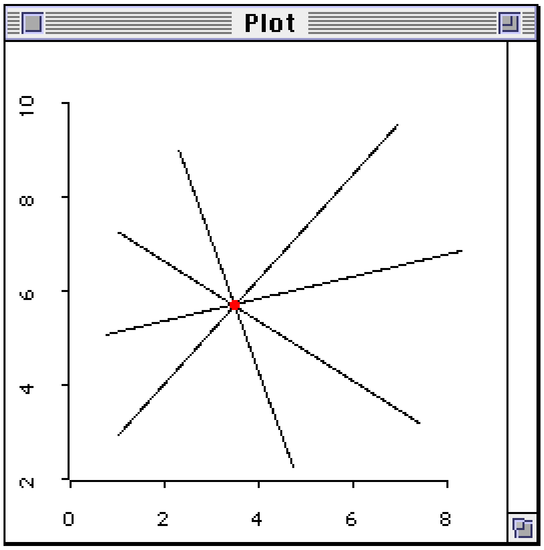


Figure 1.1. No degree of freedom with one datum point.

When the degree of freedom is zero (DF = *n* - *r* = 1 - 1 = 0), there is no way to affirm or reject the model! In this sense, there is no "freedom" to vary in the data. Freedom can be viewed as the opposite of constraint or restriction. And the constraint in this situation is absolute and you do not have any "freedom" to conduct further research with this data set. Put it bluntly, one subject is basically *ineffective*, and thus DF defines the *effective* sample size (Eisenhauer, 2008). However, zero degree of freedom is not the worst case scenario. Something even worse will be discussed in Chapter 3.

Over-determination

One can conceptualize the preceding scenario as an over-determination in a mathematical system. If there are excessive unknowns and thus there is no unique solution to the equation or the system of equations, this equation or the system is said to be over-determined. Specifically, each unknown in the equation is counted as a degree of freedom and each equation is a constraint that restricts a degree of freedom. In the above example the relationship between X and Y can be written as a simple regression equation:

$Y = a + b\text{x}$

Where
a = intercept
b = beta weight

There is one and only datum point, and thus we can substitute the numbers into X and Y:

$5 = \text{a} + \text{b}(3)$

Needless to say, it is impossible to obtain the value of the beta weight to determine the slope of the line.

Perfect fitting

In order to plot a regression line, you must have at least two data points as indicated in Figure 1.2.

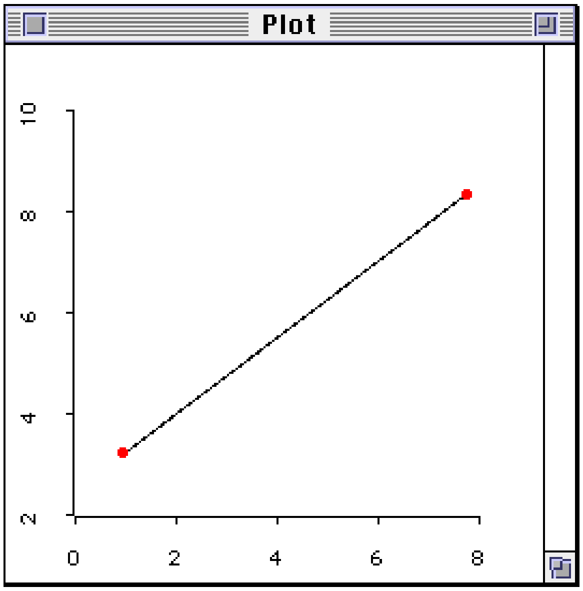


Figure 1.2. Perfect fit with two data points.

In this case, there is one degree of freedom for estimation (*n* - 1 = 1, where *n* = 2). When there are two data points only, one can always join them together to form a straight regression line and get a perfect correlation (Pearson's *r* = 1.00). Since the slope goes through all data points and there is no residual, it is considered a "perfect" fit. The word "perfect-fit" can be misleading. Novices may regard this as a good sign. Indeed, the opposite is true. When you marry a perfect man/woman, it may be too good to be true! The so-called "perfect-fit" results from the lack of useful information. Since the data points do not have much "freedom" to vary and no alternate models could be explored, the researcher has no "freedom" to further the study. Again, the effective sample size is defined by DF = *n* -1 = 1.

This point is extremely important because very few researchers are aware that perfect fitting is a sign of serious problems. For instance, during the development of the theory of evolution, one of the central questions is whether variation of a trait is inheritable. In the late 19th century Mendel gave a definite answer by introducing an elementary form of genetic theory. When Mendel conducted research on heredity, the conclusion was derived from almost "perfect" data. Later R. A. Fisher (1936) questioned that the data were too good to be true. After re-analyzing the data, Fisher found that the "perfectly-fitted" data were actually erroneous (Press & Tanur, 2001).

Another noteworthy point is that DF = 1 is detrimental to data analysis using continuous-scaled data. In categorical data analysis, such as the Chi-square test, DF = 1 is acceptable when Yates correction is employed. And thus Walker's rule should not be regarded as universal. More details on Chi-square will be discussed in Chapter 5.

Over-fitting

In addition, when there are too many variables in a regression model, i.e. the number of parameters to be estimated is larger than the number of observations, this model is said to lack degrees of freedom and thus it is over-fitted. To simplify the illustration, a scenario with three observations and two variables are presented (Figure 1.3).

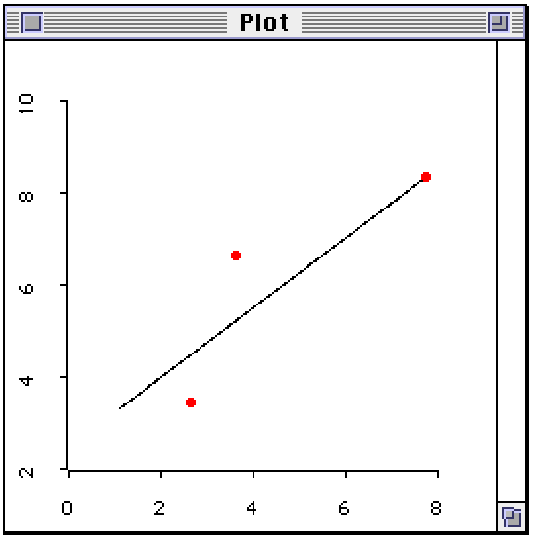


Figure 1.3. Over-fit with three data points.

Conceptually speaking, there should be four or more variables, and three or fewer observations to over-fit a model. Nevertheless, when only three subjects are used to estimate the strength of association between two variables, the situation is bad enough. Since there are just a few observations, the residuals are small and it gives an illustration that the model and the data fit each other very well. When the sample size is larger and data points scatter around the plot, the residuals are higher, of course. In this case, the model tends to have a lesser degree of fit. Nevertheless, a less fitted model ensued by more degrees of freedom carry more merits. Thus, you should see that DF consists of the number of pieces of useful information. Table 1.1 is a summary of the preceding illustration.

Table 1.1. Amount of useful information in terms of degrees of freedom

Sample size	Degree(s) of freedom	Amount of useful information
1	0	No information
2	1	Not enough information
3	2	Still not enough information

Falsifiability

To further explain why lacking useful information is detrimental to research, this author ties degrees of freedom to falsifiability. In the case of "perfect-fitting," the model is "always right." In "over-fitting," the model tends to be "almost right." Both models have a low degree of falsifiability. The concept "falsifiability" was introduced by Karl Popper (1959), a prominent philosopher of science. According to Popper, the validity of knowledge is tied to the probability of falsification. Scientific propositions can be falsified empirically. On the other hand, unscientific claims are always "right" and cannot be falsified at all. We cannot conclusively affirm a hypothesis, but we can conclusively negate it. As a theory is more specific, there is a higher possibility that the statement can be negated. For Popper, the scientific method entails "proposing bold hypotheses, and exposing them to the severest criticism, in order to detect where we have erred" (1974, p.68). If the theory can stand "the trial of fire," then we can confirm its validity. When there is no or low degree of freedom, the data could fit with any theory and thus the theory is said to be unfalsifiable.

No estimation in the population

It is noteworthy that the preceding discussion centers on the role of DF in estimation. But do we always need to perform estimations? Statistics learners can recall that the function of inferential statistics is to infer from the sample statistics to the population parameter. However, if estimation is not needed, do we still need DF? Actually DF is necessary in sample statistics only. When we obtain complete information of the population, DF becomes redundant. And that’s why DF is omitted in any formula related to the population, such as the population standard deviation formula.

Now let’s revisit some basic statistical concepts. As many readers know, there are two formulas for standard deviation (SD). One is the sample SD formula and the other is the population SD formula, as shown in the following:

Sample SD:

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{N-1}}$$

Population SD:

$$\sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{N}}$$

The following question arises in many classes: “Why do we need N – 1 to compute the sample SD, but not the population one?” As mentioned before, DF tells us our freedom or restrictions while estimating the population from the sample. By doing so, we have to reduce the total sample size to the *effective sample size*. Basically, the first piece of data points is ineffective. However, when we have access to the full population, we can obtain the exact number and thus no estimation is needed. In this case, we can simply use N instead of N – 1 to compute the *exact* SD.

What is “effective”?

Before ending the chapter, it is noteworthy that “reducing the total sample size to the effective sample size” does not mean that the researcher should exclude one subject from the analysis. After all, we cannot have the second subject without counting the first one. Many students are confused by the meaning of “effective.” Fortunately, this term is used in both statistics and engineering. Perhaps a concrete example from engineering can help explain this concept. The advertisement of a digital camera may be very boastful of the high resolution of its sensor (e.g. 12 millions). But when you read the fine print, you may see a phrase like “effective pixels: 10-million.” What does it mean? It means that only 10-million pixels on the sensor are utilized to produce the digital image, and additional 2-million pixels surrounding the effective area are used for reconstructing or interpolating the edge pixels. The image is created by the effective pixels but without the additional pixels there will be no picture at all. In other words, all pixels must be used during image processing. By the same token, “N – 1” does not imply that we toss out one observation. Rather, all subjects are used, but the result is attributed to the effective sample size.

Summary

According to Walker (1940), the degrees of freedom can be defined as the number of observations minus the number of necessary relations. This approach reduces the total sample size to the effective sample size. It tells us how many restrictions we face and how falsifiable the model is while inferring from the sample to the population. However, when the entire population becomes accessible, no estimation is needed and no model needs to be falsified. As a result, the role of DF fades away.

Chapter 2
Vector space and orthogonal projection

Degrees of freedom could be illustrated in terms of dimensionality and parameters. According to I. J. Good, degrees of freedom can be expressed as

D(K) - D(H),

whereas

D(K) = the dimensionality of a broader hypothesis,

such as a full model in regression

D(H) = the dimensionality of the null hypothesis,

such as a restricted or null model

In the next chapter, dimensionality of a model expressed as vectors will be used for illustrating df (Saville & Wood, 1991; Rawlings, 1988; Wickens, 1995). However, the vector space is not as intuitive as the subject space (e.g. plotting variable X on the X-axis and variable Y on the Y-axis), and thus it necessitates a detailed explanation. For a short moment this journey will take you away from DF, but the information presented in this chapter is essential for you to understand how DF is expressed in terms of the dimensions in vector space and the number of parameters to be estimated.

Scalar and vector

We deal with numbers every day. A mathematical object with a numeric value is called a scalar. A mathematical object that has both a numeric value and a direction is called a vector. If I just tell you to drive 10 miles to reach my home, this instruction is definitely useless. I must say something like, "From Walnut drive 10 miles North to Glendora, then turn west and drive one mile to Azusa." This example shows how essential it is to have both quantitative and directional information.

If you are familiar with computing networking, you may know that the Distance Vector Protocol is used for network routing to determine which path is the best way to transmit data. Again, the router must know two things: Distance (how far is the destination from the source?) and vector (To what direction should the data travel?)

Another example can be found in computer graphics. There is a form of computer graphics called vector-based graphics, which is used in Adobe Macromedia Flash and Paint Shop Pro. In vector-based graphics, the image is defined by the relationships among vectors instead of the composition of pixels. For example, to construct a shape, the software stores the information like "Start from Point A, draw a straight line at 45 degrees, stop at 10 units, draw another line at 35 degrees..." In short, both the scalars and vectors define the characteristics of an image.

In the context of statistical analysis, vectors help us to understand the relationships among variables. Sometimes a vector is called an eigenvector. The word “eigen,” coined by Hilbert in 1904, is a German word, which literally means "own", "peculiar", or "individual." The most common English translation is “characteristic.” An eigenvalue has a numeric property while an eigenvector has a directional property. These properties together define the characteristics of a variable, just like what happens in distance vector protocol and vector-based graphics. The original German word emphasizes the unique nature of a specific transformation in eigenvalues. Eigenvalues and eigenvectors are mathematical objects in which inputs are largely unaffected by a mathematical transformation.

Very brief overview of vector geometry and algebra

The subsequent discussion is a brief overview of vector geometry. Readers who are familiar with this subject matter can skip the following and go to the next chapter. If you want to learn from concrete examples instead of abstract mathematics, you can jump to the section “Data as matrix” in this chapter. With reference to Figure 2.1, vectors have the following properties (Hausner, 1965):

1. A vector is determined by two points. For example, if Point A and Point B are connected, it forms a vector, $\underset{AB}{\rightarrow}$.
2. Assuming $A \neq B$, $\underset{AB}{\rightarrow} = \underset{A'B'}{\longrightarrow}$ if and only if:
   (a) $\underset{AB}{\rightarrow} || \underset{A'B'}{\longrightarrow}$ (two vectors are parallel)
   (b) Length of $\underset{AB}{\rightarrow}$ = Length of $\underset{A'B'}{\longrightarrow}$
   (c) Orientation of $A \rightarrow B$ = Orientation of $A' \rightarrow B'$
3. $\underset{AA'}{\longrightarrow} = \underset{BB'}{\longrightarrow}$
4. $\underset{AB}{\rightarrow} = \underset{AB}{\rightarrow}$ (reflexivity)
5. If $\underset{AB}{\rightarrow} = \underset{A'B'}{\longrightarrow}$’ then $\underset{A'B'}{\longrightarrow} = \underset{AB}{\rightarrow}$ (symmetry)
6. If $\underset{AB}{\rightarrow} = \underset{A'B'}{\longrightarrow}$’ and $\underset{A'B'}{\longrightarrow} = \underset{A"B"}{\longrightarrow}$ then $\underset{AB}{\rightarrow} = \underset{A"B"}{\longrightarrow}$ (transitivity)

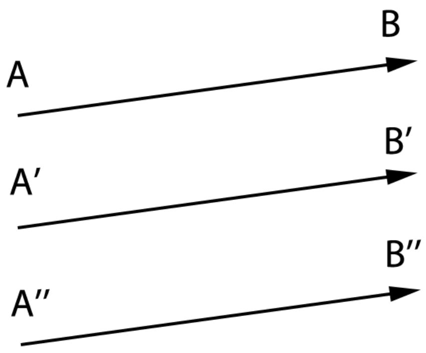


Figure 2.1. Three identical vectors

At first glance, some of the above axioms seem intuitive and redundant. However, it is important to point out that without these precise definitions, manipulation of mathematical objects would be impossible, because different people might have different ideas of how vectors should work. In the above discussion two letters are used to denote a vector. For convenience, from this point on only a single letter is used. For example, $\underset{AB}{\rightarrow}$ could be simplified as Vector *X* and $\underset{A\prime B\prime}{\longrightarrow}$ could be shortened as Vector *Y*.

With these vectors a vector space can be generated. A vector space is a non-empty set of objects, namely, vectors, which are defined on two operations, addition and multiplication by scalars (subtraction is another form of addition, which will be explained later). Vector spaces can be formed by using subsets of other vector spaces. In this case, it is called subspace. Next, we will look at how vectors "work."

Addition of vectors

Let X and Y be two vectors with the same orientation, whereas X = $\underset{AB}{\rightarrow}$ and Y = $\underset{BC}{\rightarrow}$. If two vectors are added together, it will be:

X + Y = $\underset{AC}{\rightarrow}$ (see Figure 2.2).

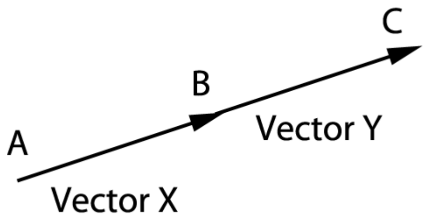


Figure 2.2. Adding two vectors with the same orientations.

When the directions of the two vectors are different, another way of addition is needed. In Figure 2.3, X and Y are two vectors with different orientations. Vector Z is generated by putting the initial point of Y at the ending point of X, and then joining the starting point of X to the ending point of Y.

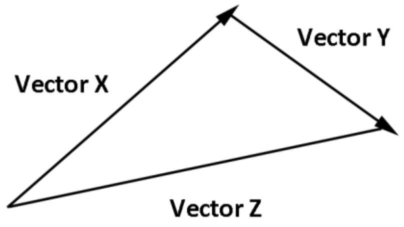


Figure 2.3 Adding two vectors with different orientations.

Subtraction of vectors

Vector subtraction can be illustrated as vector addition. For example, let X and Y be two vectors. X – Y = X + (-Y). If we reverse the direction of Y, it will be –Y. We can apply the preceding addition method, joining the head of X and the tail of –Y, to generate a new vector, Z (see Figure 2.4). Thus, vector addition/subtraction is known as the “tip to tail” method.

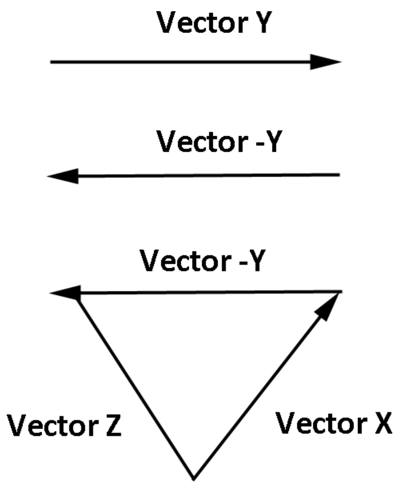


Figure 2.4. Vector subtraction as vector addition by direction reversion.

Multiplication of vectors by scalars

Any vector can be multiplied by a scalar (number) and the end product is a vector. The scalar can be viewed as a magnification factor. The following summarizes how this multiplication works:

1. If Scalar *b* is positive, Scalar *b* * Vector X becomes the vector having the same direction and orientation as Vector X, but its length is longer than the original.
2. If Scalar *b* is negative, Scalar *b* * Vector X becomes the vector parallel to the original, but with opposite direction, but the length is still longer because b is taken as an absolute value.
3. If Scalar *b* is 0, then Scalar *b* * Vector X becomes 0.

Association of vectors

Let X and Y be two vectors with the same origin at A, the association between the two vectors is expressed in terms of *Cos*(θ) whereas θ is the angel formed by the two vectors. In Figure 2.5, θ is less than 90 degree. If it is exactly 90 degree, then Vectors X and Y are said to be orthogonal. In this case, they are independent from each other and thus their association is virtually non-existent.

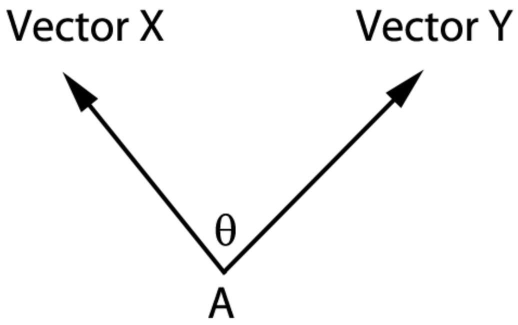


Figure 2.5. Association of two vectors in terms of *Cos*(θ)

Further discussion of vector space could get extremely complicated. Nevertheless, the purpose of this book is not to thoroughly explain the vector space. The preceding basic concepts are sufficient to carry the readers through the journey of exploring degrees of freedom in terms of dimensionality. Next, we will look at concrete examples instead of abstract mathematics.

Data as matrix

To understand the role of vectors in statistics, the data should be treated as a matrix, in which the column vector represents the subject space while the row vector represents the variable space. The function of eigenvalue can be conceptualized as the characteristic function of the data matrix. For convenience, the author will use an example with only two variables and two subjects, as shown in Table 2.1:

Table 2.1. GRE scores of two subjects

	GRE-Verbal test scores	GRE-Quantitative test scores
David	550	575
Sandra	600	580

The above data can be viewed as a matrix as the following.

$$\begin{pmatrix} 550 & 575 \\ 600 & 580 \end{pmatrix}$$

The columns of the above matrix denote the subject space, which are {550, 600} and {575, 580}. The subject space tells you that between the two subjects, David and Sandra, how GRE-Verbal and GRE-Quantitative scores are distributed, respectively. The rows reflect the variable space, which are {550, 575} and {600,580}. The variable space indicates how the scores of the subjects are distributed across the variables GRE-V and GRE-Q, respectively.

Variable space

In a scatterplot we deal with the variable space. In Figure 2.6, GRE-V lies on the X-axis whereas GRE-Q is on the Y-axis. The data points are the scores of David and Sandra. In a two data-point case, the regression line is perfect, as shown in Chapter 1.

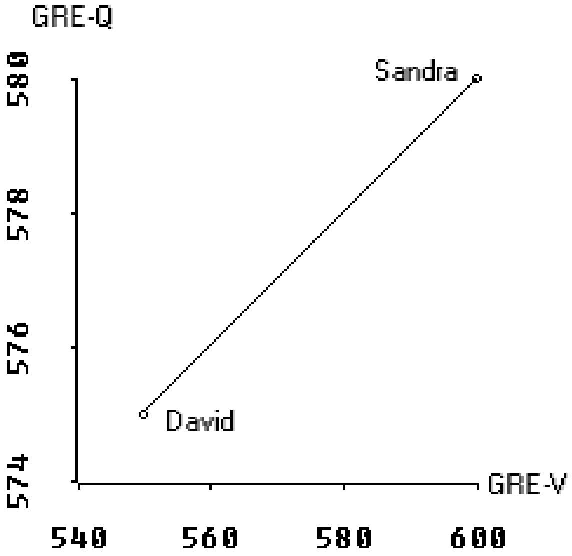


Figure 2.6. A perfectly fit regression line in the variable space

Subject space

Figure 2.7 is a plot of the subject space, also known as the vector space. In this graph the X-axis and Y-axis represent the two subjects, Sandra and David. In GRE-V David scores 550 and Sandra scores 600. A vector is drawn from 0 to the point where Sandra's and David's scores meet (the scale of the graph actually starts from 500 rather than 0 in order to make other portions of the graph visible). The vector for GRE-Q is constructed in the same manner. In reality, a research project always involves more than two variables and two subjects. In a multi-dimensional hyperspace, the vectors in the subject space can be combined to form an eigenvector, which depicts the eigenvalue. The longer the length of the eigenvector, the eigenvalue increases and the more variance it can explain. When subject space and variable space are combined, we call it “hyperspace.”

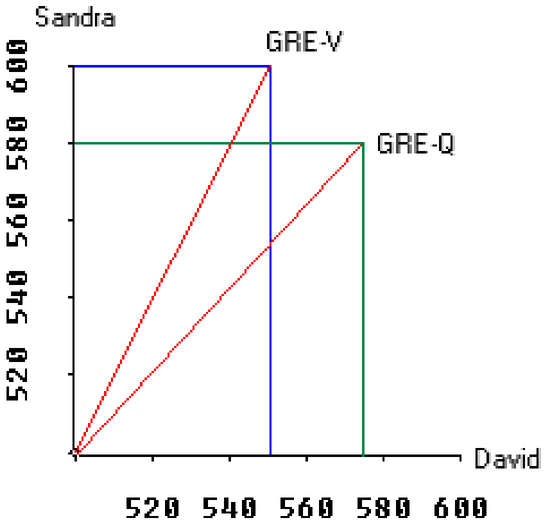


Figure 2.7. Vectors in the subject space

Proximity of vectors: Correlation between variables

Vectors can be applied into regression diagnosis and principle component analysis/factor analysis. In a regression model the independent variables should not be too closely correlated, otherwise the variance explained (R^2) will be inflated due to redundant information. This problem is commonly known as "collinearity," which means that the predictor variables are linearly dependent on each other. In this case the higher variation explained is just due to duplicated information.

For example, assume that you are questioning whether you should use GRE-V and GRE-Q together to predict GPA. In a two-subject case, you can examine the relationship between GRE-Q and GRE-V by looking at the proximity of two vectors in Figure 2.8. In the previous section, it was mentioned that the correlation between two vectors can be indicated by *Cos*(θ). Nonetheless, even if we do not compute *Cos*(θ), we can still visualize the relationship. When the angle between two vectors is large, both GRE-Q and GRE-V can be retained in the model. But if two vectors exactly overlap or almost overlap each other, then we must have a second thought of using both GRE scores as the predictors in a regression model.

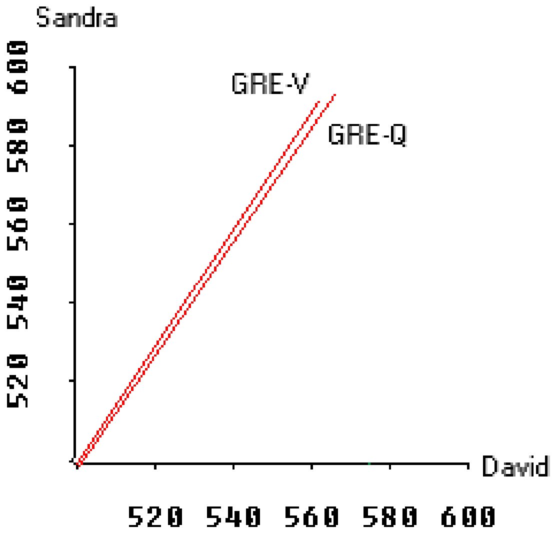


Figure 2.8. Two vectors close to each other

In the context of principal component analysis and factor analysis, vectors or eigenvectors can be used to visualize the relationships between the variables and the vectors (factors), as well as the inter-relationships between vectors. Simply put, each vector represents one *dimension*. If a test or a survey is intended to measure one single construct, then principal component analysis or factor analysis should yield a result that cluster all items of the test or the survey into one component or one factor. In this case, the scale is said to be uni-dimensional and in vector space there should be one single vector. If the vectors point to different directions and the angels between them are wide, then it is reasonable to believe that there are distinct subscales in the factor model. Figure 3.9 displays a Gabriel biplot (Gabriel, 1981) created in JMP (SAS Institute, 2011b). Biplot simply means a plot of two spaces: the subject and variable spaces. In the following example certain measures of national well-being are placed in a biplot: years of education, Gross National Income, life expectancy, life satisfaction, and happy life year. It is obvious that happy life year and life satisfaction could be collapsed as a single construct, but these two variables are far away from Gross National Income. This concurs with the conventional wisdom that money cannot buy happiness.

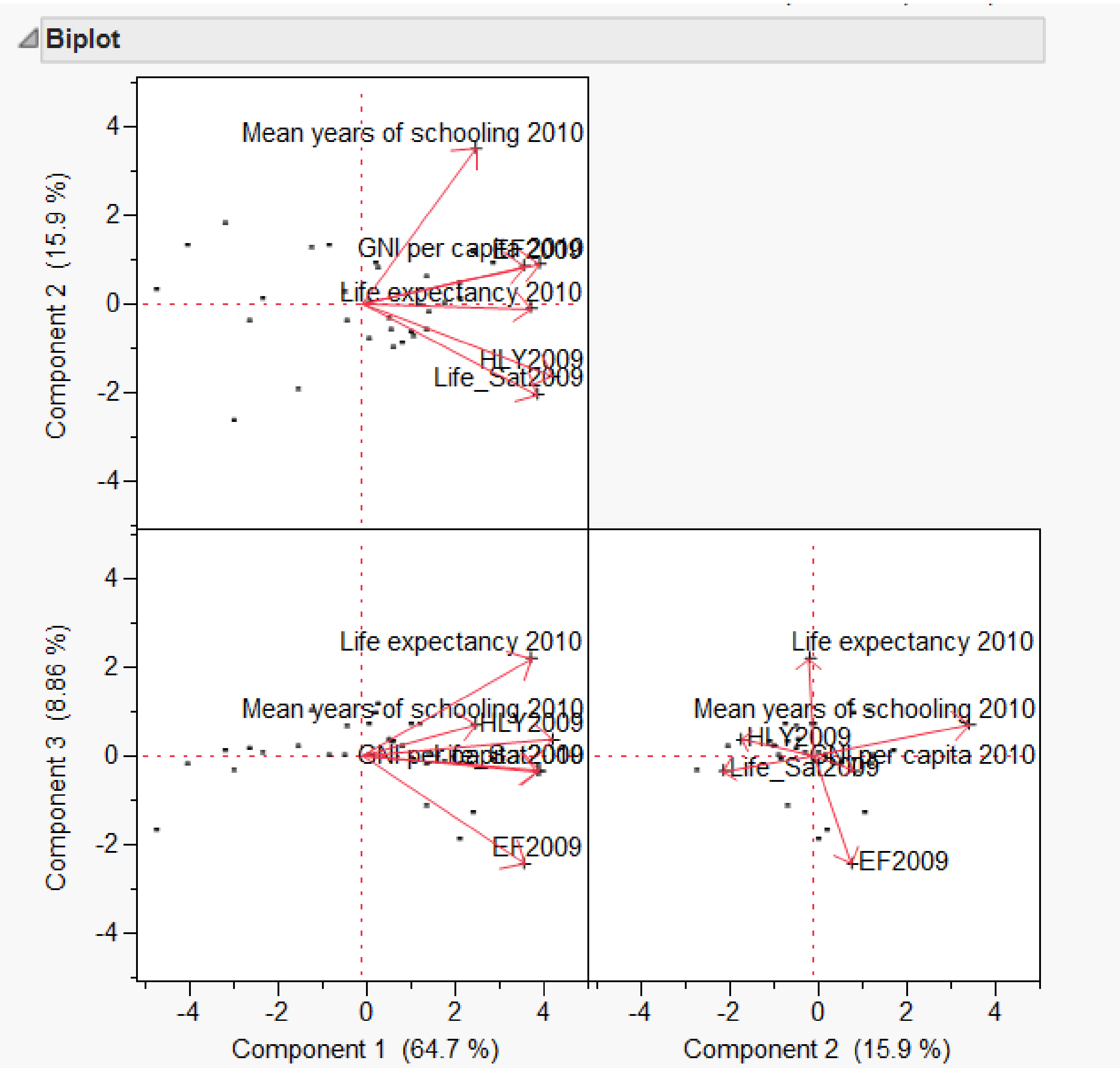


Figure 2.9. Vectors in Biplot for visualizing relationships among constructs.

What is orthogonal?

As mentioned before, if two vectors form a 90-degree angle, then they are considered orthogonal. Ideally speaking, in a two-variable regression model, the two predictors should be unrelated so that they could contribute unique variance explained to the model; otherwise, the two predictors are collinear. In the context of the subject space, the two vectors should be orthogonal (90 degree). According to Hacking (1992), orthogonality is not only a pure mathematical concept, but also a cultural concept that carries value judgment:

> Normal and orthogonal are synonyms in geometry; normal and ortho- go together as Latin to Greek. Norm/ortho has thereby a great power. On the one hand the words are descriptive. A

> line may be orthogonal or normal (at right angles to the tangent of a circle, say) or not. That is a description of the line. But the evaluative 'right' lurks in the background of right angles. It is just a fact that an angle is a right angle, but it is also a 'right' angle, a good one. Orthodonists straighten the teeth of children; they make the crooked straight. But they also put the teeth right and make them better. Orthopaedic surgeons straighten bones. Orthopsychiatry is the study of mental disorders chiefly in children. It aims at making the child-normal. The orthodox conform to certain standards, which used to be a good thing (p.163).

In the context of regression, orthogonalization can make a "good" regression model. In the subject space, "orthogonalization" can be viewed as a process of subtracting the vector from its projection. In variable space, "orthogonalization" can be explained as a process of finding the residual of the interaction term.

Figure 2.10 illustrates how a new vector, W, is made by X - Y. To subtract Y from X, as illustrated earlier, we need to reverse the orientation Y. Next, a parallel line of -Y is drawn at the end of X. Then a new vector is formed by joining the origin of X, Y and the other end of Y's parallel. In other words, subtraction creates a new vector pointing to a different direction, which is significantly far away from the original vectors. As you see, although X and Y are highly correlated, which is indicated by the small angle between the two vectors, W is uncorrelated to either X or Y. That's why vector subtraction can help to alleviate the problem of collinearity.

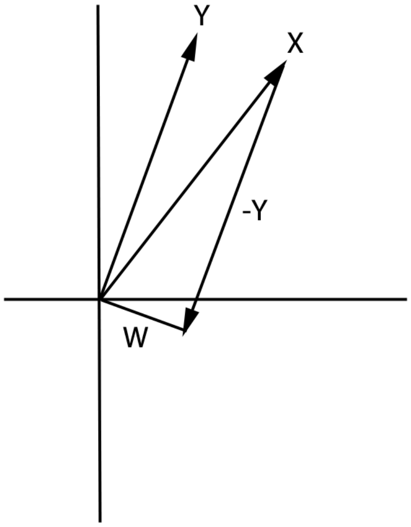


Figure 2.10. Forming orthogonal vectors by subtraction

Next, let's turn our attention to projection. Please keep in mind that the following illustration is simplified. The actual orthogonalization is not in the exact same way as described here. *Y* is omitted from the illustration because in this procedure we care about the predictors only.

We might take “psychological well-being” and “patriotism” as the independent variables, and “public opinion of the president” as the outcome variable in a regression model. Assume that all of these variables are measured in a 5-point Likert scale. With four observations, the data set shown as in Table 2.2 is obtained:

Table 2.2. A data set with an interaction term.

Observation	Well-being	Patriotism	Interaction term	Opinion of president
1	5	5	5*5=25	5
2	5	1	5*1=5	3
3	3	5	3*5=15	4
4	1	1	1*1=1	1

If we plot the data in the vector space, it will result in a graph like Figure 2.11. In Figure 2.11, X_1 and X_2 are not strongly related. You could tell by the wide angel between the two vectors. However, the product of X_1, X_2 is strongly associated with either X_1 or X_2, which is indicated by the proximity between X_1 and X_1 X_2, and between X_2 and X_1 X_2, respectively (As you notice, the product vector is longer than X_1 and X_2. In reality the interaction vector is much longer. This will be shown in the next section).

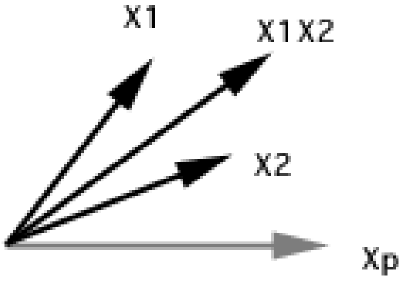


Figure 2.11. Projection of vectors in the vector space

To solve this collinearity problem, the first step is to draw a projection of X_1 X_2 vector. A projection in the subject space is equivalent to the predicted in the variable space. In Figure 2.12, X_1 X_2 is the actual vector and Xp is the predicted vector. After locating the projection, the next step is to create a new vector (new variable), which is orthogonal (not closely related) to X_1 and X_2, but is conceptually equivalent to X_1 X_2. By using the subtraction method mentioned above, we can create the new vector Xo. Xo can be viewed as a result of negotiating between what is (X_1 X_2) and what ought to be (Xp). Before orthogonalization, there exists a threat of collinearity. After orthogonalization, Xo is far away from X1 and X2 and thus collinearity is no longer a threat. The main point of this illustration is that we can construct and manipulate a model by *projecting vectors*. Later you will see why vector projection is important for us to understand the concept of DF in terms of dimensionality and parameters.

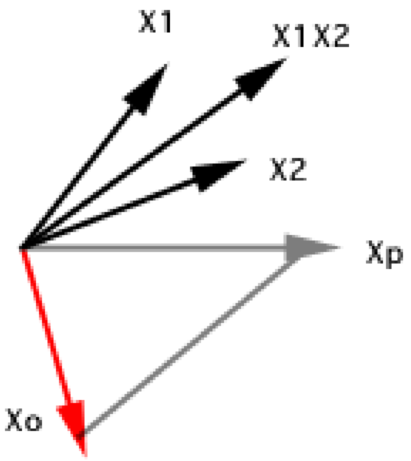


Figure 2.12. Creating a new vector in the vector space.

Orthogonal, Independence and Uncorrelated

In most cases, "orthogonality," "independence, "and "uncorrelated" are interchangeable. However, there is a slight difference among them. "Uncorrelated" is when two variables are not related and information about one of them cannot provide any information about the other. "Orthogonality" means that the two variables provide non-overlapping information. In some cases, two variables may be orthogonal but correlated. For example, let X = {-1, 1}. A new variable Y is derived by X multiplying itself (X * X), which is {-1 * -1 = 1 and 1 * 1 = 1}. Y is definitely correlated to X. But when you plot the data in the subject space, you see two orthogonal vectors as shown in Figure 2.13.

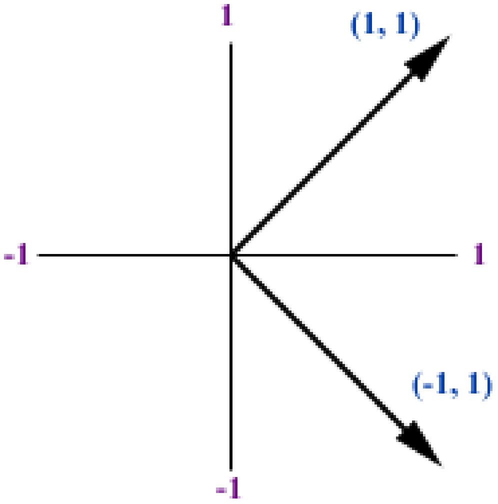


Figure 2. 13. Correlated variables but orthogonal vectors.

Rodgers, Nicewander, and Toothaker (1984) explained the difference among the three concepts using both algebraic and geometrical approaches. These explanations are complicated but the authors clearly illustrated the relationships among the three concepts in Figure 2.14.

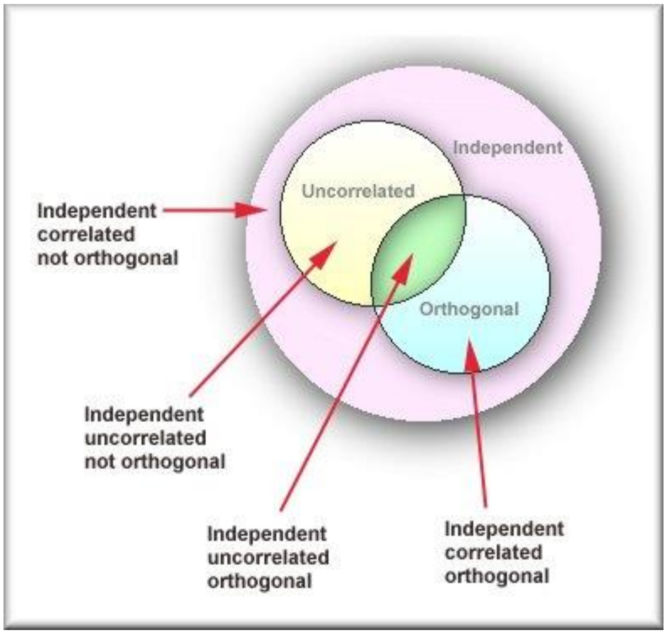

Independent
Uncorrelated
Orthogonal
Independent correlated not orthogonal
Independent uncorrelated not orthogonal
Independent uncorrelated orthogonal
Independent correlated orthogonal

Figure 2.14. Relationships among independent, uncorrelated and orthogonal variables.

Summary

Most graphical illustrations of statistical data use the variable space, but the same information can be portrayed in the subject space, also known as the vector space. A vector is composed of both numeric and directional information. Relationships between vectors could be expressed by the angles. When the angle between two vectors is 90 degree, the two vectors are said to be orthogonal. Vectors can be manipulated by addition, subtraction, multiplication, and projection. Usually a statistical model can be constructed by such manipulation. In the vector space, vectors denote the dimensionality and the relationships among variables. This background information facilitates our comprehension of Good’s approach to the degrees of freedom.

Chapter 3
DF in terms of parameters and dimensions

In this chapter, let's start with revisiting the definition of DF introduced by Good:

Df = D(K) - D(H),

whereas

D(K) = the dimensionality of a broader hypothesis,

such as a full model in regression

D(H) = the dimensionality of the null hypothesis,

such as a restricted or null model

Next, the above information will be translated into vector graphics. It is important to point out that the illustration is only a metaphor to make comprehension easier. Vectors do not behave literally as shown.

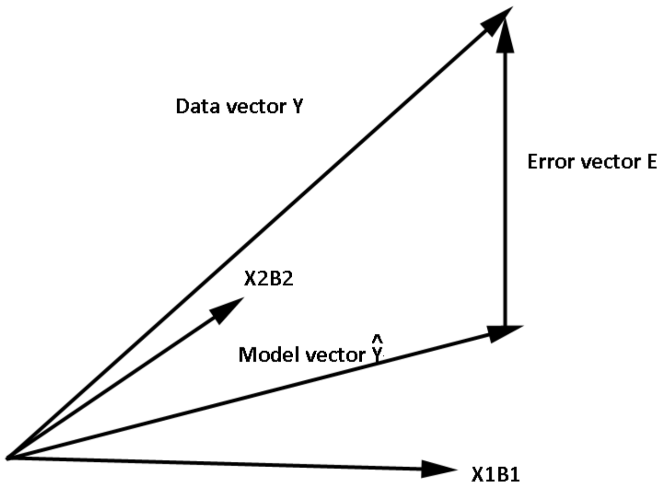


Figure 3.1. Vectors of a regression model.

Figure 3.1 depicts a regression model in the vector space. Data vector Y represents the outcome variables, which carries the observed data scores. Model vector Y denotes the predicted values of Y. Needless to say, the Error vector shows the residuals between the model and the actual data. Vectors X_1B_1 and X_2B_2 denote the dimensions of the predictors with two estimated parameters, namely, their beta weights or regression coefficients. In this example, X_1 and X_2 are two unrelated predictors and thus their vectors are said to be orthogonal (the two vectors form a 90-degree angle). In the figure it may not look orthogonal because the 3D graphic is compressed into a 2D plane.

In many typical statistics models,

Data = Model + Error

In the vector forms,

Data vector = Model vector + Error vector

As introduced in Chapter 2, adding two vectors is done by putting the vectors head to tail in sequence to create a triangle. Logically, the triangle shown in Figure 3.1 is created by this tip-to-tail method. Because the angle between the Model vector and the Error vector is exactly 90 degrees, it is said to be orthogonal.

After variables are transformed into vectors, we can visualize the dimensions of the model, and Good's approach will be comprehensible. For the time being, the intercept of the regression model will not be included. What is(are) the degree(s) of freedom when there is one variable (vector) in a regression model? First, we need to find out the number of parameter(s) in a one-predictor model. Since only one predictor is present, there is only one beta weight to be estimated. The answer is straight-forward: There is only one parameter to be estimated. How about a null model? In a null model, the number of parameters is set to zero. The expected Y score is equal to the mean of Y and there is no beta weight to be estimated.

Based upon DF = D(K) - D(H), when there is only one predictor, the degree of freedom is just one (1 - 0 = 1). This conceptualization can help us solving a mystery found in SAS (SAS Institute, 2011a). In almost all regression output tables in SAS, the user can see an unknown DF that is always set to 1. Table 3.1(a) displays the output of a regression model using two predictors only whereas Table 3.1(b) shows another output using 12 regressors. But no matter how many independent variables are involved, the same number (1) is across the entire column "DF." Actually, for this DF the null hypothesis is that there is no significant relationship between each independent variable and the outcome variable. In a pairwise perspective, there is only one parameter to be estimated. Therefore, the DF value is always set to 1 for each bivariate relationship. SAS reports this DF but SPSS does not bother to show this constant (Pandey & Bright, 2008).

Table 3.1(a). SAS regression output using two predictors.

Variable	DF	Parameter Estimate	Standard Error	*t* Value	*p* Value
2003 percentage of graduates in science	1	2033.42823	621.69946	3.27	0.0023
2003 percentage of graduates in engineering, manufacturing, and construction	1	681.85963	382.47475	1.78	0.0828

Table 3.1(b). SAS regression output using 12 predictors.

Variable	DF	Parameter Estimate	Standard Error	*t* Value	*P* Value
I usually do well in science	1	518.58730	7.18169	72.21	<.0001
I would like to take more science	1	-12.60359	1.60285	-7.86	<.0001
Science is more difficult for me	1	-1.57030	1.33396	-1.18	0.2392
I enjoy learning science	1	10.66306	1.23269	8.65	<.0001
Science is not one of my strengths	1	8.28309	1.70731	4.85	<.0001
I learn things quickly in science	1	8.96379	1.17361	7.64	<.0001
Science is boring	1	-6.11610	1.41339	-4.33	<.0001
I like science	1	-4.02016	1.14878	-3.50	0.0005
Science will help me in my daily life	1	-5.67637	1.67858	-3.38	0.0007
I need science to learn other subjects	1	0.72540	1.39843	0.52	0.6040
I need science to get into the <university> of my choice	1	8.13891	1.28825	6.32	<.0001
I need to do well in science to get the job I want	1	-13.61491	1.34709	-10.11	<.0001

The above examples have more than one predictor. But what would happen if there is one and only one predictor, which happens in a simple regression model? The implication is that there is only one piece of useful information for estimation. In this case, the model is not as well-supported as a multivariate model.

As you notice, a 2-predictor model (DF = 2 - 0 = 2) is better-supported than the 1-predictor model (DF = 1 - 0 = 1). When the number of orthogonal vectors increases, we have more pieces of *independent* information to predict Y, and thus the model tends to be more well-supported.

What does “support” mean in the context of vector space? To illustrate this, it necessitates the explanation of conditioning. Conditioning is to give initial data to express an abstract mathematical model in a specific condition (situation). When vectors are connected to form a volume, the condition of the model could be detected as to whether or not it is well-built. Figure 3.3(a) shows a cubic-like structure representing a well-conditioned model constructed of orthogonal vectors, and Figure 3.3(b) depicts a wafer-like object representing an ill-conditioned model constructed of collinear (non-orthogonal) vectors.

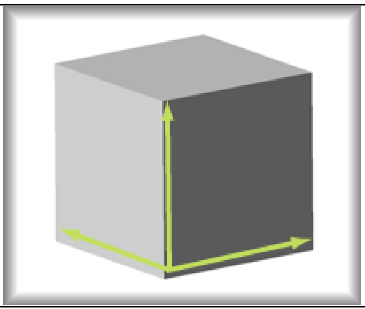


Figure 3.3(a) Well-conditioned model.

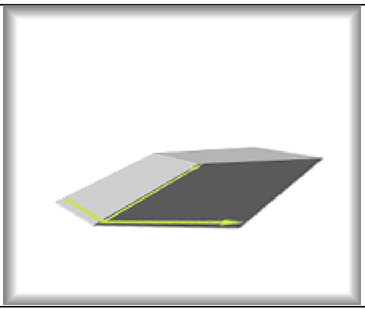


Figure 3.3(b) ill-conditioned model.

In short, DF can be defined in the context of dimensionality, which conveys the amount of *useful, independent* information. However, increasing the number of variables is not always desirable when the predictors are collinear (dependent or non-orthogonal).

Chapter 1 regarding DF as *n* - *r* mentions the problem of "overfitting," in which there are too few observations for too many variables. When you add more variables into the model, the R^2 (variance explained) will definitely increase. However, adding more variables into a model without enough observations to support the model is another way to create the problems of "overfitting." Simply, the more variables you have, the more observations you need.

It is important to note that some regression methods, such as ridge regression, linear smoothers and smoothing splines, are not based on least-squares, and thus DF defined in terms of dimensionality is not applicable to these modeling.

Chapter 1 and Chapter 3 compartmentalize DF in terms of sample size and DF in terms of dimensionality (variables). Observations (*n*) and parameters (*k*), in the context of DF, must be taken into consideration together. For instance, in regression the working definition of degrees of freedom involves the information of both observations and dimensionality: DF = *n* - *k* - 1 whereas n = sample size and k = the number of parameters. Take the 3-observation and 2-predictor case as an example. In this case, *df* = 3 - 2- 1 = 0!

In a one-way ANOVA, DF = *nk* - *k*. ANOVA aims to compare group means. If there are *k* groups, then the model will be composed of the *k* means, and there are three parameters to be estimated. If there are *n* observations in each group, then the total sample size is *nk*. It could be written as N – 1, too, where N = total sample size. In both regression and ANOVA the DF needs further partitioning, and it will be discussed in the next chapter.

Variables and parameters

It is important to point out the number of parameters is not equated with the number of variables. Let’s start with a simple example. The following is the formula of the standard error of estimate when predicting Y given X (simple regression):

$$Se = \sqrt{\frac{SSE}{N-2}}$$

where *SSE* = Sum of squared residuals

In this formula, SSE is divided by N -2 instead of N-1 even though there is one predictor only. It is because there are two parameters to be estimated, namely, slope and intercept, resulting in the deviations from the regression line with N – 2 degrees of freedom. For clarity, the illustration about the regression model in the vector space assumes that the intercept is 0. However, it could be more complicated than just including the intercept. When a non-linear model is constructed, the number of parameters to be estimated is much more than the number of variables.

Indeed, not all regression models are linear. In some situations the relationship among variables may be non-linear. A classic example is stress-performance relationship. Initially pressure could lead to better efficiency, but if the stress is too intense, performance will decrease due to physical or mental break down. Figure 3.4 is an obvious example.

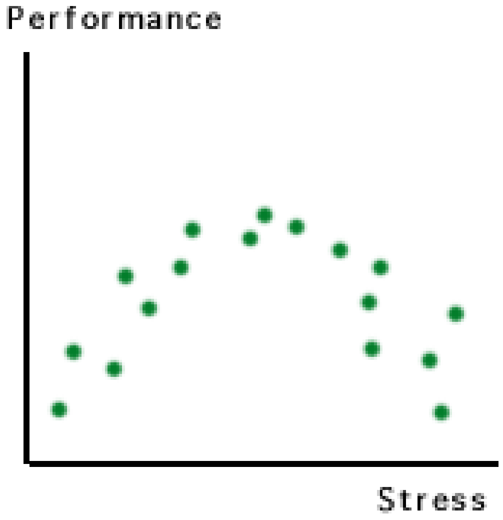


Figure 3.4. Stress-performance relationship as a non-linear model.

Another classic example is the relationship between performance and ability. Contrary to the popular belief, increasing ability in a discipline or a specific task does not lead to a linear increase in performance. Many teachers are frustrated with the phenomenon that many low achievers do not show improvement in test scores despite tremendous efforts contributed by both teachers and students. It is because low-ability learners do not have the required skills to perform even the basic function. Once they master the basic skills, their performance gain would be proportional to their ability gain. The curve hits an inflection point and turns virtually flat again when ability has matured. For example, the score difference in a writing test between a master and a Ph.D. may be minimal. The technical term for this Sshaped curve is *ogive*. Figure 3.5 depicts this relationship

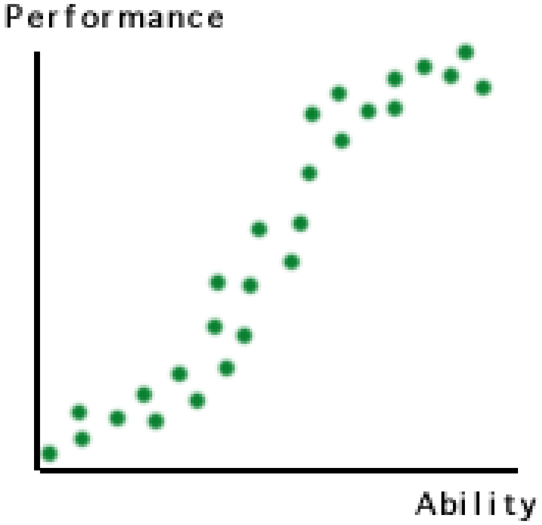


Figure 3.5. Ability-performance relationship in a non-linear model.

In curvilinear cases, polynomial regressions, which involve quadratic, cubic, or quartic terms, should be implemented. The equations of polynomial regressions are listed in the following:

Quadratic: $Y = A + B_1X + B_2X^2$

Cubic: $Y = A + B_1X + B_2X^2 + B_3X^3$

Quartic: $Y = A + B_1X + B_2X^2 + B_3 X^3 + B_4X^4$

Which term should be used depends on the number of "turns" (inflection points) on the non-linear curve. In case 1 there is only one turn on the curve and a quadratic term should be used. In case 2 there are two inflection points and thus a cubic term should be applied. Assume that the intercept is zero. In the quadratic regression model, there are two parameters to be estimated (B_1 and B_2). In the second equation there are three beta weights and in the final one there are four. Obviously, the number of parameters could be more than the number of variables.

Parameters in confirmatory factor analysis

Confirmatory factor analysis (CFA) is another way to illustrate the relationships among the degrees of freedom, the number of variables, and the number of parameters. Factor analysis is a measurement model, which specifies the relationship between observed items and latent factors. For example, when a psychologist wants to study the causal relationships between anxiety and job performance, first he/she has to define the latent constructs “anxiety” and “job performance.” To accomplish this step, the psychologist needs to develop observed items that measure the mental construct. The relationship between the abstract construct and the observed items can be expressed in terms of equations. In the context of CFA, finding a solution to these equations is called identification. When there are more unknown parameters than the number of equations, this situation is called under–identification. For example, given the equation X+Y=2, this equation may yield infinite sets of solutions i.e. (X=1, Y=1), (X=3, Y=–1), (X=2, Y=0) . . . etc. When there is just enough information to get a value for every parameter, the model is said to be just–identified. When there are more equations than unknown parameters, the model is considered over–identified. With over–identification there will be many solutions, but one can select the best or optimal solution. The following discussion will focus on the issue of under-identification. Consider the following factor model (Statistical support, 2001) (see Figure 3.6)

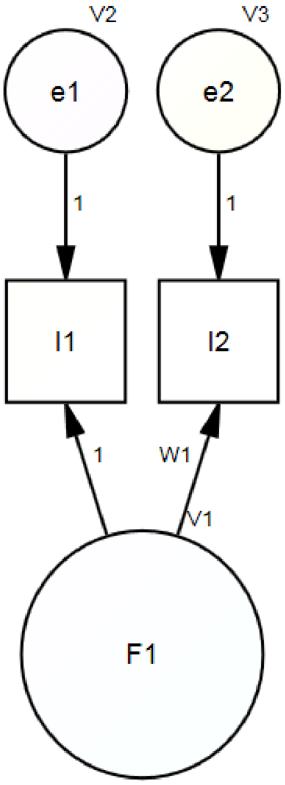
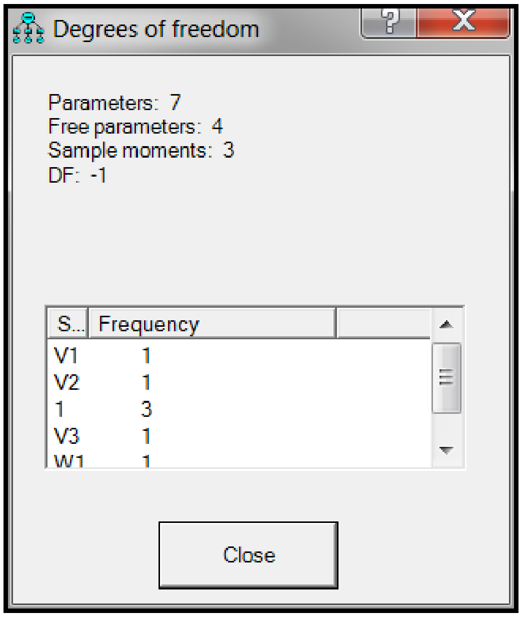


Figure 3.6. A latent factor model with two observed items.

The model has one latent factor (F1) and two observed items (I1 and I2). All measurement models contain some measurement errors, and therefore there are two error variances, namely, e1 and e2, associated with the model. Although in this factor model there are two measured variables (I1 and I2), four parameters will be estimated: the factor's variance (V1), the two error variances (e1 and e2), and one factor loading (W1), which is the relationship between the factor and I2. There is a formula to indicate the input variables, also known as sample moments: $(p(p + 1)) / 2$ where p=the number of measured variables. We can be certain that there are three input variables or distinct elements, resulted from (2(2+1))/2=3. Can we estimate four parameters given three inputs or unique elements? We cannot, of course. In this case, the DF is 3-4 = -1 (Statistical support, 2001). This calculation can be conceptualized as the number of independent components minus the number of parameters estimated (Toothaker, 1986), as mentioned at the beginning of Chapter 1. And the result can be verified by a CFA software application named AMOS (IBM SPSS, 2011a) (see Figure 3.7).



Parameters: 7
Free parameters: 4
Sample moments: 3
DF: -1

S...	Frequency
V1	1
V2	1
1	3
V3	1
W1	1

Figure 3.7 Degrees of freedom computed by AMOS.

It is worse than the scenario of zero degree of freedom mentioned in the first chapter. The example of CFA can be boiled down to one phrase: insufficient information. The number of parameters is not equated with the number of variables, and thus Walker's concept of necessary relationships must also be taken into account while addressing the issue of DF. In brief, without sufficient degrees of freedom, estimating parameters becomes a mission impossible. CFA is a part of structural equation modeling. The role of DF in SEM will be further discussed in Chapter 6.

Summary

According to Good, the degrees of freedom could be conceptualized as the difference between the dimensionality of a broader hypothesis and the dimensionality of the null hypothesis. The dimensionality of a model could be visualized in the vector space. In order to construct a well-supported and well-conditioned model, the researcher needs independent information from the predictors, meaning that the predictor vectors should be orthogonal. Again, DF tells us the number of useful, independent pieces of information in modeling. When Walker's and Good's approaches are integrated together, the freedom or restriction in statistic is tied to both the sample size and the number of parameters to be estimated. But the number of parameters is not the same as the number of variables. In a complicated model, such as a non-linear regression model, the number of parameters could far exceed the number of variables. In a factor model, the number of parameters is always more than the number of measured variables. To reduce the complexity of a regression model, degrees of freedom contribute information to model pruning, and this will be discussed in the next chapter.

Chapter 4
DF, restrictions, and penalty

Tension between simplicity and fitness

Freedom is the opposite of constraints. In other words, degrees of freedom indicate how the model should be restricted by the effective sample size and the number of independent parameters. But there is a stronger word than "constraint": penalty. Built upon the concept of DF, we can even relate DF to reducing the number of parameters and model complexity. Specifically, how a complex model should be penalized is tied to the degrees of freedom.

Researchers always face the tension between parsimony (simplicity) and fitness (complexity). When the model is so simple that it is involved with only a few variables or parameters, the model may not correspond to the phenomenon in the real world. On the other hand, a complex model, which includes many variables or parameters, may fit the data (the observed phenomenon) very well. However, the model may be too complicated to be useful (Yu, 2010). In order to obtain the optimal balance, different approaches and criteria have been developed, such as the Root Mean Square Error (RMSE), the Mallow's Cp, and the Akaike information criterion (AIC) (Akaike, 1973).

AIC is a fitness index for trading off the complexity of a model against how well the model fits the data. To reach a balance between fitness and parsimony, AIC not only rewards goodness of fit, but also gives a penalty to over-fitting and complexity. Hence, the best model is the one with the lowest AIC value. Since AIC attempts to find the model that best explains the data with a minimum of free parameters, it is considered an approach favoring simplicity. The AIC can expressed as follows (Upton & Cook, 2002):

$$AIC = 2k – 2Ln(L)$$

where
k = the number parameters to be estimated
L = the maximized value of the likelihood function (G2 - 2v)

where
G2 = the likelihood-ratio goodness-of-fit statistic
v = the number of degrees of freedom associated with the model.

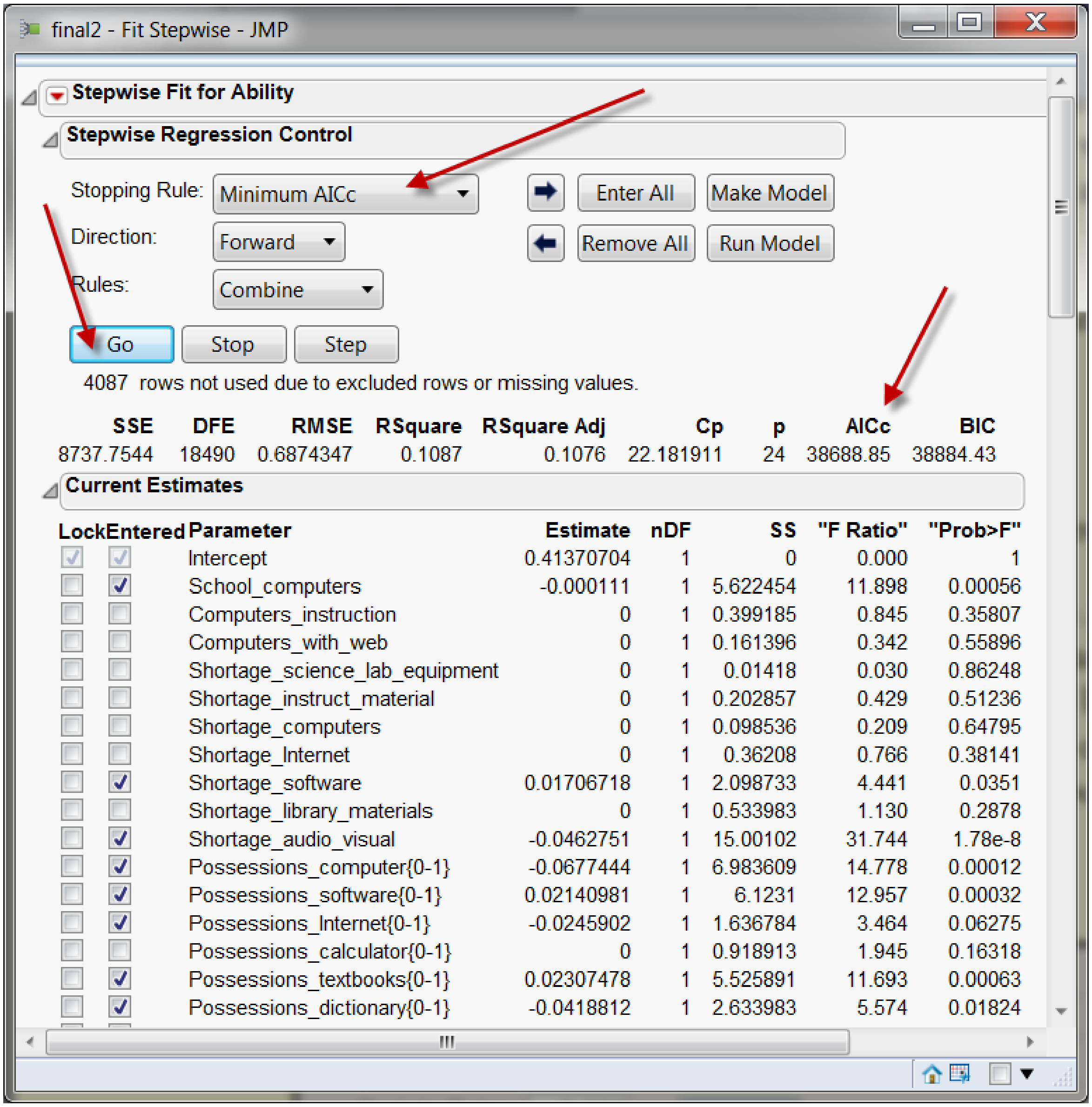
AICc is a further step beyond AIC in the sense that AICs imposes a greater penalty for additional parameters. The formula of AICs is:

$$AICc = AIC + (2K(K+1)/(n-k-1))$$

where
n = sample size
k = the number of parameters to be estimated.

If you look closely at the above formula, you can see that AICs take (n-k-1) into account, and this is the degrees of freedom in a regression model. AICc could be conceptualized as the converse of DF. While a large value of DF is more desirable, a smaller AICc is favored by researchers. Nonetheless, both of them are just two sides of the same coins, because the information from both can be used to prevent the model from over-expansion by imposing restrictions and even penalties on the model.

Figure 4.1 is a screenshot of the stepwise regression output created in JMP (SAS Institute, 2011b). In the section "Stepwise regression control," the following numbers are shown in tandem: Sum of squared errors (SSE), degree of freedom of the errors (DFE), R-square, adjusted R-square, Mallow's Cp, AICs, and Bayesian information criterion (BIC). When all these pieces of information are taken into consideration simultaneously, the role of degrees of freedom becomes clear.



Stepwise Fit for Ability

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

Rules: Combine

4087 rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
8737.7544	18490	0.6874347	0.1087	0.1076	22.181911	24	38688.85	38884.43

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	0.41370704	1	0	0.000	1
□	☑	School_computers	-0.000111	1	5.622454	11.898	0.00056
□	□	Computers_instruction	0	1	0.399185	0.845	0.35807
□	□	Computers_with_web	0	1	0.161396	0.342	0.55896
□	□	Shortage_science_lab_equipment	0	1	0.01418	0.030	0.86248
□	□	Shortage_instruct_material	0	1	0.202857	0.429	0.51236
□	□	Shortage_computers	0	1	0.098536	0.209	0.64795
□	□	Shortage_Internet	0	1	0.36208	0.766	0.38141
□	☑	Shortage_software	0.01706718	1	2.098733	4.441	0.0351
□	□	Shortage_library_materials	0	1	0.533983	1.130	0.2878
□	☑	Shortage_audio_visual	-0.0462751	1	15.00102	31.744	1.78e-8
□	☑	Possessions_computer{0-1}	-0.0677444	1	6.983609	14.778	0.00012
□	☑	Possessions_software{0-1}	0.02140981	1	6.1231	12.957	0.00032
□	☑	Possessions_Internet{0-1}	-0.0245902	1	1.636784	3.464	0.06275
□	□	Possessions_calculator{0-1}	0	1	0.918913	1.945	0.16318
□	☑	Possessions_textbooks{0-1}	0.02307478	1	5.525891	11.693	0.00063
□	☑	Possessions_dictionary{0-1}	-0.0418812	1	2.633983	5.574	0.01824

Figure 4.1. Stepwise regression output in JMP

As more and more predictors are included into the regression model, the R^2 is getting higher and higher regardless of the meaningfulness of the variables. Thus, the adjusted R^2 compensates for the problem of too many parameters. RMSE is a measure of the lack of fit while Mallow's CP is the total squared errors. The lower the RMSQ and Cp are, the better the model is. Like AIC, BIC also uses a penalty against complexity, but this penalty is much stronger than that of the AIC. However, some authors believe that AIC and AICc are superior to BIC for a number of reasons. First, AIC and AICc are based on the principle of information gain. Second, the Bayesian approach requires a prior input but usually it is debatable. Third, AIC is asymptotically optimal in model selection in terms of the least squared mean error, but BIC is not asymptotically optimal (Burnham & Anderson, 2004; Yang, 2005).

Nonetheless, the preceding indicators share a common thread. The number of parameters must be taken into account in order to find the optimal balance between over-fitting and under-fitting, though the preference is placed on simplicity. Simply put, the degrees of freedom contribute information to model pruning (e.g. AICc).

Partitioning degrees of freedom

In the previous chapter the author deliberately hides some information from the readers in order to avoid confusion. Now let's reveal it. In the JMP output "the degrees of freedom" is shown as DFE, which means "the degrees of freedom of errors." Actually, in virtually every statistical model, DF can be partitioned into the degrees of freedom of model (DFM) and the degrees of freedom of errors (DFE). In regression, DF is partitioned as the following. It is clear that AICs utilizes DFE only, and JMP also reports DFE.

DF total = n – 1
DF model = k
DF error = n – k – 1

In the context of ANOVA,

DF total = $nk - 1$
DF model = $k - 1$
DF error = $(nk - 1) - (k - 1) = k(n - 1)$

To unpack this, we can start with viewing statistics as a process of separating the signal from the noise (Mathews, 2004). The signal extracted from the data is called a model and whatever unexplained variation is considered noise. Take ANOVA as an example, the purpose of ANOVA is to compare group mean differences. Suppose that there are three groups, namely A, B, and C. But all members of each group have the same score (A = 118, B = 76, C = 55). In other words, there is no within-group variation. We can easily conclude that A is better than B, and B is better than C. Is it possible to see a result like this? Yes. According to Consumer Reports (2012), the fuel efficiency of Nissan Leaf is measured as 118 mile-per-gallon (MPG), the score of Chevrolet Volt is 76 MPG, and the score of Toyota Prius Four is 55 MPG. Could we conclusively assert that Nissan Leaf is by far the most fuel-efficient vehicle, Chevrolet Volt is the second best, and Toyota Prius Four is outperformed by Nissan and GM. We may expect just a very slight variation, such as plus and minus 2 MPG. If I purchased a Nissan Leaf from the Los Angeles Nissan dealer and you bought the same car from the Phoenix dealer, but your vehicle outperforms mine by 10 MPG, I would file a complaint to the Tokyo headquarter! In engineering any within-group variation is considered a serious error, and it must be reduced or even eliminated at all cost.

But variations are expected in any studies with human subjects, such as assessment scores. As these within-subject variations may overlap each other, the "noisy" within-group variation muddles the between-group difference. In this case, the group differences are no longer clear-cut. Thus, we need to construct a model to estimate *k* means and the group differences. In other words, the model is the signal that we extract from the data, and the within-group variability is considered noise or error. The rationale of partitioning DF is based on the idea that we want to filter noise from signal, or to construct a model by filtering errors.

The DFs in a regression model can be conceptualized and partitioned in the same fashion. If all data points fall on top of the regression lines and thus there is no residual, then the model is simply the observed, or the predicted is the same as the actual. In this sense, individual deviations from the model are considered "errors" or noise. Again, it rarely happens in social sciences using human subject data. Hence, we follow the same line of reasoning to partition DF into DFM and DFE.

As mentioned before, when using DF, we are concerned with the restrictions. Given the restrictions, we want to know how well we can obtain a valid model. This concern is addressed by DFE, not DFM. Therefore, the focal interest of AICs is put on DFE. Specifically, given the noise-level, AICc computes how much penalty the model should be received.

Summary

Degrees of freedom should not be treated as a standalone concept. It is more meaningful when we place it in a broader context, along with R^2, adjusted R^2, Mallow's Cp, AICs, and BIC. There is always a tension between building a parsimonious model and a fit model. Different criteria, such as R^2, adjusted R-square, Mallow's Cp, AICs, and BIC, have been used by researchers to optimize models. Degrees of freedom, as an index of constraints, contribute information to AICc, which go one step further than DF by imposing penalties for extra parameters to be estimated. AICc utilizes the degree of freedom of errors, which is one of the three components of DF total. Because our interest concentrates on how noise restricts our ability in model building, AICc uses DFE to determine the severity of penalties.

Chapter 5
DF in the context of Chi-square analysis

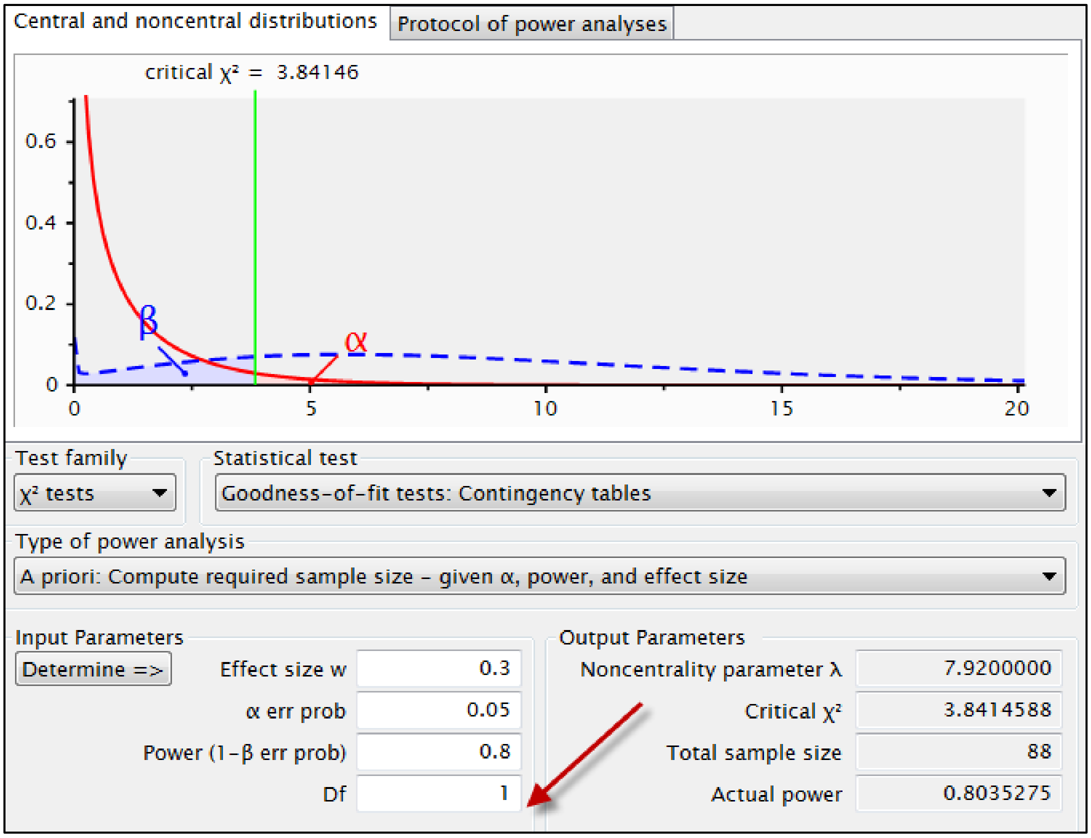
“Degrees of freedom” is a complicated concept because on some occasions, such as Chi-square analysis, the DF is tied to the number of dimensions and the number of categories in each dimension, not the number of observations (sample size). The Chi–squared procedure is a test of goodness of fit between the expected and the observed frequency of categorical data in a table. It is a well-known fact that significance of Chi-square test results is affected by the sample size (Besag, 1980). But it is also noteworthy that Chi-square’s significance also depends on the degrees of freedom. And different variants of Chi-square have different ways of calculating DF. In this chapter the relationships between DF and different data structures in Chi-square analysis will be discussed. But before going into the procedural aspect of Chis-square analysis, we will look at the historical root of the Chi-square test. During the development of this procedure there were controversies and debates. At the end the issue was resolved by adjusting the degrees of freedom.

History of Chi-square

Some students misperceive that Chi-square is a model-free approach because it is a non-parametric procedure, which is equated with distribution-free methods (Marascuilo & McSweeney, 1977). Indeed, there is a Chi-square sampling distribution. Although the original version of Chi-square is empirical-based, Chi-square is indeed distribution-dependent with regard to power. Figure 5.1(a)-(c) display the power analysis results when DF values range from 1 to 30 (Buchner, Faul, & Erdfelder, 2012). The sampling distributions of Chi-square change according to the varying DF values. Readers who are interested in the relationship between DF and Chi-square distribution are encouraged to read Chapter 10.

The model-based or distribution-based character of Chi-square is attributed to the British scientist R. A. Fisher (1922) though the original Chi-square test was invented by another British scientist Karl Pearson (1900). E. S. Pearson (1938), son of Karl Pearson, praised the Chi–square test as “a powerful new weapon in the hands of one who sought to battle with the myths of a dogmatic world” (p. 31). Pearson presented the Chi–square distribution in place of the normal distribution to solve the goodness of fit for multinomial distributions. In Pearson’s view, there is no “true” chi–square in the Platonic or absolute sense, and the so–called “true” chi–square cannot be estimated even if it exists. Rather, the focal point of the Chi–square test is the exact frequency of the sample and thus there is no probabilistic property in the data.

In 1913 Pearson and his assistants (Elderton et al., 1913) published a paper regarding the association between family size and the place in birth order using the Chi-square-based reasoning. This paper was a response to a hot topic of that era: nature vs. nurture. The position taken by Pearson is eugenics. He believed that fertility and big family size was a consequence of low social value; eventually this trend, if unchecked, could lead to de-evolution of the national population (Stiger, 1992). Research utilizing Chi-square was considered a tool to expose the problem.


Central and noncentral distributions
Protocol of power analyses
critical χ² = 3.84146
0.6
0.4
0.2
0
0
5
10
15
20
β
α
Determine =>

Test family: χ² tests

Statistical test: Goodness-of-fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size w: 0.3

α err prob: 0.05

Power (1–β err prob): 0.8

Df: 1

Output Parameters

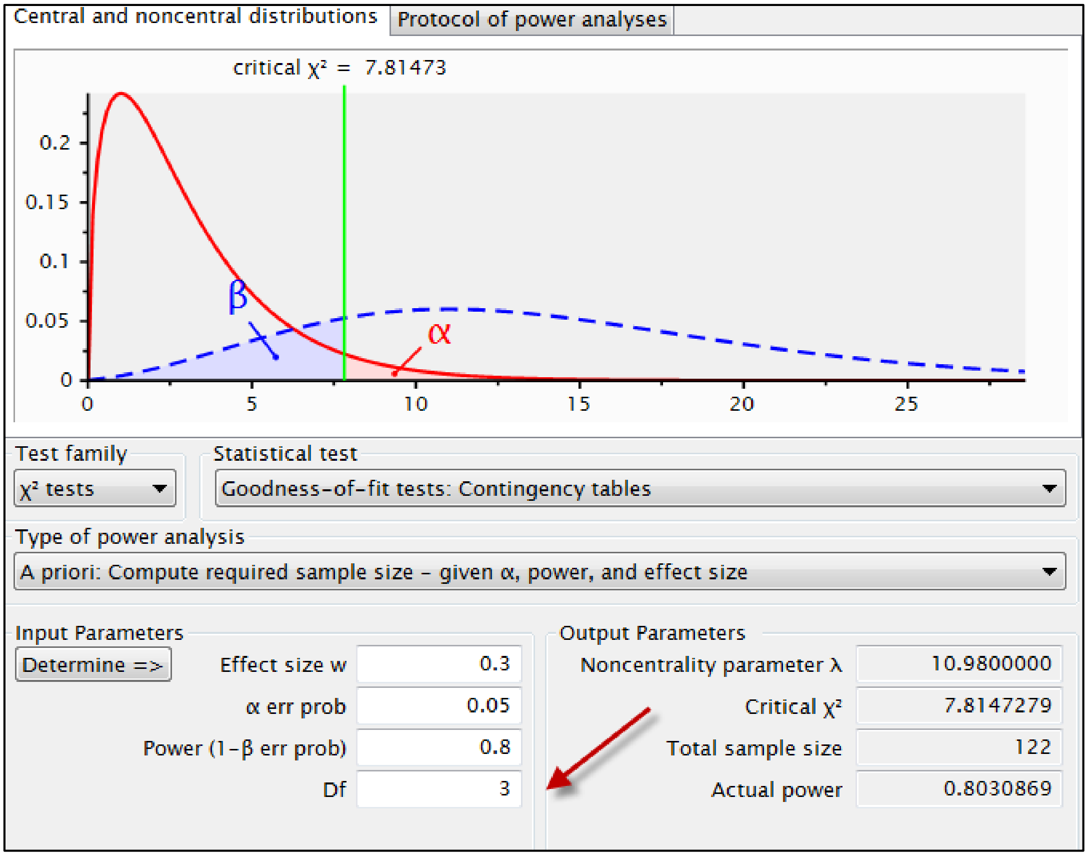
Noncentrality parameter λ: 7.9200000

Critical χ²: 3.8414588

Total sample size: 88

Actual power: 0.8035275

Figure 5.1(a). Power analysis for Chi-square analysis when DF=3.


Central and noncentral distributions
Protocol of power analyses
critical χ² = 7.81473
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
β
α
Determine =>

Test family: χ² tests

Statistical test: Goodness-of-fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size w 0.3

α err prob 0.05

Power (1–β err prob) 0.8

Df 3

Output Parameters

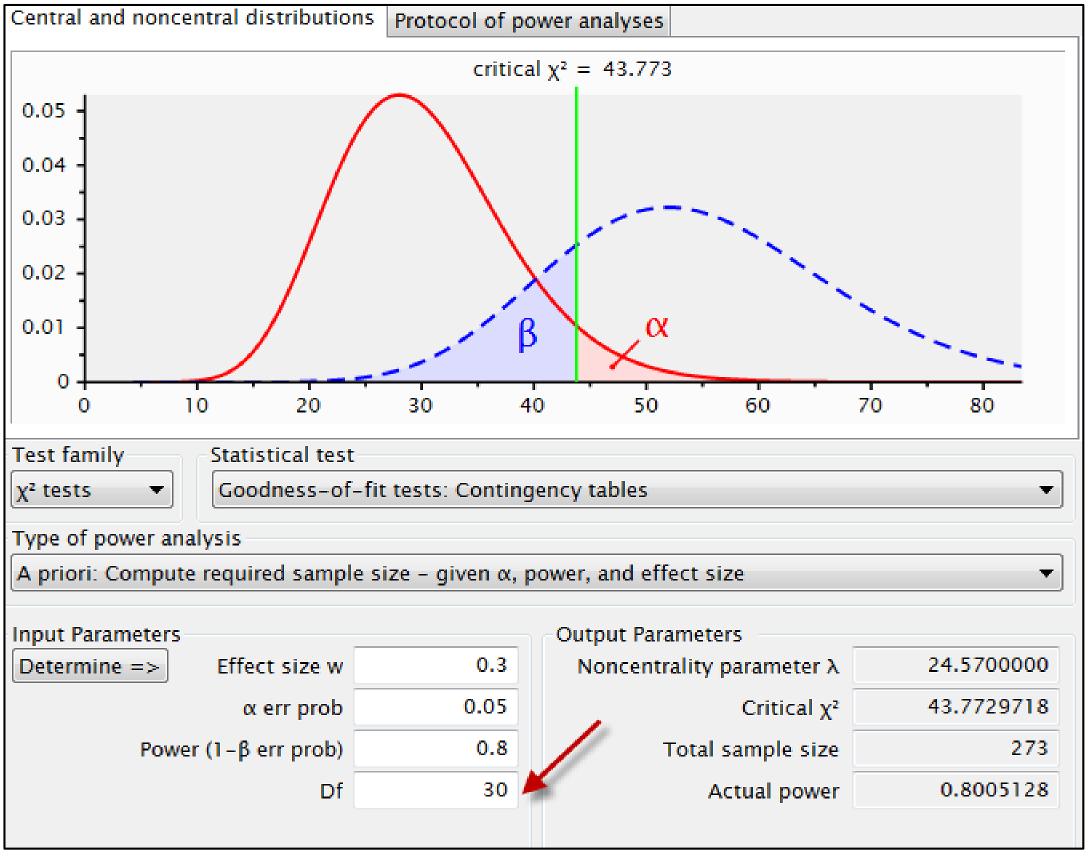
Noncentrality parameter λ 10.9800000

Critical χ² 7.8147279

Total sample size 122

Actual power 0.8030869

Figure 5.1(b). Power analysis for Chi-square analysis when DF=3.



Test family: χ² tests

Statistical test: Goodness-of-fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size w	0.3
α err prob	0.05
Power (1–β err prob)	0.8
Df	30

Output Parameters

Noncentrality parameter λ	24.5700000
Critical χ²	43.7729718
Total sample size	273
Actual power	0.8005128

Figure 5.1(c). Power analysis for Chi-square analysis when DF=30.

In many parametric tests (e.g. *t*-test, *F*-test) the user must look up the critical value in the distribution tables based upon the degrees of freedom associated with the statistics. However, Pearson disliked this model-based approach. To be specific, for Pearson the so–called probabilities associated with the test do not represent a model–based attribute such as the frequency of incorrectly rejecting the hypothesis. Rather, it is just a convenient way to describe the fit between the hypothesis and the data (Baird, 1983). Pearson emphasized that the model must fit the data, but not the other way around. When data seemed to contradict calculation, he doubted the mathematics (Porter, 2004). However, Greenwood and Yule (1915) realized that Pearson's Chi-square was limited by a lack of compensation for estimated parameters.

Pearson's model-free reasoning was questioned by his contemporary R. A. Fisher. Fisher was not opposed to the use of Chi–square. Indeed he applied this to expose the errors made by Gregor Mendel, the father of genetics (Press & Tanur, 2001; Fisher, 1936). The clash between Fisher and Pearson on Chi–square happened when Fisher introduced "degrees of freedom" to modify Chi–square in 1922. Fisher argued that in terms of causal explanation every free parameter reduces one degree of freedom. Pearson was strongly opposed to Fisher's criticism (Baird, 1983).

Fisher was not the only one that was critical of Pearson's Chi-square. In 1927 and 1929 Harris and his colleagues published two papers highlighting the limitations of Pearson's coefficient of contingency. They bluntly asserted that an extremely naive use of Pearson's Chi-square would produce misleading conclusions. In 1930 Pearson replied to the challenge by saying "I venture to think that my friend Professor Harris has overlooked the essential nature of contingency…I do not think those methods are at all applicable…I do not think they form any real limitation to the use of contingency methods" (cited in Stiger, 1992, p.573). Simply put, Pearson rejected their suggestions and insisted upon the validity of the original Chi-square analysis.

In contrast, Fisher's criticism was well–taken by Yule (1922). He attributed Pearson's stubbornness to his personality, an unwillingness to admit errors. But Porter (2004) argued that perhaps there is also something in Pearson's attitude that reflects a long standing notion of favoring data at hand over model. Obviously, Pearson was on the wrong side of history. Chi–square analysis is now applied as Fisher argued it ought to be (Baird, 1983). The degree of freedom is a measure of the informativeness of a hypothesis. Today for detecting misfits in Item Response Theory, it is a common practice to divide the Chi–square by degrees of freedom (χ^2/DF), which will be explained in a later section of this chapter.

More importantly, Fisher's interpretation of Chi–square represents a very different philosophy from Pearson's. As mentioned before Pearson did not accept the notion of true Chi–Square; the meaning of "fit" between the expected and the observed, to him, was nothing more than constructing a convenient model to approximate the observed frequencies in different cells of a contingency table. However, to Fisher a true Chi–square could be obtained even when expected cell frequencies must be estimated. In this sense, the meaning of "fit" is the closeness to the truth of a hypothesis or a model (Baird, 1983). This brief review of the history of Chi-square conveys an important message: the concept "degrees of freedom" transforms Chi-square analysis from a mere classification to a model-based approach.

Chi-square significance in one-dimensional classification

In previous chapters we learn that DF = 1 is not desirable because there is insufficient information to support a meaningful estimation. However, in a one-way Chi-square test it is not common to see DF = 1. When this situation occurs, the *Yates correction for continuity* must be employed as a remedy. The one-dimensional chi-square analysis is also known as one-way classification. The aim of the analysis is to examine whether the numbers of observations that belong to two or more categories happen by chance alone or meet the expected pattern. One may argue that when there are only two possible outcomes, the binomial distribution is a better way to conceptualize the problem. Thus, usually many instructors use rolling die (six possible outcomes) to explain one-dimensional Chi-square. Nonetheless, for clarity the following illustration is based on a two-outcome example.

If a researcher flips a coin 100 times and the coin is believed to be a fair one, by chance alone the ideal outcomes should be 50 heads and 50 tails. However, in reality sampling fluctuations may occur and thus the outcomes may be 51:49, 52:48, 54:46, and so on, instead of exactly 50:50. Hence, a tantalizing question arises: at what point could we still accept that the coin is fair? Is it acceptable at 60:40, 61:39, or even more asymmetrical outcomes? The computational procedure is simple, but the tricky part is about the degrees of freedom. As mentioned before, the DF in Chi-square tests is not tied to the number of observations, and thus in this example the DF is not 100-1=99. Rather, in a one-way classification the DF is: the number of categories in the dimension – 1. A dimension could be conceptualized as an array in which the categories are located. In this analysis there is only one dimension with two possible outcomes (Head or tail). Hence, DF = 2 – 1 = 1.

The formula of computing the Chi-square is as follows:

$$\chi^2 = \Sigma((\text{Observed frequency} - \text{Expected frequency})^2/\text{Expected frequency}))$$

When DF = 1, the Yates correction is added into formula as follows:

$$\chi^2 = \Sigma((\text{Observed frequency} - \text{Expected frequency} - 0.5)^2/\text{Expected frequency}))$$

Table 5.1. Yates correction when DF = 1.

	Observed	Expected	(Expected-Observed-0.5) ^{2}	χ^2
Head	56	50	(56-59)^2=12.25	0.245
Tail	46	50	(46-50)^2=20.25	0.405
				0.65

Table 5.1 illustrates the computational steps of Yates correction. Let's start with 56 heads and 46 tails. For the outcome of the head, the difference between the expect count and the observed count is 4. After inserting the Yate's correction (-0.5) into the formula, the adjusted difference turns into 3.5 and the squared adjusted difference is 3.5*3.5=12.25. The Chi-square value for the head is the squared adjusted difference divided by the expected frequency, which is 12.25/50=0.245. The Chi-square value for the tail is obtained by the same method, which is 0.405. The sum of both is 0.65. To determine the *p* value, one can look up the Chi-square critical value from a table or running the analysis in a statistical software package, such as SAS or SPSS. Table 5.2 is a summary of SAS (SAS Institute, 2011a) output for different combinations of head and tail outcomes.

Table 5.2. SAS one-dimensional Chi-square analysis output of head-tail combination with 100 trials.

Heads	Tails	DF	χ^2	*p* value
54	46	1	0.65	0.4237
55	45	1	1.00	0.3173
56	44	1	1.44	0.2301
57	43	1	1.96	0.1615
58	42	1	2.56	0.1096
59	41	1	3.24	0.0719
60	**40**	1	4.01	**0.0455**
61	39	1	4.84	**0.0278**
62	38	1	5.76	**0.0164**
63	37	1	6.76	**0.0093**
64	36	1	7.84	**0.0051**
65	35	1	9.00	**0.0027**

Interesting enough, the threshold is at the combination of 60 heads and 40 tails. Before 60:40 all other combinations yield a *p* value larger than .05, meaning that the null hypothesis is not rejected. In other words, there is no significant difference between the expected frequency and the observed frequency, and thus a slight deviation from 50:50 does not necessarily imply that the coin is not fair. But when the ratio becomes more imbalanced (beyond 59:41), the fairness of the coin is questionable. It is important to point out that in all of the preceding Chi-square tests the DF value remains the same (DF = 1).

Even if the number of trails is reduced to 10, the DF is not affected. What would happen if the number of trials is 10 instead of 100? Based on the preceding result, one may draw the conclusion that the threshold happens at the ratio of 6:4. If the number of coin-flipping decreases to 10 and the head-tail ratio is 6 to 4, the result should be the same. The answer may shock you. Table 5.3 shows the results of head-tail combinations from 6:4 to 9:1.

Table 5.3. SAS one-dimensional Chi-square analysis output of head-tail combination with 10 trials.

Heads	Tails	DF	χ^2	*p* value
6	4	1	0.4	0.5271
7	3	1	1.6	0.2059
8	2	1	3.6	0.0578
9	1	1	6.4	**0.0114**

Actually when the frequency of heads is 6 and the frequency of tails is 4, the *p* value is found to be 0.5271, which is not significant. With ten trials only, the one-dimensional Chi-square test refuses to declare that the coin is unfair unless the ratio is as extreme as 9 to 1. Although this is counter-intuitive, it is still possible that in a short run even a fair coin could produce very imbalanced results. Thus, one may not be able to tell whether a coin is unbiased or not by ten trials only. According to the law of large numbers, the average of the results obtained from a large number of trials should approximate the expected value. Obviously, the results yielded from 100 trials of coin-tossing would be more informative.

Many authors (e.g. Besag, 1980) pointed out that the Chi-square test is sensitive to a large sample size and thus a significant Chi-square test result based upon a large n may be meaningless. But the coin-flipping example indicates that on some occasions a large number of observations is needed in order to obtain a meaningful result.

More importantly, in both sets of trials, the DF is 1 no matter if 10 or 100 observations are used. In Chapter 1 DF is explained as the number of pieces of useful information, but in one-dimensional Chi-square DF does not contribution much useful information to the model. Rather, the usefulness and meaningfulness of the test is tied to the sample size.

Chi-square significance and the numbers of row and columns

It is more common for researchers to run two-dimensional Chi-square tests. As the name implies, in this test there are two dimensions: one row and one column. Thus, this data structure is also known as a crosstab table. In the following an example from psychometrics will be used to illustrate the role of DF in a two-dimensional Chi-square test.

Psychometrics is a discipline aiming to validate instruments so that the assessment tool can measure what it intends to measure. For example, if the objective of the test is to measure mathematics skill but some items are written as a passage that requires reading skills to comprehend the questions, the psychometrician might detect that the response pattern of those items by the students does not fit into the overall pattern of the entire test. Similarly, if a test developer attempts to create an exam pertaining to American history, but accidentally an item about European history is included in the exam, again it is expected that the response pattern for the item on European history will substantially differ from that of other items. In psychometrics there is a specific term for this type of mistakes: Misfit.

In classical test theory, misfits are typically detected by either point-biserial correlation or factor analysis. In Item Response Theory (IRT) it is identified by examining the fit (or misfit) indices. Two of the popular fit indices are infit mean square and outfit mean square. The mean square is the Chi-square divided by the degrees of freedom. Because the focus of the chapter is not psychometrics, the meanings of infit and outfit will not be discussed here. Readers who are interested in psychometrics are encouraged to consult Yu (2012a).

According to Weiss (1968), the degrees of freedom used with a Chi-square statistics is equal to the number of independent components that entered into the calculation. Each cell in the Chi-square statistics represents a single component. For an independent component, both observed and expected values are not determined by the frequencies of other cells. Thus, in a crosstab table one row and one column are fixed while all the rest are free to vary. Therefore, the DF for a Chi-square test is obtained by (the number of row - 1) X (the number of column - 1) or (r - 1)(c - 1).

Another way to conceptualize the DF value for a two-way Chi-square starts from the fact that a two-way crosstab table has r rows and c columns and thus the number of all cells in the table is r X c. If the count of all but one cell is known, the frequency of the last cell will have no freedom to vary. When the analyst wants to estimate the values of the cells, the overall restriction is $rc – 1$. The null Chi-square hypothesis is that there is no relationship between the row variable and the column variable. Alternatively, it could be expressed by the independence between the row and the column. Under the assumption of independence, the analyst estimates $(r - 1) + (c - 1)$. Taking all of the above into account:

$$\begin{aligned} \text{DF} &= (rc - 1) – [(r - 1)+(c - 1)] \\ &= rc – 1 – r + 1 – c + 1 \\ &= \text{rc} – r – c +1 \\ &= (r – 1)(c -1) \end{aligned}$$

Table 5.4 is a 3X3 crosstab table showing the number of correct and incorrect answers to an item categorized by the skill level of test takers. At first glance this item seems to be problematic because while only 10 skilled test-takers were able to answer this item correctly, 15 less skilled test-takers answered the question correctly. Does this mean the item is a misfit? To answer this question, we will break down how chi-square is calculated. The following illustration is for novices only. If you already know the computational procedure of Chi-square, please skip the illustration and directly go to the discussion on DF.

Table 5.4. 3X3 table of answer and skill level.

	More skilled (theta > 0.5)	Average (theta between -0.5 and +0.5)	Less skilled (theta < -0.5)	Row total
Answer correctly (1)	10	5	15	30
Answer incorrectly (0)	5	10	5	20
Column total	15	15	20	Grand total: 50

Like many other statistical tests, we address this issue by starting from a null hypothesis: There is no relationship between the skill level and test performance. If the null hypothesis is true, then what percentage of less skilled students would you expect to answer the item correctly? Regardless of the skill level, 30 out of 50 students could answer the item correctly, and thus the percentage should be 30/50 = 60%.

Table 5.5. 3X3 table showing one expected frequency and one actual frequency.

	More skilled	Average	Less skilled	Row total
Answer correctly (1)	10	5	**15 (20*60%=12)**	30
Answer incorrectly (0)	5	10	5	20
Column total	15	15	**20**	Grand total: 50

Because 20 students are classified as low skilled and if 60% of them can answer the item correctly, then the expected count (E) for students who gave the right answer belong to the low skilled group is 12 (20 X 60%) (see Table 5.5). In Table 5.6, the number inside the bracket is the expected count assuming the null hypothesis is correct.

Table 5.6. 3X3 table showing two expected counts and two actual counts.

	More skilled	Average	Less skilled	Row total
Answer correctly (1)	10	5	15 (12)	30
Answer incorrectly (0)	5	10	5 (8)	20
Column total	15	15	20	Grand total: 50

You may populate the entire table using the preceding approach, but you can also use a second approach, which is a short cut found by using the following formula:

Expected count= [(Column total) X (Row total)]/Grand total

For example, the expected count cell of (less skilled, answer correctly) is: 20 X 20/50 = 8.

Table 5.7. 3X3 table showing all expected counts and all actual counts.

	More skilled	Average	Less skilled	Row total
Answer correctly (1)	10 (9)	5 (9)	15 (12)	30
Answer incorrectly (0)	5 (6)	10 (6)	5 (8)	20
Column total	15	15	20	Grand total: 50

Table 5.7 shows the expected count in all cells. You can see that there is a discrepancy between what is expected (E) and what is the observed (O) in each cell. To measure the fit between the E and O, we use the formula:

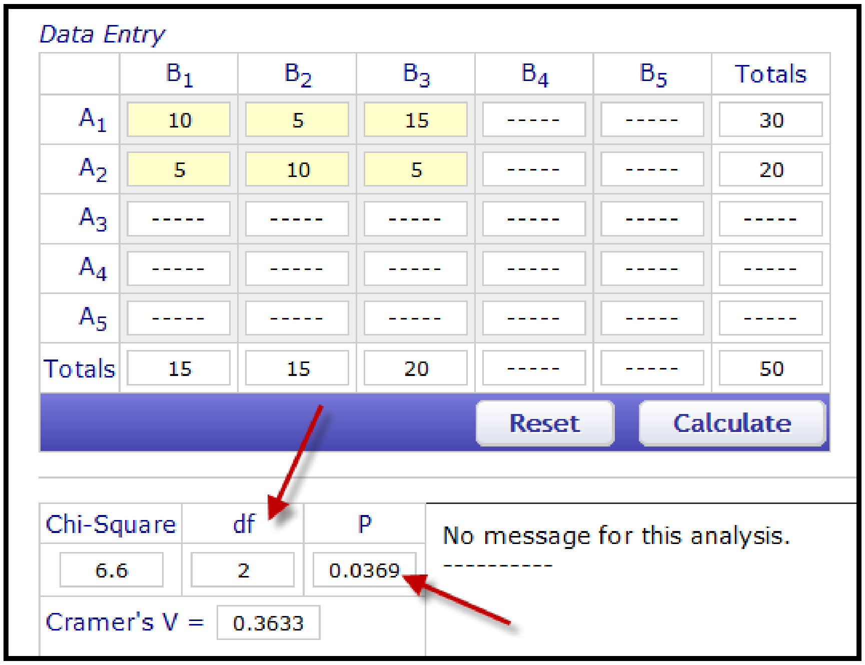
$(O-E)^2 / E$

For example, for the cell (more skilled, answer correctly): $(10-9)^2 / 9 = 0.111$

Table 5.8. 3X3 table showing all expected counts and all actual counts.

	More skilled	Average	Less skilled
Answer correctly (1)	0.111	1.778	0.750
Answer incorrectly (0)	0.167	2.667	1.125

Table 5.8 shows the computed Chi-square in all cells. The number in each cell indicates the value of the discrepancy (residual). The bigger the number is, the worse the discrepancy is. The sum of all $(O-E)^2 / E$ is called the Chi-square, which is the sum of all residuals that shows the overall discrepancy. If the Chi-square is big, it indicates that the item is likely to be a misfit. With the advent of computer technology, today the analyst could save time from performing hand calculation by employing statistical software packages, such as SAS and SPSS. Alternatively, one can simply use an online Chi-square calculator, such as the one at http://vassarstats.net/newcs.html. Figure 5.2(a) is the output of computing the preceding data set using this online tool. As mentioned earlier, the degrees of freedom for a Chi-square test is (r – 1) X (c – 1). In this example DF = (2 - 1)*(3-1) = 2. When the data are configured as a 2X3 table (competency is divided into three levels), the DF is 2 and the p value is 0.0369, which is considered significant. The mean square in this example is the chi-square divided by the degrees of freedom, which is 6.6/2 = 3.3.



Data Entry

	B_1	B_2	B_3	B_4	B_5	Totals
A_1	10	5	15	-----	-----	30
A_2	5	10	5	-----	-----	20
A_3	-----	-----	-----	-----	-----	-----
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	15	15	20	-----	-----	50

Chi-Square	df	P
6.6	2	0.0369

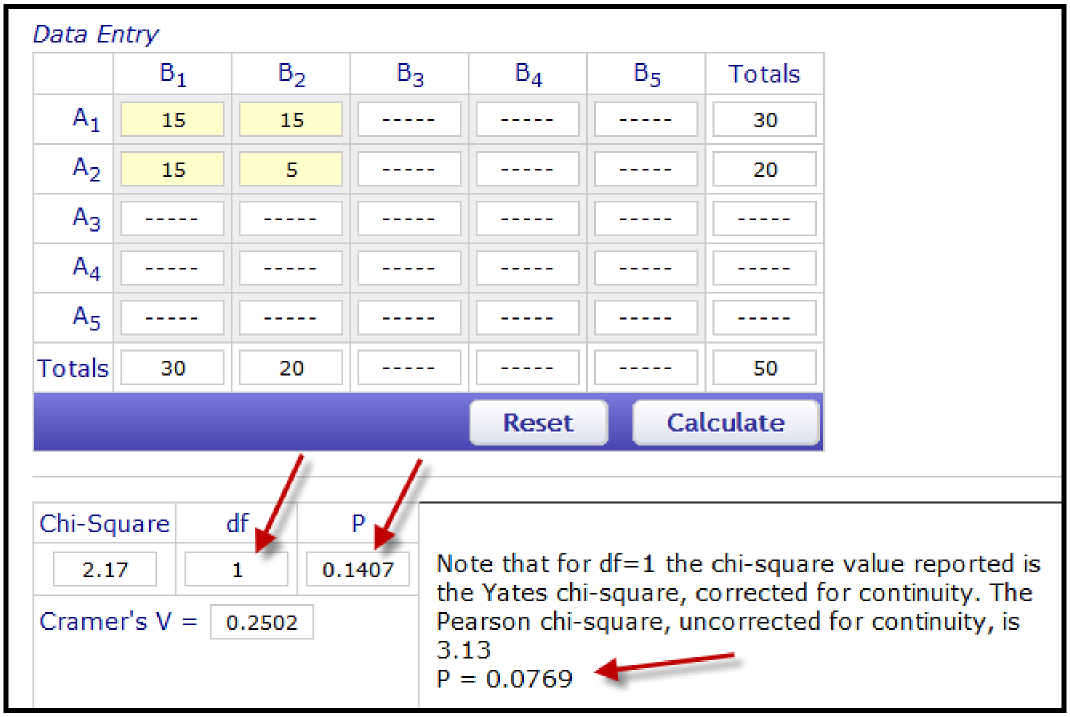
Cramer's V = 0.3633

No message for this analysis.

Figure 5.2(a). Chi-square and degrees of freedom from a 2X3 table.

The above illustration is over-simplified. In the actual computation of misfit, examinees are not typically divided into only three groups. Instead, many more levels may be used. There is no common consent about the optimal numbers of intervals. Yen (1981) suggested using 10 grouping intervals. Some item analysis software modules, such as (RUMM, Rasch Uni-dimensional Measurement Model) adopts 10-level grouping as the default. It is important to point out that the number of levels is tied to the degrees of freedom, which affects the significance of a Chi-square test.

What would happen if the psychometrician decides to use two levels only (competent, not competent) by collapsing "more skilled" and "average" into one group ("competent")? Figure 5.2(b) shows the result from a 2X2 table, in which the DF is 1 and the *p* value based on the Yates Chi-square is 0.1407. Additionally, the Pearson chi-square is 3.13 whereas *p* = 0.0769. But neither one is significant. In addition, the mean square is 3.13/1 = 3.13, which is smaller than the mean square using three levels (3.3). If the Yates correction is used, then the mean square is 2.17. In short, whether the Chi-square is significant or not highly depends on the degrees of freedom and the number of rows/columns (the number of levels chosen by the psychometric software package). Obviously, altering the classification of skill levels would definitely change the degrees of freedom, and subsequently affecting the mean square value and its significance.



Data Entry

	B_1	B_2	B_3	B_4	B_5	Totals
A_1	15	15	-----	-----	-----	30
A_2	15	5	-----	-----	-----	20
A_3	-----	-----	-----	-----	-----	-----
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	30	20	-----	-----	-----	50

Chi-Square	df	P
2.17	1	0.1407

Cramer's V = 0.2502

Note that for df=1 the chi-square value reported is the Yates chi-square, corrected for continuity. The Pearson chi-square, uncorrected for continuity, is 3.13
P = 0.0769

Figure 5.2(b). Chi-square and degrees of freedom from a 2X2 table.

Degrees of freedom for Mantel-Haenszel Chi-square

It is noteworthy that Chi-square statistics have many variants. One version of Chi-square is called the Mantel-Haenszel Chi-square statistic. This method tests the null hypothesis that the relationship between the row variable and the column variables is linear. The degrees of freedom is calculated as (N - I) x r^2, where r^2 is the Pearson product-moment correlation between the two variables. One of the useful applications of Mantel-Haenszel Chi-square is detecting test bias (Holland & Thayer, 1988). Although in psychometrics specific procedures have been developed to detect test bias (e.g. Differential item functioning, DIF), Mantel-Haenszel Chi-square is much easier to compute and interpret. Figure 5.3 shows a test item that favors males of low ability while favoring females of higher ability. As mentioned before, the Mantel-Haenszel Chi-square statistic aims to examine a linear relationship, as shown in Figure 5.3. Thus, its degrees of freedom entail the Pearson product-moment correlation.

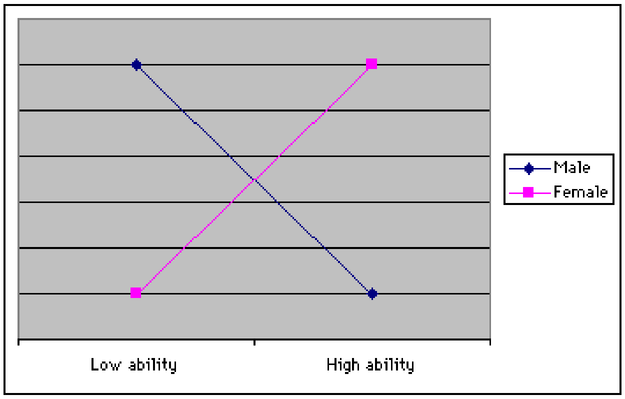


Figure 5.3 Potential test bias.

Summary

The original version of Chi-square analysis invented by Karl Pearson is a method of classifying data without referencing any model. But later Fisher transformed the test to be a model-based approach by adjusting the degrees of freedom. Today we use the Fisher's version of Chi-square, not the Pearson's one.

When there are only two possible outcomes in a one-dimensional Chi-square test, the degree of freedom is 1 and only 1. In many other tests this problem is insurmountable because one degree of freedom is considered insufficient for conducting any meaningful test procedure. Nevertheless, the researcher can still proceed with this type of Chi-square test when the Yates correction is employed as a remedy. In this situation, the DF is a constant no matter how many observations are made in the study. And the meaningfulness of the test result is tied to the sample size, not the DF.

In a two-dimensional Chi-square test, whether the test result is significance or not depends on the degrees of freedom, which is defined by $(r-1)(c-1)$. If the grouping method is natural (e.g. gender) or the cut-off for grouping is clear (e.g. sick/healthy), there will be no debate in grouping. However, on some occasions (e.g. psychometrics) the number of levels in a group is highly subjective or even arbitrary. In this case, Chi-square-based fit indices are strongly affected by the degrees of freedom.

Chi-square statistics have many variants. The Mantel-Haenszel Chi-square approach tests the null hypothesis whether the row and the column variables are linearly dependent, and thus the degrees of freedom associated with this test include the Pearson's r into the calculation.

Chapter 6
DF in the context of structural equation modeling

In the previous chapter we discussed the role of DF in Chi-square analysis as standalone tests. In some situations Chi-square statistics is a part of a larger-scale analysis, such as structural equation modeling (SEM). The role of Chi-square in SEM is to inform the researcher to what extent the data and the model fit each other.

The following is a very brief overview of SEM for those readers who are not familiar with this type of modeling. If you have learned the basic concepts of SEM, you can proceed to the next section. A typical SEM is composed of one or more measurement models and a structural model (path model). A measurement model is also known as a factor model. A factor model identifies the relationship between observed items and latent factors. Let's refresh what you have learned in Chapter 3. For example, when a psychologist wants to study the causal relationships between anxiety and job performance, first he/she has to define the constructs "anxiety" and "job performance." To accomplish this step, the psychologist needs to develop items that measure the defined construct. The relationship between factors and observed variables is indicated in Figure 6.1. The ellipse represents a latent construct, and the rectangles represent observed variables, which are individual items in a scale. The circles denote measurement errors.

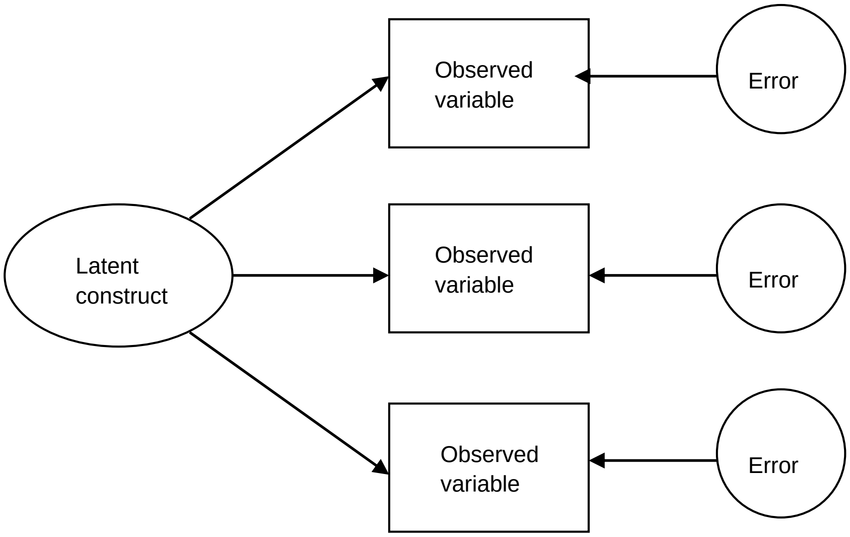


Figure 6.1. Measurement model.

A path model is similar to a regression model. In both models the relationships between variables or/and factors are examined. Indeed, the path coefficient in SEM is virtually the same as the regression coefficients. However, in regression only one dependent variable is allowed even though the researcher could enter as many independent variables as possible into the model. The versatility of SEM is that a variable could be a dependent and an independent variable at the same time when a causal structure, also known as a causal path, is configured. Figure 6.2 depicts a very simple path model. In this example, A is said to be a cause of B, and B is said to be a cause of C. Hence, B is an independent variable that influences the outcome C, but at the same time B is a dependent variable that responds to the input from A.

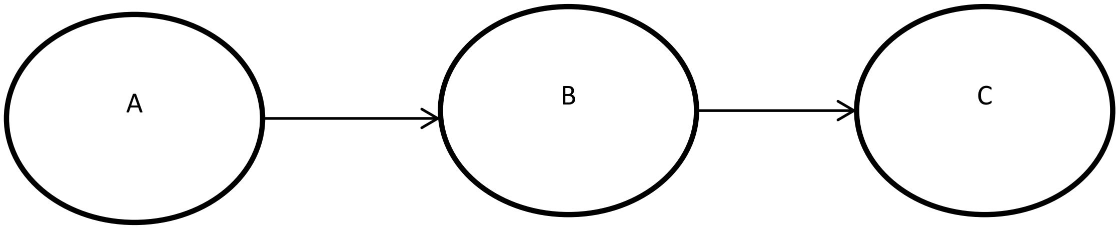


Figure 6.2. A simple path model.

When a measurement model and a path model are combined together, the fused product is a complete structural equation model. The example illustrated in Figure 6.3 is given by Lomax (1992). Based upon literature review, a researcher hypothesizes that "home background" could be a predictor to "school achievement," and "school achievement" could predict "career success", he defines such vague concepts as home background, school achievement, and career success by the factor model. Afterwards, a chain (path) of cause and effect is drawn among constructs. Then he/she employs SEM techniques to examine the fitness between the data and the model.

Careful readers might observe that the directions of arrows in Figure 6.1 and 6.2 are different. In the example of factor modeling the arrows point from the latent to observed but in the example of combining factor and path modeling the direction of the arrows is opposite. Actually in SEM directionality depicts causality. It is reasonable to believe that our mental state (e.g. anxiety) causes our observed behaviors, and therefore the arrows are drawn from latent to observed. But in the second example career success is defined by net worth, income, and job satisfaction. In other words, the construct is a result of the observed items, not the other way around. The same line of reasoning also applies to home background and school achievement. And therefore the arrows start from the observed items and end at the latent constructs. Please keep in mind that this example is simplified. A real–life SEM could be much more complicated. Because of the complexity of SEM, there are numerous possible ways to fit the data with the model. The fit indices suggest the potential causal structure in SEM.

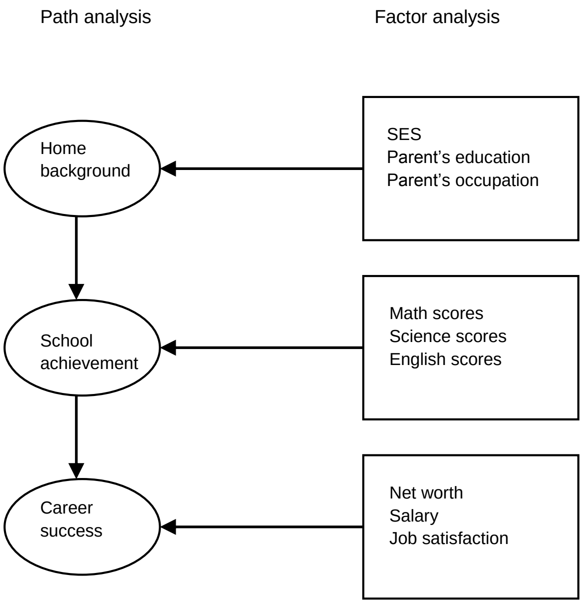


Figure 6.3. Example of a simple SEM consisting of a path model and a factor model.

Chi-square-based fitness and zero degrees of freedom

This section illustrates the role of Chi-square statistics in SEM and why sometimes we obtain zero degrees of freedom while computing Chi-square-based fitness tests. In this example only observed items are used.

The data source for this example is the archival data set entitled World Development Indictors (WDI) and Global Development Finance (GDF), which is downloadable from the World Bank (2012). This comprehensive data set contains indicators for each country of national well-being, including data concerning a country's education, environment, economic policies, financial sector, health, infrastructure, labor force, social protection, poverty, and international trade. The variables chosen for this example are as follows (Yu, 2012b):

1. SCI 2003: the percentage of people who graduated from college or university in 2003 with a major in science.
2. EMC 2003: the percentage of people who graduated from college or university in 2003 with a major related to engineering, manufacturing, or construction (EMC).
3. Paper 2005: the number of scientific and technical papers published in peer-review journals in 2005.
4. Patent 2005: the number of patents applied for by residents in 2005.
5. Productivity 2007: Gross domestic product per person employed in 2007.

Based on prior research and other background information, the initial conjecture proposed by the author is that the number of graduates in science and EMC in a given year might positively influence the number of scientific papers published in peer-review journals and the number of patents applied by residents two years later. Patents held by non-residents are not taken into account because their accomplishment might not be attributed to local education. Subsequently, new ideas and new innovations manifested in research papers and patents could eventually improve productivity.

Figure 6.4(a) and 6.5(a) depict the untransformed distributions of 2005 scientific papers and patents. Both are extremely skewed distributions because scientific research and innovations tend to concentrate on very few developed nations, such as the US and Japan. As a remedy, natural logarithm transformation was utilized to normalize the distributions (see Figure 6.4(b) and 6.5(b)).

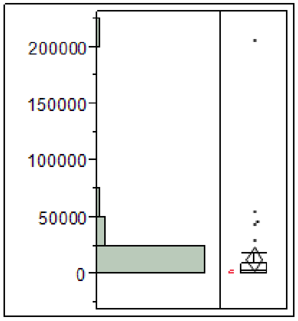
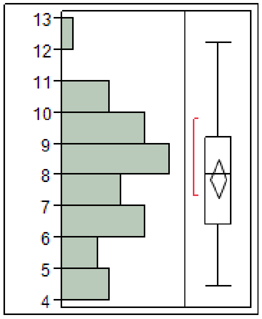
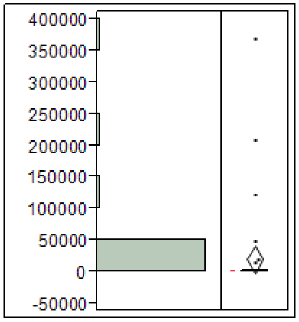
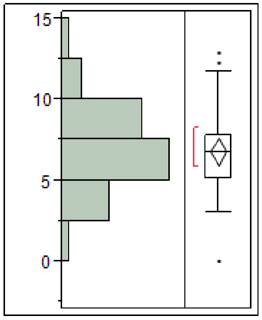
			

Figure 6.4(a). Untransformed distribution of 2005 scientific articles.	Figure 6.4 (b). Natural log distribution of 2005 scientific articles.	Figure 6.5(a). Untransformed distribution of 2005 patents by residents.	Figure 6.5(b). Natural log distribution of 2005 patents by residents.

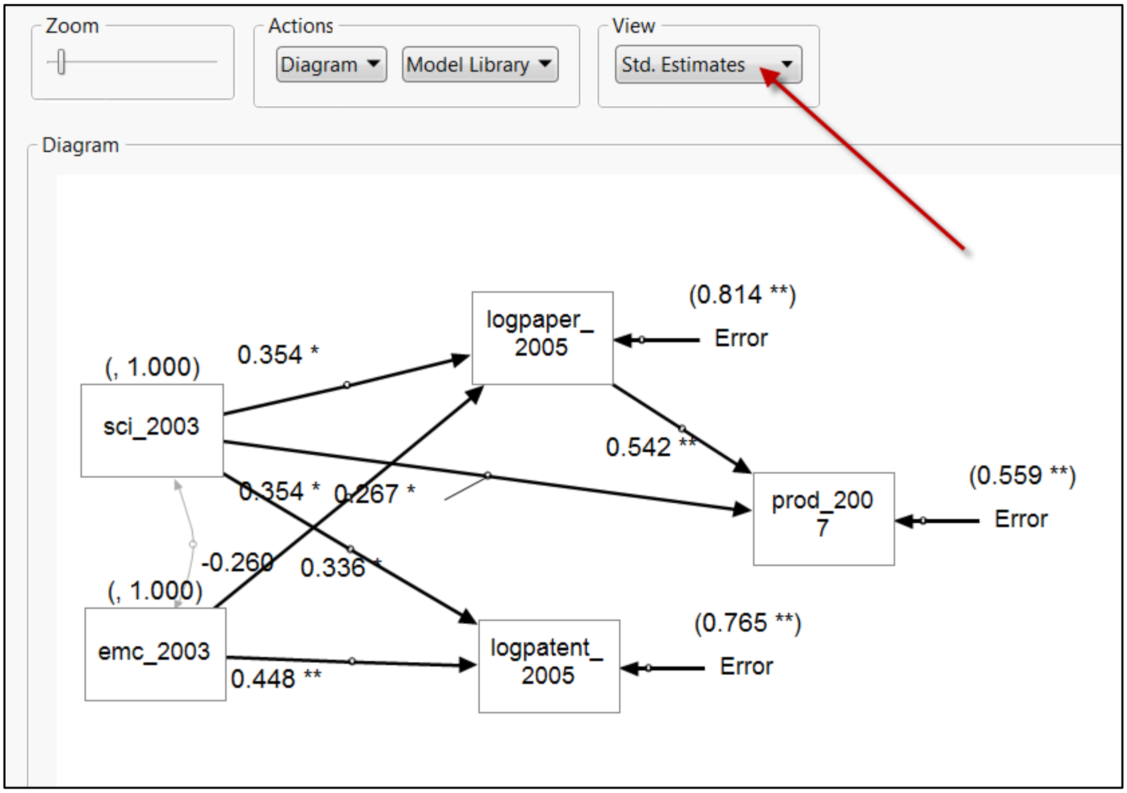
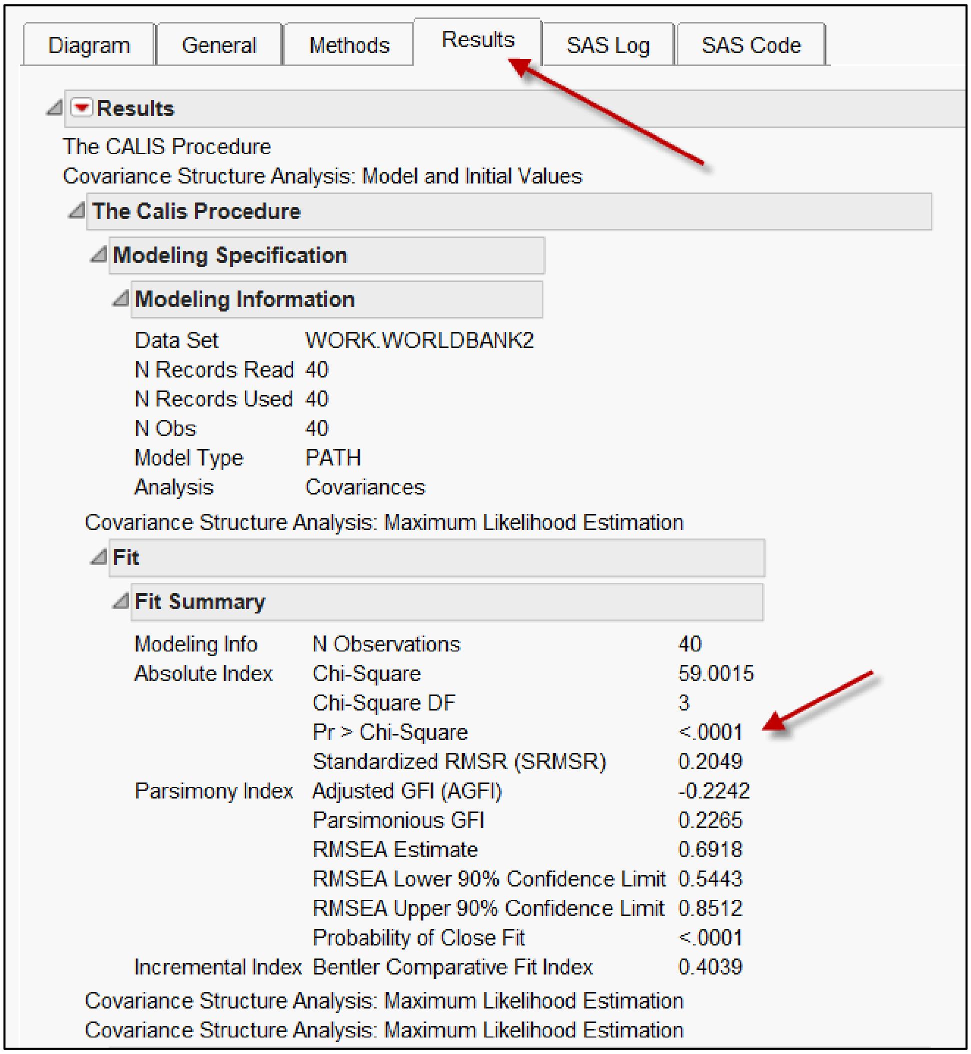


Figure 6.6. Using JMP interface to run SEM in SAS.

The author utilized the interface in JMP (SAS Institute, 2011b) to run a path model in SAS. Figure 6.6 depicts the standardized estimates of the coefficients along the paths between the variables. The parameters with the sign "*" or "**" are considered significant. One asterisk indicates that p < 0.05 whereas two asterisks indicate that p < 0.01.

To unveil more details, the analyst can check the table output. In this model, the chi-square statistics suggest that the model does not seem to be promising ($\chi^2 = 59.0015$, $p < .0001$). The Chi-square test is a measure of the goodness of fit between the observed and the expected. The null hypothesis is that there is no significant discrepancy between the covariance matrix generated by the model and that observed in the data. A *p* value that is small enough to reject the null signifies that the data-model fit is questionable. It is a well-known fact that the significance of the Chi-square is subject to the sample size (Besag, 1980). If the sample size is very large, it is more likely that the model will be unfairly rejected, However, in this small data set (n=40) this is not a concern at all. In addition to the Chi-square, there are many other fitness indices, such as AGFI, RMSEA, Bentler Comparative Fit Index...etc. (see Figure 6.6).



```
Results
The CALIS Procedure
Covariance Structure Analysis: Model and Initial Values
  The Calis Procedure
    Modeling Specification
      Modeling Information
        Data Set         WORK.WORLDBANK2
        N Records Read   40
        N Records Used   40
        N Obs            40
        Model Type       PATH
        Analysis         Covariances
  Covariance Structure Analysis: Maximum Likelihood Estimation
    Fit
      Fit Summary
        Modeling Info     N Observations                      40
        Absolute Index    Chi-Square                          59.0015
                          Chi-Square DF                       3
                          Pr > Chi-Square                     <.0001
                          Standardized RMSR (SRMSR)           0.2049
        Parsimony Index   Adjusted GFI (AGFI)                 -0.2242
                          Parsimonious GFI                    0.2265
                          RMSEA Estimate                      0.6918
                          RMSEA Lower 90% Confidence Limit    0.5443
                          RMSEA Upper 90% Confidence Limit    0.8512
                          Probability of Close Fit            <.0001
        Incremental Index Bentler Comparative Fit Index       0.4039
  Covariance Structure Analysis: Maximum Likelihood Estimation
  Covariance Structure Analysis: Maximum Likelihood Estimation
```

Figure 6.6. Poor fitness indices in the initial model.

To rectify the poor fit, an alternate model should be proposed and thoroughly examined. JMP provides the user with a “Copy” button and thus model revision in JMP is relatively easy. The existing model is retained for model comparison in a later stage. Rather than starting over from scratch, the user can drops some variables and redraw the paths in the cloned model (see Figure 6.7).

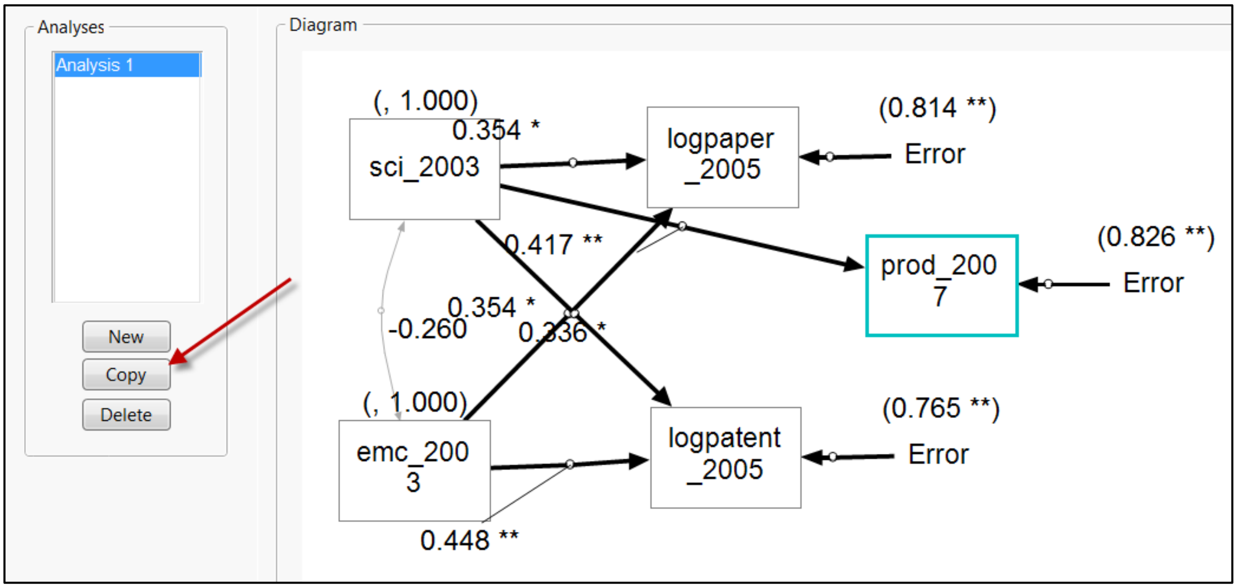


Figure 6.7. Copy and paste to revise the existing model.

The new model is put into a new analysis (see Figure 6.8). Because the ultimate goal is to identify the variables that could make contributions to productivity, the natural log of 2005 patents is redundant and thus it is taken out of the equation.

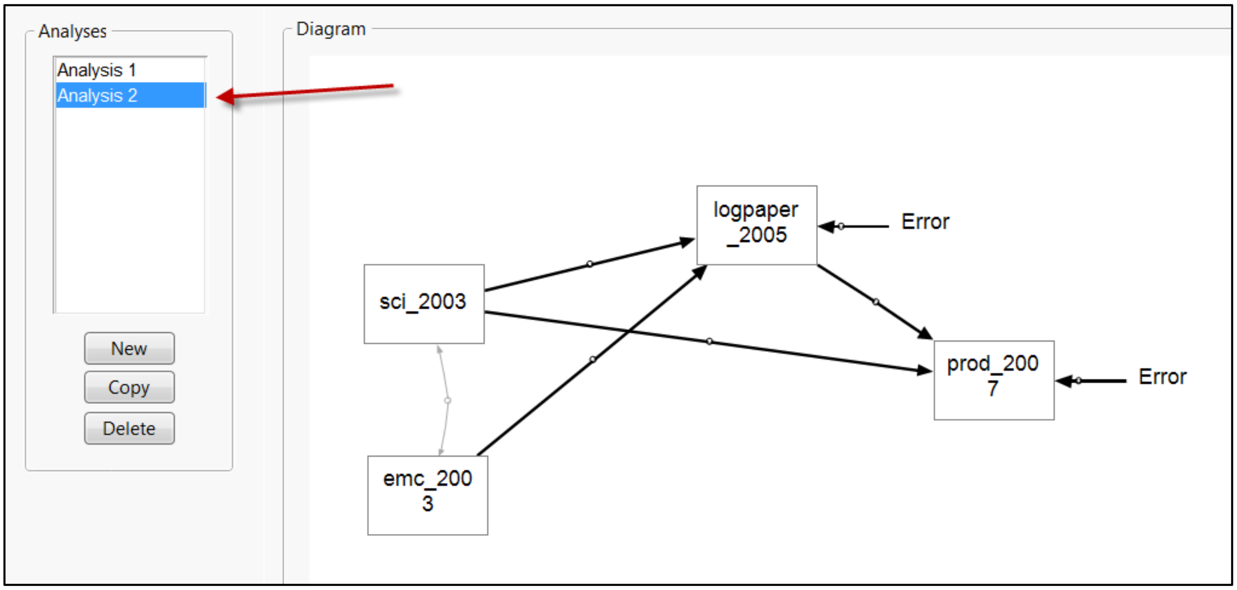


Figure 6.8. A new model without natural log of 2005 patents.

Figure 6.8 depicts the new model with standardized estimates. Figure 6.9 shows that the fitness has been substantively improved. Some researchers might want to stop at this point and accept this one as the final. In the past this decision was understandable because running SEM is very involved and time-consuming. However, being equipped with the JMP interface, today the analyst can afford further exploration with minimal efforts.

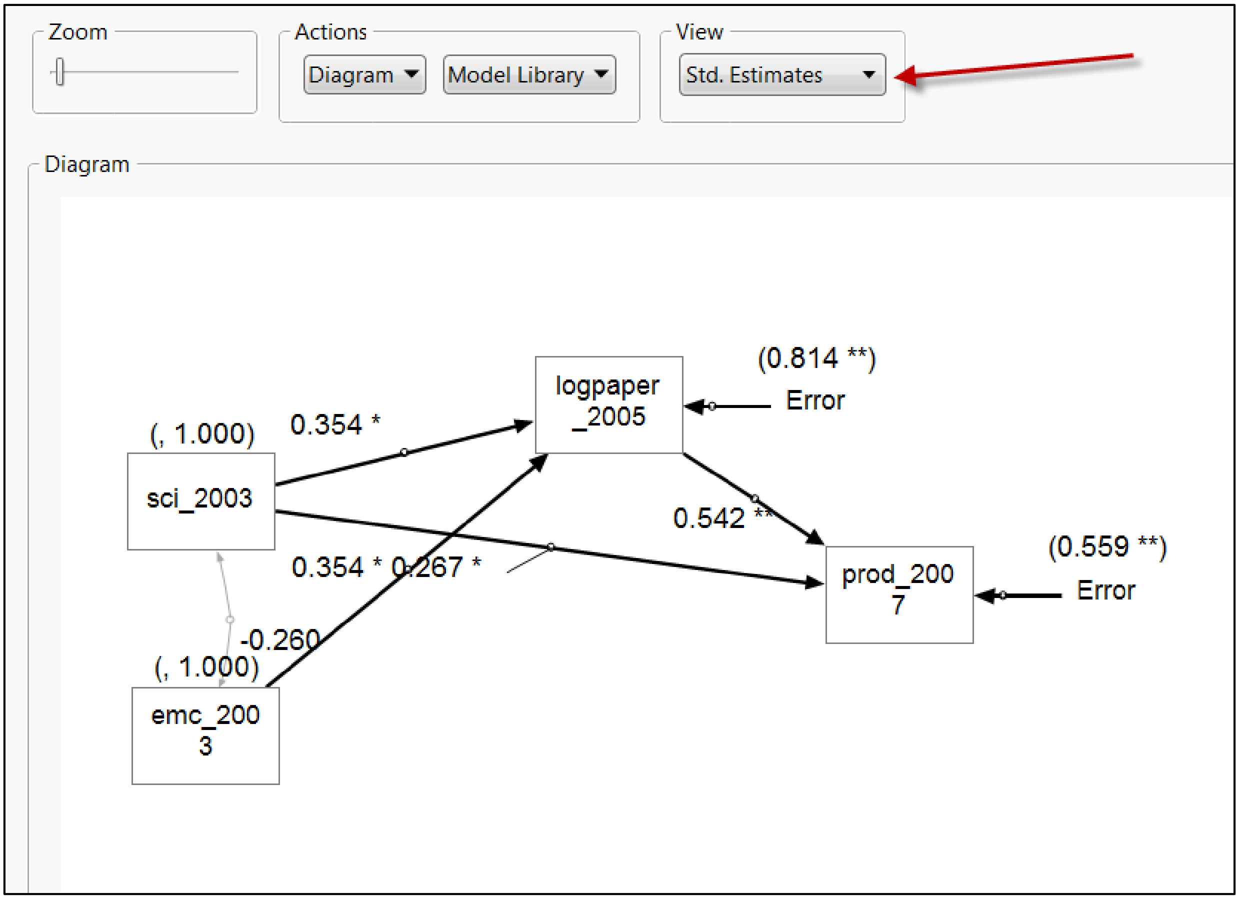
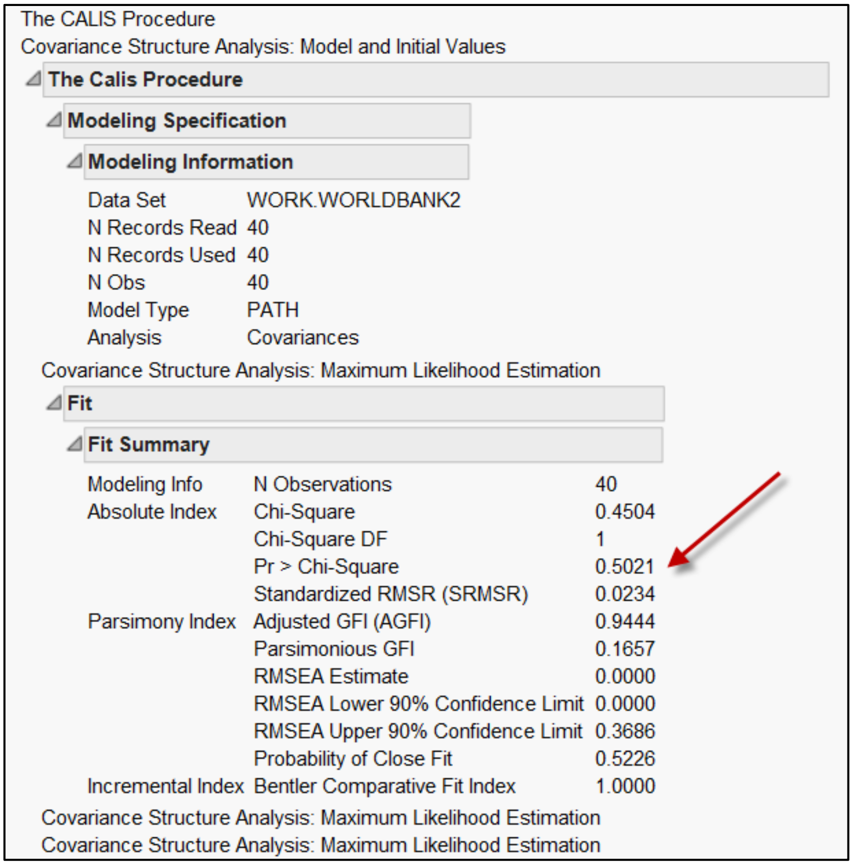


Figure 6.8. A new model without natural log of 2005 patents.



```
The CALIS Procedure
Covariance Structure Analysis: Model and Initial Values
 The Calis Procedure
   Modeling Specification
     Modeling Information
       Data Set          WORK.WORLDBANK2
       N Records Read    40
       N Records Used    40
       N Obs             40
       Model Type        PATH
       Analysis          Covariances
  Covariance Structure Analysis: Maximum Likelihood Estimation
   Fit
     Fit Summary
       Modeling Info      N Observations                      40
       Absolute Index     Chi-Square                          0.4504
                          Chi-Square DF                       1
                          Pr > Chi-Square                     0.5021
                          Standardized RMSR (SRMSR)           0.0234
       Parsimony Index    Adjusted GFI (AGFI)                 0.9444
                          Parsimonious GFI                    0.1657
                          RMSEA Estimate                      0.0000
                          RMSEA Lower 90% Confidence Limit    0.0000
                          RMSEA Upper 90% Confidence Limit    0.3686
                          Probability of Close Fit            0.5226
       Incremental Index  Bentler Comparative Fit Index       1.0000
  Covariance Structure Analysis: Maximum Likelihood Estimation
  Covariance Structure Analysis: Maximum Likelihood Estimation
```

Figure 6.9. A new model without natural log of 2005 patents.

Again, using the copy button the user can create Analysis 3 in a new canvas. If EMC graduates have no significant effects on productivity, could this variable be dropped, too? This time only three variables remained in the model, as shown in Figure 6.9. The standardized estimates and the fitness indices are shown in Figure 6.10 and Figure 6.11, respectively.

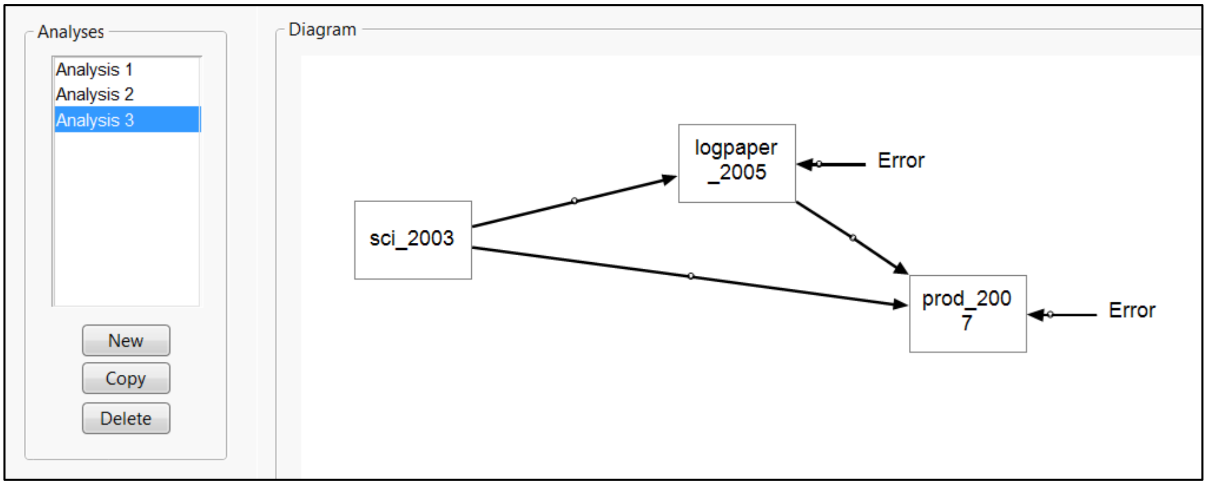


Figure 6.9. A highly parsimonious model with three variables.

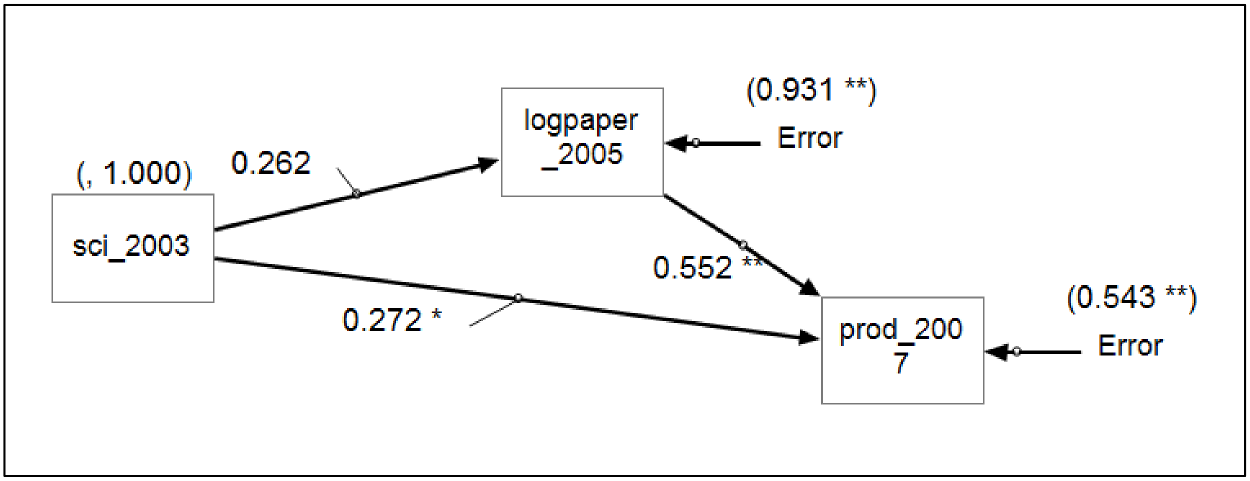
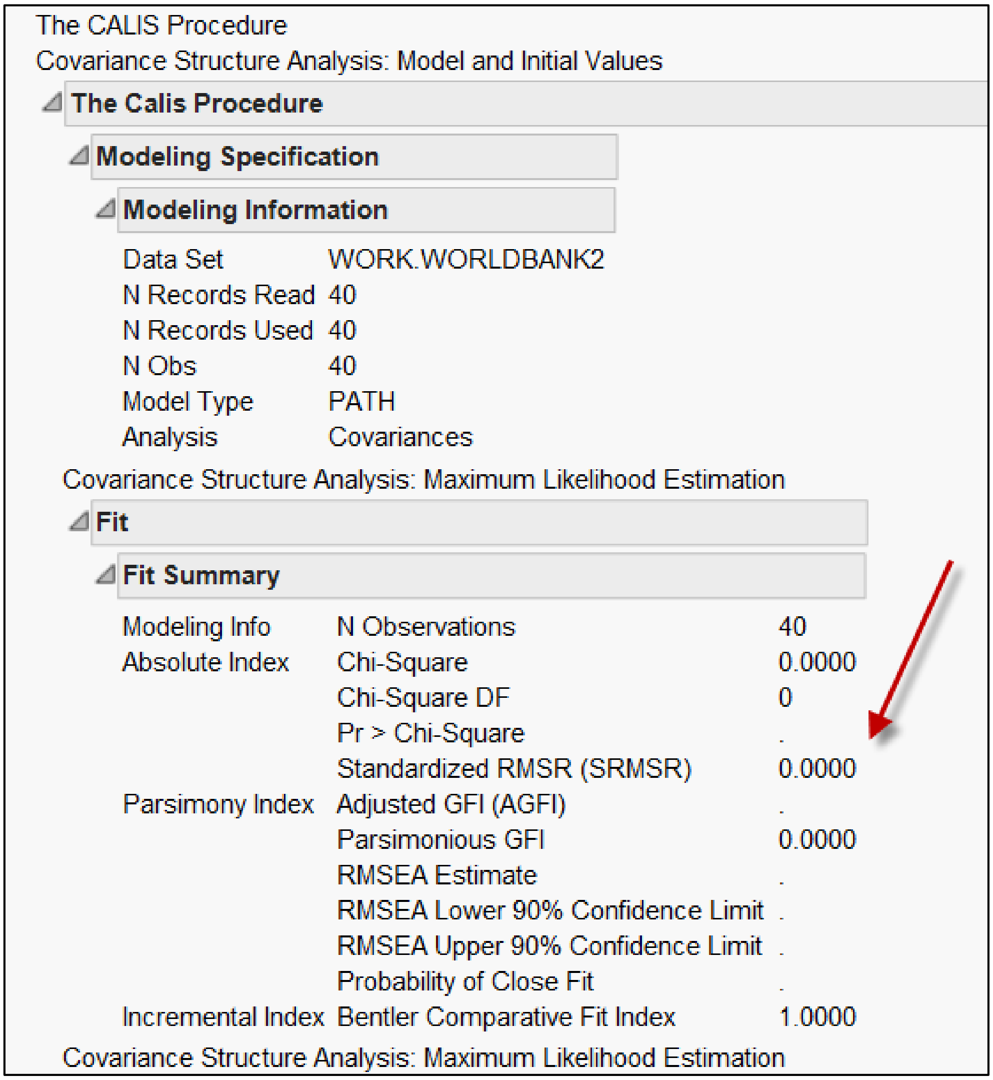


Figure 6.10. Standardized estimates of the parsimonious model.



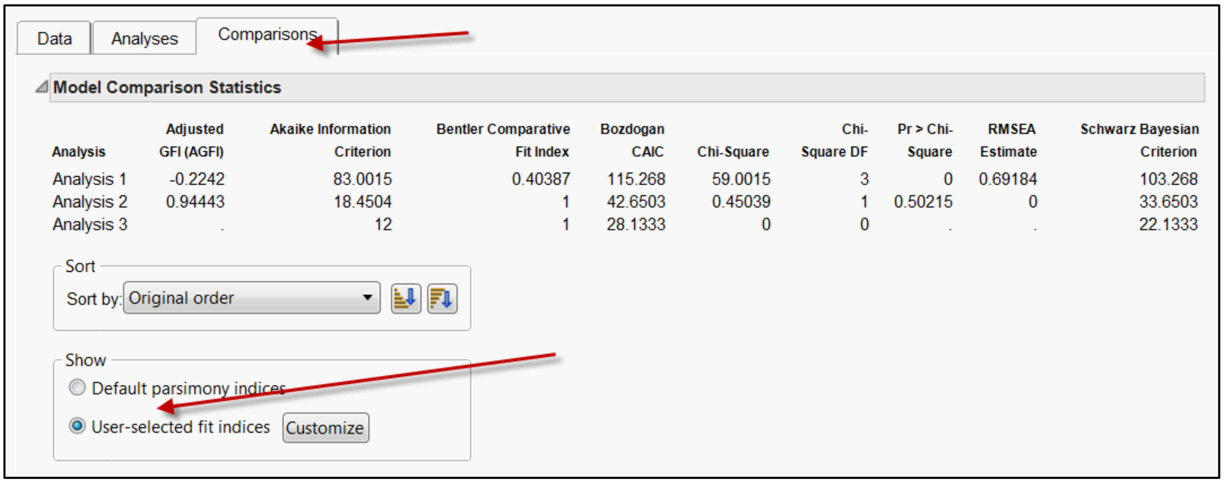
```
The CALIS Procedure
Covariance Structure Analysis: Model and Initial Values
 ◿ The Calis Procedure
   ◿ Modeling Specification
     ◿ Modeling Information
        Data Set          WORK.WORLDBANK2
        N Records Read    40
        N Records Used    40
        N Obs             40
        Model Type        PATH
        Analysis          Covariances
  Covariance Structure Analysis: Maximum Likelihood Estimation
   ◿ Fit
     ◿ Fit Summary
        Modeling Info      N Observations                      40
        Absolute Index     Chi-Square                          0.0000
                           Chi-Square DF                       0
                           Pr > Chi-Square                     .
                           Standardized RMSR (SRMSR)           0.0000
        Parsimony Index    Adjusted GFI (AGFI)                 .
                           Parsimonious GFI                    0.0000
                           RMSEA Estimate                      .
                           RMSEA Lower 90% Confidence Limit    .
                           RMSEA Upper 90% Confidence Limit    .
                           Probability of Close Fit            .
        Incremental Index  Bentler Comparative Fit Index       1.0000
  Covariance Structure Analysis: Maximum Likelihood Estimation
```

Figure 6.11. Fitness indices of the saturated model.

The information portrayed in Figure 6.11 seems to be strange. In this panel, both the Chi-square and the DF show a "zero" value while the *p* value is missing. In the context of a bivariate analysis, zero degree of freedom is problematic. It means that there is no useful independent information to do any meaningful estimation of the relationship. Is this the case here? As mentioned before, the Chi-square test is a measure of the goodness of fit between the expected and the observed. If there is no discrepancy between the expected model and the observed data, the Chi-square is zero, of course. In this case, the model is said to be saturated, meaning that the model can perfectly reproduce all of the variances, covariance, and means. Some researchers (e.g. Savalei & Bentler, 2006) argue that such a "perfect" model has *no explanatory value* at all. Nonetheless, a saturated model is still useful for functioning as a baseline model with which other non-saturated models are compared.

The modeler can perform a model comparison by choosing the "Comparisons" tab (see Figure 6.12). By default JMP shows Akaike Information Criteria, Bozdogan CAIC, Schwartz Bayesian Criterion, and RMSEA. The user can check the box "User-selected fit indices" to reveal more options. Please review Chapter 4 for the details of AIC and BIC. The last model, which has a much smaller AIC (12), is the simplest but we might not like a saturated model that would have no explanatory value. Thus, we might want to settle down with the middle one.

Based on the adopted model, it is concluded that a high percentage of graduates majoring in science and EMC could lead to better scientific research, indicated by a higher volume of research papers. And better research might eventually benefit productivity.



Model Comparison Statistics

Analysis	Adjusted GFI (AGFI)	Akaike Information Criterion	Bentler Comparative Fit Index	Bozdogan CAIC	Chi-Square	Chi-Square DF	Pr > Chi-Square	RMSEA Estimate	Schwarz Bayesian Criterion
Analysis 1	-0.2242	83.0015	0.40387	115.268	59.0015	3	0	0.69184	103.268
Analysis 2	0.94443	18.4504	1	42.6503	0.45039	1	0.50215	0	33.6503
Analysis 3	.	12	1	28.1333	0	0	.	.	22.1333

Sort

Sort by: Original order

Show

○ Default parsimony indices

● User-selected fit indices

Figure 6.12. Model comparison in SEM.

Using DF to help model specification and calculate needed sample size in a path model

Ridgon (1994) observed that many students perceive structural equation modeling (SEM) to be a cluster of mysteries, in which the origin and calculation of the degrees of freedom of a SEM are not meaningful. As a result, many students ignored this task because apparently there is no practical value of computing the DF of a SEM. But Ridgon asserted that the DF calculation provides an important check on the accuracy of the model specification.

In addition to check model specification, the DF can also be used to compute the required sample size for the proposed SEM. There are many “rules of thumb” regarding the proper sample size for SEM, but most of these rules are nothing more than conventional wisdom or guesswork. Utilizing the information of DF is definitely a better approach.

Like many statistical tests, SEM requires unique pieces of information. However, these pieces of information are not the raw data. Rather, the data points are the elements in the covariance matrix. The variance of one single variable indicates the dispersion or the distribution of the values. When we put two variables together, we have covariance to indicate their relationships. When there are multiple variables, there is a covariance matrix, which is a summary of the inter-relationships among many variables. The goal of SEM is to investigate whether the proposed causal relationships is the same as the actual covariance matrix. The former is symbolized by ***S*** whereas the latter is represented by Σ.

In SEM the DF is the number of distinct elements in the covariance matrix minus the number of free parameters to be estimated. The formula for calculating the number of distinct elements is: $p(p+1)/2$. Savalei and Bentler (2006) used a simplified example to illustrate the calculation. In their example only five variables are considered. Based on the formula and also the structure depicted in Figure 6.13, there are 15 distinct elements ($p(p+1)/2 = 5(6)/2 = 15$) and 12 parameters to be estimated, as shown in the following:

- Four regression (path) coefficients: A→D, B→D, C→D, D→E
- Five variances: A, B, C, D, E
- Three pairs of covariance among the variables: A <-> B, A <-> C, B <-> C.

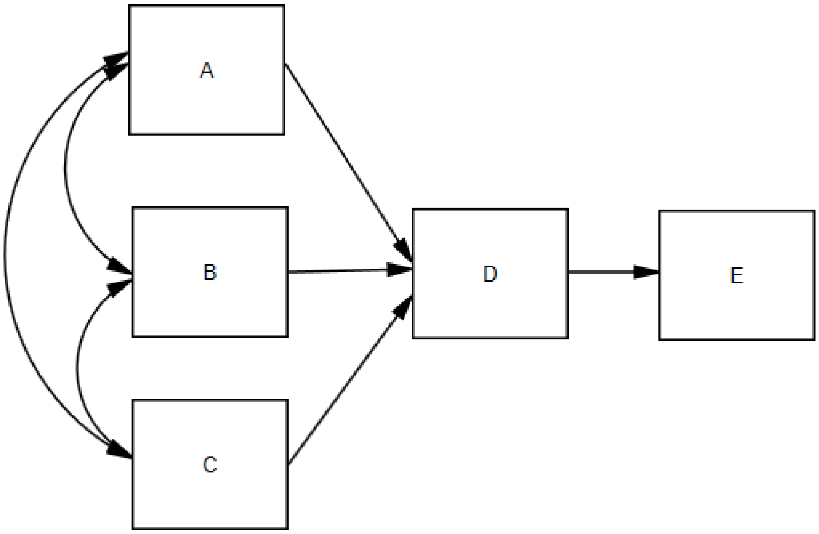
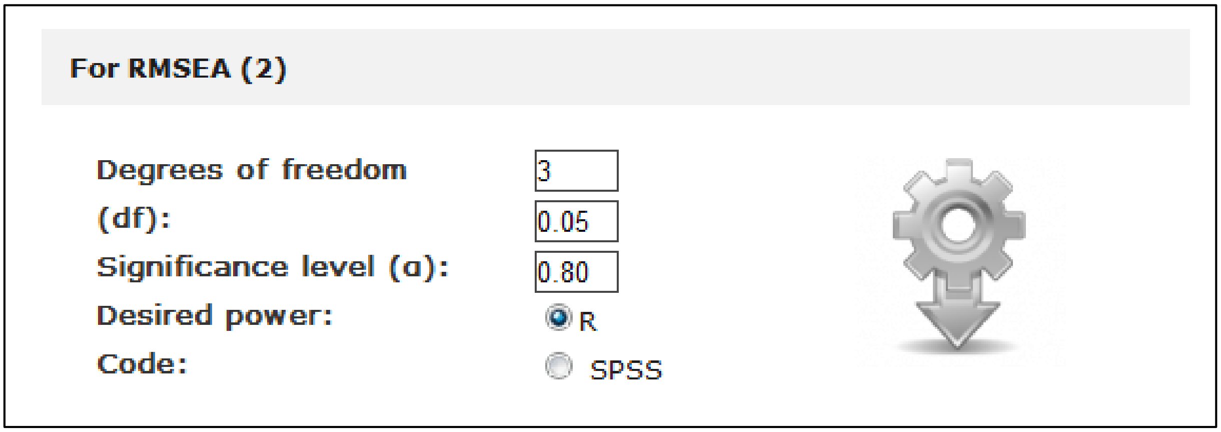


Figure 6.13. A simple path model with four variables.

In this simple example, the DF value is the number of unique elements minus the number of model parameters, which is DF = 15 -12 = 3

Next, we need to know how many subjects are required to test this model. In the past the computation could be very tedious and error-prone. Today we can simply go to a website (http://timo.gnambs.at/en/scripts/powerforsem) maintained by Gnambs (2008) to perform this task. The website provides researchers with different options of calculating the required sample size for SEM. One of the easiest computation approaches is based on the fitness index named Root Mean Square Error of Approximation (RMSEA) (Kim, 2005). Mean Square Error (MSE) is the difference between the estimated and the true values. Root Mean Square Error (RMSE) is derived from taking the squared root of MSE. Because an exact fit between the data and the model is almost impossible and thus the null hypothesis tends to be false, RMSEA aims to measure only the approximate fit instead of the exact fit (Schermelleh-Engel, Moosbrugger, & Muller, 2003). To compute the recommended sample size using Gnambs's program, what it takes is entering the alpha level, the desired power, and the DF into the online calculator (see Figure 6.14).



For RMSEA (2)

Degrees of freedom (df): 3

Significance level (α): 0.05

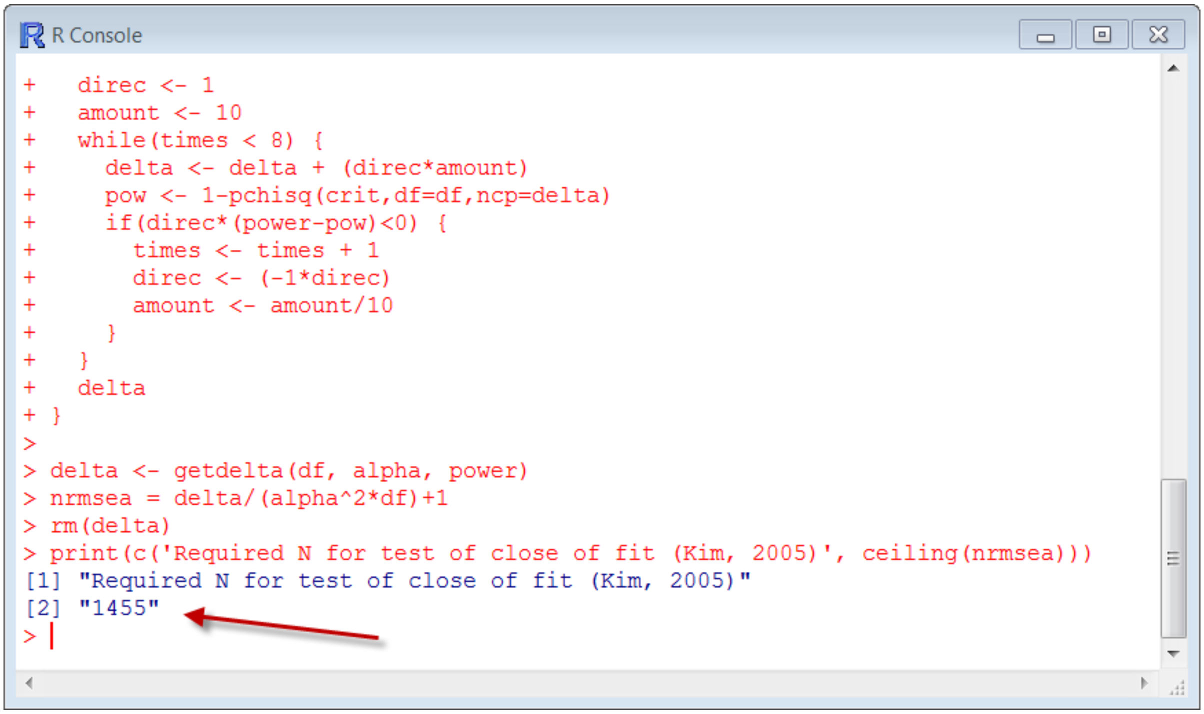
Desired power: 0.80

Code: ● R ○ SPSS

Figure 6.14. Sample size calculation for SEM based on RMSEA.

After the information is entered, the user can choose to generate a SPSS program or an R program for the calculation. R is an open source statistical application that can be freely downloaded (Institute for Statistics and Mathematics, 2012). The researcher can copy and paste the program into the R console to obtain the result. The R code suggests that the desired sample size is 1,455, and it may be out of reach by many researchers. Indeed, usually SEM demands a very large sample. If a large and complex model is tested with a sample size of 100, it is likely that there are estimation problems, and also test statistics have extremely low power (Savalei & Bentler (2004). To rectify the situation, the researcher might consider lowering the power level slightly or re-specifying the model.

What would happen if we give up estimating the three pairs of covariance and also reducing the lower level to 0.7? One may argue that these are unwise decisions, but the purpose of this exercise is just to demonstrate how altering the DF value would change the sample size requirement. If there are 9 model parameters only, the degrees of freedom become: 15 – 9 = 6, and the required n drops to 744.



```
+   direc <- 1
+   amount <- 10
+   while(times < 8) {
+     delta <- delta + (direc*amount)
+     pow <- 1-pchisq(crit,df=df,ncp=delta)
+     if(direc*(power-pow)<0) {
+       times <- times + 1
+       direc <- (-1*direc)
+       amount <- amount/10
+     }
+   }
+   delta
+ }
>
> delta <- getdelta(df, alpha, power)
> nrmsea = delta/(alpha^2*df)+1
> rm(delta)
> print(c('Required N for test of close of fit (Kim, 2005)', ceiling(nrmsea)))
[1] "Required N for test of close of fit (Kim, 2005)"
[2] "1455"
>
```

Figure 6.15. R code for calculating the desired sample size for SEM.

Using DF to help model specification and calculate needed sample size in a factor model

The previous example is a path model. In a measurement model the calculation of degrees of freedom is slightly different. Figure 6.16 depicts a two-factor model. In this model there are two factors, namely, F1 and F2. Each factor has three observed items or indicators (X1-X6) and each observed item has a measurement error term (*e*). In a typical factor model, DF is computed as (Rigdon, 1994):

DF = (*p* * (*p* + 1)/2)-(2 * *p*)-(F * (F - 1)/2)

where

p = the number of observed items
F = the number of latent factors

Hence, the first term, *p* * (*p* + 1), represents the number of distinct elements. Because there are 6 observed variables, the number of unique elements is 6*(6*1)/2 = 21. The second term, 2 * *p*, denotes the number of model parameters to be estimated. In each observed item there are two free parameters to be estimated: the measurement error terms and the factor loadings, which indicate the relationship between the observed items and the latent factor. Thus, 2 * *p* = 12. The third term, F(F - 1)/2, represents the free parameters associated with the covariance of the latent factors. And the result is 2 * (2 - 1)/2 = 1.

When the symbols in the formula are substituted with the numbers, the result is:

DF = 21 – 12 – 1 = 8

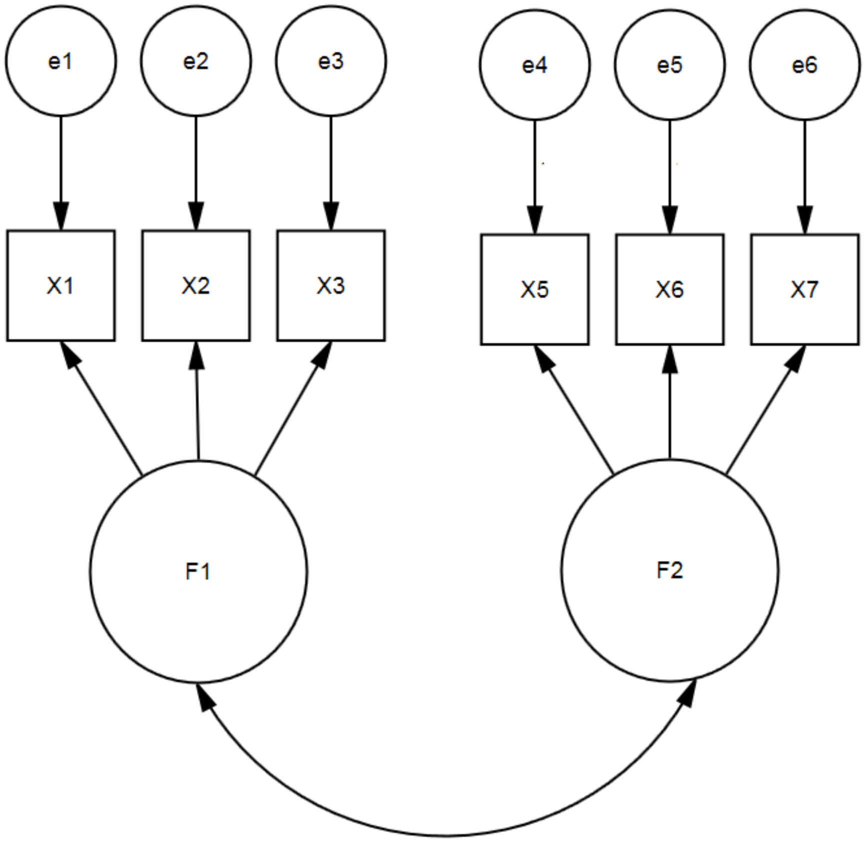


Figure 6.16. A two-factor model.

Again, we can plug the DF into the R program to find out the needed sample size. The R code indicates that 753 observations are needed to confirm the factor structure given that the power level is .08 and the alpha level is .05.

Summary

In the context of SEM, Chi-square analysis functions as a fitness index to inform the modeler whether there is a significance discrepancy between the sample and the model covariance matrices. When both the Chi-square statistic and the DF are zero, the model is said to be saturated and thus non-informative. It is very rare for the analyst to obtain a saturated model in the first run. SEM is more an iterative process than a one-shot analysis. By pushing further and further, the modeler may eventually see a saturated model with DF = 0. And a saturated model can be used as a baseline for model comparison.

In SEM the DF is the number of unique elements minus the number of parameters to be estimated. With the DF value the researcher could easily compute the proper sample size and also determine whether the specified model is viable.

Chapter 7
DF in the context of repeated measures

Assumption of sphericity

Chapter 3 illustrates the degrees of freedom of an ANOVA model based upon a between-subject design. However, in a within-subject design the rule of DF may be different, depending upon whether the data structure conforms to a certain parametric assumption. ANOVA repeated measures are vulnerable to the violation of sphericity, which is the condition that the variances of the differences between all combinations of levels are equal. Conversely, a violation of sphericity happens when the variances of the differences between all combinations of levels are unequal.

Compound symmetry is a concept related to sphericity. If the assumption of compound symmetry is met, then it also satisfies the sphericity assumption, but not vice versa. Compound symmetry requires homogeneity of variances and covariances within a group. In other words, the variances are supposed to be equal and the covariances between the pairs of levels are also equal (Warner, 2013).

The consequence of violating sphericity or compound symmetry is that the significance test becomes too liberal. In other words, the Type I error rate might be inflated and a significant effect would be incorrectly declared by the researcher even though the so-called significant effect is trivial. The Mauchly test is often used to determine whether such assumption is violated. If the data structure cannot pass the Mauchly test, then certain corrections (e.g. Greenhouse-Geisser, Huynh-Feldt) should be used in order to produce a more conservative critical *F* value. Specifically, these correctional procedures estimate the magnitude to which sphericity has been violated. And then a correction factor is applied to the degrees of freedom of the *F* distribution, so that the Type I error rate is under control. It is important to point out the adjustment in the DF does not affect the sum of square, the *F* ratio, and the effect size. Rather, this affects the *p* value only. When many students saw that the output of the *F* ratio remained the same in spite of the correction, they tried to debug the computation. Needless to say, this effort is futile.

Mauchly's test of sphericity is not necessary when there are two levels only (e.g. 2 paired-sample t-test), because it takes at least two sets of differences for variance and covariance comparison. For example, if the within-subject factor has three levels (Week 1, Week 2, and Week 3), there will be three sets of differences as shown below:

Week 1 - Week 2 = Difference 1

Week 1 - Week 3 = Difference 2

Week 2 - Week 3 = Difference 3

However, when there are two levels only, what we can obtain is one and only one set of difference:

Week 1 - Week 2 = Difference 1

It is important to point out that the corrections mentioned above might not be the best solution. The ANOVA repeated measures approach accepts only one type of covariance matrix. On the other hand, Hierarchical Linear Modeling (HLM), also known as multi-level modeling, allows different covariance matrix structures, such as auto-regressive, Toeplitz, heterogeneous AR, heterogeneous compound symmetry, and many others. In addition, multi-level modeling uses maximum likelihood estimation, which is more accurate than the sum of squares approach in General Linear Model (GLM), the foundation of many ANOVA-based procedures. Further, the efficacy of a HLM could be evaluated by multiple fitness criteria, namely, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), AAIC, and so on (Shin, 2009; Shin, Epsin, Deno, & McDonnell, 2004). Nevertheless, the ANOVA repeated measures procedure is still widely used today and thus it is noteworthy to illustrate how DF is corrected when sphericity is not fulfilled.

Example of within-subject design: Recovery from disease

Table 6.1 shows a hypothetical data set adopted from Warner's (2013) text. The hypothetical results show the weekly progress of recovery from disease and the purpose of the study is to evaluate the effectiveness of the disease treatment. In this case, a lower number signifies a lesser degree of the symptoms, and vice versa.

Table 6.1. Weekly progress of recovery as a result of disease treatment.

Subject #	Week 1	Week 2	Week 3	Week 4	Week 5
1	22	23	9	7	7
2	21	20	11	5	10
3	8	6	6	5	6
4	26	31	14	13	5
5	31	34	11	9	7
6	20	28	9	8	5
7	27	17	6	3	6
8	14	5	9	2	6
9	27	25	15	9	18

Table 6.2 is the result of the Mauchly's test of sphericity returned by SPSS (IBM SPSS, 2011b). The null hypothesis for this test is that the variances of the differences are not significantly different from each other. In this output, the *p* value ("sig" in the SPSS output) is .009, which is lower than the alpha cutoff, .05. Obviously, the null hypothesis is rejected and the assumption is found to be violated.

Table 6.2. Mauchly's Test of Sphericity.

Measure:disease

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.030	22.516	9	.009	.422	.522	.250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

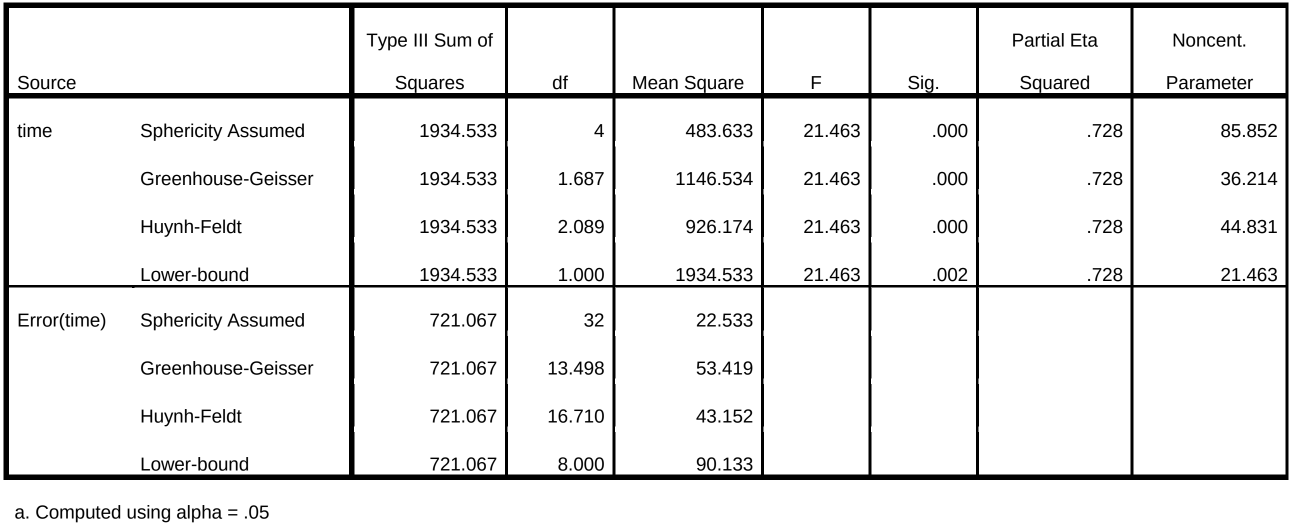
Adjusting DF by Epsilon

On the left hand side of Table 6.2, the header named “Epsilon” (ε) carries a pointer to the footnote “a”: “May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.” SPSS is very explicit about the necessity of correcting DF. The epsilon value indicates the degree of the violation of sphericity. When ε = 1, the condition of sphericity is perfect. When ε < 1, it means that violation has happened. As the epsilon value is further away from 1, the violation of sphericity becomes more severe. However, the Greenhouse-Geissser, Huynh-Feldt, and lower bound methods estimate epsilon in different ways.

According to the lower bound approach, the lowest possible of the epsilon value is 1/(*k* - 1) where *k* = the number of level. In this data set the time factor consists of five weeks, thus the lower bound method yields (1/(5 - 1)) = .25. No doubt this number is the lowest among the three approaches. In other words, the lower bound method looks for the worst case scenario and therefore its adjustment might be over-conservative. On the other hand, the Greenhouse-Geisser method tends to underestimate the epsilon value when it is close to 1, whereas the Huynd-Feldt approach tends to overestimate the epsilon.

Table 6.3 shows four types of significance tests. The first test assumes sphericity and it should not be trusted due to the assumption violation, as stated earlier. Hence, we need to look into the other three types of test statistics: Greenhouse-Geisser, Huynh-Feldt, and Lower bound. The third column clearly indicates that the degrees of freedom in other three tests are substantially reduced. However, the Type III sum of squares, the *F* statistics, and the partial eta squares (the effect size) remain the same regardless of whether any correction is applied or not. This seemingly strange result will be explained in the following.

Figure 6.3. Tests of within-subjects effects.



Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter
time	Sphericity Assumed	1934.533	4	483.633	21.463	.000	.728	85.852
	Greenhouse-Geisser	1934.533	1.687	1146.534	21.463	.000	.728	36.214
	Huynh-Feldt	1934.533	2.089	926.174	21.463	.000	.728	44.831
	Lower-bound	1934.533	1.000	1934.533	21.463	.002	.728	21.463
Error(time)	Sphericity Assumed	721.067	32	22.533				
	Greenhouse-Geisser	721.067	13.498	53.419				
	Huynh-Feldt	721.067	16.710	43.152				
	Lower-bound	721.067	8.000	90.133				

a. Computed using alpha = .05

Before discussing how these three approaches adjust DF, we need to understand how DF is used for computing the *F* statistics in an ANOVA repeated measures design. The formula for the DF of the model is:

$DF_{\text{time/condition}} = k – 1$

The formula for the DF of the error is:

$DF_{\text{error}} = (k – 1)(n - 1)$

where

K = the number of levels (time factor in repeated measures)

n = sample size

All of the above three corrections put the epsilon as a weight into the formulas of DF:

$DF_{\text{time/condition}} = (\varepsilon)\ \mathrm{k} – 1$

$DF_{\text{error}} = (\varepsilon)\ (k – 1)(n -1)$

The *F* statistics is the ratio between the mean square of the model and the mean square of the error. The mean square of the model is the sum of square of the model divided by the degrees of freedom of the model, whereas the mean square of the error is the sum of square of the error divided by the degrees of freedom of the error. These relationships are expressed by the formulas below:

$F = MS_{\text{model}}/MS_{\text{error}}$

$MS_{\text{model}} = SS_{\text{model}}/df_{\text{model}}$

$MS_{\text{error}} = \text{SS}_{\text{error}}/\text{DF}_{\text{error}}$

As mentioned earlier, ε is applied to both DF_{model} and DF_{error}. When the same coefficient appears on both the numerator and the denominator, the added constants cancel out each other. That's why the Type III sum of squares, the *F* statistics, and the partial eta squares (the effect size) remain the same even though the degrees of freedom are adjusted by the epsilon value.

When the epsilon value estimated by Greenhouse-Geissser is plugged into the DF formulas, the following results are obtained:

$\text{DF}_{\text{time/condition}} = (.422)\ (5 - 1) = 1.688$

$\text{DF}_{\text{error}} = (.422)\ (5 - 1)(9 - 1) = 13.504$

The following results are based on

$\text{DF}_{\text{time/condition}} = (.522)\ (5 - 1) = 2.088$

$\text{DF}_{\text{error}} = (.522)\ (5 - 1)(9 - 1) = 16.704$

Below are the results using the lower bound epsilon:

$\text{DF}_{\text{time/condition}} = (.25)\ (5 - 1) = 1$

$\text{DF}_{\text{error}} = (.25)\ (5 - 1)(9 - 1) = 8$

The slight discrepancy between the preceding set of numbers and those in Figure 6.3 is due to rounding errors. These degrees of freedom are different from the one without corrections. If the researcher looks up the *F* table using a set of smaller degrees of freedom, the critical value of *F* is definitely different. In this example, although all four tests yield significant results, the corrected *p* values are increased. Unfortunately, SPSS does not use e-notation to indicate very small *p* values and therefore it is difficult for the readers to tell the differences. Nonetheless, the most conservative method, lower bound, shows a *p* value of .002, which is larger than the uncorrected *p* value.

Summary

In an ANOVA repeated measures design, the logic of employing DF remains unchanged no matter if the assumption of sphericity is violated or not. When the assumption is not met, a weighing factor called epsilon is applied to both the formulas of DF_{model} and DF_{error}. This constant does not affect the sum of square, the *F* ratio, and the effect size, because the weight is added to both the numerator and the denominator. However, the degrees of freedom become smaller and thus the *p* value becomes larger, resulting in a reduction of the Type I error rate.

Chapter 8
DF in the context of multi-stage sampling

Conventional usage of degrees of freedom can be well-applied to many statistical procedures that are designed for simple random samples. However, in recent years more and more researchers employ multi-stage sampling for complex populations, which pose a challenge to simple random sampling. There are at least two types of multi-stage sampling, namely, stratified sampling and cluster sampling. In the former the entire population is divided into one or several levels (strata), and each level carries several segments, but information about the strata must be known. For example, a researcher who wants to obtain a national sample in the America must be familiar with the US geography and population composition in order to divide meaningful sampling levels and segments. If this type of information is not available or is too costly to cover too many segments, clustering sampling is another option. In this case the researcher might simply divide the population into certain relatively homogeneous units (clusters) and then subjects are selected from there (Martin, 2010). Although these sampling methods improve the quality of the data, the conventional DF is no longer applicable to these situations.

The myth of equal chances in simple random sampling

In Chapter 1 DF is defined as the effective sample size. This definition is based on the implicit assumption that all observations are equal. If we trace one step further, this assumption is built upon another hidden assumption: every element in the sampling space has equal chance to be sampled. However, this reasoning becomes problematic in the situation of multi-stage sampling. Before discussing the role of DF in statistical analysis that utilizes data collected from multi-stage sampling, it is necessary to debunk the myth of equal chance in simple random sampling.

Many textbooks define random sampling as a sampling process that each element within a set has equal chances to be drawn (e.g. Pagano, 2010). In other words, equality is associated with fairness. Indeed, “equal chance” is a theoretical ideal that is hardly observed in reality. For example, if I randomly throw a ball in a public area, children who have smaller bodies do not have equal chances to be hit by the ball as taller adults. If you observe any lucky drawing process carefully, you can see that usually the drawer reaches the middle or the bottom of the box. It is extremely rare for the host to grab a number from the top pile. Take putting “random” dots on a piece of paper as another example. If you ask any person to randomly draw 50 dots on a sheet of paper, would the dots randomly scatter around the paper? Do all the coordinates on the paper have equal chances to receive a dot? The answer is “no.” Almost all people would not draw on the edges of the paper and the so-called random dots actually form systematic clusters (Gardner, 2008). Jaynes (1995) fully explained this problem:

> The probability of drawing any particular ball now depends on details such as the exact size and shape of the urn, the size of balls, the exact way in which the first one was tossed back in, the elastic properties of balls and urn, the coefficients of friction between balls and between ball and urn, the exact way you reach in to draw the second ball, etc.. It is deliberating throwing away relevant information when it becomes too complicated for us to handle...For some,

> declaring a problem to be 'randomized' (sic, randomization is different from random sampling) is an incantation with the same purpose and effect as those uttered by an exorcist to drive out evil spirits...The danger here is particularly great because mathematicians generally regard these limit theorems as the most important and sophisticated fruits of probability theory. (pp. 319-320)

Before Jaynes, Poincare (1988) also made a similar criticism in an even more radical tone: "Chance is only the measure of our ignorance" (p.1359). Phenomena appear to occur according to equal chances, but indeed in those incidents there are many hidden biases and thus observers assume that chance alone would decide. Since authentic equality of opportunities and fairness of outcomes are not properties of randomness, a proper definition of random sampling should be a sampling process that each member within a set has *independent* chances to be drawn. In other words, the probability of one being sampled is not related to that of others.

At the early stage of the development of randomness, the essence of randomness was believed to be tied to independence rather than fair representation. It is important to note that when R. A. Fisher and his coworkers introduced randomization into experiment, their motive was not trying to obtain a representative sample. Instead they contended that the value of an experiment depends on the valid estimation of error (Cowles, 1989). In other words, the errors must be independent rather than systematic.

Simple random sampling vs. multi-stage sampling

Hence, the belief that random sampling gives every member in the population an equal chance of being samples could lead to erroneous results. Consider this scenario: A US researcher would like to obtain a representative sample across the entire nation. But if she blindly believes that simple random sampling treats every American equally, and takes the entire US population as a single sampling space, it is more likely that many people from larger states, such as California, Texas, and New York, will be sampled. But residents in Idaho and Wyoming might never appear on her radar screen.

To rectify this situation, she might start with randomly selecting several states out of 50 (first stage). Next, each state is divided into non-overlapping segments, and then certain counties in the chosen states are randomly drawn (second stage). In the last stage, subjects are randomly chosen from each county.

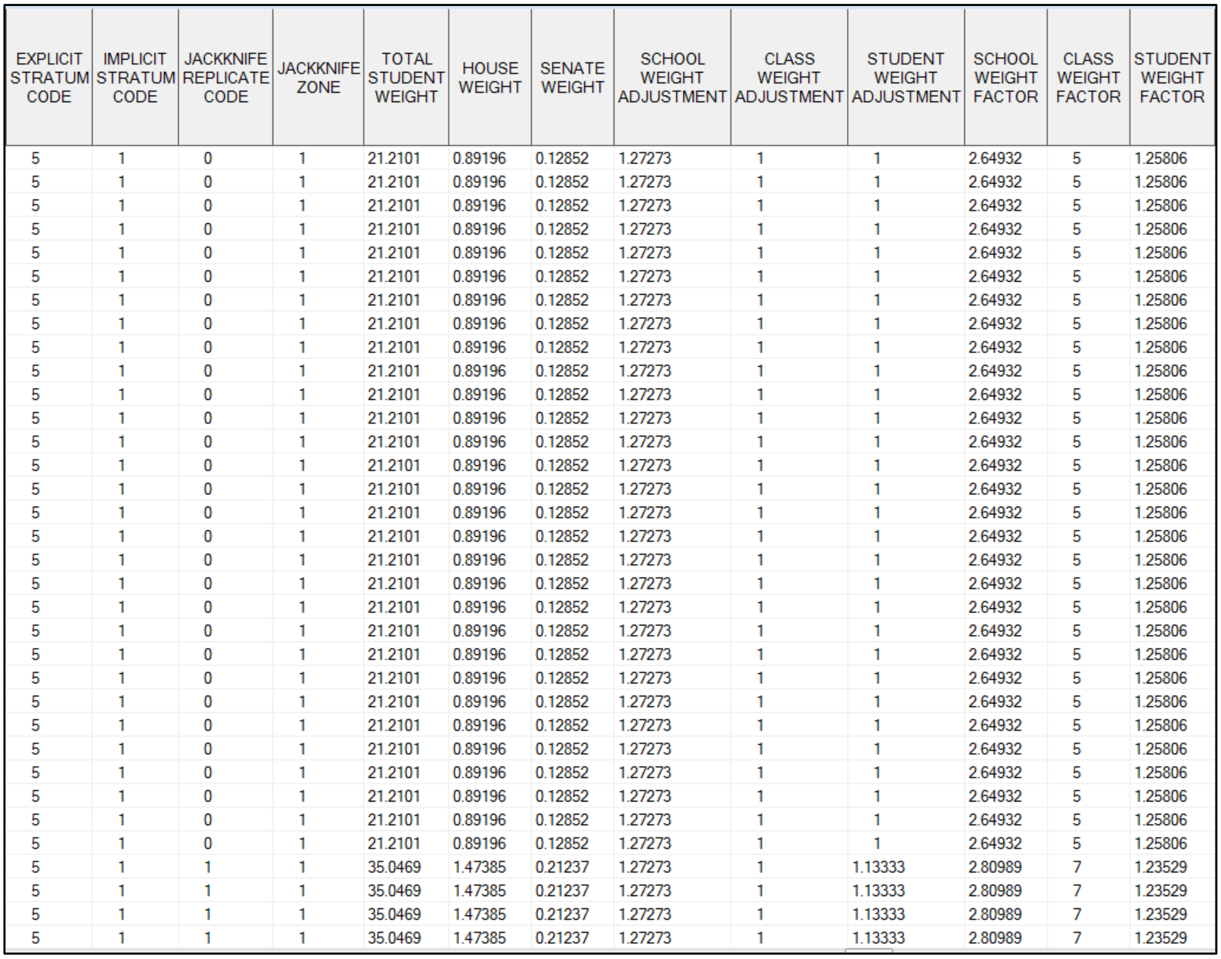
However, sometimes it is necessary to oversample certain smaller subsets. For instance, the researcher may include 10% Rhode Islanders (105,130) but only 1% Californians (376,919) into her sample. In this case, a sampling weight, also known weighting factor, is required to compensate for the over- or under-sampling segments of the population. If the sampling scheme entails a multi-stage design, then there will be several sampling weights.

Table 8.1. Population coverage and the sample size of Grade 8 in TIMSS 2007.

Country/Region	Population coverage	Participated schools	Students assessed
Hong Kong	100	120	3470
Japan	100	146	4312
Singapore	100	164	4599
South Korea	100	150	4240
Taiwan	100	150	4046
United States	100	239	7377

In the international arena, the *Trends for International Mathematics and Science Study* (TIMSS), which is a transnational assessment administered every four years to fourth and eighth grade students, also adopted a multi-stage sampling scheme. In the first stage, schools are sampled with probability proportional to size. Next, one or more intact classes of students from the target grades were drawn at the second stage. Because of its large population size, the Russian Federation added an additional stratum: regions. Singapore also had a third sampling stage, from which students were sampled within classes (Joncas, 2008).

Table 8.1 shows the population coverage and the sample size of six participating countries/regions in TIMSS 2007. In each of these countries/regions the population coverage is 100% because using a multi-stage sampling scheme instead of simple random sampling warrants that no particular segment of the population is ignored. However, readers may realize that comparing across these countries/regions might not be “fair” because the US is a large country but Hong Kong and Singapore are just cities. N = 3,470 is considered a large sample size for Hong Kong, but n = 7,377 is just like a drop in the ocean for the US (the Grade 8 student population in the US is about 4,300,000). In other words, Hong Kong is over-sampled relative to the US or the US is under-sampled relative to Hong Kong. The same issue happens to other strata, too (school, class). Again, we need sampling weights for compensation in order to obtain accurate computations and fair comparisons. Figure 7.1 shows a typical TIMSS data set, which indicates different strata and their sampling weights (e.g. school, class…etc.).



EXPLICIT STRATUM CODE	IMPLICIT STRATUM CODE	JACKKNIFE REPLICATE CODE	JACKKNIFE ZONE	TOTAL STUDENT WEIGHT	HOUSE WEIGHT	SENATE WEIGHT	SCHOOL WEIGHT ADJUSTMENT	CLASS WEIGHT ADJUSTMENT	STUDENT WEIGHT ADJUSTMENT	SCHOOL WEIGHT FACTOR	CLASS WEIGHT FACTOR	STUDENT WEIGHT FACTOR
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	0	1	21.2101	0.89196	0.12852	1.27273	1	1	2.64932	5	1.25806
5	1	1	1	35.0469	1.47385	0.21237	1.27273	1	1.13333	2.80989	7	1.23529
5	1	1	1	35.0469	1.47385	0.21237	1.27273	1	1.13333	2.80989	7	1.23529
5	1	1	1	35.0469	1.47385	0.21237	1.27273	1	1.13333	2.80989	7	1.23529
5	1	1	1	35.0469	1.47385	0.21237	1.27273	1	1.13333	2.80989	7	1.23529

Figure 8.1. Parts of TIMSS data sets showing sampling weights and strata.

Errors in standard errors

Ideally speaking, in simple random sampling every member in the sample pool has equal chance to be drawn. But as mentioned before, this is a theoretical construct. Suppose that we accept this theory, we can use conventional methods to calculate various statistics obtained by simple random sampling, such as the mean or the frequency. However, when the researcher wants to infer the sample statistics to the population parameters, how accurate would it be? Consider this hypothetical situation: The analyst draws 1,000 different samples for a study on TIMSS. These 1,000 data sets can yield 1,000 estimates of the statistics of interest, such as the mean. But these means are different and thus there is a dispersion of the means, of course. This dispersion constitutes the sampling variance of the mean whereas the square root of this sampling variance is known as the standard error. As the sample size increases, the variance is reduced and the accuracy of the estimation becomes better. The following formula is used to compute the sampling variance:

σ^2/n

where

σ^2= the population standard deviation
n = sample size

Any estimation entails uncertainty. Thus, the standard error is used for computing the confidence interval. We can estimate the population standard deviation by using the sample standard deviation. However, while it can apply to the case of simple random sampling, it does not work for a sample obtained from multi-stage sampling. It is due to the fact that in a multi-stage sampling not all members have equal chances to be drawn. Thus, if we employ conventional statistical methods, the standard error would be incorrect (Williams, 2004).

DF, strata and clusters

Specific adjustments based on weighting must be made to data collected from multi-stage sampling. To alleviate the problem, SAS (SAS Institute, 2011b) equips the users with procedures that utilize the degrees of freedom for adjusting different statistics, such as SYRVEYMEAN, SURVEYFREQ, and SURVETREG. When the sampling weights for the strata are provided, the degrees of freedom would become the number of replicates (also known as subsamples, segments, or subsets). If the stratum information is not provided, the DF is the number of clusters minus one. Take TIMSS 2007 as an example. There are 239 schools or clusters in the US sample, and thus the DF value is 239-1 = 238 (To simplify the illustration, the following examples use clusters instead of strata).

SURVEYMEANS

In the following the TIMSS 2007 science assessment data of US Grade 8 students are utilized for illustration. Table 8.2 shows the estimation of the mean, the standard error, and the confidence intervals by the traditional method and the SURVEYMEANS approach, which adjusts the degrees of freedom by using the number of clusters. There are 239 schools in the US data set. The first row of Table 8.2 shows the mean, the standard error, and the confidence interval obtained by the conventional method: sum of all values divided by the number of observations. This method assumes that every observation has equal chance to be drawn from the population, and thus the calculation has a hidden weight as 1. In the SURVEYMEANS method there are 239 clusters (schools) and students in each school have different chances to be sample. Thus, their sampling weights and the adjusted degrees of freedom are taken into account while the mean, the standard error, and the confidence band are calculated.

Table 8.2. Estimations from conventional method and survey means

Method	Mean	Std Error of Mean	95% CL for Mean	
Conventional method: Sum/n	517.269	0.9738230	515.360125	519.1780675
Survey Means	522.174	1.4306130	519.369248	524.9780690

SURVEYFREQ

Next, we turn our attention to frequency count. Table 8.3 shows both the traditional frequency and the weighted frequency of responses to the question “Do you use a computer at school?” The traditional frequency is descriptive whereas the weighted version is the estimate of the value in the population. The standard errors computed yielded from this procedure are based on the multistage sampling scheme with DF = 239 – 1 = 238. This is different from the traditional estimation, which assumes simple random sampling from an infinite population. Like SURVEYMEANS, the SUVEYFREQ approach yields a more trustworthy result.

Table 8.3. Comparing traditional frequency and weighted frequency.

	Traditional Frequency	Weighted Frequency	SD of Weighted Frequency	Percent	Standard Error of Percent
Yes	5356	566493	4759	79.2559	0.6657
No	1857	148271	4759	20.7441	0.6657
Total	7213	714764	0.0001703	100.000	

SURVEYREG

The TIMSS data set has many variables and different weighting factors, and thus it could be very confusing to use TIMSS to illustrate the role of DF in SURVEYREG. In the following a hypothetical data set sourced from SAS Institute (2012) is used instead. Suppose that in a junior high school there are 4,000 students spanning across Grades 7, 8, and 9. The researcher wanted to use household income and the number of children in a household to explain or predict students’ average weekly spending for ice cream. The researcher tried his best to reduce bias in the sampling process. If simple random sampling is employed, there might be more students from a particular grade level. To avoid this potential flaw, the researcher divided the sampling space into three segments by the grade level. She knew the population size of each grade level and tried to sample the same proportion across all grades. However, some invited students refused to participate in the survey. At the end she obtained 40 observations, but the sample-population ratios are not the same in all grades. Table 8.4 provides the information about the sample size and the population size in each grade.

Table 8.4. Sample size and population size in each grade.

Grade	Sample size	Population size
7	20	1824
8	9	1025
9	11	1151
Total	40	4000

The researcher could ignore the grade-level information, treat every observation as equal, and proceed with a regular regression analysis. Table 8.5 shows the regression output of using household income and the number of kids as predictors of spending for ice cream. There is an overall significant effect for $p <$ .0001. In Chapter 4, we learned that

DF total = $n – 1$
DF model = k
DF error = $n – k – 1$

Not surprisingly, in this example DF model is 2 whereas DF error is 37. And DF total is 39.

DF total = 40 – 1 = 39
DF model = 2
DF error = 40 – 2 – 1 = 37

Table 8.5. Significance of the overall regression model.

Source	DF	Sum of Squares	Mean Square	*F* Value	*P* value
Model	2	903.327717	451.663859	75.22	<.0001
Error	37	222.172283	6.004656		
Corrected Total	39	1125.500000			

Table 8.6 shows the significance or insignificance of each predictor in terms of its contribution to variance explained. Apparently, household income is a significant predictor ($p <$ .0001) while the number of kids is not (p = .2486). In Chapter three we learned that in a pairwise perspective, there is only one parameter to be estimated. Therefore, the degree of freedom is always set to 1 for each bivariate relationship.

Table 8.6. Significance of each predictor.

Source	DF	Type III SS	Mean Square	*F* Value	P Value
Income	1	883.1189594	883.1189594	147.07	<.0001
Kids	1	8.2502675	8.2502675	1.37	0.2486

No doubt the researcher can obtain more insights from the data if the grade-level information is taken into account. Table 8.7 demonstrates how the population information could be utilized to compute the probability of being sampled in each grade and the sampling weight. In each grade, the probability that an individual is sampled yields from this formula: sample size/population size. It is obvious that students in different grades do not have equal chances to be sampled. Students in Grade 7 has the highest probability of being sampled (p = 0.01096), students in Grade 9 has the second highest probability (p = 0.00956), and students in Grade 8 has the lowest one (p = 0.00878). At most we could say the probability that any student is drawn from the population does not depend on or is not correlated with others. To be more specific, no student can do anything to increase the sampling probability of another student.

The sampling weight is the inverse of the probability because the over-sampled group should be down-weighted while under-sampled group should be compensated by increasing the weight. After the inversion, students in Grade 7 that have the highest probability of being sampled receive the lowest weight. Conversely, students in Grade 9, the under-sampled group, receive the highest weight. In the subsequent computation, every score is adjusted by the weight (score*weight).

Table 8.7. Probability of being sampled and sampling weight.

Grade	Sample size	Population size	Probability of being sampled	Sampling weight
7	20	1824	20/1824=0.01096	1/probability=91.2
8	9	1025	9/1025=0.00878	1/probability=113.889
9	11	1151	11/1151=0.00956	1/probability=104.636

Table 8.8 shows the summary output produced by SURVEYREG. Instead of reporting the arithmetic mean of money spent for ice cream (sum/sample size), SAS reports the weighted mean (9.1413).

Table 7.8. Summary output of SURVEYREG.

Number of Observations	40
Sum of Weights	4000.0
Weighted Mean of Spending	9.14130
Weighted Sum of Spending	36565.2

Table 8.9 informs us the model effects. The denominator degrees of freedom for the F tests is still 37 (DF error = *n* – *k* – 1 = 40 – 2 – 1 = 37). However, the numerator DF (DF model) is no longer 2. Because we have four parameters to be estimated (the income effect and the effect in each grade), DF model = k = 4. For the number of kids effect, the DF is the number of segments. The sample is stratified by the grade level and therefore DF = 3. There is no stratification for household income and thus DF = 1. In this result the Income effect is significant but the number of children effect is not.

Table 8.9 Tests of Model Effects.

Effect	DF	*F* value	*P* value
Model	4	124.85	<.0001
Intercept	1	150.95	<.0001

Effect	DF	*F* value	*P* value
Income	1	326.89	<.0001
The number of kids	3	0.99	0.4081

Table 8.10 displays the regression coefficients. The number of children effect was examined in each grade level, but none of them yielded significant result.

Table 8.10. Estimated Regression Coefficients.

Parameter	Estimate	Standard error	*t* value	*P* value
Intercept	-26.086882	2.44108058	-10.69	<.0001
Income	0.776699	0.04295904	18.08	<.0001
Kids 1 (Grade 7)	0.888631	1.07000634	0.83	0.4116
Kids 2 (Grade 8)	1.545726	1.20815863	1.28	0.2087
Kids 3 (Grade 9)	-0.526817	1.32748011	-0.40	0.6938

Summary

Conventional procedures treat every observation equally and thus DF could be defined as the effective sample size. However, the notion of equal chance in random sampling is more a theoretical ideal than an observed fact. The general rule of degrees of freedom can be applied to data sets collected from simple random sampling. When multi-stage sampling is employed, conventional procedures that treat every observation equally would produce misleading results. As a remedy, sampling weights must be taken into account and also the degrees of freedom must be adjusted according to the number of strata or clusters.

Chapter 9
DF, normality, and sampling distributions

Walker (1940) found that some of her contemporary texts seem to imply the prevalence of normal curves. But she added that if the normal distribution can adequately describe all sampling distributions, then the concept of DF would be relatively unimportant, because no matter what the sample size and the DF value are, the shape of the distribution would remain the same. However, this is not the case. Thus, in constructing various sampling distributions, the DF functions as a parameter and probability tables built from these distributions must be associated with the correct values of n and DF.

In many introductory statistics class, the first sampling distribution that students encounter is the standard normal distribution, and the first or second computing-intensive exercise is transforming the raw scores to z scores assuming a normal curve. In a test and measurement class students learn more about standard scores under the umbrella of normal distribution, such as IQ, SAT, T scores, and Stanines (see Figure 9.1(a) and (b)). Therefore, it is not surprising to see that many students regard the normal curve is *the* sampling distribution.

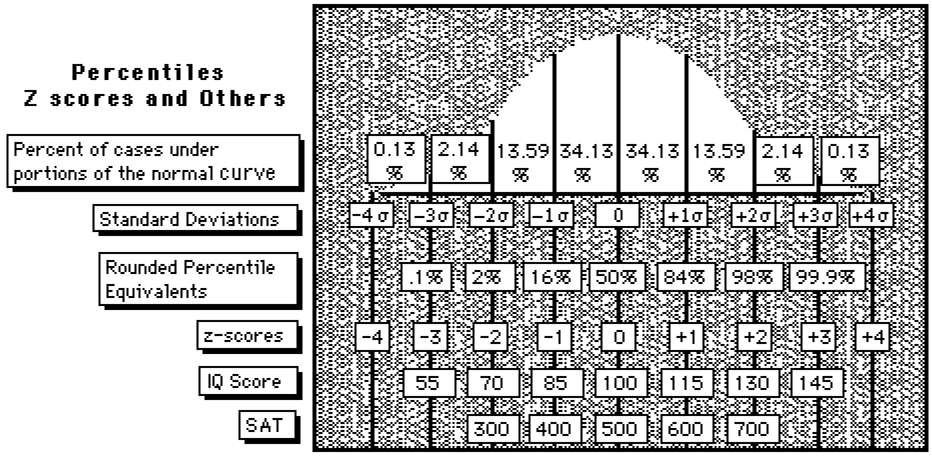


Figure 9.1(a). Standard normal curve, SD, percentile rank, z scores, IQ, and SAT.

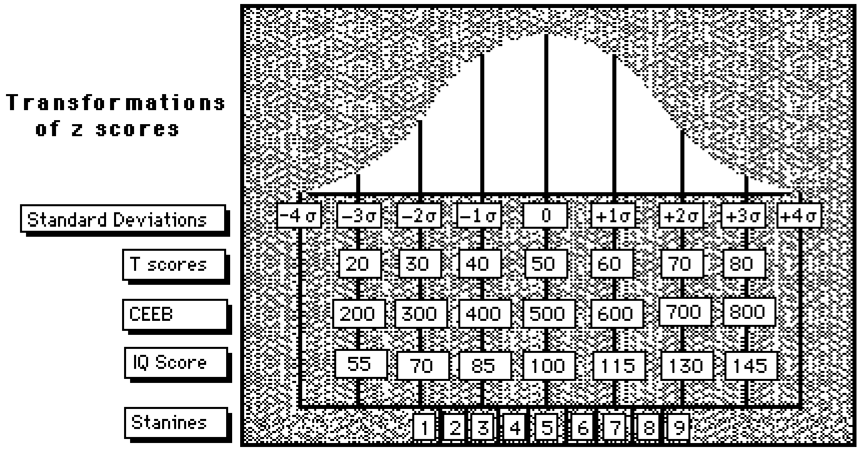


Figure 9.1(b). Standard normal curve, T scores, CEEB, IQ, and Stanines.

Before showing the relationship between DF and various sampling distributions, it is important to debunk the myth of the universality of normal curves. As Walker said, if normality is ubiquitous, the role of DF would virtually vanish.

Myth of normality

The belief that most distributions are normal is hardly an empirical fact. As early as 1900, Pearson was critical of normal curves because in his view the normal curve possesses no special fitness for describing errors or deviations. French physicist Lippmann pointed out the circular logic of proving normality: "Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact" (cited in Thompson, 1959, p.121). In a similar vein to Lippmann, Stigler (1986) also criticized the circular logic employed by Gauss, who developed the Guassian (normal) distribution. Gauss conceptualized the mean in terms of "least squares": the mean could be used to summarize a data set, because when more observations are closer to the mean and fewer observations are farther from the mean, the sum of squares of the deviation is minimal. The mean is only "most probable" if the errors (deviations) are normally distributed; and the supposition that errors are normally distributed leads back to least squares. In response to the lack of proof of universal normal distributions, Geary (1947) stated that normality could be viewed as a special case of many distributions rather than a universal property. However, universal normality has been favored and interest in non-normality has retreated to the background. In conclusion, Geary suggested that future editions of all existing textbooks and new textbooks should include this warning: "Normality is a myth; there never was, and never will be, a normal distribution." (p.241). Needless to say, this warning has been ignored for over six decades.

Chi-square distribution

To be fair, some sampling distributions approach normality as the degrees of freedom increases, such as the Chi-square distributions. But as shown in Chapter 5, it is common that researchers work with a small number of degrees of freedom in Chi-square analysis, and thus very often the underlying distribution is by no means normal.

The Chi-square distribution is constructed by the following formula:

$Y = Y_0 * (\chi^2)^{(DF/2 - 1)} * e^{-\chi 2 / 2}$

where

Y_0 = a constant that is tied to the degrees of freedom
χ^2 = the chi-square statistic
DF = degrees of freedom
e = the base of the natural logarithm, also known as Euler's *e* (approximately 2.718).
Y_0 is defined, so that the area under the chi-square curve is equal to one.

The Chi-square distribution has the following attributes:

1. The expected mean of the Chi-square distribution is equal to the degrees of freedom.
2. The expected variance is equal to two multiplies the degrees of freedom: σ2 = 2 * DF
3. When DF is greater than or equal to 2, the maximum value for Y occurs.
4. As the degrees of freedom increase, the chi-square distribution approaches normality.

The above formula and attributes might mean nothing to many readers. Visualization might be helpful. As shown in Figure 9.2, when DF = 1 or 2, the Chi-square distribution is highly skewed. It gets improvement when DF = 3 or 5. It becomes fairly normal and the degree of skewness is trivial when DF = 10. Walker (1940) said that the Chi-square curve becomes approximately normal when n is 30 or so. Actually if you experiment with R, you can see that approximation of normality occurs before n = 30. This widespread myth will be discussed in the next section.

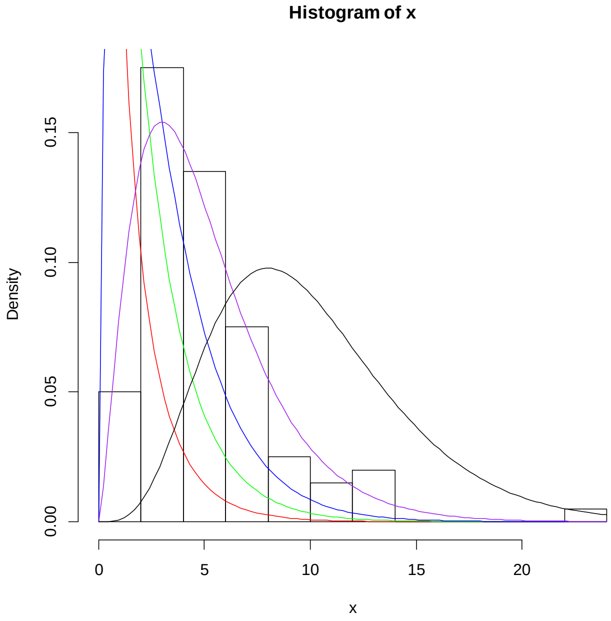


Figure 9.2. Chi-square distributions with different DF values.

Student's t distribution

When the population variance is unknown, we need the *t*-distribution to estimate the population parameters. The formula of computing the *t* statistics is:

t = [x - μ] / [s / SQRT(n)]

where

x = sample mean
μ = population mean
s = standard deviation of the sample,
n = sample size.

There is a family of *t*-distributions. The appearance of a specific *t*-distribution depends on the degrees of freedom, which is *n* – 1. Figure 9.3 shows various *t*-distributions when the DF values are set to 1, 3, 8, and 30, meaning that the numbers of observations are 2, 4, 9, and 31, respectively.

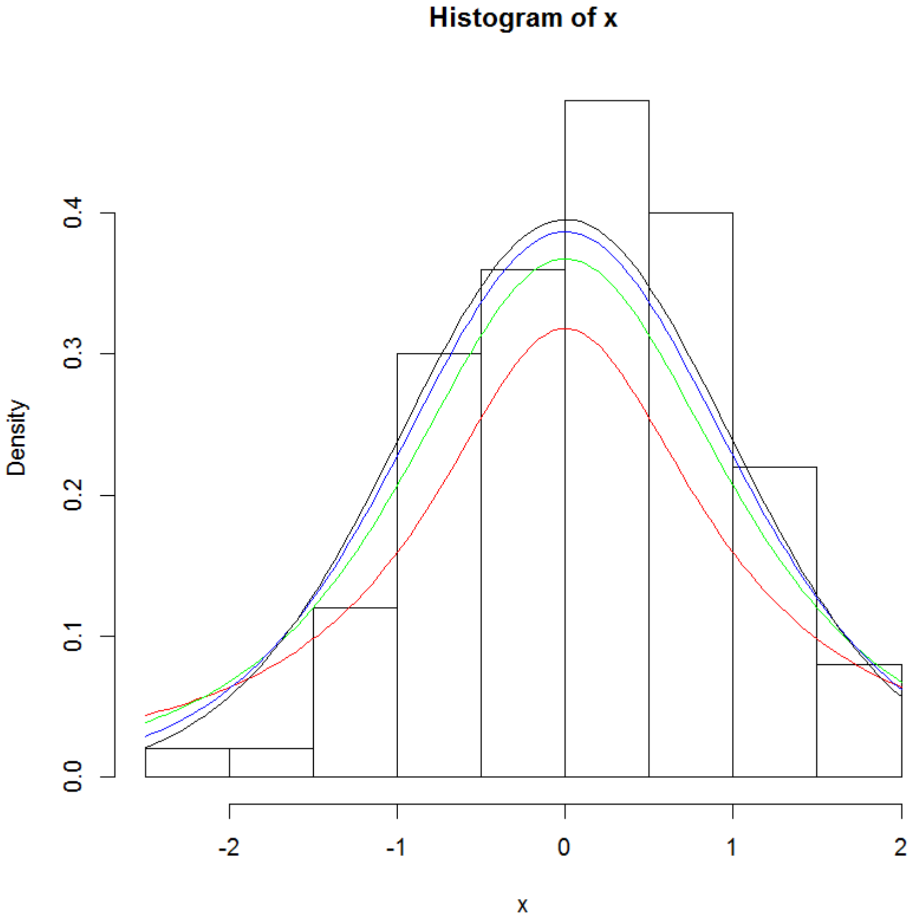


Figure 9.3. Student's *t*-distributions with different DF values from R code 1.

The *t*-distribution has the following characteristics:

1. Like the *z*-distribution, the expected mean of the *t*-distribution is equal to 0.
2. The expected variance is equal to DF / (DF - 2) and DF > 2.
3. The variance is always greater than 1. It is close to 1 when there are many degrees of freedom.
4. When DF is infinite, the *t*-distribution is the same as the standard normal distribution.

There is an urban legend that when n = 30 or DF > 30, the t-distribution becomes normal. This misconception is popularized by some statisticians who tried to set a cut-off point of large sample size for generating a bell-shaped sampling distribution out of a non-normal population (Saddler, 1971). Gordon and Gordon (1989) argued that if the sample size is less than 30 and the population is not normal, the sampling distribution will be a *t*-distribution, otherwise, a normal distribution. Velleman (1997), the inventor of the statistical software package *DataDesk*, traced this misconception back to the pre-computer age: When high power computers were not available, statisticians had to use a *t*-table only. The *t*-family goes on and on for any number of degree of freedom (DF) and *t*-distribution is really normal if and only if the degree of freedom is infinite. However, it was not practical to have such a long *t*-table. As a result, a compromise was made at around 30 DF because it could fit nicely in one page.

This urban myth has been debunking by many authors over and over for several decades (Hesterberg, 2008). In an article entitled "Why is n = 30 'magic'? … and other frequently asked questions," again Miller (2006) reminded us that 30 is not a magical number for switching from *t* to *z* in inference about a mean; indeed the distinction between *t* and *z* procedures for a mean has nothing to do with sample size. Like Velleman, Miller attributed this misconception to the pre-computer "dark ages": "In the dark ages, when we used tables with one line for each sample size, about 30 lines would fit on a page. So, after n = 30, we were advised to switch from exact t to approximate z" (p.8).

F distribution

The *F*-distribution is a continuous probability distribution, which is often used in Analysis of Variance. An *F*-distribution can be formulated as the following:

$$F = \frac{V1/DF1}{V2/DF2}$$

where

V1 and V2 = two independent random variables having the Chi-Square distribution
DF1 = numerator degrees of freedom
DF2 = denominator degrees of freedom

In other words, the *F* value is the ratio of two Chi-squares. The F-distribution has the following properties:

1. The expected mean of the *F*-distribution is equal to DF1 / (DF2 - 2).
2. The variance is equal to [2*DF2^2 * (DF1 + DF2 - 2)] / [DF1 * (DF2 - 2)^2 * (DF2 - 4)]

3. The F-distribution is positively skewed. Its shape is like a Chi-square with small degrees of freedom.
4. The *F*-ratio is always positive because the variance is squared. The range of the *F* value is from zero to infinity.
5. The *F* curve reaches a peak near 0, and then approaches, but never touches the horizontal axis.
6. $F = t^2$ with numerator DF = 1

Again, the shape of the distribution responds to the degrees of freedom (see Figure 9.4).

There are many other sampling distributions, namely, Binominal distribution, Poisson distribution, Negative Binomial Distribution, Geometric Distribution, Gamma Distribution, Weibull Distribution, Log-Normal Distribution, and Beta Distribution. But some are rarely used. Nonetheless, you can learn about how DF affects the appearance of these distributions by altering the DF values in the R code.

Summary

As Walker said, if the normal curve can sufficiently describe all sampling distributions, then DF would not play an important role in statistics. However, the universality of normality is an urban legend. The appearance of different sampling distributions is defined by the DF values. And there is no magic cut-off number. It is advisable to put aside the belief that n = > 30 or DF = > 30 would automatically turn a distribution into a normal curve.

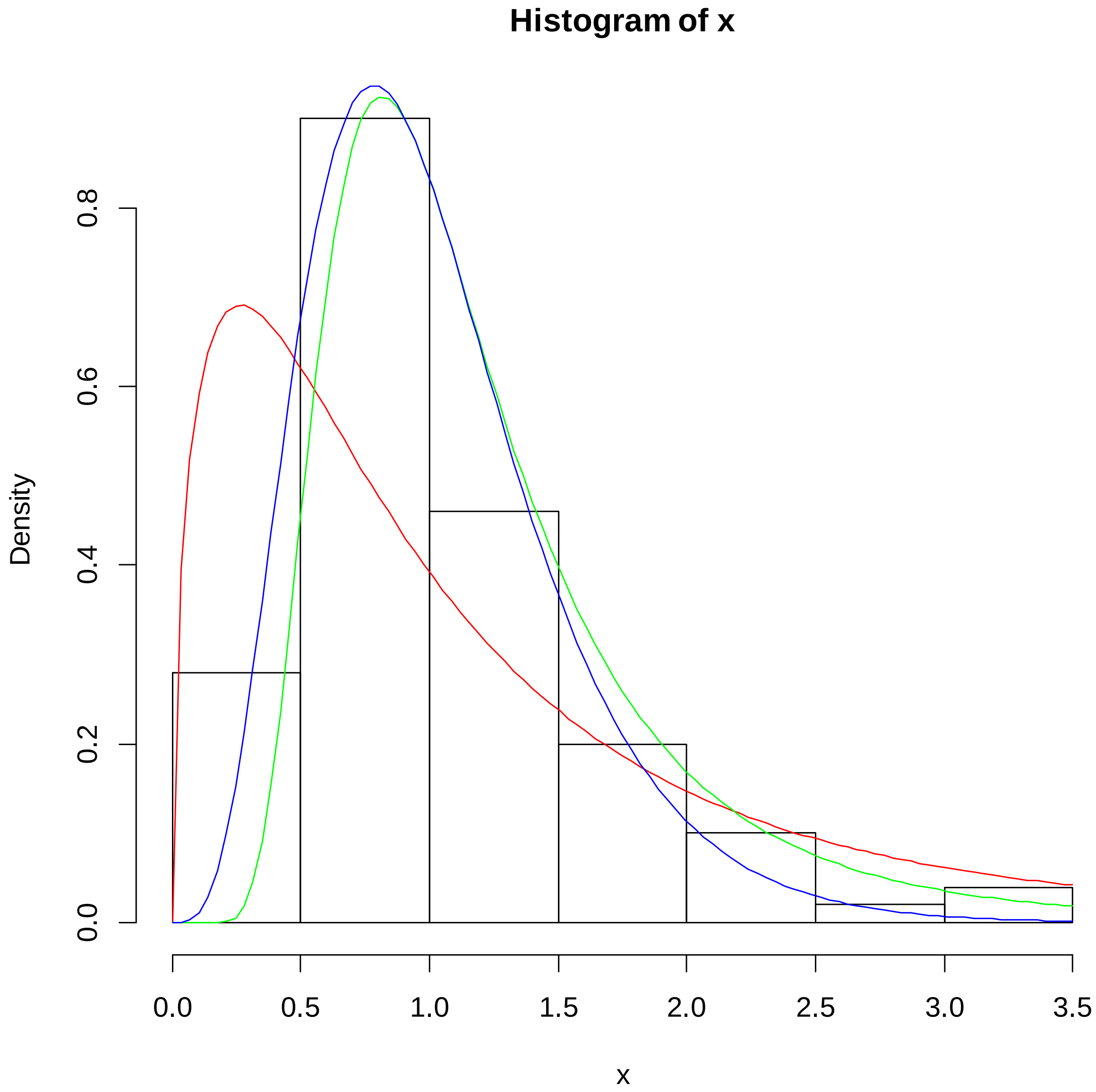


Figure 9.4. *F*-distributions with different DF values.

Epilogue

This is the end of the journey. The author hopes that by now the concept “degrees of freedom” is no longer an “intimate stranger” to you. From this point on, when you report any statistics with the degrees of freedom, you might appreciate its importance. Simply put, DF is situated in the context of model-based or distribution-based methodology. In exploratory data analysis or data mining, which does not requires hypothesis testing, you don’t have to report any degrees of freedom. When you have access to the population-level data, there is no need to reduce the total sample size to the effective sample size. In a mere descriptive study that does not demand making inferences from the sample statistics to the population parameter, again DF ceased to play any role. In all of the three preceding scenarios your life as an analyst would be much easier. To be more specific, in those situations you are a *data* analyst, not a *statistical* analyst. Indeed long time ago Karl Pearson (1900) had embraced the idea of drawing conclusions based upon the data at hand, but it was counteracted by R. A. Fisher (1922) and his followers. Today the Fisherian legacy is the dominant paradigm in quantitative methodologies. Thus, Chi-square, ANOVA (*F* test), and many other statistics must be compared against certain theoretical sampling distributions to determine whether the null hypothesis should be rejected. If all distributions are normal, we don’t need DF at all. But normality is just a special case of many different distributions and the shape of these distributions depends on DF. Therefore, we have to take DF into account while computing the data, and at the end it is necessary to report the DF values associated with the statistics.

We do not directly infer from the sample to the population. Rather, the sampling distributions function as a bridge between the two, and these intangible theoretical distributions make DF a mystery. By far the Walker’s (1940) and the Good’s (1973) are the best two to demystify DF. But no approach is perfect. As mentioned before, Good commented that Walker’s notion of “necessary relationships” is very vague. Interestingly enough, Good is not immune against criticism. Dallal (2009) asserted that the Good’s approach to DF is a partial answer because “it explains what degrees of freedom is for many Chi-square tests and the numerator degrees of freedom for F tests, but it doesn't do as well with *t* tests or the denominator degrees of freedom for *F* tests” (p. 1). As a matter of fact, both approaches left some cases unexplained. For example, for data collected from multi-stage sampling, the conventional sense of DF falls apart because Fisher, Walker, and Good were alienated from the concepts of strata and clusters. As statistical knowledge keeps evolving, it is the conviction of the author that this book will also be criticized as vague and inadequate. Nevertheless, if you are patient enough to walk through the exercises and also the “side trips” in the book, at this point the DF concept should be your friend, not your “intimate stranger”.

References

Agresti, A., & Finlay, B. (1986). *Statistical methods for the social sciences*. San Francisco, CA: Dellen.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *International Symposium on Information Theory* (pp.267–81). Budapest: Akademia Kiado.

Baird, D. (1983). The Fisher/Pearson chi–squared controversy: A turning point for inductive inference. *British Journal for the Philosophy of Science, 34*, 105–118.

Besag, F. (1980). Academic science, policy decision, and Chi-square. *Urban Education, 15*, 215-230. DOI: 10.1177/0042085980152006.

Buchner, A., Faul, F, & Erdfelder, E. (2012). G*Power. Retrieved from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Burnham, K. P., and Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed). New York, NY: Springer-Verlag.

Burnham, K. P., & Anderson, D.R. (2004), Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261-304.

Consumer Reports. (2012, June). The most fuel-efficient cars. Retrieved from http://www.consumerreports.org/cro/2012/02/the-most-fuel-efficient-cars/index.htm

Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, New Jersey: LEA.

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Dallal, G. E. (2009). Degrees of freedom. Retrieved from http://www.tufts.edu/~gdallal/dof.htm

Eisenhauer, J. G. (2008). Degrees of Freedom. *Teaching Statistics, 30*(3), 75–78.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables and the calculation of *P*. *Journal of Royal Statistical Society, 85*, 87-94.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science, 1*, 115–117.

Flatto, J. (1996, May 3). Degrees of freedom question. *Computer Software System-SPSS Newsgroup* (comp.soft-sys.spss).

Galfo, A. J. (1985). Teaching degrees of freedom as a concept in inferential statistics: An elementary approach. *School Science and Mathematics, 85*(3), 240-247.

Gardner, D. (2008). *The science of fear: Why we fear the things we shouldn't--and put ourselves in greater danger*. New York, NY: Dutton Adult.

Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In V. Barnett (Ed.) *Interpreting multivariate data*. London: John Wiley & Sons.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34*, 209-241.

Gnambes, T. (2008). Required sample size and power for SEM. Retrieved from http://timo.gnambs.at/en/scripts/powerforsem

Good, I. J. (1973). What are degrees of freedom? *American Statisticians, 27*, 227-228.

Gordon, F. S., & Gordon, S. P. (1989). Computer graphics simulations of sampling distributions. *Mathematics and Computer Education, 21*, 48-55.

Greenwood, M. & Yule, G. U. (1915). The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Proceedings of Royal Society of Medicine, Section of Epidemiology and State Medicine, 8*, 113-90.

Hacking, I. (1992). *The taming of chance*. Cambridge, UK: Cambridge University Press.

Harris, J. A. & Treloar, A. E. (1927). On a limitation in the applicability of the contingency coefficient. *Journal of American Statistics Association, 22*, 460-472.

Harris, J. A. & Tu, C. (1929). A second category of limitations in the applicability of the contingency coefficient. *Journal of American Statistics Association, 24*, 367-375.

Hausner, M. (1965). *A vector space approach to geometry*. Mineola, NY: Dover Publications.

Hays, W. L. (1981). *Statistics*. New York: Holt, Rinehart and Winston.

Hesterberg, T. (2008, August). *It's time to retire "n >= 30".* Paper presented at the Joint Statistical Meeting, Denver, CO.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Howell, D. C. (1992). *Statistical methods for psychology*. (3rd ed.). Belmont, CA: Duxberry.

IBM SPSS. (2011a). Amos 20. [Software]. Armonk, NY: Author.

IBM SPSS. (2011b). SPSS 20. [Software]. Armonk, NY: Author.

Institute for Statistics and Mathematics. (2012). The R Project for statistical computing. Retrieved from http://www.r-project.org/

Jaccard, J. & Becker, M.A. (1990). *Statistics for the behavioral sciences*. (2nd ed.). Belmont, CA: Wadsworth.

Jaynes, E. T. (1995). Probability theory: The logic of science. Retrieved from http://www.math.albany.edu:8008/JaynesBook.html

Johnson, R. A. & Wichern, D. W. (1998). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

Joncas, M. (2008). TIMSS 2007 sample design. In J. F. Olson, M. O. Martin, & I. V. S. Mullis, (Eds.). *TIMSS 2007 technical report* (pp. 77-92). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kabacoff, R. (2012). Quick R: Probability plots for teaching and demonstration. Retrieved from http://www.statmethods.net/advgraphs/probability.html

Kim, K. H. (2005). The relation among fit, indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling, 12*, 368-390.

Lomax, R. G. (1992). *Statistical concepts: A second course for education and the behavioral sciences*. White Plains, NY: Longman.

Mathews, P. (2004). Degrees of freedom. Retrieved from http://www.mmbstatistical.com/Notes/DegreesOfFreedomPri.pdf

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. New York, NY: Brooks/Cole Publishing Co.

Martin, K. (2010). Analyzing data from complex samples using PROC SURVEYFREQ in SAS 9.2. *Proceedings of Western Users of SAS Software Conference*. Retrieved from http://www.wuss.org/proceedings10/

Maxwell, S., & Delany, H. (2003). *Designing experiments and analyzing data* (2nd ed).Belmont, CA: Wadworth.

Moore, D. S. & McCabe, G. P. (1989). Introduction to the practice of statistics. New York: W. H. Freeman and Company.

Pandey, S., & Bright, C. L. (2008). What are degrees of freedom? *Social Work Research, 32*, 119-128.

Pagano, R. (2010). *Understanding statistics in the behavioral sciences*. Belmont, CA: Wsadsworth.

Pearson, E. S. (1938). *Karl Pearson: An appreciation of some aspects of his life and work*. Cambridge: The University Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*, 157–175.

Pearson, K. (1930). On the theory of contingency: I. Note on Professor J. Arthur Harris' papers on the limitation in the applicability of the contingency coefficient (with a Reply by J. A. Harris, A. E. Treloar and M. Wilder, and a Postscript by Karl Pearson). *Journal of American Statistics Association, 25*, 320-7.

Poincare, H. (1988). Chance. In J. Newman (Ed.), *The world of mathematics* (Vol. 2). (pp. 1359-72). Redman, WA: Tempus.

Popper, K. R. (1959). *Logic of scientific discovery*. London : Hutchinson.

Popper, K. R. (1974). Replies to my critics. In P. A. Schilpp (Eds.), *The philosophy of Karl Popper* (pp.963-1197). La Salle: Open Court.

Press, S. J., & Tanur, J. M. (2001). *The subjectivity of scientists and the Bayesian approach*. New York: John Wiley & Sons.

Public Broadcasting System. (1999). Intimate strangers: Unseen life on earth. Retrieved from http://www.pbs.org/opb/intimatestrangers/

Rawlings, J. O., (1988). *Applied regression analysis: A research tool*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Rigdon, E. E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling, 1*, 274-278.

Saddler, D. R. (1971). The central limit theorem: An empirical investigation. *The Australian Mathematics Teacher, 27*(3), 90-94.

SAS Institute. (2011a). SAS 9.3. [Software]. Gary, NC: Author.

SAS Institute. (2011b). JMP Pro 10 [Software]. Gary, NC: Author.

SAS Institute. (2012). Getting started: SURVEYREG procedure. Retrieved from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveyreg_sect002.htm

Savalei V, & Bentler P. M. (2006). Structural equation modeling. In Grover R, Vriens M, (eds.) *The Handbook of marketing research: Uses, Misuses, and future advances* (pp. 330–364). Thousand Oaks, CA: Sage.

Saville, D. & Wood, G. R. (1991). *Statistical methods: The geometric approach*. New York: Springer-Verlag.

Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74. Retrieved from http://user.uni-frankfurt.de/~kscherm/schermelleh/mpr_Schermelleh.pdf

Shin, J. (2009). Application of repeated-measures Analysis of Variance and Hierarchical Linear Model in nursing research. *Nursing Research, 58*, 211-217.

Shin, J., Espin, C. A., Deno, S., McConnell, S. (2004). Use of hierarchical linear modeling and curriculum-based measurement for assessing academic growth and instructional factors for students with learning difficulties. *Asia Pacific Education Review, 5*, 136-148.

Snow, G. (2008). How to plot Chi-square distribution in the graph. Retrieved from http://r.789695.n4.nabble.com/how-to-plot-chi-square-distribution-in-the-graph-td872614.html

Statistical Support. (2001). *Structural equation modeling using AMOS: An introduction*. Retrieved from http://www.utexas.edu/cc/stat/tutorials/amos/index.html

Steinberg, W. (2008). *Statistic alive!* Los Angeles, CA: Sage.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.

Stigler, S. M. (1992). Studies in the history of probability and statistics XLIII Karl Pearson and quasi-independence. *Biometrika, 79*, 563-575.

Thompson, D. W. (1959). *On growth and form*. Cambridge: Cambridge University Press.

Toothaker, L. E., & Miller, L. (1996). *Introductory statistics for the behavioral sciences*. (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Velleman, P. (1997, Feb.) Is n=30 large enough? Retrieved from http://jse.stat.ncsu.edu/11/edstat

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage Publications.

Walker, H. W. (1940). Degrees of Freedom. *Journal of Educational Psychology, 31*, 253-269.

Weiss, R. S. (1968). *Statistics in social research: An introduction*. New York, NY: John Wiley & Sons.

Weisstein, E. W. (2011). Degree of freedom. Retrieved from http://mathworld.wolfram.com/DegreeofFreedom.html

Wickens, T. (1995). *The geometry of multivariate statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Williams, J. D., & Smith, T. (2004). *A manual for conducting analyses with data from TIMSS and PISA*. Retrieved from http://www.unb.ca/crisp/pdf/Manual_TIMSS_PISA2005_0503.pdf

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

World Bank. (2012). World Development Indictors (WDI) and Global Development Finance (GDF). Retrieved from http://databank.worldbank.org/ddp/home.do

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika, 92*, 937-950.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yu, C. H. (2010). A model must be wrong to be useful: The role of linear modeling and false assumptions in theoretical explanation. *Open Statistics and Probability Journal, 2*, 1-8. Retrieved from http://www.bentham.org/open/tospj/openaccess2.htm

Yu, C. H. (2012a). A simple guide to Item Response Theory (IRT) and Rasch Modeling. Retrieved from http://www.creative-wisdom.com/computer/sas/IRT.pdf

Yu, C. H. (2012, September). *Blurring the line between confirmation and exploration: Model comparison of structural equation modeling in JMP/SAS*. Paper presented at Western Users of SAS Software Conference, Long Beach, CA.

Yule, G. U. (1922). On the application of the χ^2 Method to association and contingency tables with experimental illustration. *Journal of Royal Statistical Society, 85*, 95-104.

Appendix
Computing exercises

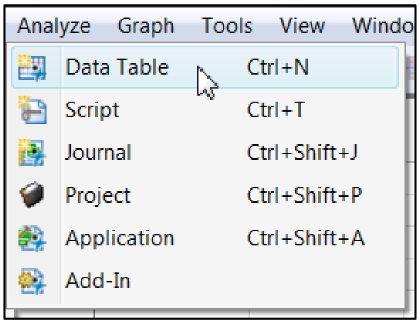
Chapter 1 Exercise

Objective: These exercises can help you understand DF in terms of effective sample size and minimum information for meaningful estimations.

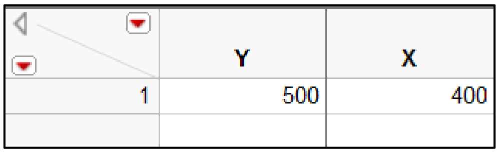
Tool: You can use either JMP (SAS Institute, 2011b) or SPSS (IBM SPSS, 2011b) to run the analysis. The academic license of JMP costs $29.95 per six months or $49.95 per year. You can go to www.jmp.com to download a 30-day trial version. Please download JMP, not JMP Pro.

JMP Steps

1. Open JMP and create a new data table

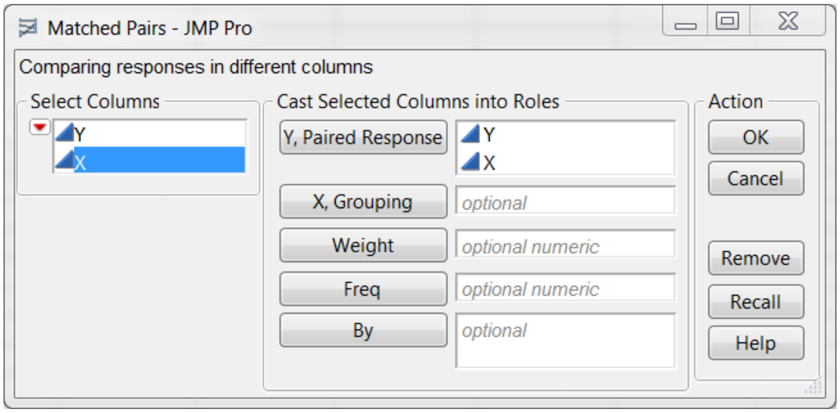


2. Enter “500” in the first column of the data table and “400” in the second column. Double-click the column headers and rename them to be “Y” and “X”.



	Y	X
1	500	400

3. Open **Match Pairs** from the pull down menu **Analyze**. Put both Y and X into **Y, Paired Response**. Then click **OK**.



Matched Pairs - JMP Pro

Comparing responses in different columns

Y, Paired Response: Y, X

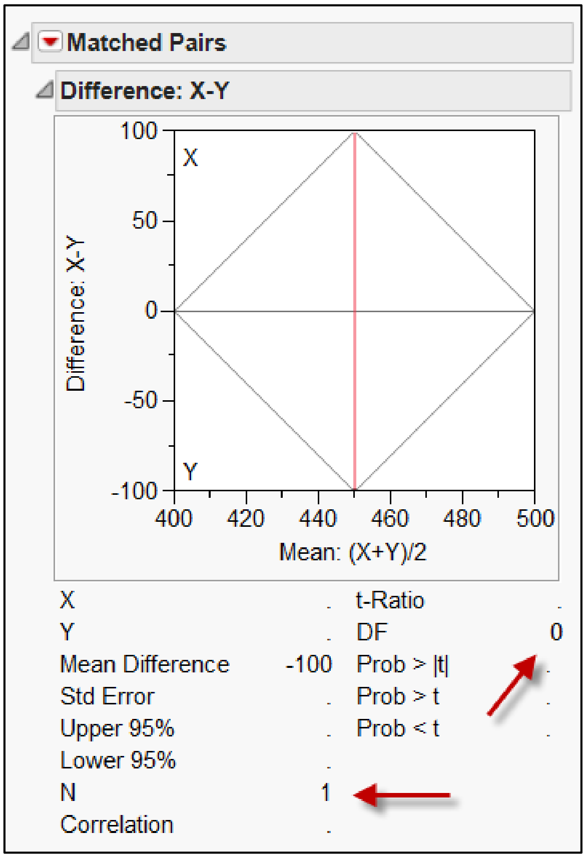
X, Grouping: optional

Weight: optional numeric

Freq: optional numeric

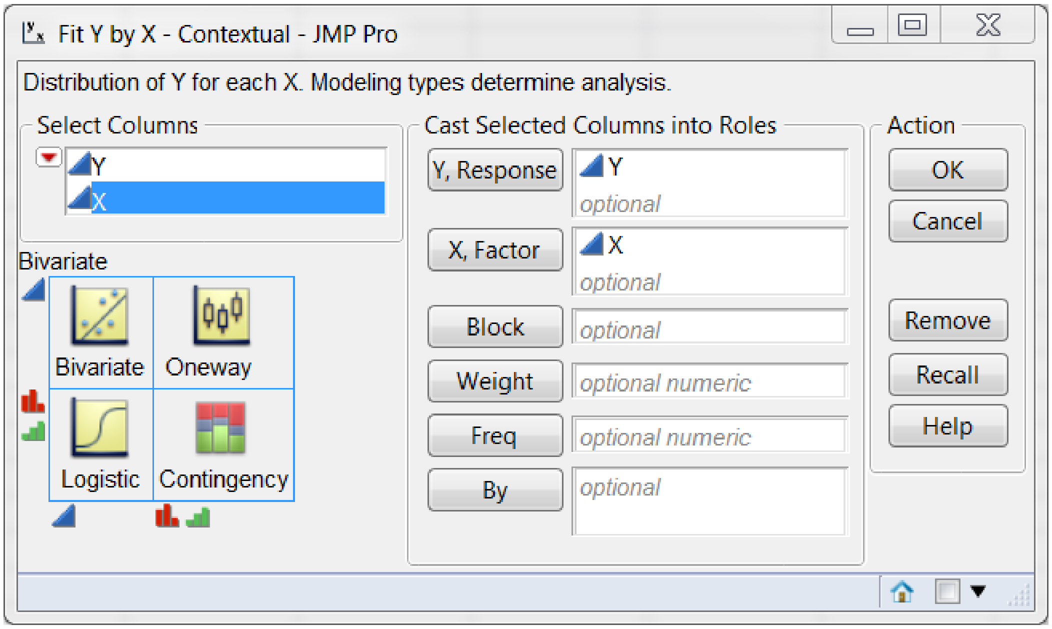
By: optional

4. Observe the output that when N = 1, DF = 0 and no statistics is reported.



```
X                        .    t-Ratio        .
Y                        .    DF             0
Mean Difference       -100    Prob > |t|     .
Std Error                .    Prob > t       .
Upper 95%                .    Prob < t       .
Lower 95%                .
N                        1
Correlation              .
```

5. Select **Fit Y by X** from **Analyze**. Put Y into **Y, Response** and X into **X, Factor**. Then click **OK**.



Distribution of Y for each X. Modeling types determine analysis.

Select Columns: Y, X

Y, Response: Y
optional

X, Factor: X
optional

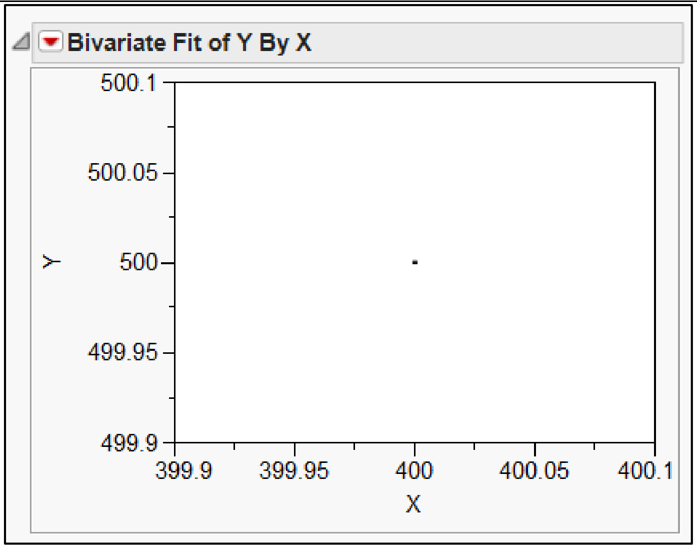
Block: optional

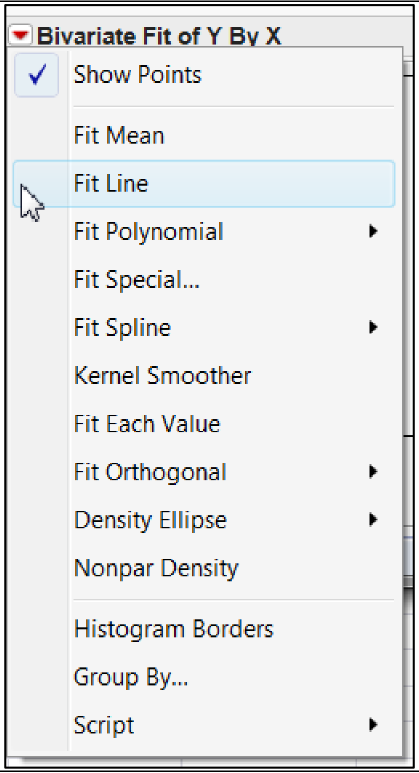
Weight: optional numeric

Freq: optional numeric

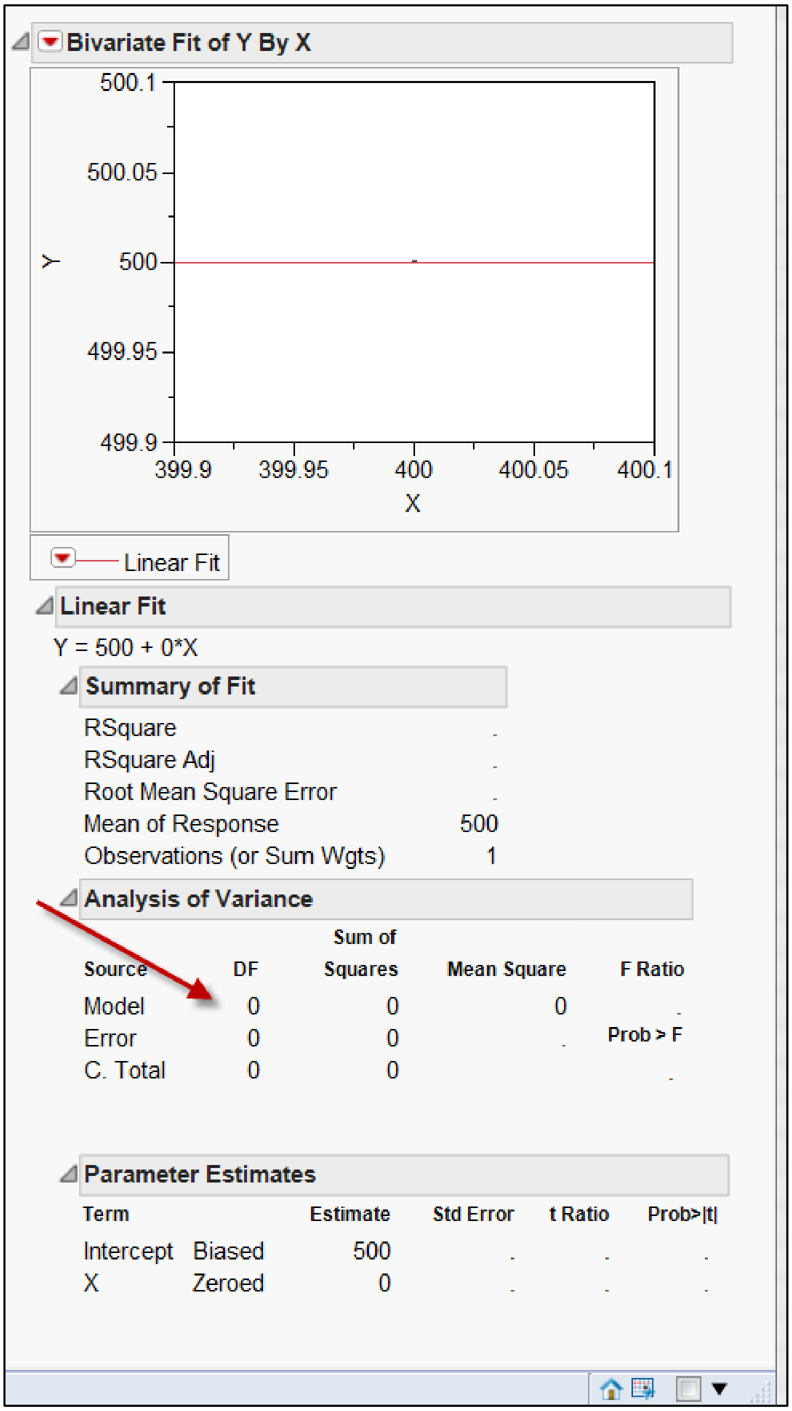
By: optional

6. A panel depicting a scatterplot with one data point will pop up. Choose **Fit Line** from the red triangle.





7. Observe the output that when DF = 0, all other statistics are either missing or zero. In other words, no meaningful computation could be performed. Although the program fitted a line on the scatterplot, this line does not mean anything. It simply tells you that the mean of Y is 500. No matter what the X value is, the Y value is always 500. In other words, no predictive model is proposed.



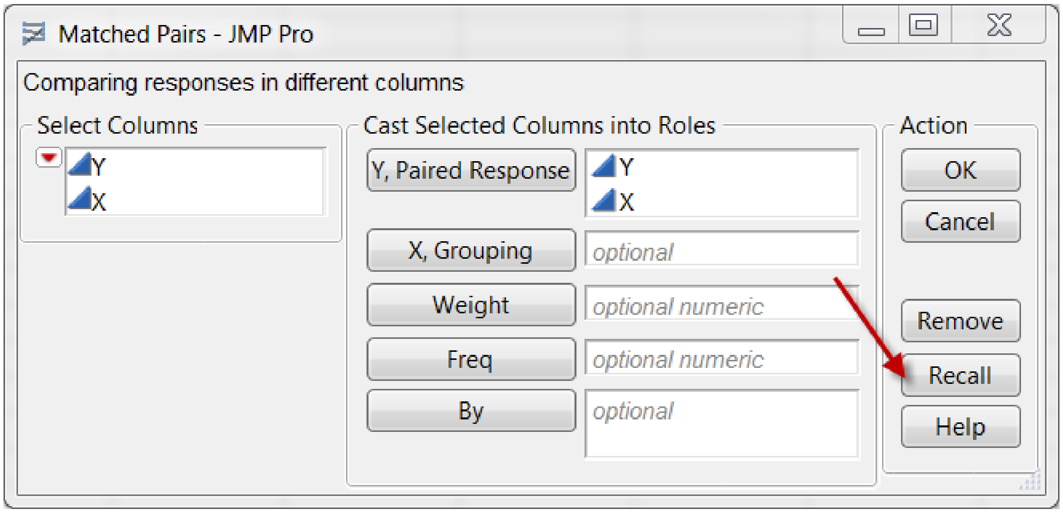
```
Linear Fit
Y = 500 + 0*X

Summary of Fit
RSquare                              .
RSquare Adj                          .
Root Mean Square Error               .
Mean of Response                   500
Observations (or Sum Wgts)           1

Analysis of Variance
                      Sum of
Source       DF      Squares    Mean Square    F Ratio
Model         0            0              0          .
Error         0            0              .   Prob > F
C. Total      0            0                         .

Parameter Estimates
Term                 Estimate   Std Error   t Ratio   Prob>|t|
Intercept  Biased         500           .         .          .
X          Zeroed           0           .         .          .
```

8. Go back to the data table. Enter 350 into Y as the second observation and enter 150 into X. Repeat **Matched Pairs**. Because you have done this before, this time you can hit the **Recall** button and the same variables will be selected. Then click **OK**.



Matched Pairs - JMP Pro

Comparing responses in different columns

Select Columns: Y, X

Cast Selected Columns into Roles

Y, Paired Response: Y, X

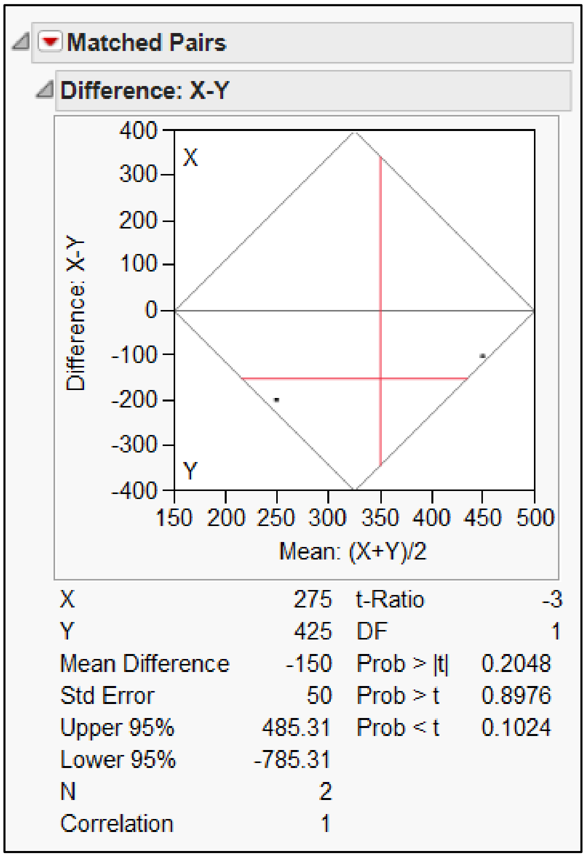
X, Grouping: optional

Weight: optional numeric

Freq: optional numeric

By: optional

9. Observe the output that N = 2, DF = 1, and the correlation coefficient = 1. It is a “perfect” but unfalsifiable model. Now the panel is populated with numbers. It gives us an illusion that it is “OK.”

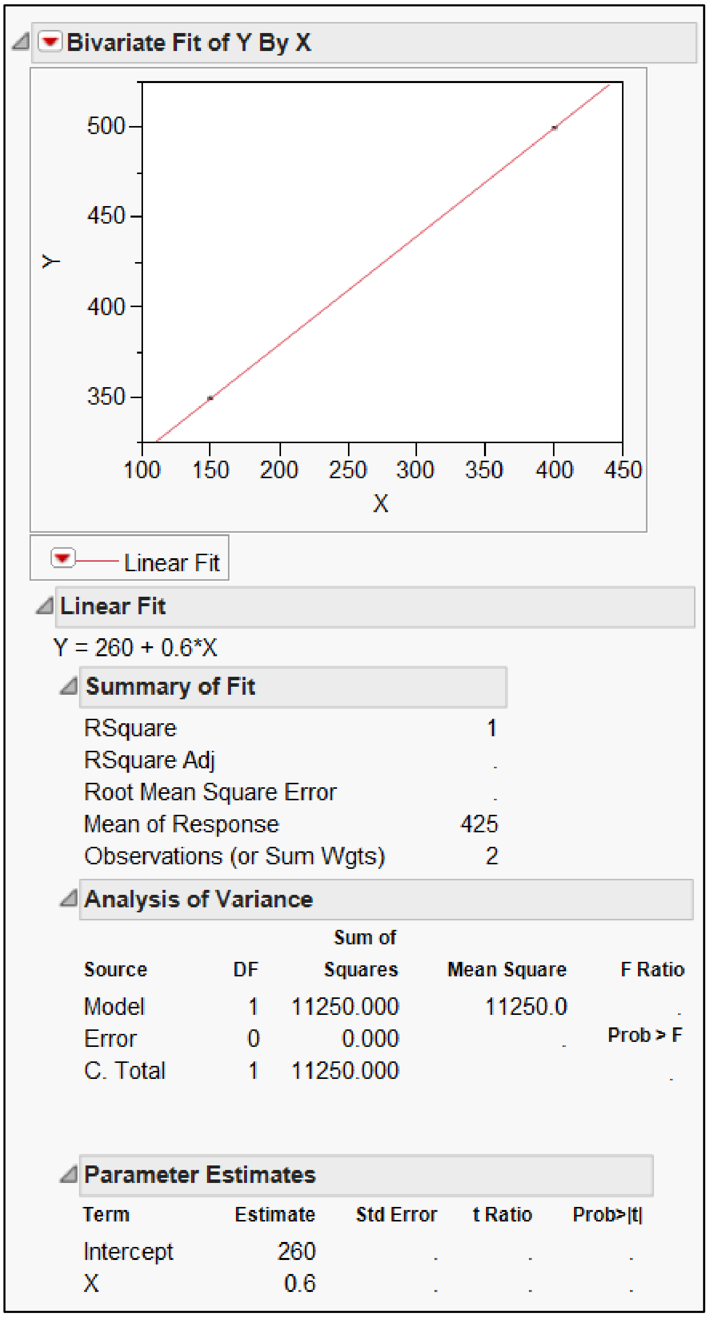


Matched Pairs

Difference: X-Y

```
X                  275   t-Ratio        -3
Y                  425   DF              1
Mean Difference   -150   Prob > |t|  0.2048
Std Error           50   Prob > t    0.8976
Upper 95%       485.31   Prob < t    0.1024
Lower 95%      -785.31
N                    2
Correlation          1
```

10. Repeat **Fit Y by X** and **Fit Line**. Again you can use the **Recall** button. The scatterplot shows a perfectly fitted straight line. But this time we can see that most statistics are missing (e.g. adjusted *R* square, *F* ratio, *p* value, standard error, *t* ratio…etc.). This implies that no meaningful analysis can be done with DF = 1.



Linear Fit

Y = 260 + 0.6*X

Summary of Fit

RSquare	1
RSquare Adj	.
Root Mean Square Error	.
Mean of Response	425
Observations (or Sum Wgts)	2

Analysis of Variance

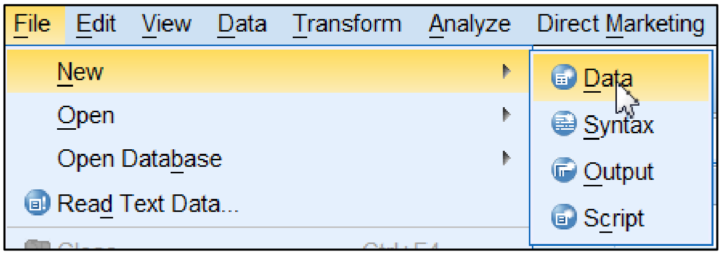
		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	1	11250.000	11250.0	.
Error	0	0.000	.	Prob > F
C. Total	1	11250.000		.

Parameter Estimates

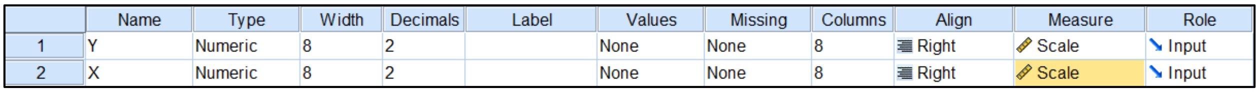
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	260	.	.	.
X	0.6	.	.	.

SPSS steps

1. Open **New Data** from the pull down menu **File**.

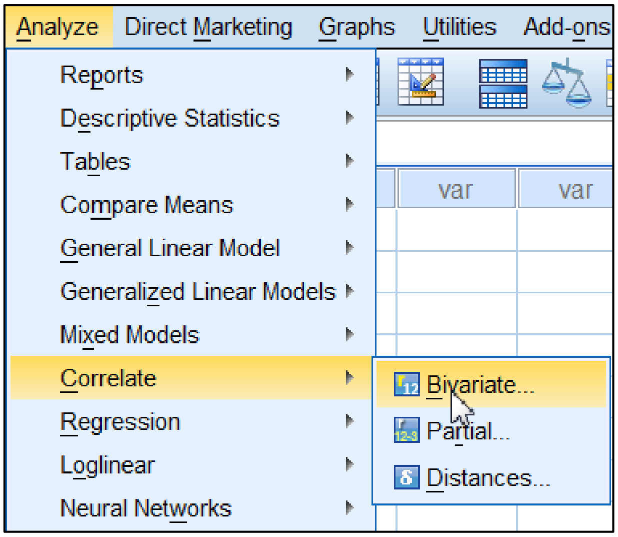


2. Enter “500” in the first column of the data table and “400” in the second column. Double-click the column headers and rename them to be “Y” and “X”. SPSS does not automatically recognize the data type. Choose **Scale** from **Measure** to assign the data type as continuous.

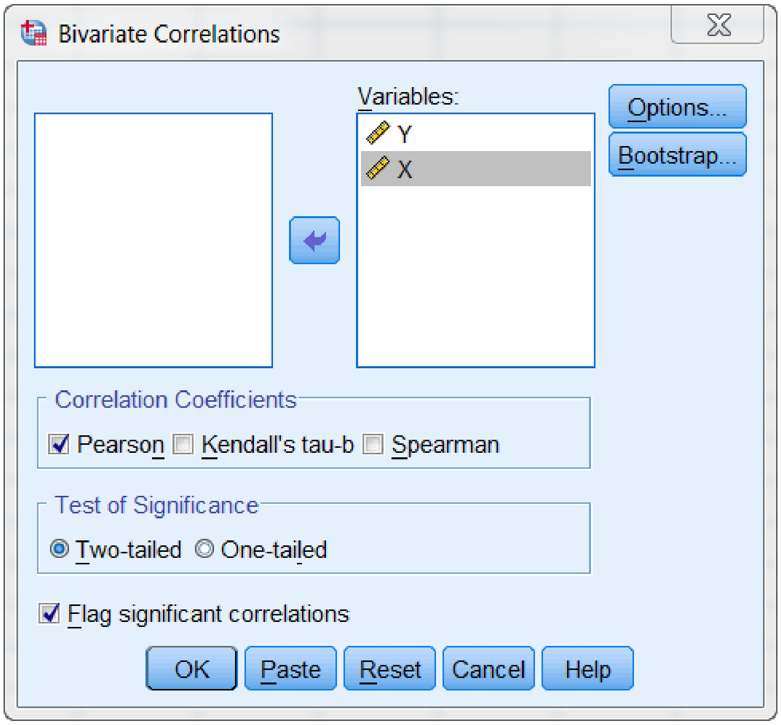


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Y	Numeric	8	2		None	None	8	Right	Scale	Input
2	X	Numeric	8	2		None	None	8	Right	Scale	Input

3. Select **Correlate Bivariate** from **Analyze**.



4. Put both Y and X into **Variables**. Then click **OK**.



Bivariate Correlations

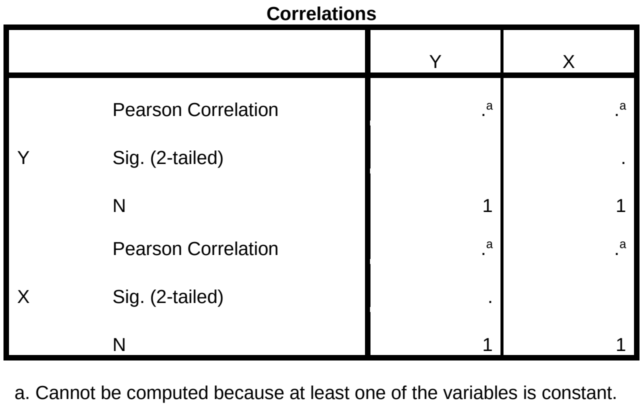
Variables:
Y
X

Correlation Coefficients
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

5. Observe the output that no statistics can be reported. The footnote “a” indicates: “Cannot be computed because at least one of the variables is constant.

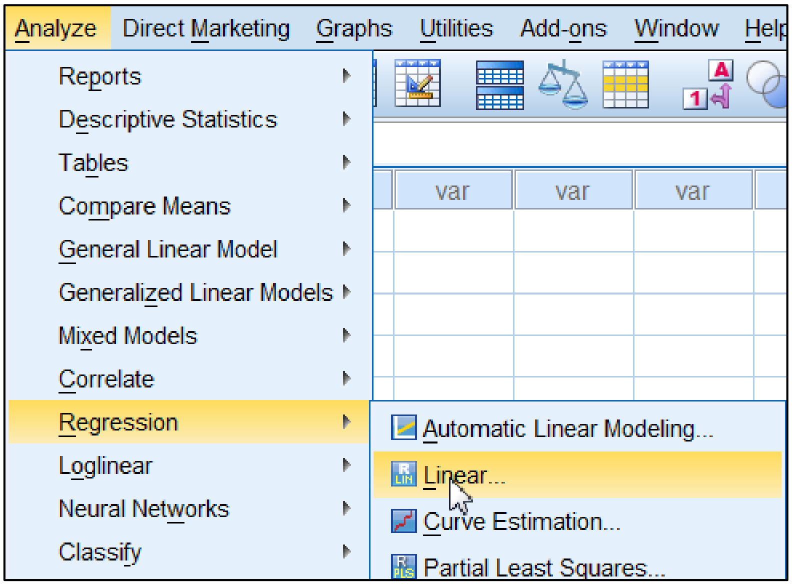


Correlations

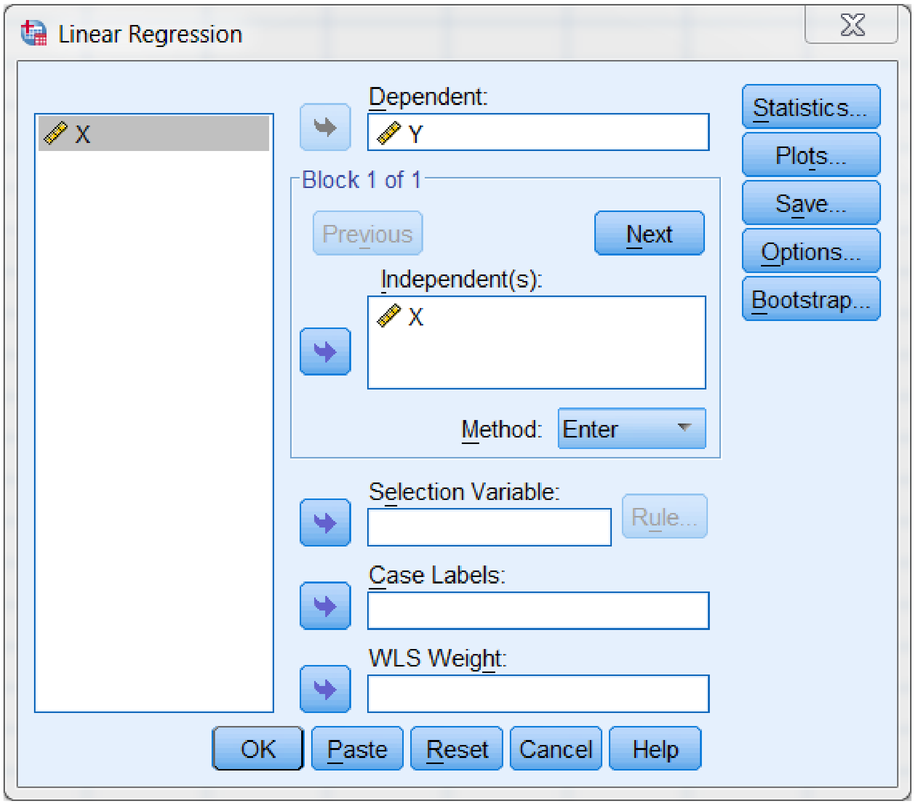
		Y	X
Y	Pearson Correlation	.[a]	.[a]
	Sig. (2-tailed)		.
	N	1	1
X	Pearson Correlation	.[a]	.[a]
	Sig. (2-tailed)	.	
	N	1	1

a. Cannot be computed because at least one of the variables is constant.

6. Choose **Linear regression** from **Analyze**.



7. Put Y into **Dependent** and X into **Independent(s)**. Then click **OK**.

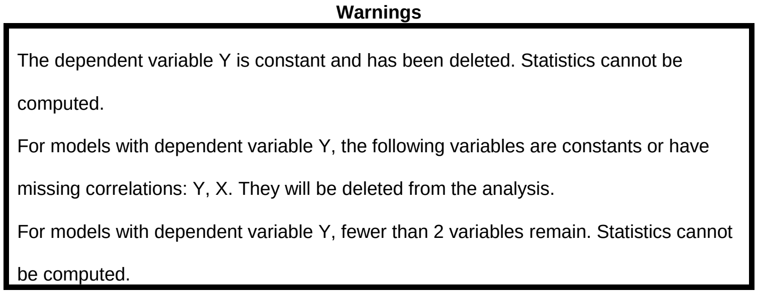


Dependent: Y

Independent(s): X

Method: Enter

8. Read the warnings. It is self-explanatory.



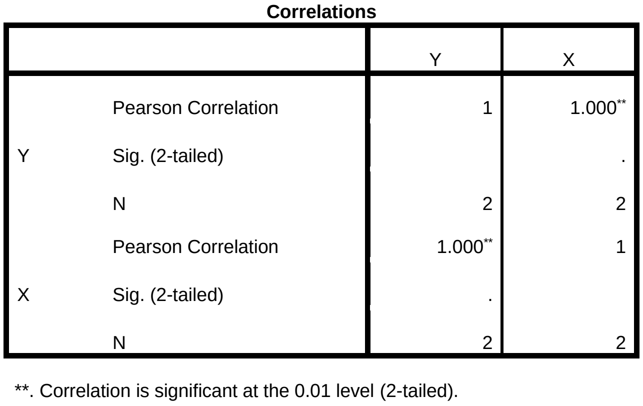
Warnings

The dependent variable Y is constant and has been deleted. Statistics cannot be computed.

For models with dependent variable Y, the following variables are constants or have missing correlations: Y, X. They will be deleted from the analysis.

For models with dependent variable Y, fewer than 2 variables remain. Statistics cannot be computed.

9. Go back to the data table. Enter 350 into Y as the second observation and enter 150 into X. Repeat **Correlation**. This time SPSS computed the data and yielded a Pearson correlation coefficient (1.000). But the *p* value (sig.) is missing. *P* value is the probability of observing the statistics in the long run given that the null hypothesis (there is no significant correlation between X and Y) is true. In other words, this is an inference from the sample to the population. When N = 1, no inference or estimation can be made.



Correlations

		Y	X
Y	Pearson Correlation	1	1.000**
	Sig. (2-tailed)		.
	N	2	2
X	Pearson Correlation	1.000**	1
	Sig. (2-tailed)	.	
	N	2	2

**. Correlation is significant at the 0.01 level (2-tailed).

10. Repeat **Regression**. You can see that most statistics are missing when DF = 1 in regression. But DF = 1 is acceptable in Chi-square analysis when Yates's correction is used. We will have exercises on Chi-square later.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	.	.

a. Predictors: (Constant), X

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11250.000	1	11250.000	.	.[b]
	Residual	.000	0	.		
	Total	11250.000	1			

a. Dependent Variable: Y

b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	260.000	.000		.	.
	X	.600	.000	1.000	.	.

a. Dependent Variable: Y

Chapter 2 exercises

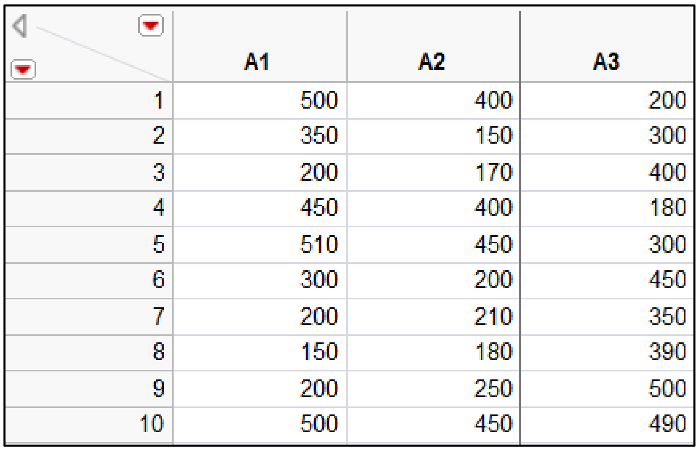
Objective: These exercises can help you understand that

1. Variables could be visualized as vectors.
2. The relationships between the variables can be signified by the angle formed by the vectors.

Tool: JMP

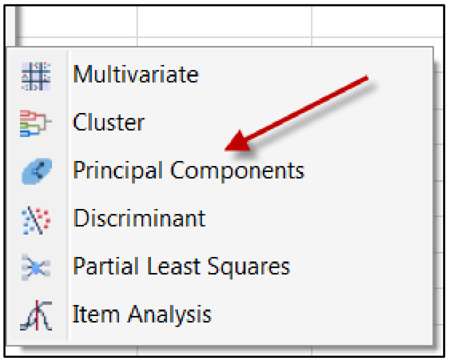
Steps

1. Enter the data below in the data table. Please save the file because we will reuse it in Chapter 4 exercises.

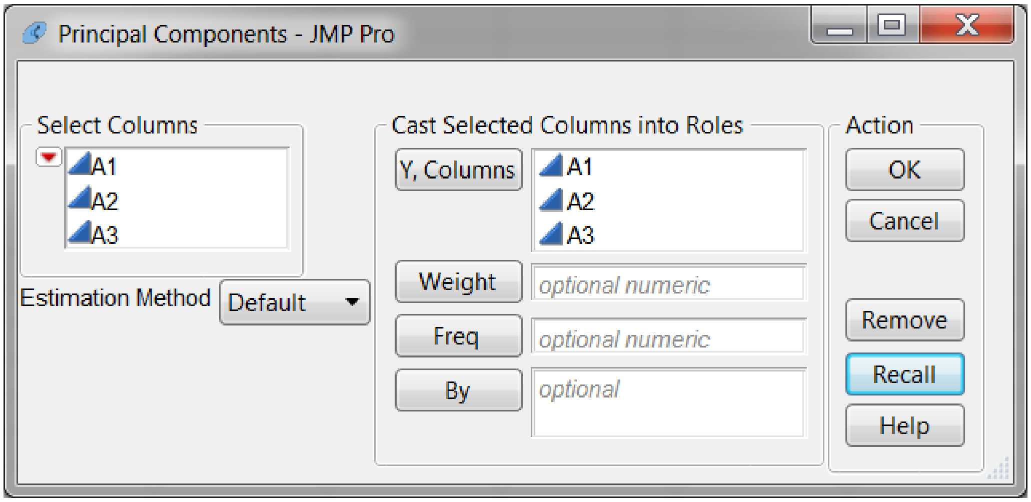


	A1	A2	A3
1	500	400	200
2	350	150	300
3	200	170	400
4	450	400	180
5	510	450	300
6	300	200	450
7	200	210	350
8	150	180	390
9	200	250	500
10	500	450	490

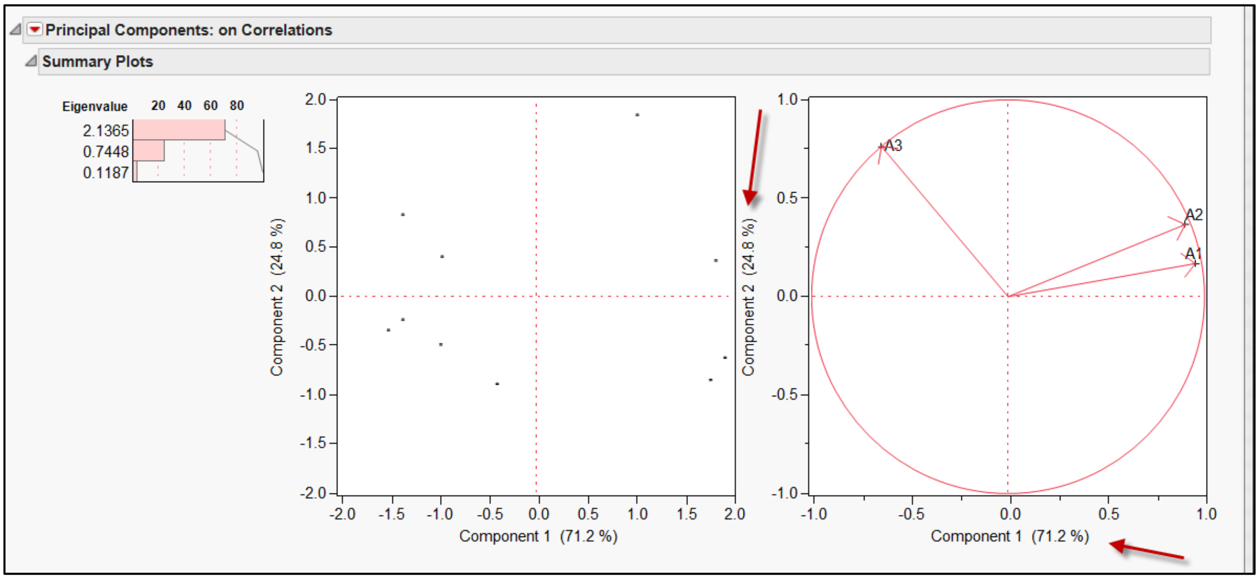
2. Select **Multivariate, Principal Components** from the pull down menu **Analyze**.



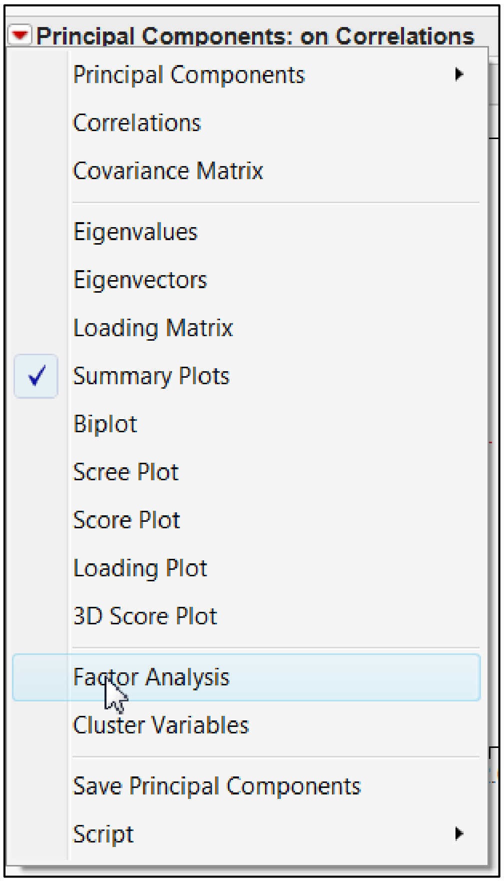
3. Put all variables into **Y, Columns**. Then click **OK**.



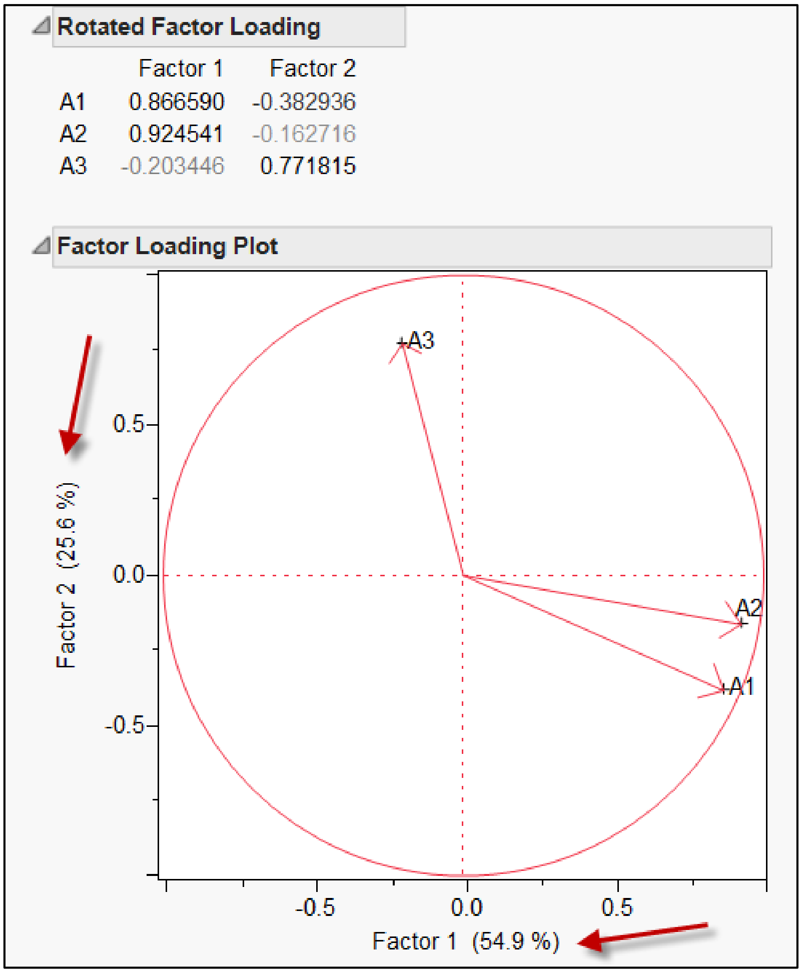
4. The summary plot displays the three variables as three vectors. A1 and A2 form a narrow angle but A3 departs from A1 and A2, and points to a vastly different direction. The first component extracts 71.2% variance whereas the second component extracts 24.8%.



5. Select Factor Analysis from the red triangle. When the pop up window appears, accept the default and click **OK**. Principal component analysis (PCA) is for data reduction whereas factor analysis is for identifying latent constructs. Additionally, the former takes all variances into account while the latter extract shared variances.



6. The vector plot in factor analysis returns the same result as that in PCA. Based on the proximity of the vectors, it seems that the variables can be loaded into two factors. The factor loading table suggests that A1 and A2 belong to Factor 1 whereas A3 can be put into Factor 2. Factor 1 extracts 54.9% variance while Factor 2 extracts 25.6%. A real factor model needs more than 1-2 items, of course. The purpose of this exercise is to learn that variables can be expressed in terms of vectors and thus don’t worry about the psychometric soundness of this “factor model.” In a real data set that has many observed items, the researcher might need to visualize the vectors using the biplot. But in this simple exercise it is not necessary to do so.



Rotated Factor Loading

	Factor 1	Factor 2
A1	0.866590	-0.382936
A2	0.924541	-0.162716
A3	-0.203446	0.771815

Factor Loading Plot

Chapter 3 exercises

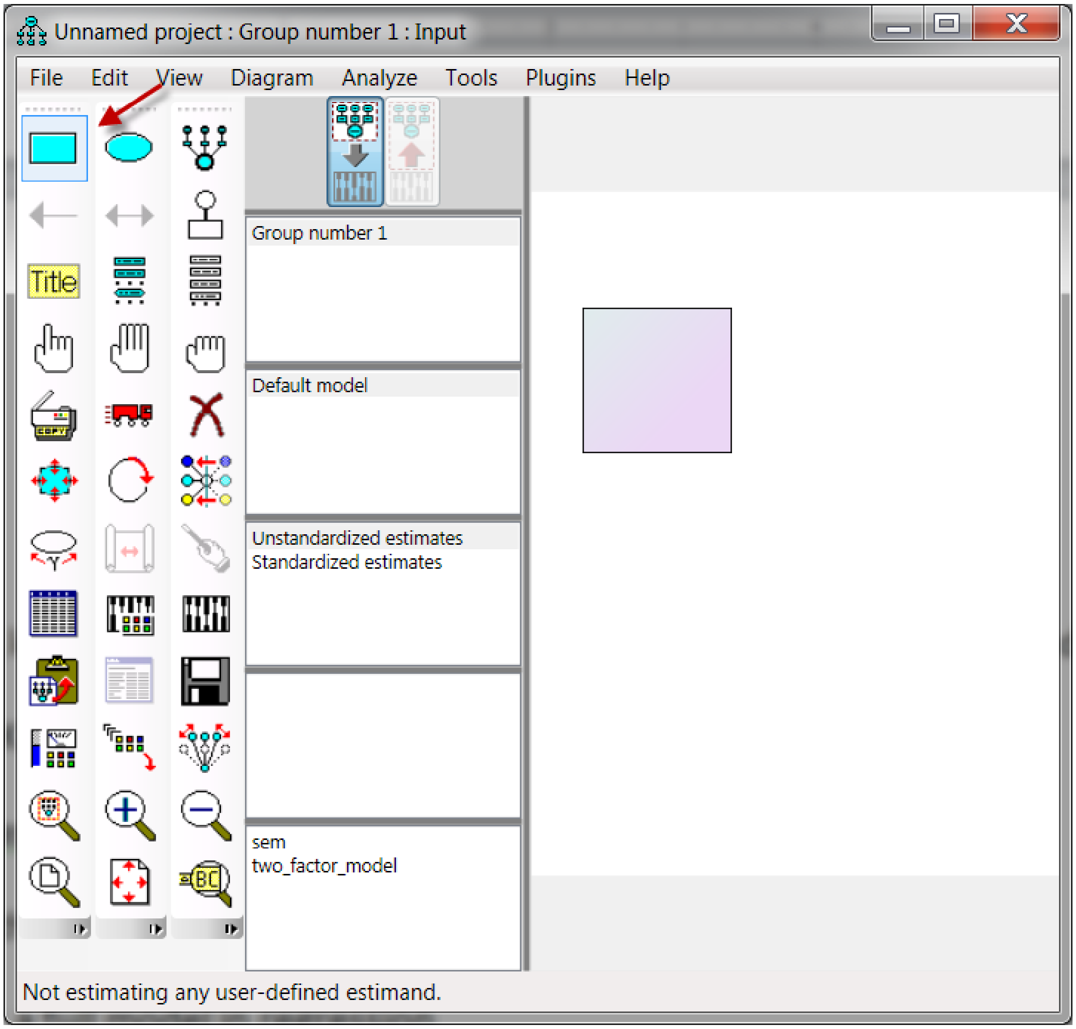
Objective: These exercises can help you understand that

1. The degrees of freedom can be conceptualized as the difference between the dimensionality of a broader hypothesis and that of a null hypothesis,
2. The relationships between DF and the number of parameters to be estimated. When there are too many parameters and data are insufficient, the DF value could be negative.

Tool: AMOS (IBM SPSS, 2011a) is required to go through this exercise. If you don't have access to AMOS, you can download a student version from http://amosdevelopment.com/download/. The free Amos student version has the same features as that of the full commercial version except that the student version is limited to eight observed variables and 54 parameters.

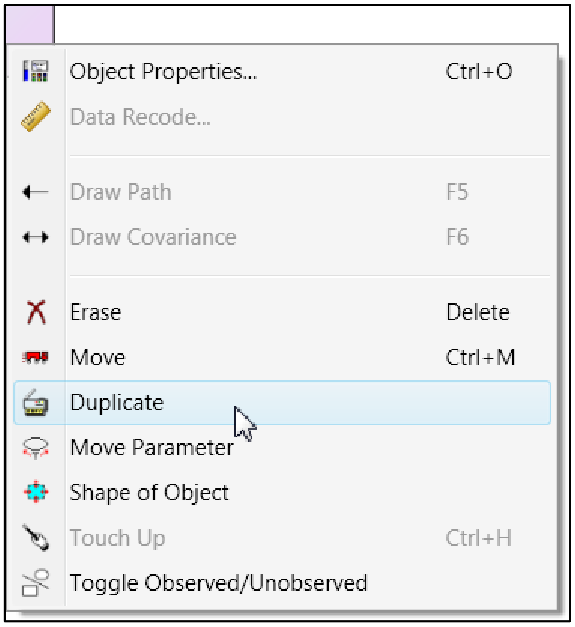
Steps

1. Open AMOS. Double click the rectangle icon, which symbolizes an observed variable. Then draw a rectangle on the canvas.

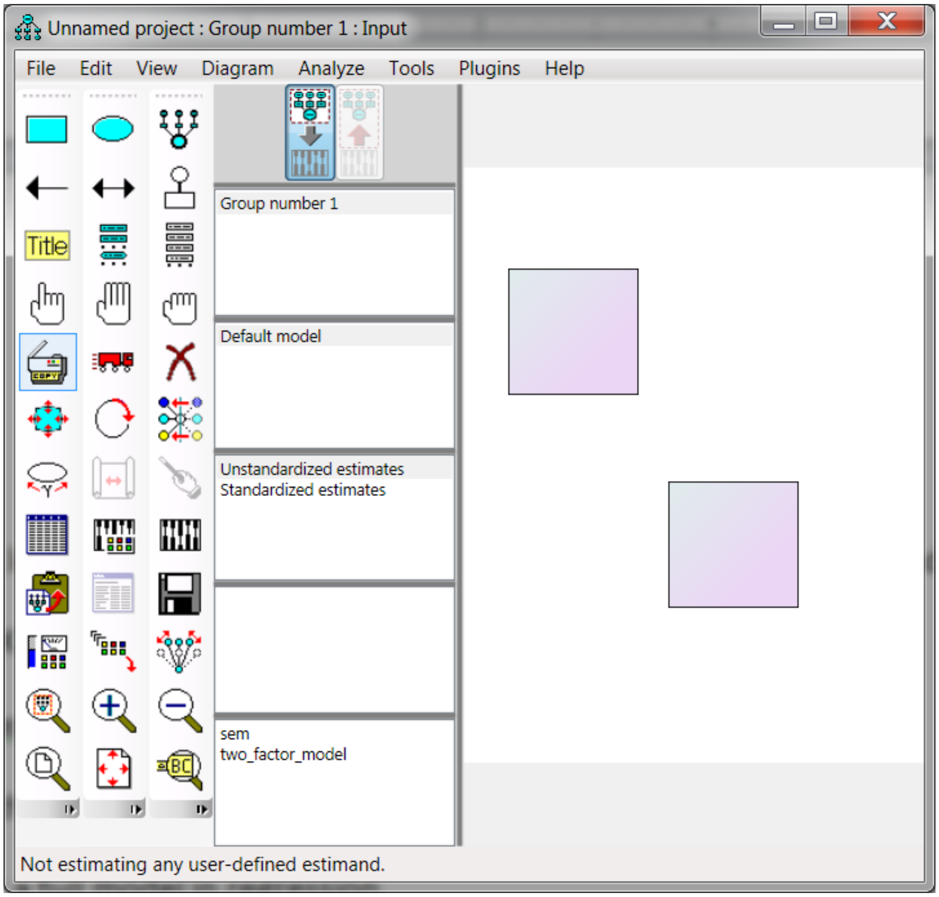


Not estimating any user-defined estimand.

2. Right-click on the object and choose **Duplicate**.

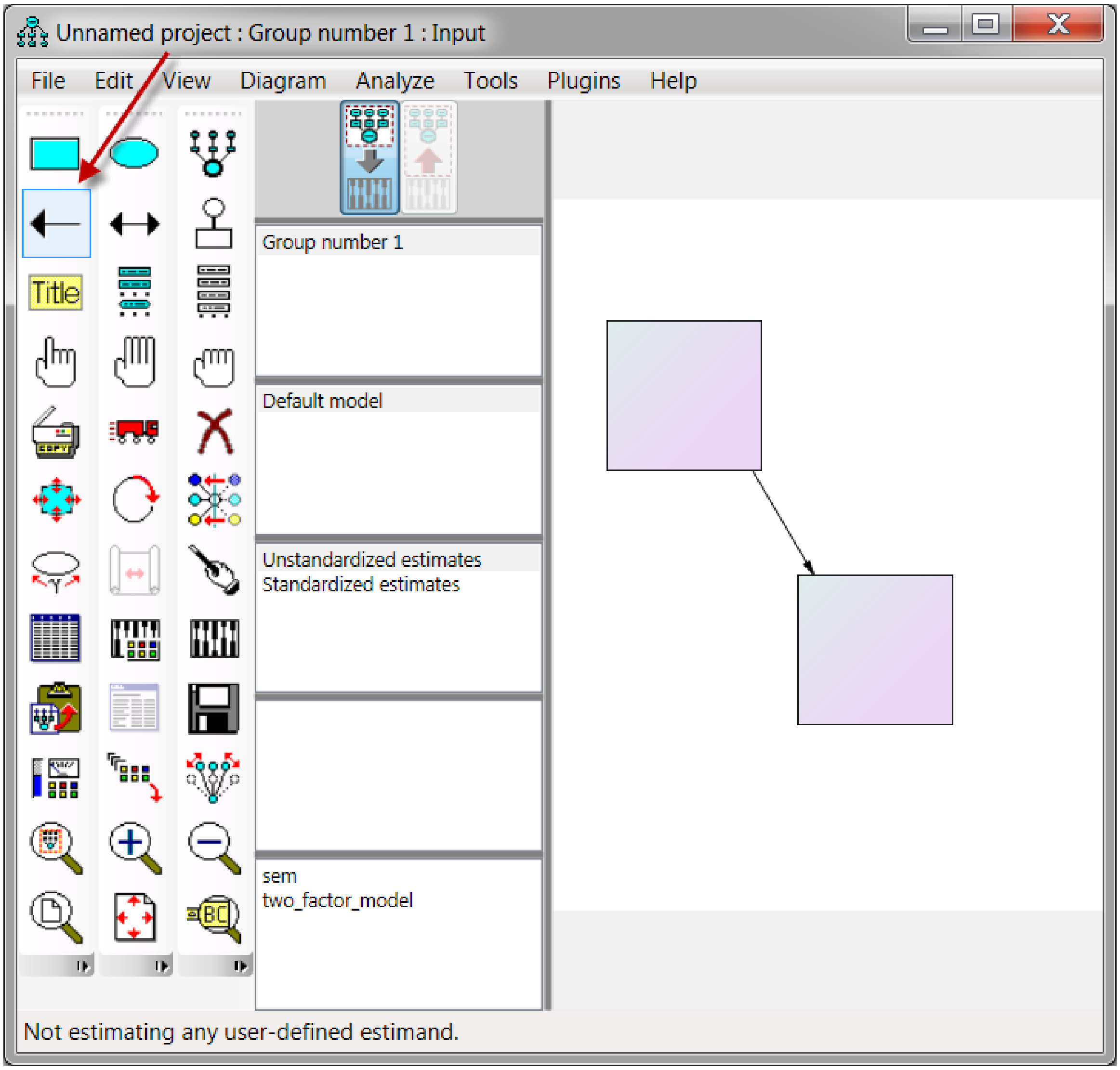


3. A duplicated rectangle is on top of the original one. Drag it away.

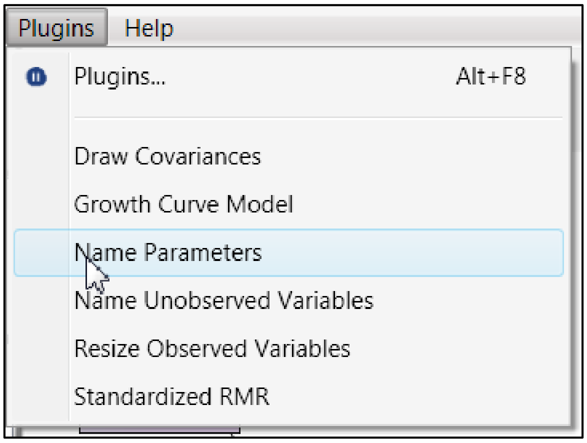


Not estimating any user-defined estimand.

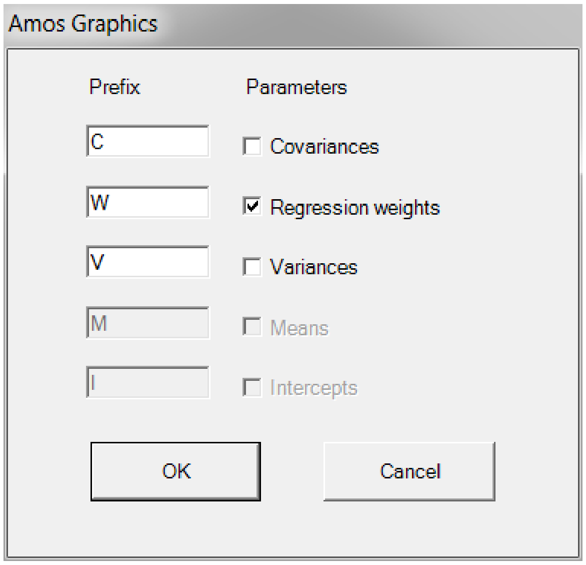
4. Click on the arrow icon. Then draw an arrow to connect the two variables.



5. Choose **Name Parameters** from **Plugins**.



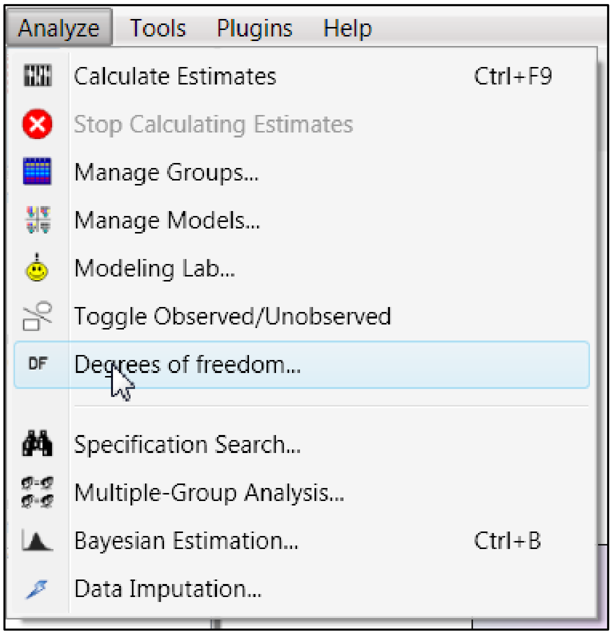
6. In the pop-up window, check the box **Regression weights**. Then click **OK**.



Amos Graphics

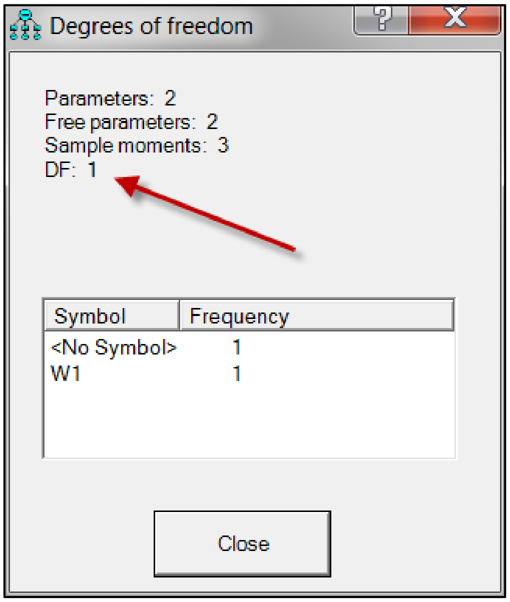
Prefix	Parameters
C	☐ Covariances
W	☑ Regression weights
V	☐ Variances
M	☐ Means
I	☐ Intercepts

7. Choose **Degrees of freedom** from the pull down menu Analyze.



Analyze
Tools
Plugins
Help
Calculate Estimates Ctrl+F9
Stop Calculating Estimates
Manage Groups...
Manage Models...
Modeling Lab...
Toggle Observed/Unobserved
DF
Degrees of freedom...
Specification Search...
Multiple-Group Analysis...
Bayesian Estimation... Ctrl+B
Data Imputation...

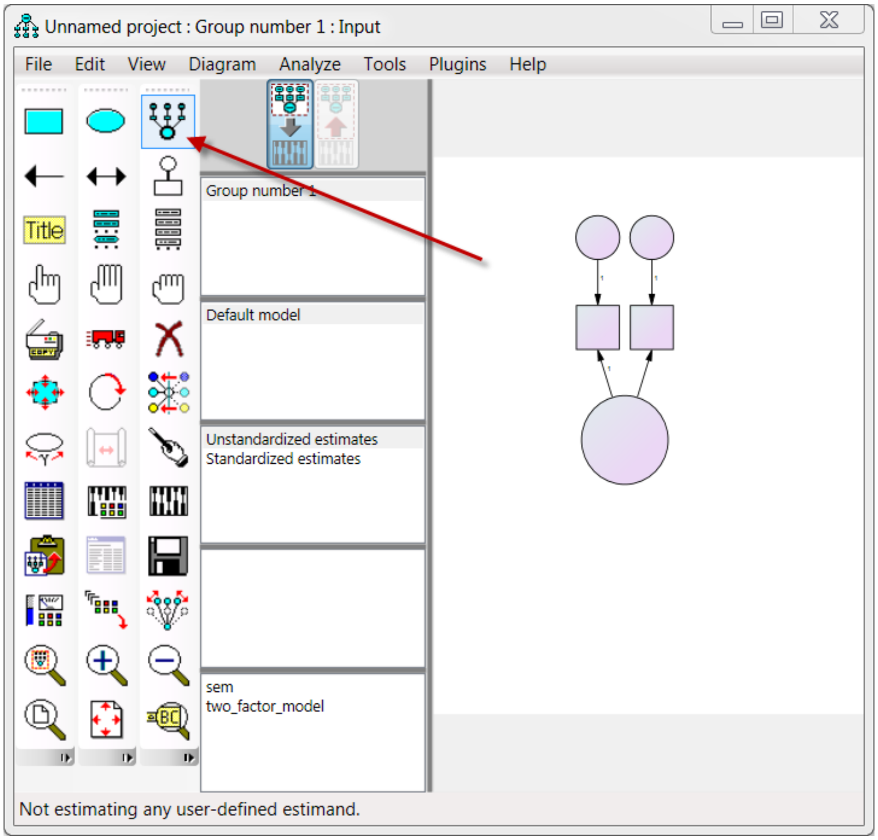
8. Observe the output that DF = 1. If the intercept is included, in a simple regression model there are two parameters to be estimated: Intercept and slope. If we omit the intercept, there is only one parameter: slope.



Parameters: 2
Free parameters: 2
Sample moments: 3
DF: 1

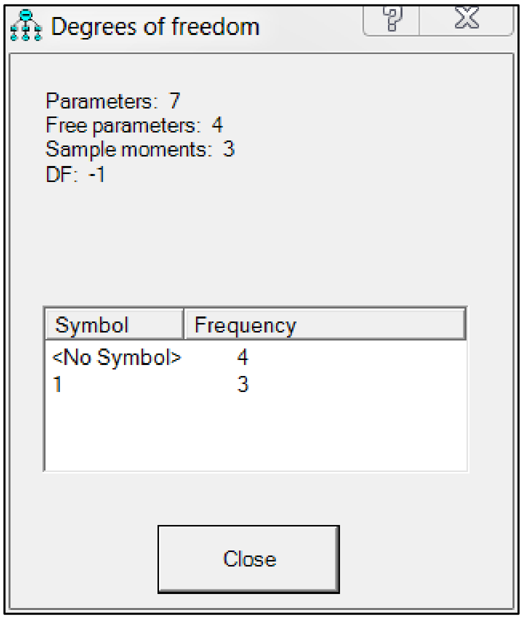
Symbol	Frequency
<No Symbol>	1
W1	1

9. Open a new file. Double-click the rightmost icon in the first row of icons. Click three times on the canvas to create a two-item factor model.



Not estimating any user-defined estimand.

10. Choose **Degrees of freedom** from **Analyze**. Observe the output that DF = -1. It is worse than DF = 0.



Parameters: 7
Free parameters: 4
Sample moments: 3
DF: -1

Symbol	Frequency
<No Symbol>	4
1	3

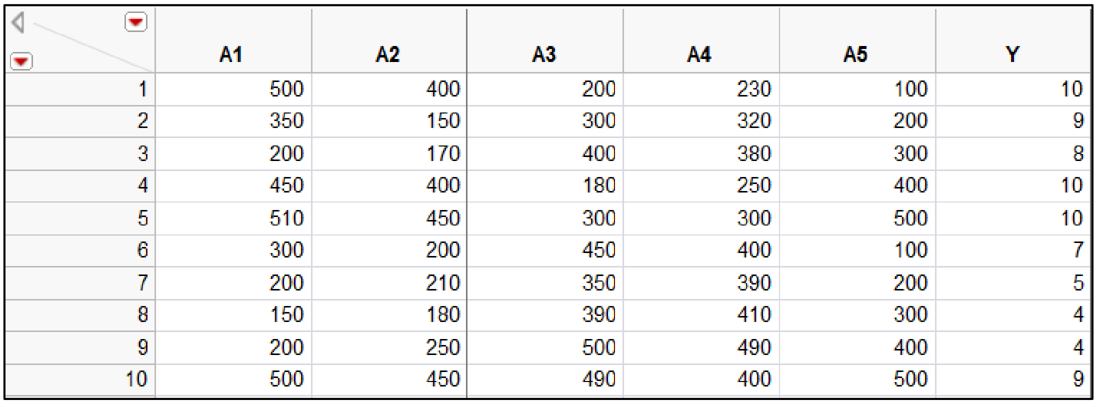
Chapter 4 exercises

Objective: These exercises can help you understand that DF is the inverse of AICs. The former is concerned with freedom and informativeness of the data whereas the latter is imposing restrictions and penalty on modeling.

Tool: JMP

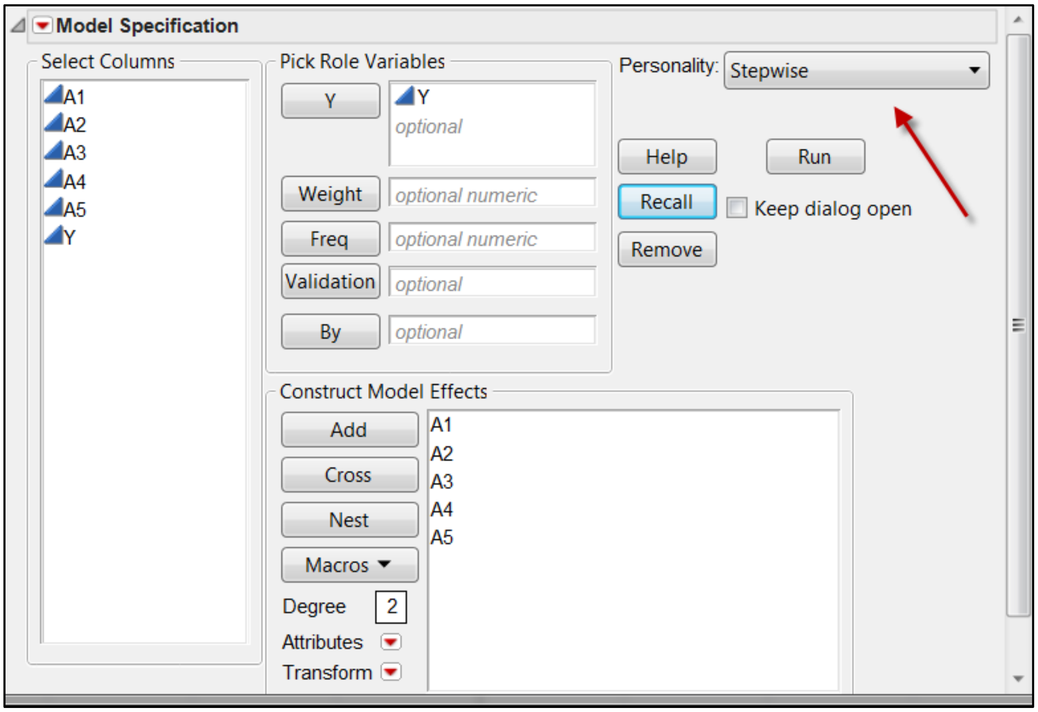
Steps

1. Open the file used in Chapter 2 exercises. This time add three more variables (A4, A5, and Y) and enter the data as shown below.



	A1	A2	A3	A4	A5	Y
1	500	400	200	230	100	10
2	350	150	300	320	200	9
3	200	170	400	380	300	8
4	450	400	180	250	400	10
5	510	450	300	300	500	10
6	300	200	450	400	100	7
7	200	210	350	390	200	5
8	150	180	390	410	300	4
9	200	250	500	490	400	4
10	500	450	490	400	500	9

2. Select **Fit Models** from the pull down menu **Analyze**. In the pop-up window, put Y into **Y**. Add A1-A5 into **Construct Model Effects**. Select Stepwise for **Personality**. Then click **Run**.



Model Specification

Select Columns: A1, A2, A3, A4, A5, Y

Pick Role Variables

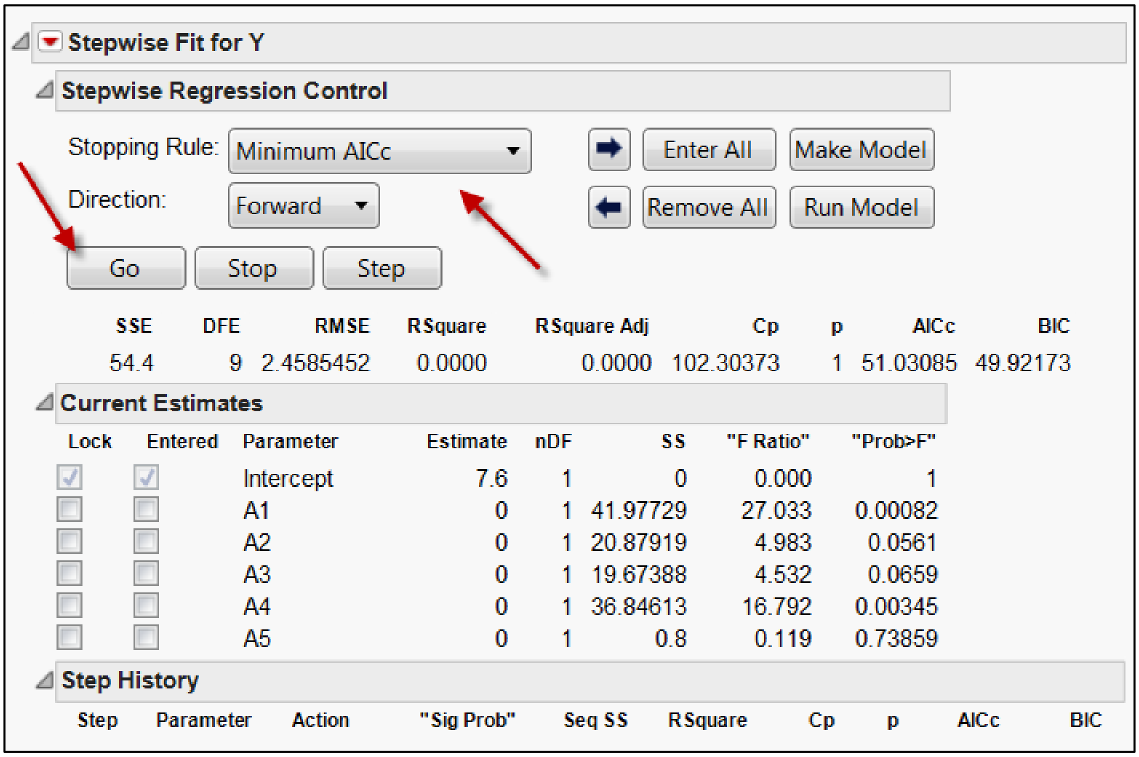
Y: Y

Personality: Stepwise

Construct Model Effects: A1, A2, A3, A4, A5

Degree 2

3. Choose **Minimum AICc** for Stopping Rule. Then click on **Go**.



Stepwise Fit for Y

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
54.4	9	2.4585452	0.0000	0.0000	102.30373	1	51.03085	49.92173

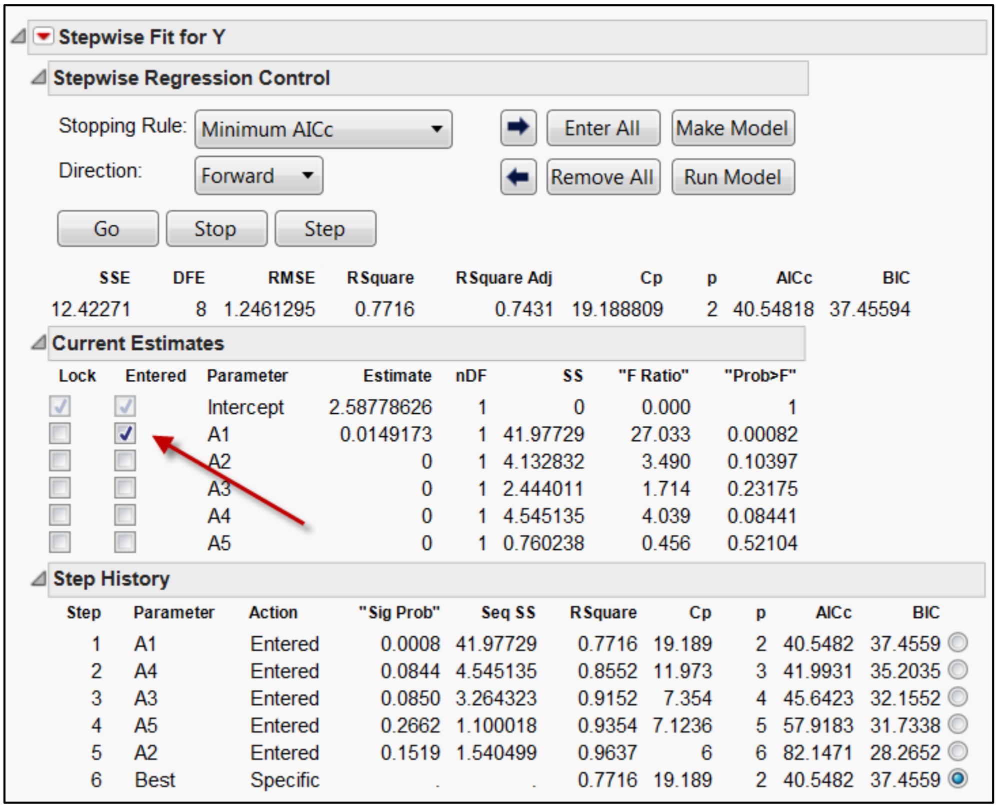
Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	7.6	1	0	0.000	1
□	□	A1	0	1	41.97729	27.033	0.00082
□	□	A2	0	1	20.87919	4.983	0.0561
□	□	A3	0	1	19.67388	4.532	0.0659
□	□	A4	0	1	36.84613	16.792	0.00345
□	□	A5	0	1	0.8	0.119	0.73859

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC

4. Based on the minimum AICc criterion, JMP selected only A1 into the regression model.



Stepwise Fit for Y

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
12.42271	8	1.2461295	0.7716	0.7431	19.188809	2	40.54818	37.45594

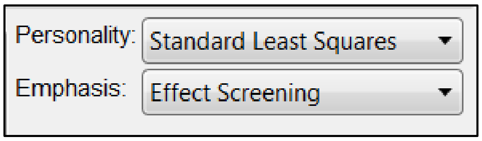
Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	2.58778626	1	0	0.000	1
□	☑	A1	0.0149173	1	41.97729	27.033	0.00082
□	□	A2	0	1	4.132832	3.490	0.10397
□	□	A3	0	1	2.444011	1.714	0.23175
□	□	A4	0	1	4.545135	4.039	0.08441
□	□	A5	0	1	0.760238	0.456	0.52104

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	A1	Entered	0.0008	41.97729	0.7716	19.189	2	40.5482	37.4559	○
2	A4	Entered	0.0844	4.545135	0.8552	11.973	3	41.9931	35.2035	○
3	A3	Entered	0.0850	3.264323	0.9152	7.354	4	45.6423	32.1552	○
4	A5	Entered	0.2662	1.100018	0.9354	7.1236	5	57.9183	31.7338	○
5	A2	Entered	0.1519	1.540499	0.9637	6	6	82.1471	28.2652	○
6	Best	Specific	.	.	0.7716	19.189	2	40.5482	37.4559	●

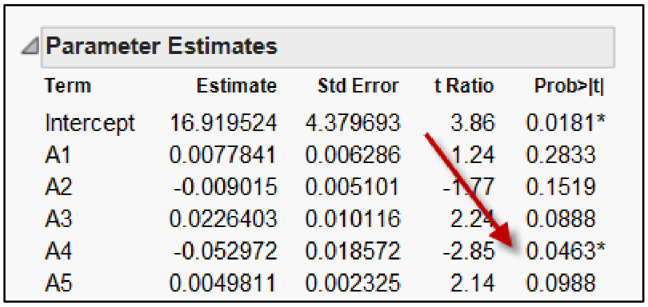
5. Choose **Fit Models** again. Use the same variables. But this time keep the default Personality: **Standard Least Squares**. Then click **OK**.



Personality: Standard Least Squares

Emphasis: Effect Screening

6. The standard least squares method suggests keeping A4 in the model, which is different from what AICc recommends. AICc takes DF into account while trying out all possible combinations in the stepwise procedure, and thus the result is more trustworthy.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	16.919524	4.379693	3.86	0.0181*
A1	0.0077841	0.006286	1.24	0.2833
A2	-0.009015	0.005101	-1.77	0.1519
A3	0.0226403	0.010116	2.24	0.0888
A4	-0.052972	0.018572	-2.85	0.0463*
A5	0.0049811	0.002325	2.14	0.0988

Chapter 5 exercises

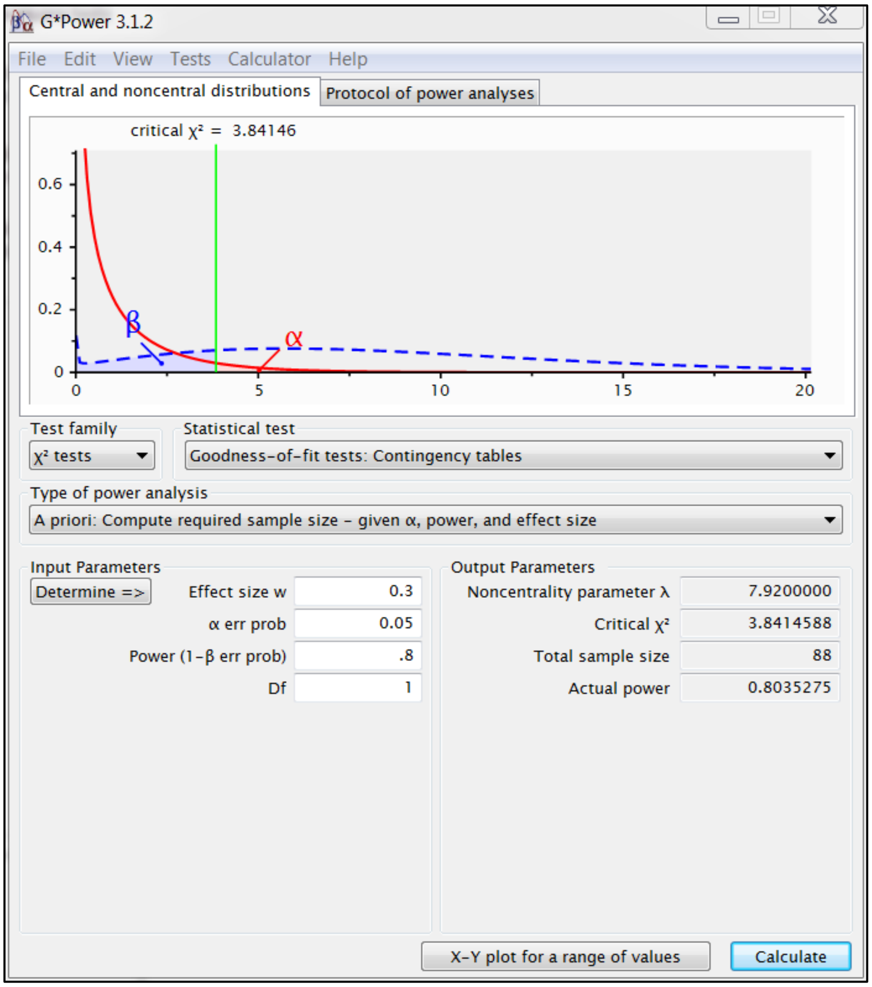
Objective: These exercises can help you understand that

1. Chi-square analysis is distribution-dependent with regard to power.
2. Whether the Chi-square statistics is significant or not is tied to the number of levels and the DF value.
3. DF = 1 is acceptable in Chi-square analysis when Yates correction is employed.

Tools: G*Power and Online Chi-square calculator. G*Power is a freeware that can be downloaded from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/. The online Chi-square calculator can be accessed at http://vassarstats.net/newcs.html

G*Power steps

1. Open G*Power. Select χ^2 from **Test Family**. Select Goodness-of-fit tests: Contingency tables from **Statistical test**. Enter .8 for **Power** and 1 for **DF**. The click **Calculate**. Observe the shapes of the null and the alternate distributions.


G*Power 3.1.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical χ² = 3.84146
0.6
0.4
0.2
0
0
5
10
15
20
β
α
Determine =>
X-Y plot for a range of values
Calculate

Test family: χ² tests

Statistical test: Goodness-of-fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size w: 0.3

α err prob: 0.05

Power (1-β err prob): .8

Df: 1

Output Parameters

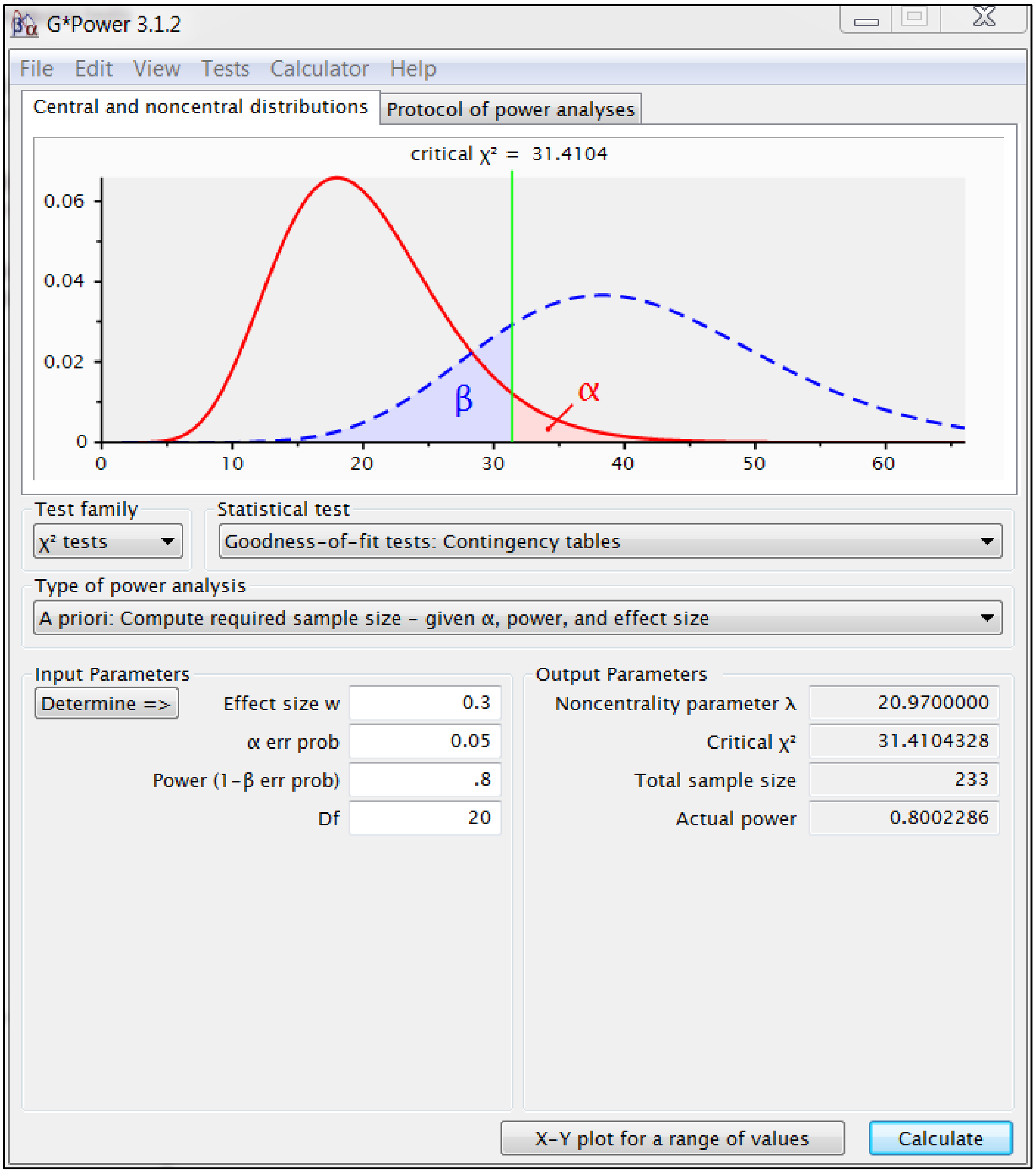
Noncentrality parameter λ: 7.9200000

Critical χ²: 3.8414588

Total sample size: 88

Actual power: 0.8035275

2. Change **DF** to 20 and click **Calculate**. Observe the shape of the Chi-square distribution. Is it fairly normal?



Test family: χ² tests

Statistical test: Goodness–of–fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size w: 0.3

α err prob: 0.05

Power (1–β err prob): .8

Df: 20

Output Parameters

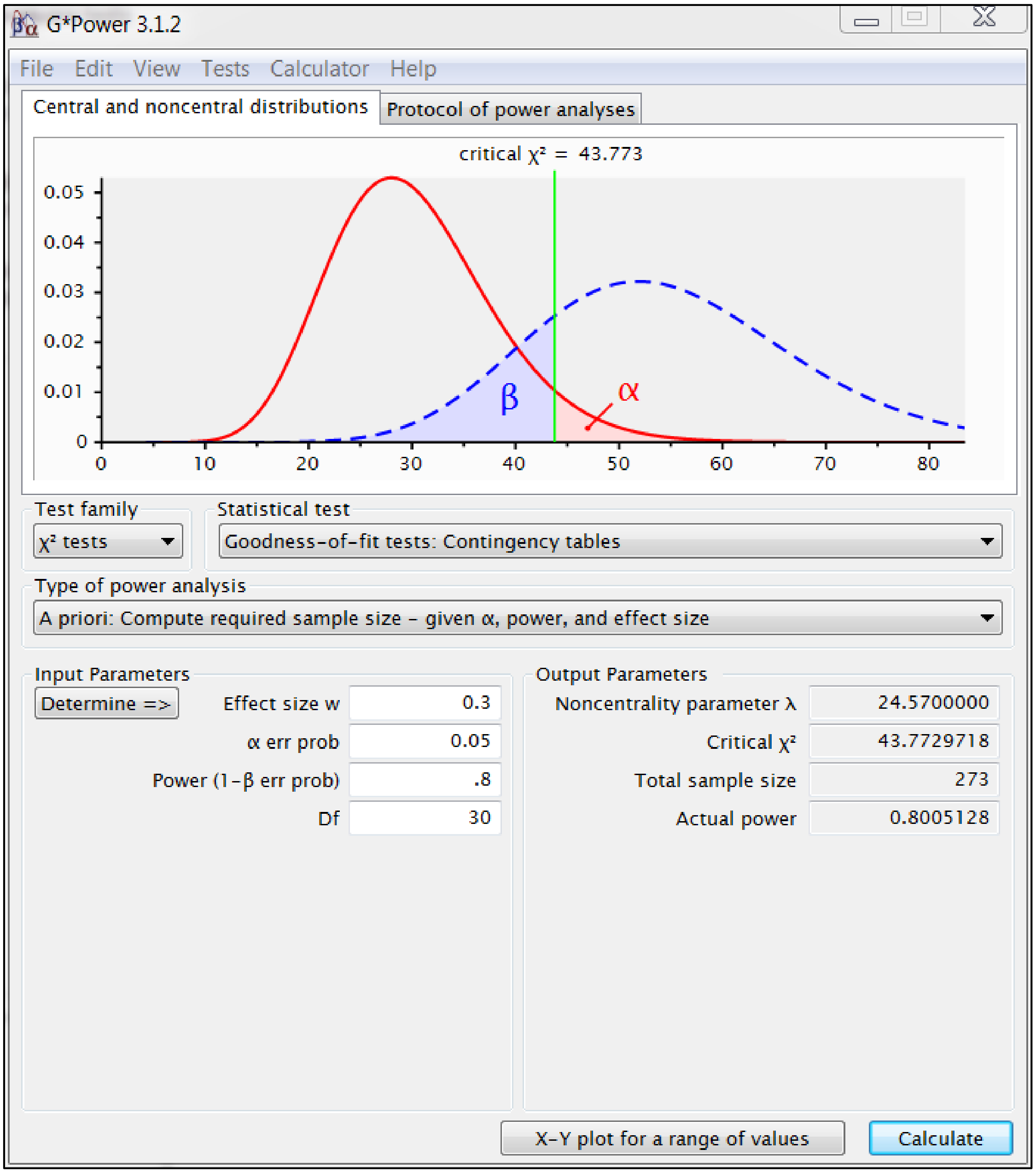
Noncentrality parameter λ: 20.9700000

Critical χ²: 31.4104328

Total sample size: 233

Actual power: 0.8002286

3. Change **DF** to 30 and click **Calculate**. Observe the appearance of the Chi-square distribution. Is it fairly normal? Compare this distribution with the previous one. Is there any substantial difference?


G*Power 3.1.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical χ² = 43.773
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
80
β
α
Determine =>
X-Y plot for a range of values
Calculate

Test family: χ^2 tests

Statistical test: Goodness-of-fit tests: Contingency tables

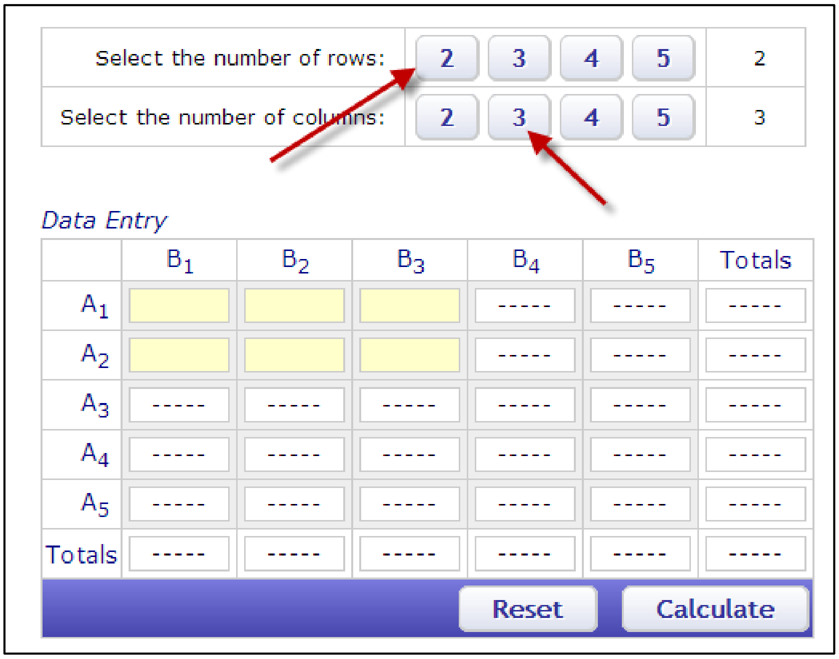
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters	
Effect size w	0.3
α err prob	0.05
Power (1−β err prob)	.8
Df	30

Output Parameters	
Noncentrality parameter λ	24.5700000
Critical χ^2	43.7729718
Total sample size	273
Actual power	0.8005128

Online Chi-square calculator steps

1. Go to the website http://vassarstats.net/newcs.html. Click on the buttons 2 rows and 3 columns.

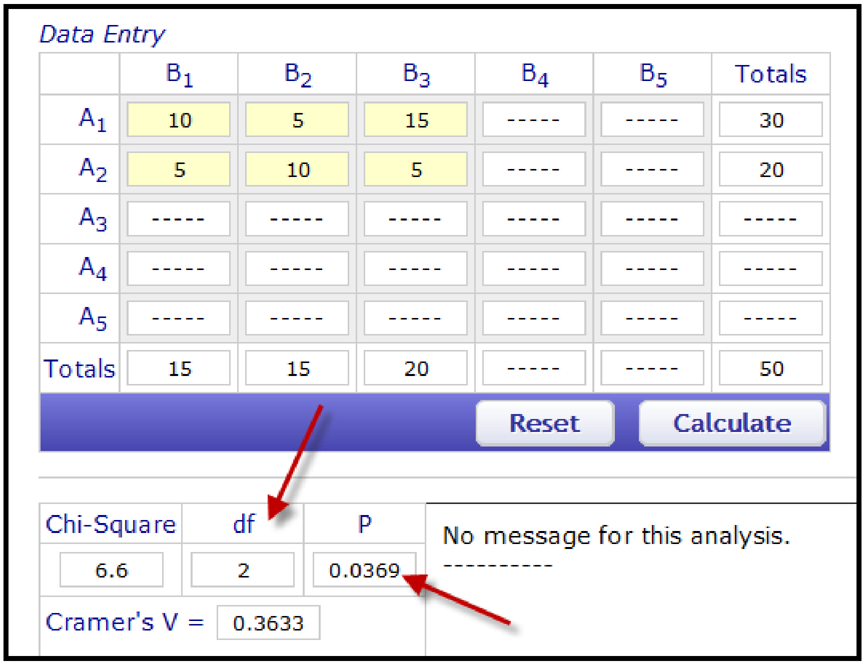


Select the number of rows:	2	3	4	5	2
Select the number of columns:	2	3	4	5	3

Data Entry

	B_1	B_2	B_3	B_4	B_5	Totals
A_1				-----	-----	-----
A_2				-----	-----	-----
A_3	-----	-----	-----	-----	-----	-----
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	-----	-----	-----	-----	-----	-----

2. Enter the data as shown below. Then click **Calculate**. Observe the DF and *p* value.



Data Entry

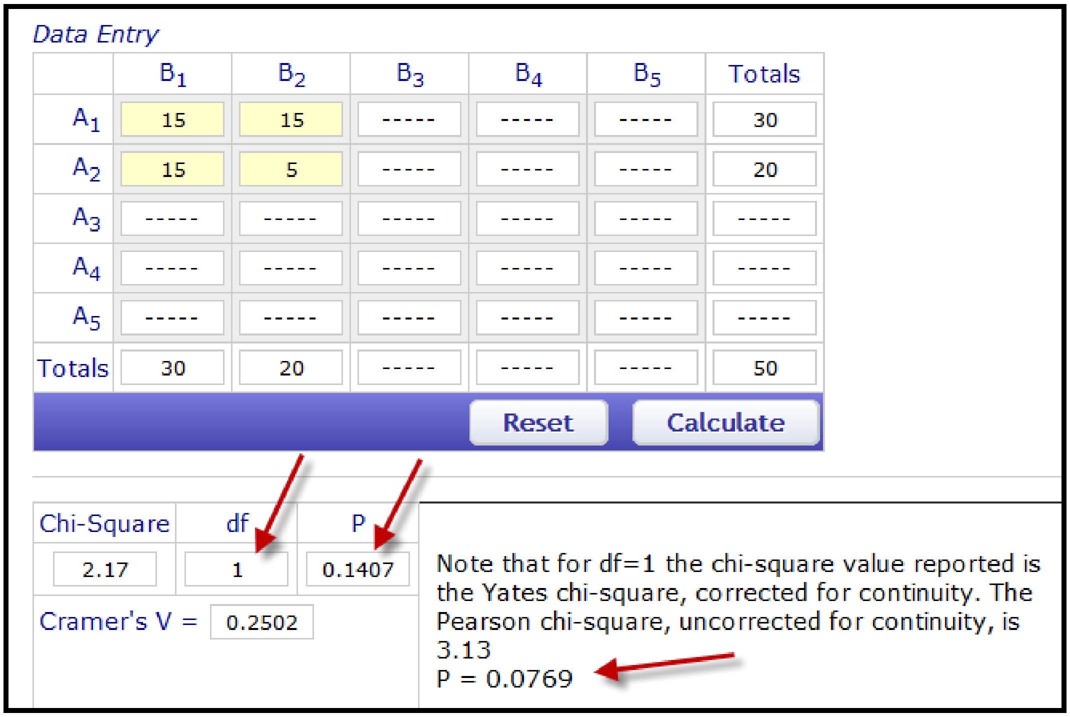
	B_1	B_2	B_3	B_4	B_5	Totals
A_1	10	5	15	-----	-----	30
A_2	5	10	5	-----	-----	20
A_3	-----	-----	-----	-----	-----	-----
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	15	15	20	-----	-----	50

Chi-Square	df	P
6.6	2	0.0369

Cramer's V = 0.3633

No message for this analysis.
----

3. Collapse B1 and B2 into one column by entering the data into a 2X2 table. Click **Calculate**. Observe the DF and *p* value. Also read the note about using the Yates Chi-square when DF = 1.



Data Entry

	B_1	B_2	B_3	B_4	B_5	Totals
A_1	15	15	-----	-----	-----	30
A_2	15	5	-----	-----	-----	20
A_3	-----	-----	-----	-----	-----	-----
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	30	20	-----	-----	-----	50

Chi-Square	df	P
2.17	1	0.1407

Cramer's V = 0.2502

Note that for df=1 the chi-square value reported is the Yates chi-square, corrected for continuity. The Pearson chi-square, uncorrected for continuity, is
3.13
P = 0.0769

4. Calculate the **mean square** using this formula: **Chi-square / DF**. It can be done by Excel or a hand-held calculator. When DF = 1, mean square = Chi-square. Compare the two mean squares from the 2X3 crosstab table and the one from a 2X2 table.

Chapter 6 exercises

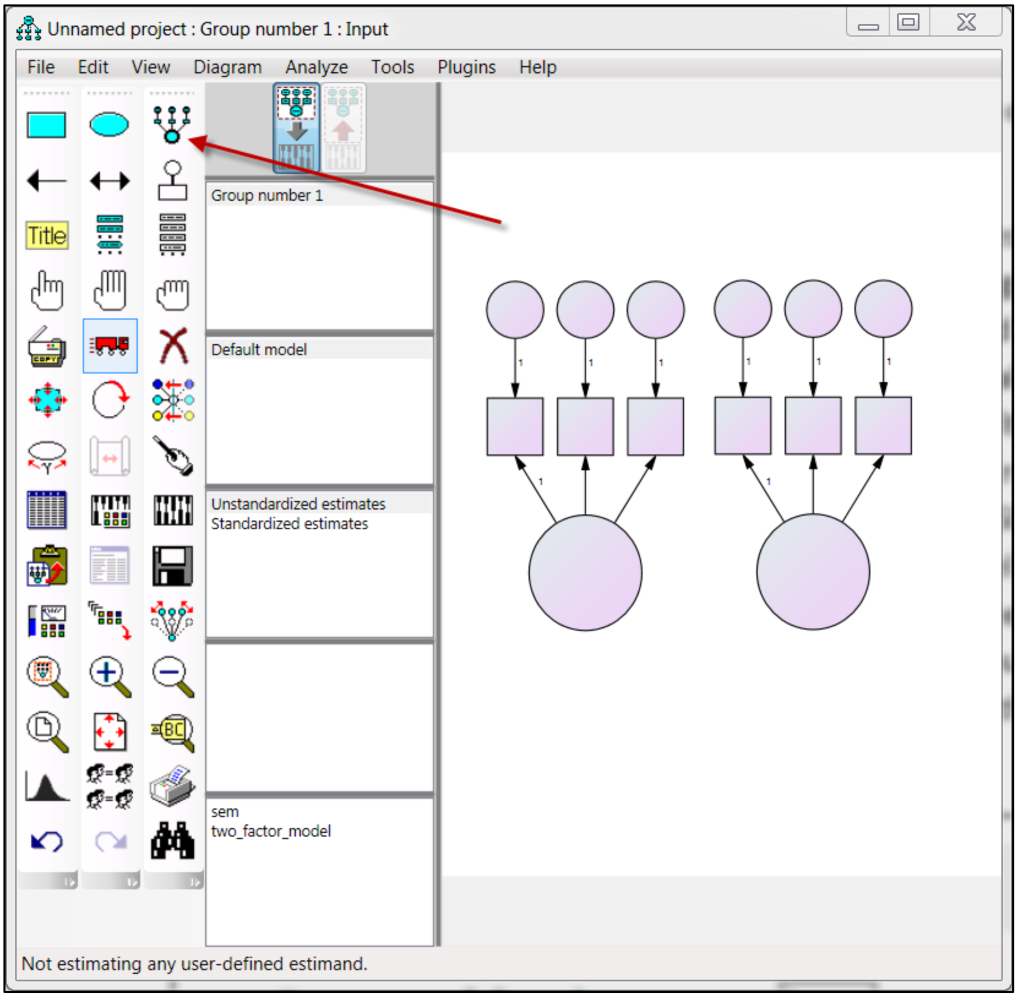
Objective: These exercises can help you understand that

1. In SEM the DF value is tied to the number of distinct elements and the model parameters to be estimated.
2. DF can be used for estimating the appropriate sample size in SEM

Tools: AMOS, online SEM sample size calculator, and R. The second resource can be accessed at http://timo.gnambs.at/en/scripts/powerforsem. R is an open source statistical package that can be freely downloaded from http://www.r-project.org/.

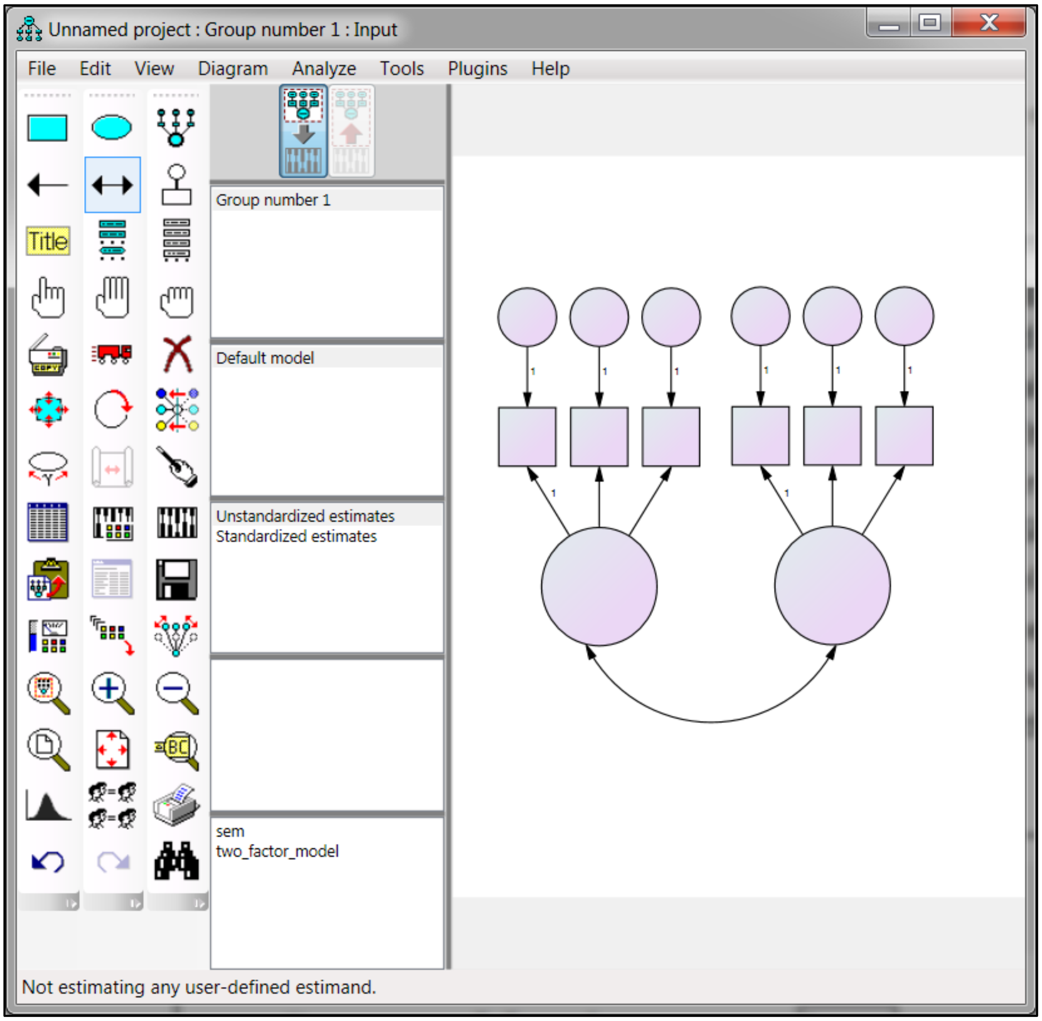
AMOS steps:

1. Open AMOS. Use the rightmost icon on the first row of icons to draw two factor models. Each model has three observed items.


Unnamed project : Group number 1 : Input

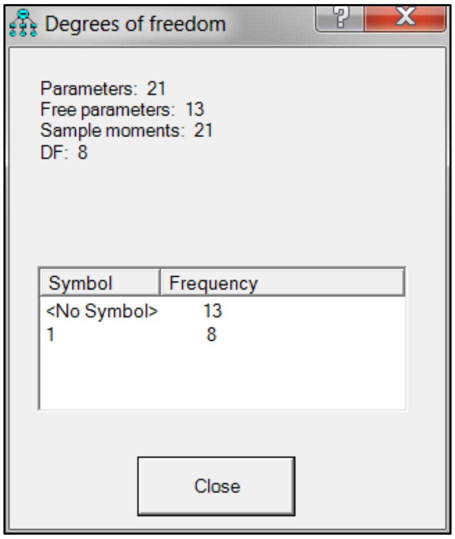
Not estimating any user-defined estimand.

2. Use the **bi-directional arrow** to connect the two circles. This indicates their covariance.



Not estimating any user-defined estimand.

3. Choose **Degrees of Freedom** from **Analyze**. Observe the output that parameters = 21 and DF = 8.

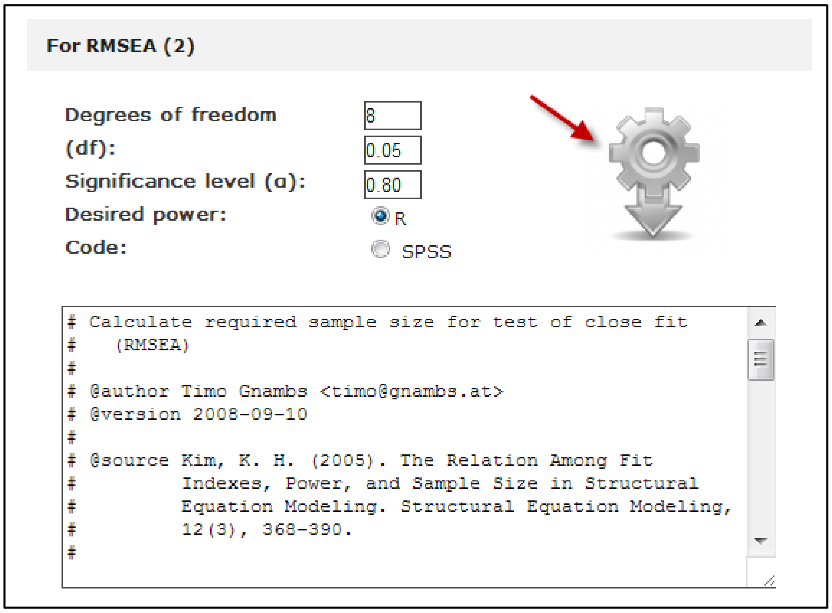


```
Parameters:  21
Free parameters:  13
Sample moments:  21
DF:  8
```

Symbol	Frequency
<No Symbol>	13
1	8

Online sample size calculator steps:

Go to the website http://timo.gnambs.at/en/scripts/powerforsem. Select **RMSEA(2)**. RMSEA stands for Root Mean Square Error of Approximation. Enter 8 into **Degrees of freedom**. Leave alpha level and desired power at the default (.05, 0.80). Click on the icon on the right. It will generate the R code for sample size calculation.



For RMSEA (2)

Degrees of freedom (df): 8

Significance level (α): 0.05

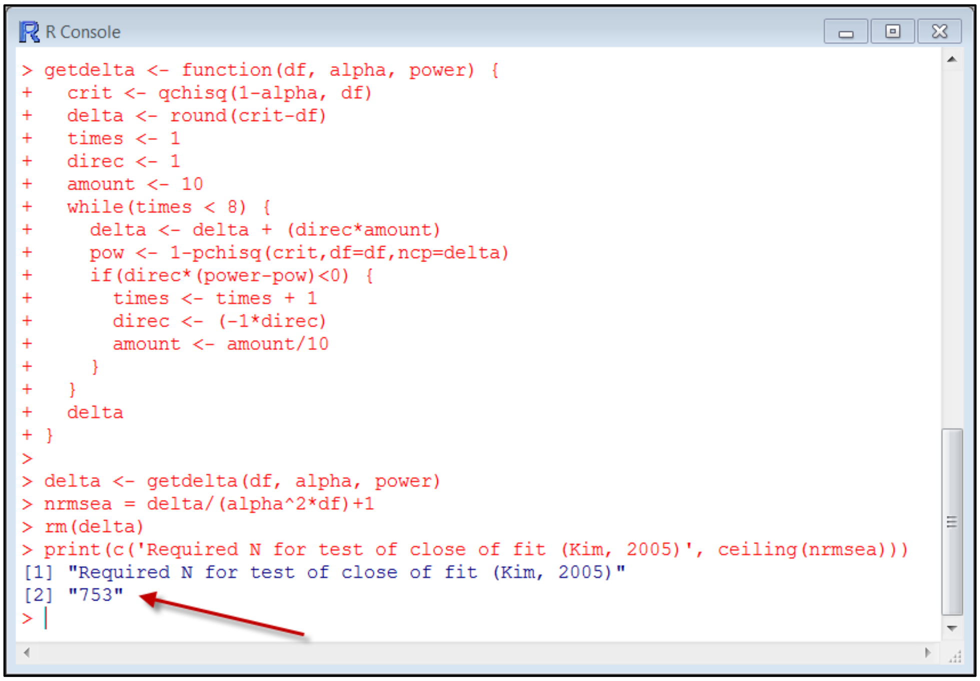
Desired power: 0.80

Code: ● R ○ SPSS

```
# Calculate required sample size for test of close fit
#    (RMSEA)
#
# @author Timo Gnambs <timo@gnambs.at>
# @version 2008-09-10
#
# @source Kim, K. H. (2005). The Relation Among Fit
#         Indexes, Power, and Sample Size in Structural
#         Equation Modeling. Structural Equation Modeling,
#         12(3), 368-390.
#
```

R steps

Copy and paste the codes into R. Press Enter. It will return the sample size recommendation.



```
> getdelta <- function(df, alpha, power) {
+   crit <- qchisq(1-alpha, df)
+   delta <- round(crit-df)
+   times <- 1
+   direc <- 1
+   amount <- 10
+   while(times < 8) {
+     delta <- delta + (direc*amount)
+     pow <- 1-pchisq(crit,df=df,ncp=delta)
+     if(direc*(power-pow)<0) {
+       times <- times + 1
+       direc <- (-1*direc)
+       amount <- amount/10
+     }
+   }
+   delta
+ }
>
> delta <- getdelta(df, alpha, power)
> nrmsea = delta/(alpha^2*df)+1
> rm(delta)
> print(c('Required N for test of close of fit (Kim, 2005)', ceiling(nrmsea)))
[1] "Required N for test of close of fit (Kim, 2005)"
[2] "753"
> |
```

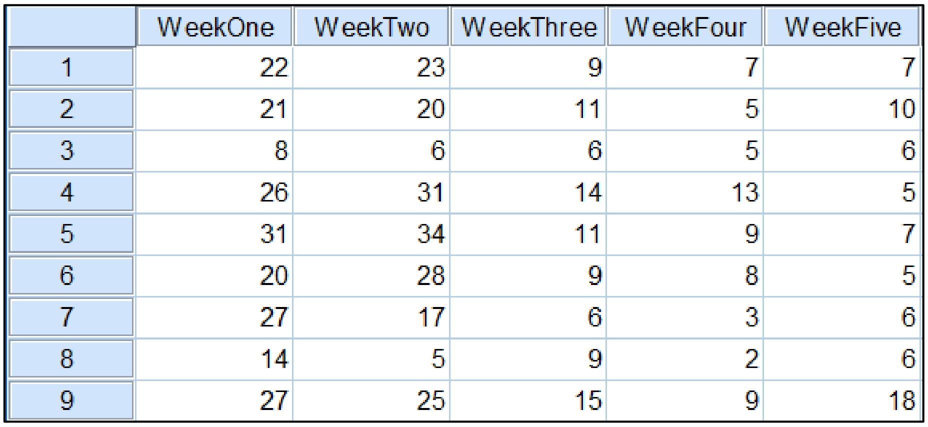
Chapter 7 exercises

Objective: These exercises can help you understand that in ANOVA repeated measures when the data structure cannot meet the assumption of sphericity, certain correctional procedures must be used to estimate the magnitude that sphericity has been violated. And then a correction factor is applied to the degrees of freedom of the *F* distribution, so that the Type I error rate is under control.

Tool: SPSS

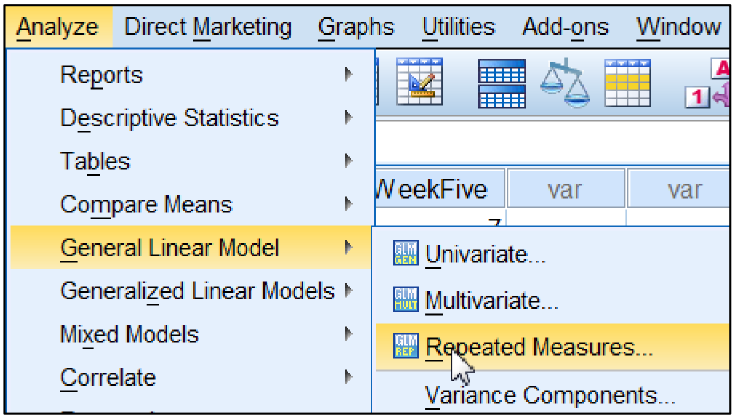
Steps

1. Open a new SPSS data table. Enter the data as below.

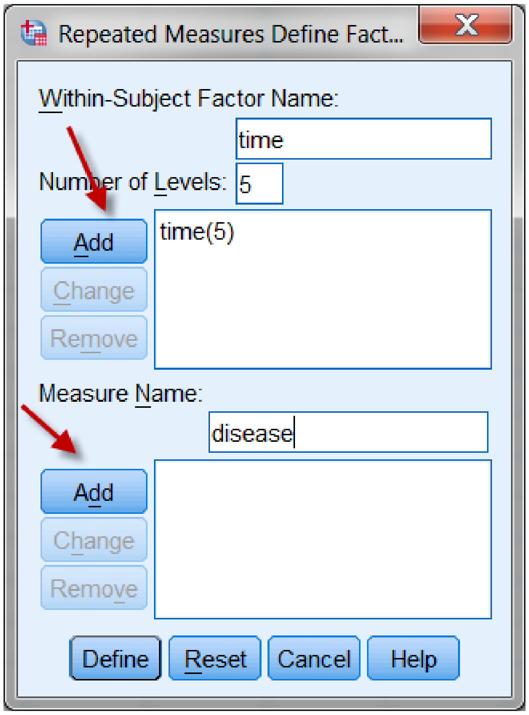


	WeekOne	WeekTwo	WeekThree	WeekFour	WeekFive
1	22	23	9	7	7
2	21	20	11	5	10
3	8	6	6	5	6
4	26	31	14	13	5
5	31	34	11	9	7
6	20	28	9	8	5
7	27	17	6	3	6
8	14	5	9	2	6
9	27	25	15	9	18

2. Select **General Linear Model/Repeated Measures** from **Analyze**.



3. In the pop-up panel, enter time into **Within-subject Factor Name**. Enter 5 into **Number of Levels**. Then click **Add**. Enter disease into **Measure Name**. Then click **Add**. Next, press **Define**.



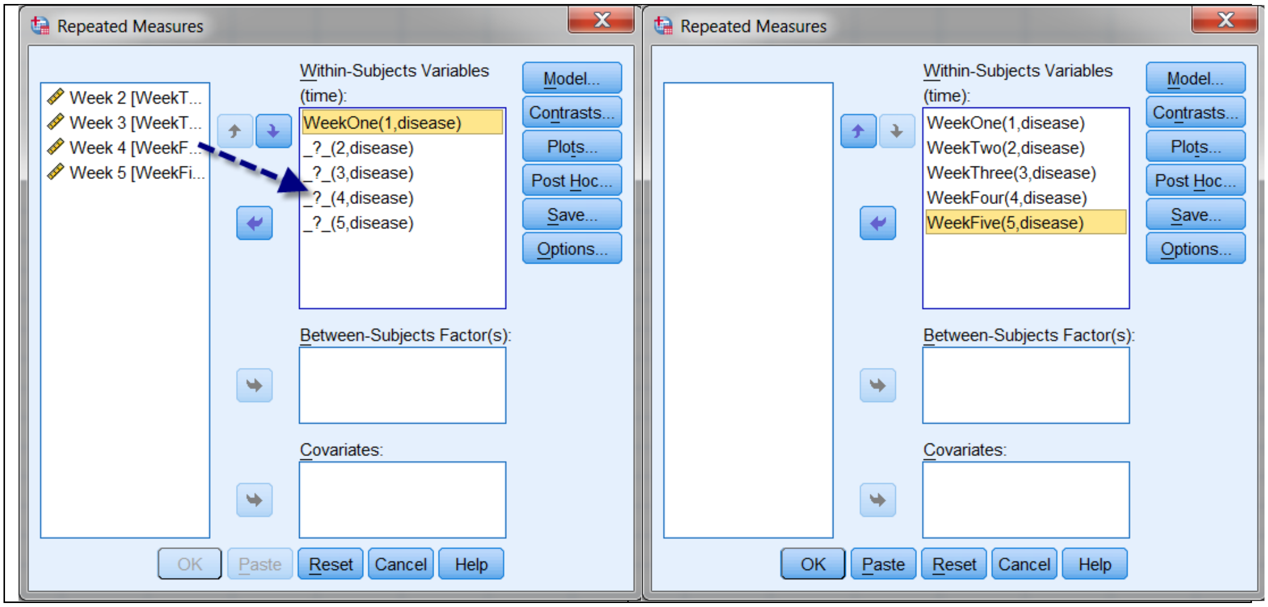
Within-Subject Factor Name: time

Number of Levels: 5

time(5)

Measure Name: disease

4. Assign each week to each within-subject variable. Then click **OK**.

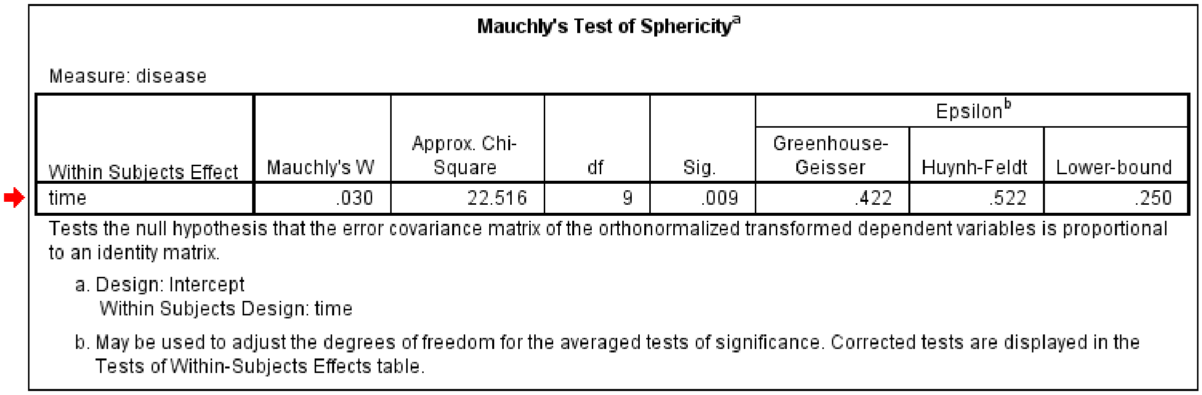


Week 2 [WeekT...
Week 3 [WeekT...
Week 4 [WeekF...
Week 5 [WeekFi...

Within-Subjects Variables (time):
WeekOne(1,disease)
_?_(2,disease)
_?_(3,disease)
_?_(4,disease)
_?_(5,disease)

Within-Subjects Variables (time):
WeekOne(1,disease)
WeekTwo(2,disease)
WeekThree(3,disease)
WeekFour(4,disease)
WeekFive(5,disease)

5. Check the output of Mauchly’s test of sphericty.



Mauchly's Test of Sphericity[a]

Measure: disease

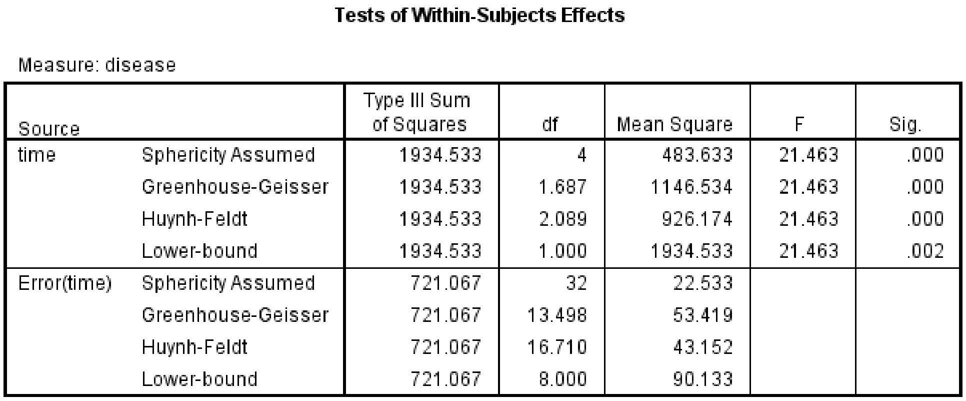
					Epsilon[b]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.030	22.516	9	.009	.422	.522	.250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: time

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

6. Observe the Type III sum of squares, degrees of freedom, *F* ratio, and *p* value. The DF values are different across the four methods (sphericity assumed and three corrections) but the F values remain constant.



Tests of Within-Subjects Effects

Measure: disease

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
time	Sphericity Assumed	1934.533	4	483.633	21.463	.000
	Greenhouse-Geisser	1934.533	1.687	1146.534	21.463	.000
	Huynh-Feldt	1934.533	2.089	926.174	21.463	.000
	Lower-bound	1934.533	1.000	1934.533	21.463	.002
Error(time)	Sphericity Assumed	721.067	32	22.533		
	Greenhouse-Geisser	721.067	13.498	53.419		
	Huynh-Feldt	721.067	16.710	43.152		
	Lower-bound	721.067	8.000	90.133		

Chapter 8 exercises

Objective: These exercises can help you understand that when data collected from multi-stage sampling are analyzed, conventional procedures treating every score equally cannot yield accurate estimations of the population parameters. Special procedures, such as SURVEYMEANS, SURVEYFREQ, and SURVEYREG, in which the degrees of freedom are defined by the number of strata or clusters, must be employed as a remedy.

Tool: SAS. If you do not have access to SAS but you are affiliated with a school, you can request an account to access *SAS on Demand*, which is totally free for academics. You can do the following exercises using either SAS programming environment or SAS Enterprise Guide. The data set and the source code for this exercise can be found at http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveyreg_sect002.htm

Steps

1. Enter the data below in SAS or copy it from the preceding website. In a junior high school there are 4,000 students spanning across Grades 7, 8, and 9 (Grade). The researcher wanted to study the relationships among household income (Income), the number of kids in a household (Kids), and students' average weekly spending for ice cream (Spending). After entering the data and the code, submit the program by clicking on the **Run** icon.

```
data IceCream;
   input Grade Spending Income Kids @@;
   if (Spending < 10) then Group='less';
   else Group='more';
   datalines;
7   7  39  2   7   7  38  1   8  12  47  1
9  10  47  4   7   1  34  4   7  10  43  2
7   3  44  4   8  20  60  3   8  19  57  4
7   2  35  2   7   2  36  1   9  15  51  1
8  16  53  1   7   6  37  4   7   6  41  2
7   6  39  2   9  15  50  4   8  17  57  3
8  14  46  2   9   8  41  2   9   8  41  1
9   7  47  3   7   3  39  3   7  12  50  2
7   4  43  4   9  14  46  3   8  18  58  4
9   9  44  3   7   2  37  1   7   1  37  2
7   4  44  2   7  11  42  2   9   8  41  2
8  10  42  2   8  13  46  1   7   2  40  3
9   6  45  1   9  11  45  4   7   2  36  1
7   9  46  1
;
```

2. Enter the following code to compute the probability of being sampled in each grade level and the sampling weights, given that the population size in each grade is known by the researcher. Next, submit the program.

```
data StudentTotals;
      input Grade _TOTAL_;
      datalines;
   7 1824
   8 1025
   9 1151
   ;
data IceCream;
      set IceCream;
      if Grade=7 then Prob=20/1824;
      if Grade=8 then Prob=9/1025;
      if Grade=9 then Prob=11/1151;
      Weight=1/Prob;
```

3. Enter the following codes to run compute the means.

```
proc surveymeans data=IceCream total=StudentTotals;
      stratum Grade / list;
      var Spending Group;
      weight Weight;
   run;
```

4. Observe the output in which stratum information and sampling weights were taken into account, and DF is adjusted by the number of strata.

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	3
Number of Observations	40
Sum of Weights	4000

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	Level	N
1	**7**	1824	1.10%	20	**Spending**		20
					Group	**less**	17
						more	3
2	**8**	1025	0.88%	9	**Spending**		9
					Group	**less**	0
						more	9
3	**9**	1151	0.96%	11	**Spending**		11
					Group	**less**	6
						more	5

Statistics

Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending		40	9.141298	0.531799	8.06377052	10.2188254
Group	**less**	23	0.544555	0.058424	0.42617678	0.6629323
	more	17	0.455445	0.058424	0.33706769	0.5738232

5. Enter the following code to run a SURVEYREG analysis

```
proc surveyreg data=IceCream total=StudentTotals;
      strata Grade /list;
      class Kids;
      model Spending = Income Kids / solution;
      weight Weight;
   run;
```

6. Observe the output in which stratum information and sampling weights are taken into account, and DF is adjusted by the number of strata.

The SURVEYREG Procedure
Regression Analysis for Dependent Variable Spending

Data Summary	
Number of Observations	40
Sum of Weights	4000.0
Weighted Mean of Spending	9.14130
Weighted Sum of Spending	36565.2

Design Summary	
Number of Strata	3

Fit Statistics	
R-square	0.8219
Root MSE	2.4185
Denominator DF	37

Stratum Information

Stratum Index	Grade	N Obs	Population Total	Sampling Rate
1	**7**	20	1824	1.10%
2	**8**	9	1025	0.88%
3	**9**	11	1151	0.96%

Class Level Information

Class Variable	Levels	Values
Kids	4	1 2 3 4

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	4	124.85	<.0001
Intercept	1	150.95	<.0001
Income	1	326.89	<.0001
Kids	3	0.99	0.4081

Note: The denominator degrees of freedom for the F tests is 37.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-26.086882	2.44108058	-10.69	<.0001
Income	0.776699	0.04295904	18.08	<.0001
Kids 1	0.888631	1.07000634	0.83	0.4116
Kids 2	1.545726	1.20815863	1.28	0.2087
Kids 3	-0.526817	1.32748011	-0.40	0.6938
Kids 4	0.000000	0.00000000	.	.

Chapter 9 exercises

Objective: These exercises can help you understand that

1. Normality is not a prevalent feature in most sampling distributions, otherwise we don’t need DF.
2. The appearance of sampling distributions depends on the degrees of freedom.

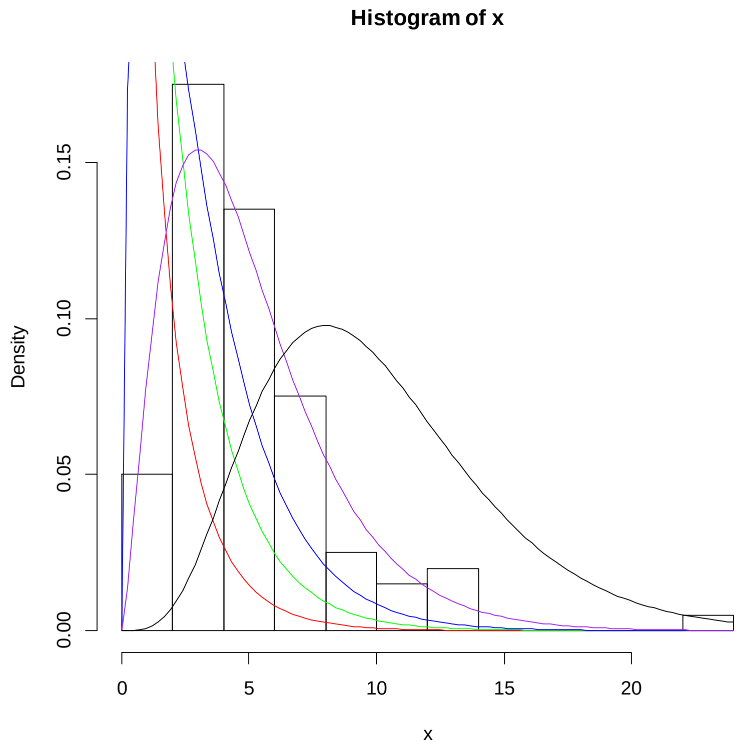
Tool: R. This can be freely downloaded from the Internet.

Steps:

1. Copy the following code (Snow, 2008) for plotting Chi-square distributions into the R console. Press Enter.

```
x <- rchisq(100, 5)
hist(x, prob=TRUE)
curve( dchisq(x, df=1), col='red', add=TRUE)
curve( dchisq(x, df=2), col='green', add=TRUE)
curve( dchisq(x, df=3), col='blue', add=TRUE)
curve( dchisq(x, df=5), col='purple', add=TRUE)
curve( dchisq(x, df=10), col='black', add=TRUE)
```

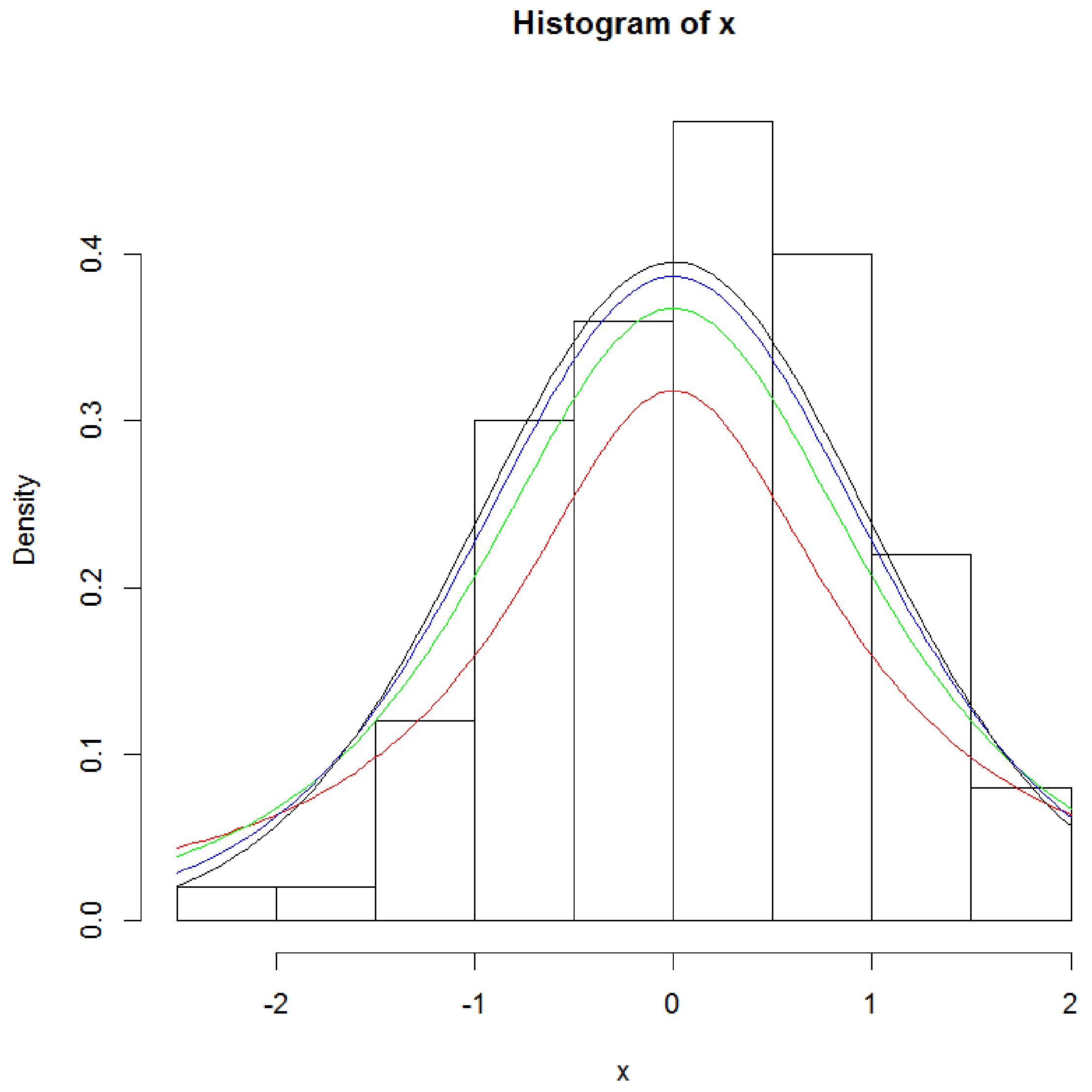
2. Observe how the shape of the Chi-square distributions corresponds to the degrees of freedom. To explore other possibilities, alter the DF values in the R code and resubmit the program.



3. Copy the following code for plotting *t*-distributions into the R console. Then press Enter.

```
x <- rt(100, 30)
hist(x, prob=TRUE)
curve( dt(x, df=1), col='red', add=TRUE)
curve( dt(x, df=3), col='green', add=TRUE)
curve( dt(x, df=8), col='blue', add=TRUE)
curve( dt(x, df=30), col='black', add=TRUE)
```

4. Observe how the shape of the *t*-distributions is affected by the degrees of freedom. Alter the DF values in the R code and resubmit the program.

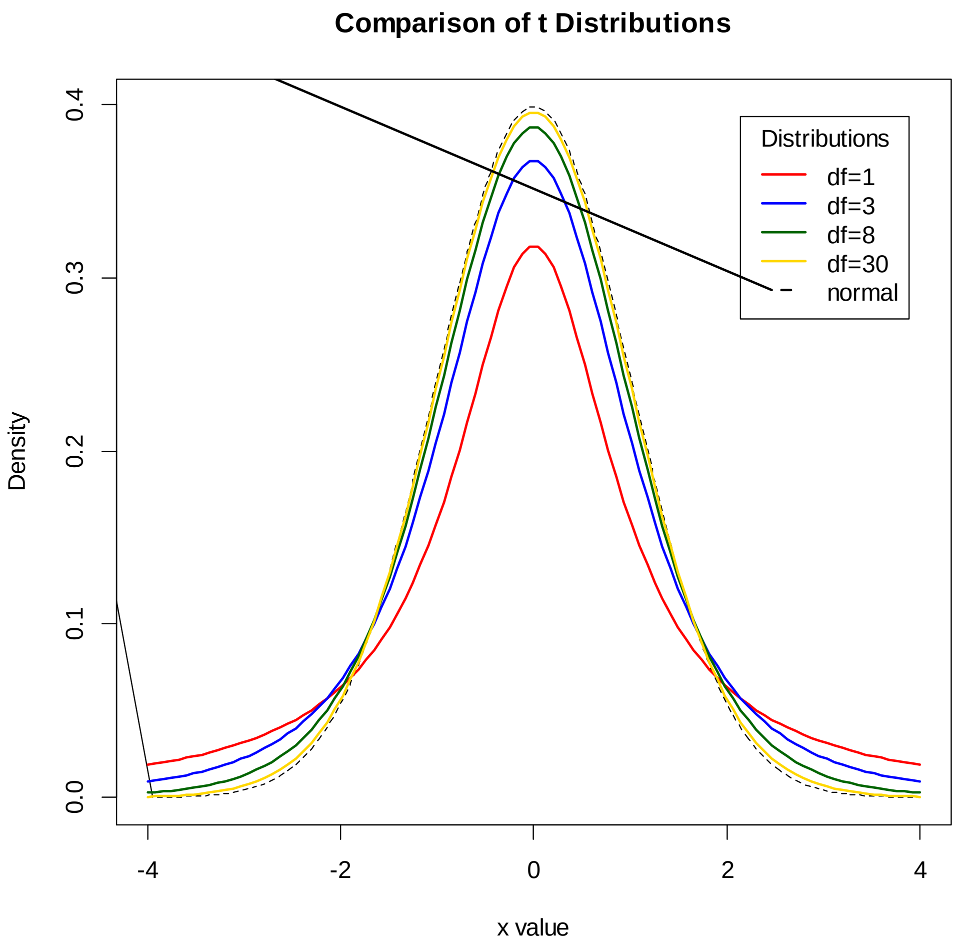


5. The following is a variant of R code for plotting *t*-distributions (Kabacoff, 2012). Copy it into the R console. Then press Enter.

```
x <- seq(-4, 4, length=100)
hx <- dnorm(x)
degf <- c(1, 3, 8, 30)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=30", "normal")
plot(x, hx, type="l", lty=2, xlab="x value", ylab="Density", main="Comparison of t Distributions")

for (i in 1:4){
lines(x, dt(x,degf[i]), lwd=2, col=colors[i])
}
legend("topright", inset=.05, title="Distributions", labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```

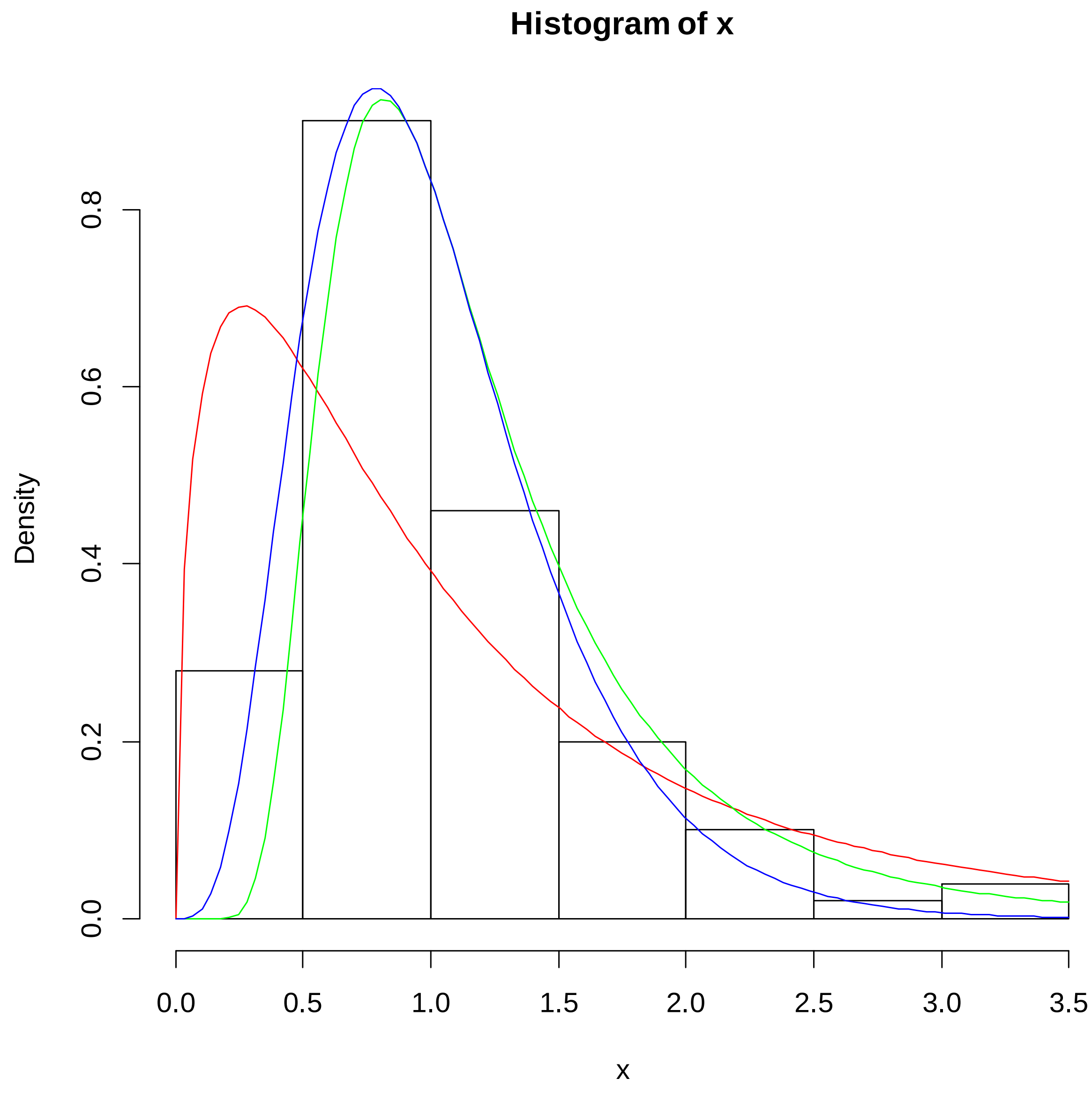
6. Observe how the shape of the *t*-distributions depends on the degrees of freedom. Alter the DF values in the R codes and resubmit the program.



7. Copy the following code for plotting the *F*-distributions into the R environment. Press Enter.

```
x <- rf(100, 10, 100)
hist(x, prob=TRUE)
curve( df(x, df1=3, df2=8), col='red', add=TRUE)
curve( df(x, df1=90, df2=10), col='green', add=TRUE)
curve( df(x, df1=10, df2=90), col='blue', add=TRUE)
```

8. Observe how the shape of the *F*-distributions corresponds to the degrees of freedom. You can explore different possibilities by altering the combinations of numerator DF and denominator DF.


Histogram of x
Density
x