A Note on Karl Pearson's 1900 Chi-Squared Test: Two Derivations of the Asymptotic Distribution, and Uses in Goodness of Fit and Contingency Tests of Independence, and a Comparison with the Exact Sample Variance Chi-Square Result

Timothy Falcon Crack,[1]

November 14, 2018

Karl Pearson's chi-squared test is widely known and used, both as a goodness-of-fit test for hypothesized distributions or frequencies, and in tests of independence in contingency tables. The test was introduced in Pearson (1900), but the derivation in that paper is almost incomprehensible. Two derivations of the asymptotic distribution are given here. The first uses joint characteristic functions, and the second uses a multivariate central limit theorem. Goodness-of-fit tests and contingency table tests of independence are discussed, and the asymptotic chi-square distribution result for Pearson's test statistic is compared and contrasted with the exact chi-square result for the sample variance estimator.

[1] University of Otago, School of Business, tcrack@otago.ac.nz, Tel: +0064 (0)3 479 8310.

Contents

1 Introduction

Pearson (1900) introduces the chi-squared test. Unfortunately, his paper is almost incomprehensible. For example, he writes $S_1\left(\frac{R_{pp}}{R}\frac{x_p^2}{\sigma_p^2}\right)$ whereas most of us would write $\sum_{i=1}^{n}\left(\frac{R_{ii}}{R}\frac{x_i^2}{\sigma_i^2}\right)$, or something similar. He writes $S_2\left(\frac{R_{pq}}{R}\frac{x_p x_q}{\sigma_p\sigma_q}\right)$ whereas most of us would write $\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{R_{ij}}{R}\frac{x_i x_j}{\sigma_i\sigma_j}\right)$, or something similar.

Let $\vec{x}$ be an $n \times 1$ vector of de-meaned random variables. Let T be their correlation matrix, and V be their variance-covariance matrix. Let $R = det(T)$, and let R_{pq}, be the minor obtained by striking out row p and column q of T. Then, noting my comments above, Pearson writes

$$S_1\left(\frac{R_{pp}}{R}\frac{x_p^2}{\sigma_p^2}\right) + S_2\left(\frac{R_{pq}}{R}\frac{x_p x_q}{\sigma_p\sigma_q}\right), \tag{1}$$

whereas most of us would write $\vec{x}'V^{-1}\vec{x}$, or something similar.

To make matters slightly more complicated for the modern reader, Pearson (1900) seems to be the first publication to use the notation χ^2 in reference to what we now call the chi-squared distribution (though the distribution itself is roughly 25 years older). It is doubly confusing, because Pearson (1900) labels the summation in Equation 1 as χ^2, and then goes on to look at the properties of χ, and then, argues that asymptotically, his χ^2 (i.e., Equation 1) has the distribution we now call a chi-squared distribution. It is confusing to the modern reader because we are used to χ^2 being the name of a family of distributions, rather than it being a term that we can take the square-root of.

Pearson also misses out very important information that would help the reader, for example, he does not couch the correlation of errors (his Equation (viii)) in terms of the multinomial distribution. Plackett (1983) discusses other sources of confusion in Pearson (1900).

After spending hours struggling through Pearson's (1900) paper, and ignoring several distracting small errors, I admit that I am not all convinced that Pearson's original arguments are correct. So, I turned to other derivations of his results, many of which also contain considerable gaps in logic or explanation. This note is my summary of what I have read, presented using modern notation, and with the gaps in logic filled in. I point to my sources as I go through.

2 Literature

First, Pearson introduces the goodness-of-fit chi-squared test in Pearson (1900). Given a hypothesized distribution, and an empirical distribution, the two may be compared using a chi-squared test. See Section 3 for a description of the test and Section 5 and Section 6 for derivations of its distribution.

Second, Pearson (1904) introduces the notion of "contingency." He shows that if you classify a population based on two attributes, then the expected frequencies in the intersection of the attributes (assuming that the attributes be independent) may be compared with the actually observed frequencies using the goodness-of-fit test from Pearson (1900). See Section 4 for a description of the test.

Third, a test of homogeneity can also be executed using the chi-squared framework from Pearson (1900), but it is not covered here. See Cramér (1946, p. 445) and DeGroot (1989, p. 542) for more details on the test of homogeneity.

3 The Chi-Squared Test of Goodness of Fit

Assumptions: Let us assume that our data sample, of size n, is drawn from a random variable with hypothesized population probability function $P(\cdot)$. If so, then for large n, we expect that our empirical distribution, obtained by placing probability $1/n$ in each observed point, may be regarded as a statistical image of (i.e., an approximation to, with some natural statistical variation) the population distribution described by $P(\cdot)$.

Let us partition the sample space into r non-overlapping and non-trivial exhaustive parts $S_1, \ldots, S_r$. The population probabilities of the sample space partition are given by $p_i = P(S_i)$ for $i = 1, \ldots, r$. Note that the non-trivial S_i assumption means that $P(S_i) > 0$ for $i = 1, \ldots, r$. The exhaustive assumption means that $\sum_{i=1}^{r} P(S_i) = 1$. The non-overlapping assumption means that $P(S_i \cap S_j) = 0$ for any $i \neq j$, where $i, j \in \{1, \ldots, r\}$.

We now note how the empirical sample falls into the sample space partition $S_1, \ldots, S_r$.

Recall first how a binomial random variable $X \sim Bin(n, p)$ is the summation of n Bernoulli variables: $X = \sum_{i=1}^{n} X_i$, where the X_i are IID, and

$$X_i = \begin{cases} 1, & \text{with probability } p, \text{ and} \\ 0, & \text{with probability } 1 - p. \end{cases} \tag{2}$$

Clearly $E(X_i) = 1 \cdot p = p$, and $V(X_i) = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = (1-p)[(1-p)p + p^2] = p(1-p)$. It follows (via linearity) that

$$E(X) = np,$$

and (via independence of trials; variance of sum equals sum of variances)

$$V(X) = np(1-p).$$

In our case, we think of our sample as the result of a set of n independent trials. If we let $\vec{X}$ be an $r \times 1$ counter of observed outcomes in the $r \times 1$ partition of the sample space, then $\vec{X} = \sum_{i=1}^{n} \vec{X}_i$, where $\vec{X}_i$ is an $r \times 1$ indicator of the placement of the i^{th} observed outcome out of the n trials. So, without loss of generality,

$$\underset{(r\times 1)}{\vec{X}_i} = \begin{cases} [1, 0, 0, \ldots, 0, 0]', & \text{with probability } p_1, \\ [0, 1, 0, \ldots, 0, 0]', & \text{with probability } p_2, \\ [0, 0, 1, \ldots, 0, 0]', & \text{with probability } p_3, \\ \vdots & \vdots \\ [0, 0, 0, \ldots, 1, 0]', & \text{with probability } p_{r-1}, \text{ and} \\ [0, 0, 0, \ldots, 0, 1]', & \text{with probability } p_r, \end{cases} \tag{3}$$

and the $\vec{X}_i$ are statistically independent, for $i = 1, \ldots, n$.

Now, let ν_i count the number of observations falling into S_i, for $i = 1, \ldots, r$. So that

$$\vec{X} = \sum_{i=1}^{n} \vec{X}_i = [\nu_1, \nu_2, \ldots, \nu_r]'. \tag{4}$$

Then $\nu_j = (\vec{X})_j = (\sum_{i=1}^n \vec{X}_i)_j = \sum_{i=1}^n (\vec{X}_i)_j$. From Equation 3, however, $(\vec{X}_i)_j$ (i.e., the j^{th} element of $\vec{X}_i$) takes the following values:

$$(\vec{X}_i)_j = \begin{cases} 1, & \text{with probability } p_j, \text{ and} \\ 0, & \text{with probability } \sum_{k=1, k\neq j}^r p_k = 1 - p_j, \end{cases} \tag{5}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, r$, and the $(\vec{X}_i)_j$ are independent for $i = 1, \ldots, n$. Given these n independent trials, it follows that $(\vec{X}_i)_j$ is a Bernoulli variable, and

$$\nu_j = \sum_{i=1}^{n} (\vec{X}_i)_j \sim Bin(n, p_j), \tag{6}$$

for each $j = 1, ..., r$. Thus, it follows from the argument just given, concerning the binomial distribution, that

$$\begin{aligned} E(\nu_j) &= np_j, \text{ and} && (7) \\ V(\nu_j) &= np_j(1 - p_j). && (8) \end{aligned}$$

Although the n Bernoulli trials with outcomes $(\vec{X}_i)_j$ are statistically independent for any fixed j and for $i = 1, \ldots, n$, because the $\vec{X}_i$ are statistically independent across i, it is not the case that the r binomial variables ν_j, for $j = 1, \ldots, r$ are independent. Indeed, the opposite is true.

For some simple intuition, consider $n = 10$ independent rolls of an $r = 6$-sided die. Let ν_j count the number of js obtained. Then, for example, ν_2 is distributed binomial $n = 10$ and $p = 1/6$: $\nu_2 \sim Bin(10, 1/6)$. It is also true that $\nu_3 \sim Bin(10, 1/6)$ for the number of threes obtained, but ν_2 and ν_3 cannot be independent. Indeed, the more twos you get, the fewer threes you are likely to get, because $\nu_1 + \nu_2 + \nu_3 + \cdots + \nu_6 = 10$. For example, if you get $\nu_2 = 0$ twos, how many threes can you get? Whereas, if you get $\nu_2 = 9$ twos, how many threes can you get? With a fixed number of trials, more of one outcome leaves less opportunity for the other outcomes to occur. So, we expect the different counts of outcomes ν_j to covary negatively.

That is, although each binomial outcome ν_j is built from the outcomes of n independent trials, the binomial variables should covary negatively as a result of the restriction/constraint that $\sum_{j=1}^{r} \nu_j = n$. Let us now prove this, for $j \neq k$.

$$\begin{aligned}
cov(\nu_j, \nu_k) &= E[\nu_j \nu_k] - E(\nu_j)E(\nu_k) \\
&= E\left[\sum_{i=1}^{n} (\vec{X}_i)_j \sum_{i=1}^{n} (\vec{X}_i)_k\right] - (np_j)(np_k) \\
&= E\left[\sum_{i=1}^{n} (\vec{X}_i)_j (\vec{X}_i)_k + \sum_{\substack{i=1 \\ i\neq l}}^{n} \sum_{l=1}^{n} (\vec{X}_i)_j (\vec{X}_l)_k\right] - (np_j)(np_k) \\
&\overset{*}{=} \sum_{\substack{i=1 \\ i\neq l}}^{n} \sum_{l=1}^{n} E\left[(\vec{X}_i)_j\right] E\left[(\vec{X}_l)_k\right] - (np_j)(np_k) \\
&= \sum_{\substack{i=1 \\ i\neq l}}^{n} \sum_{l=1}^{n} p_j p_k - (np_j)(np_k) \\
&= n(n-1)p_j p_k - (np_j)(np_k) \\
&= -np_j p_k, \qquad (9)
\end{aligned}$$

where step-$*$ follows because $\sum_{i=1}^{n} (\vec{X}_i)_j (\vec{X}_i)_k$ must be identically zero (because for fixed i, at least one of $(\vec{X}_i)_j$ or $(\vec{X}_i)_k$ must be zero), and because independence of the trials implies that $E\left[(\vec{X}_i)_j (\vec{X}_l)_k\right] = E\left[(\vec{X}_i)_j\right] E\left[(\vec{X}_l)_k\right]$, and $1 \leq j \neq k \leq r$.

Equations 7, 8, and 9 describe the vector $\vec{\nu} = [\nu_1, \ldots, \nu_r]'$ as being distributed *multinomial* with parameters $n, p_i, \ldots, p_r$ (Evans et al., 1993).

You can deduce from Equations 8 and 9 that for $j \neq k$,

$$corr(\nu_j, \nu_k) = \frac{cov(\nu_j, \nu_k)}{\sqrt{V(\nu_j) \cdot V(\nu_k)}} = \frac{-np_j p_k}{\sqrt{[np_j(1-p_j)][np_k(1-p_k)]}} = -\frac{\sqrt{p_j p_k}}{\sqrt{(1-p_j)(1-p_k)}}, \qquad (10)$$

which is not a function of sample size n.

With all of this statistical machinery behind us, we can now move on to look at the chi-squared test. Remember that the hypothesis is that our data are a random sample of size n from the population with probability function $P(\cdot)$.

Cramér (1946, p. 416) points out that we are assuming that $P(\cdot)$ is completely specified, and that we can calculate $P(S)$ for any S.

To measure how close the empirical distribution is to the hypothesized population distribution, we need a measure of the deviation between the two. The realized proportion of the sample that appears in partition S_j is given by ν_j/n, whereas the expected proportion is p_j. Cramér (1946, p. 417) argues that we should use a squared deviation measure of form

$$C = \sum_{j=1}^{r} c_j(\nu_j/n - p_j)^2, \tag{11}$$

and that the choice of c_j may be taken almost arbitrarily.

Cramér (1946, p. 417) points out that Pearson (1900) chooses $c_j = n/p_j$ in Equation 11. This choice yields

$$\begin{aligned} C &= \sum_{j=1}^{r} n/p_j \; (\nu_j/n - p_j)^2, \\ &= \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \qquad (12) \\ &= \sum_{j=1}^{r} \frac{(O_j - E_j)^2}{E_j}, \qquad (13) \end{aligned}$$

where $O_j = \nu_j$ is an observed frequency count and $E_j = np_j$ is the population expected count of the number of observations in partition S_j.

In Section 5 and Section 6, we derive the distribution of C (i.e., Equations 12 and 13) and show that under our multinomial assumptions,

$$C = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \overset{A}{\sim} \chi^2_{r-1}, \tag{14}$$

where the A denotes an asymptotic result, as $n \to \infty$.

3.1 Comparison With Sample Variance Result I

Let us now compare the result in Equation 14 with the scaled sample variance chi-square distribution result obtained as part of the derivation or the Student-t test of the mean.

Recall that in the sample variance case, if we assume that a random sample $X_1, \ldots, X_N \sim N(\mu, \sigma^2)$, then for unbiased sample variance estimator $s^2 = \frac{1}{n-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$,

$$\frac{(N-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2_{N-1}, \tag{15}$$

as proved by Fisher (1925). Although superficially similar, the results in Equation 14 and Equation 15 differ along multiple dimensions, as follows.

First, the result in Equation 15 is an *exact* result whereas the result in Equation 14 is an *asymptotic* result—and thus an *approximation* in any finite sample.

Second, the result in Equation 15 is based on a set of data $X_1, \ldots, X_N$ that are statistically *independent* of each other, whereas the result in Equation 14 is based upon a set of data $\nu_1, \ldots, \nu_r$ that are *dependent*, as shown in Equation 9.

Third, the result in Equation 15 is based on a set of data $X_1, \ldots, X_N$ that are *normally distributed*, whereas the result in Equation 14 is based upon a set of data $\nu_1, \ldots, \nu_r$ that are *multinomially distributed*, as shown in Equations 7, 8, and 9.

Fourth, the result in Equation 15 is based on a set of data $X_1, \ldots, X_N$ that are *identically distributed*, whereas the result in Equation 14 is based upon a set of data $\nu_1, \ldots, \nu_r$ that *non-identically distributed*. We can see from Equations 7 and 8 that the means and variances of the ν_j are potentially all different. Now, sure enough, there is a demeaning and some scaling in Equation 12, but even so, the summands $(\nu_j - np_j)^2/np_j$ appearing in Equation 12 are still non-identically distributed. For example,

$$E[(\nu_j - np_j)^2/np_j] = E(\nu_j - np_j)^2/np_j = V(\nu_j)/np_j = 1 - p_j, \tag{16}$$

using Equations 8 and 9, and this result potentially differs for all j.

Fifth, from Equation 14 we see that the squared terms in ν_j are deviations from the *population mean* np_j, whereas in Equation 15 we see that the squared terms in X_i are deviations from the *sample mean* $\bar{X}$.

Sixth, an additional difference related to estimation of common variances is discussed in Section 3.2, following.

So, in summary, although Equation 14 and Equation 15 both give chi-square distributed results, the assumptions and context are very different.

3.2 Comparison With Sample Variance Result II

Is it possible to reformulate our goodness-of-fit test in Equation 14 in some way so that it looks more like the sample variance result in Equation 15? After all, binomially distributed random variables are asymptotically normally distributed. Well, a significant problem is that although the ν_j are asymptotically normally distributed, they are still going to be *dependent*, because the correlation dependence structure shown in Equation 10 is not a function of sample size n, and they are still going to be non-identically distributed, as discussed in Section 3.1.

What does the sample variance of the ν_j look like? It would be

$$s^2 = \frac{1}{r-1}\sum_{j=1}^{r}(\nu_j - n/r)^2, \tag{17}$$

because the sample mean is given by

$$\bar{\nu} = \frac{1}{r}\sum_{j=1}^{r}\nu_j = n/r \ ,$$

because of the constraint that $\sum_{j=1}^{r}\nu_j = n$. Another way to think about the sample mean $\bar{\nu}$ is that the n observations are distributed over the r partitions $S_1, \ldots, S_r$, and so, on average, there are n/r observations per part S_j. In addition, note that $1/r$ is the average value of the p_j (that is, $\bar{p} = \frac{1}{r}\sum_{j=1}^{r}p_j = 1/r$, because the p_j sum to 1). So, n/r is the average value of the population mean np_j terms as well as being the average value of the sample observations, neither of which is used in Equation 14 to calculate differences.

Now ask yourself what Equation 17 is actually estimating. Given that the ν_j are non-identically distributed, Equation 17 is not estimating some common variance. So, Equation 17 is nonsense. It would make more sense to standardize the terms before summing them, by calculating

$$s^2 = \frac{1}{r-1}\sum_{j=1}^{r}\left(\frac{\nu_j - n/r}{\sqrt{np_j(1-p_j)}}\right)^2 , \tag{18}$$

using Equation 8. The summands are, however, still non-identically distributed, because $E(\nu_j - n/r) = n(p_j - 1/r)$, which potentially differs for every j.

What if we calculate Equation 19 instead

$$s^2 = \frac{1}{r} \sum_{j=1}^{r} \left(\frac{\nu_j - np_j}{\sqrt{np_j(1-p_j)}} \right)^2 , \tag{19}$$

recognizing that we know the population mean, $E(\nu_j) = np_j$, and we know that the ratio $(\nu_j - np_j) \Big/ \sqrt{np_j(1-p_j)}$ is asymptotically standard normal, assuming that $np_j > 5$ and $n(1-p_j) > 5$ (Zwillinger and Kokoska, 2000, p. 85)? Although the summands in Equation 19 are now identically distributed, the equation still does not make sense, because we already know that each summand is the square of an asymptotically standard normally distributed random variable. So, we are still in no sensible way estimating any common variance, as we were in Equation 15.

Finally, see Section 6 for a derivation of the asymptotic distribution of Pearson's chi-squared test statistic that is reminiscent of the sample variance derivation, at least insofar as it adds up squared deviations scaled by their theoretical variance (in the form of the square root of the variance-covariance matrix).

The bottom line is that there is no point in using a sample variance estimator for the sample $\nu_1, \dots, \nu_r$ because the data have heterogeneous variances (so what are you even estimating?), and if you standardize the data first, then there is no need to estimate variance (because you already know what it should be).

4 The Contingency Table Test of Independence

As mentioned already, Karl Pearson introduces the notion of "contingency" in Pearson (1904). He shows that if you classify a population based on two attributes, then the observed frequencies in the intersection of the attributes (assuming that the attributes be independent) may be compared with the expected observed frequencies using the goodness-of-fit test from Pearson (1900).

Pearson (1904) assumes that A is some attribute of a population partitioned into groups $A_1, \dots, A_s$. For a sample of size N, suppose there are empirical observations of the numbers falling into the A groupings given by $n_1, \dots, n_s$, yielding empirical probabilities of falling into the groups given by $n_1/N\,, \dots, n_s/N$. Similarly, assume that there is another attribute of the population partitioned into groups $B_1, \dots, B_t$, and empirical observations of the same N individuals falling into these groups given by $m_1, \dots, m_t$, yielding empirical probabilities of falling into the groups given by $m_1/N\,, \dots, m_t/N$.

Each individual in the sample has a pair of attributes, A_u and B_v.[2] With N such pairs of attributes (one pair for each sample observation), the theory of independent probability says that the (sample or estimated) probability that observations fall into the intersection of attributes A_u and B_v must be the product of the marginal probabilities: $(n_u/N) \cdot (m_v/N)$.

With N observations in the sample, we thus expect to find

$$N \cdot (n_u/N) \cdot (m_v/N) = \frac{n_u \cdot m_v}{N} = n_{u,v}, \tag{20}$$

say, observations in the sample intersection of attributes A_u and B_v (note that I have altered Pearson's notation here to make it consistent with Cramér's notation used in my previous section). If we actually see $\nu_{u,v}$ observations in the intersection of attributes A_u and B_v, then $\nu_{u,v} - n_{u,v}$ is the deviation from independent probability in the sample intersection of attributes A_u and B_v. This deviation must be cumulated somehow over the whole table.

"I term any measure of the total deviation of the classification from independent probability a measure of its *contingency*. Clearly the greater the contingency, the greater must be the amount of association or of correlation between the two attributes, for such association or correlation is solely a measure from another standpoint of the degree of deviation from independence of occurrence (Pearson, 1904, pp. 4–5)."

Pearson (1904) goes on to calculate

$$\begin{aligned} \mathcal{C} &= \sum_{u=1}^{s}\sum_{v=1}^{t} \frac{(\nu_{u,v} - n_{u,v})^2}{n_{u,v}} && (21) \\ &= \sum_{u=1}^{s}\sum_{v=1}^{t} \frac{(O_{u,v} - \hat{E}_{u,v})^2}{\hat{E}_{u,v}}, && (22) \end{aligned}$$

where $O_{u,v} = \nu_{u,v}$ is an observed frequency count, and $\hat{E}_{u,v} = n_{u,v}$ is the expected count of the number of observations in the sample intersection of attributes A_u and B_v.

This application differs from that seen in Section 3 (goodness-of-fit), because here sample estimates $\hat{E}_{u,v}$ are used in Equation 22 for the expected counts, rather than population parameters E_j seen in Equation 13 (i.e., the "hat" denotes a sample estimate).

The term "degrees of freedom" did not exist in 1904. Student (1908) and Fisher (1915) both use the concept of degrees of freedom without naming it. It was Fisher who introduced the term in a discussion of chi-squared contingency tables (Fisher, 1922).

[2] An "individual" need not be a solitary individual (e.g., a human with attributes height and IQ). In Section 4.1, I give an example where each individual is an individual sample of size N carrying attributes scaled sample mean M_i and scaled sample variance V_i.

It is difficult to find a clear statement in Pearson (1904) of the degrees of freedom for Equations 21 and 22. We now know that[3]

$$\mathcal{C} \overset{A}{\sim} \chi^2_{(s-1)\cdot(t-1)}, \tag{23}$$

where s is the number of categories of the first attribute and t is the number of categories of the second attribute (Cramér, 1946, 441–445). This assumes that $\hat{E}_{u,v} = n_{u,v} > 5$ in all cells (Zwillinger and Kokoska, 2000, pp. 233–234) and that $(s-1)\cdot(t-1) > 1$.

4.1 An Example and A Possible Source of Confusion

After conducting many contingency table tests of independence, I have noticed a potentially confusing phenomenon. Let me give an example. Fisher (1925) is the first to prove that given IID normally distributed data, the sample mean and sample variance are statistically independent. Indeed, *normality* of the underlying IID data is the *only* case in which the sample mean and sample variance are independent (Geary, 1936; Cramér 1946, p. 382), though these sample moments will be *uncorrelated* for any underlying distribution for which the third central moment equals zero (Cramér 1946, p. 348–349).

Suppose that you use Matlab to simulate 10,000 samples of five IID normal observations $X_1, \ldots, X_5 \sim N(\mu, \sigma^2)$, for a given μ and σ. For each sample, calculate the standardized sample mean and the scaled sample variance, as in Equation 24 and Equation 25, where $N = 5$.

$$\begin{aligned} M_i &= \frac{\bar{X}-\mu}{\sigma\Big/\sqrt{N}} && (24)\\ V_i &= \frac{(N-1)s^2}{\sigma^2}\Bigg/(N-1) = \frac{s^2}{\sigma^2}, \text{ where} && (25)\\ \bar{X} &= \frac{1}{N}\sum_{i=1}^{N} X_i, \text{ and} \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2. \end{aligned}$$

[3] Given the row and column sums, let *df* be the minimum number of cells that can be given in the table and still allow all other cell contents to be deduced. If you strike out any one row and any one column, the remaining *df* $= (s-1)(t-1)$ cells allow you to deduce those that you struck out. It is not true for *any* $(s-1)(t-1)$ cells; only if they remain after striking out a row and a column.

Figure 1 shows data from one such simulation. Table 1 counts observations and expected counts in a two-way contingency table. The chi-square test of independence calculated in the caption of Table 1 fails to reject the hypothesized independence.[4]

Now look closely more at the plot. Are these data really independent? Does it look as though the range of values taken by V_i is much wider when $-0.50 \leq M_i \leq 0.50$ than when $3 \leq M_i \leq 4$? If this were true, then V and M would not be independent. This is a possible source of confusion. So, let me explain.

In fact, the two variables M_i and V_i are independent in theory and also in the simulation. There is *no evidence* in the data or in the plot to suggest that the range of V_i differs in any way with varying M_i. Suppose I now restart my simulation code and leave it running, adding new observations to my data set only when $3 \leq M_i \leq 4$, and deleting them otherwise. If I do this until I have as many observations in the band $3 \leq M_i \leq 4$ as I do in the band $-0.50 \leq M_i \leq 0.50$, then the two bands appear indistinguishable on the plot. Possible confusion arises only because of the relative paucity of the observations (M_i, V_i) when $3 \leq M_i \leq 4$ compared with when $-0.50 \leq M_i \leq 0.50$.

5 Asymptotic Distribution I: Via Joint CFs

In Crack et al. (2018) we derive the exact chi-squared distribution shown in Equation 15 using a canonical form of an orthogonal decomposition due to Cochrane (1934) and Cramér (1946). Here I would like to derive the asymptotic distribution of the chi-squared test statistic shown in Equations 12 and 13 and 14, based upon the multinomial framework described already in Section 3. I will follow Cramér (1946, pp. 416–419), but with much more detail than Cramér gives.

In Section 3 we established that the vector $\vec{\nu} = [\nu_1, \ldots, \nu_r]'$ is distributed multinomial with parameters $n, p_i, \ldots, p_r$. This means that the probability of seeing exactly the outcome $\vec{\nu} = [\nu_1, \ldots, \nu_r]'$ is given by

$$P(\vec{\nu}; n, p_i, \ldots, p_r) = \frac{n!}{\nu_1! \cdots \nu_r!} p_1^{\nu_1} \cdots p_r^{\nu_r} = \binom{n}{\nu_1, \nu_2, \ldots, \nu_r} p_1^{\nu_1} \cdots p_r^{\nu_r}, \tag{26}$$

where $\binom{n}{\nu_1, \nu_2, \ldots, \nu_r} = \frac{n!}{\nu_1! \cdots \nu_r!}$ is the multinomial coefficient (Johnson, Kotz, and Kemp, 1993, p. 4). Note that Equation 26 is just the general term in the expansion of $(p_1 + \cdots + p_r)^n$.

[4] Why did I choose this form for M_i and V_i? It is because $t_{N-1,i} = M_i / \sqrt{V_i}$ is the Student-t test statistic for the mean for sample i. So, M_i and $\sqrt{V_i}$ are the numerator and denominator of this common test statistic. Their independence, when the underlying data are assumed to be normally distributed, is one of the requirements for the ratio to be distributed Student-t.

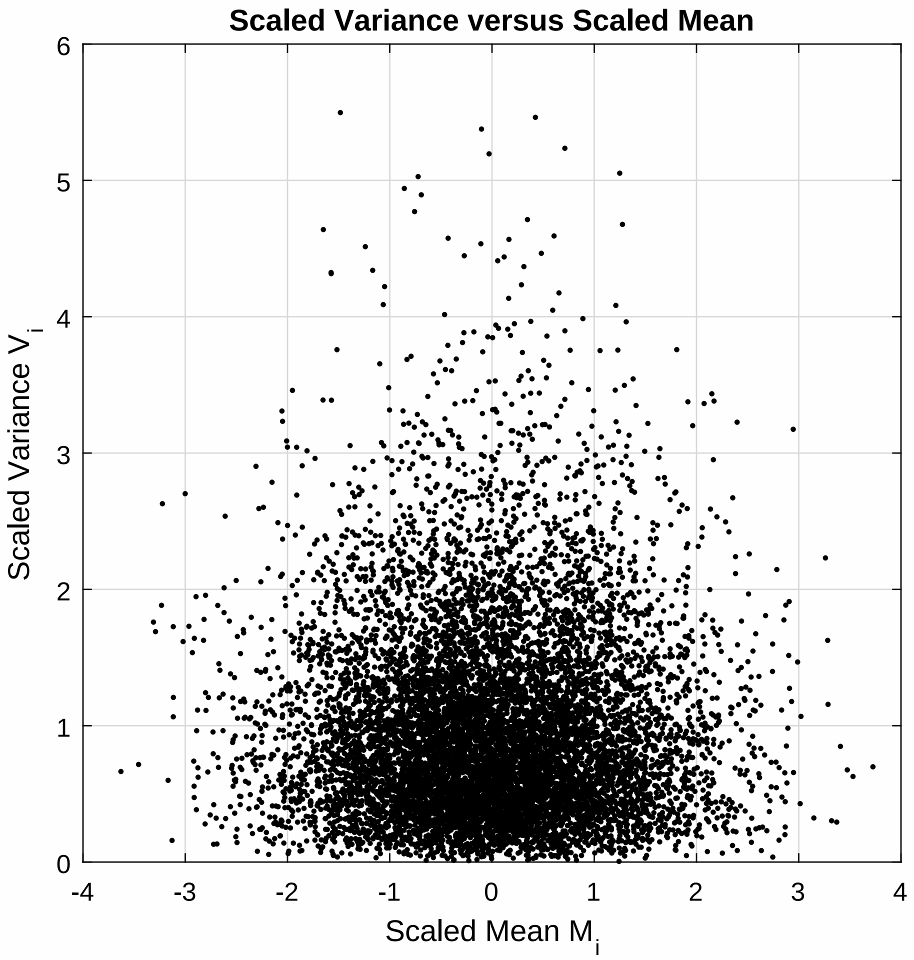

Figure 1: Scaled Variance versus Scaled Mean

The figure plots scaled sample variance $V_i = \frac{s^2}{\sigma^2}$ (Equation 25) on the vertical axis versus standardized sample mean $M_i = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}}$ (Equation 24) on the horizontal axis, for a simulation of 10,000 samples of IID data of sample size $N = 5$ where $X_1, \ldots, X_5 \sim N(\mu, \sigma^2)$. I used $\mu = 0$ and $\sigma = 0.01$, but the results presented are immune to these choices (assuming $\sigma > 0$). All 10,000 observations appear on the plot (i.e., there are no observations where $|M_i| > 4$ or $V_i > 6$). Note that under these assumptions $M_i \sim N(0, 1)$, $V_i \sim \chi^2_{N-1}/(N-1)$, and $M_i/V_i \sim t_{N-1}$. See also Table 1 for a tabulation of the data.

Observed ($O_{u,v}$)	$0 \leq V_i < 1$	$1 \leq V_i < 2$	$2 \leq V_i < 3$	$3 \leq V_i \leq 6$	Row Total
$-4 < M_i \leq -2$	137	80	15	4	**236**
$-2 < M_i \leq -1$	809	436	99	20	1364
$-1 < M_i \leq 0$	2009	1070	271	58	3408
$0 < M_i \leq 1$	2049	1049	254	71	3423
$1 < M_i \leq 2$	802	407	96	23	1328
$2 < M_i < 6$	142	80	14	5	241
Column Total	5948	**3122**	749	181	10000

Expected ($\hat{E}_{u,v}$)	$0 \leq V_i < 1$	$1 \leq V_i < 2$	$2 \leq V_i < 3$	$3 \leq V_i \leq 6$	Row Total
$-4 < M_i \leq -2$	140.4	**73.7**	17.7	4.3	236
$-2 < M_i \leq -1$	811.3	425.8	102.2	24.7	1364
$-1 < M_i \leq 0$	2027.1	1064.0	255.3	61.7	3408
$0 < M_i \leq 1$	2036.0	1068.7	256.4	62.0	3423
$1 < M_i \leq 2$	789.9	414.6	99.5	24.0	1328
$2 < M_i < 6$	143.3	75.2	18.1	4.4	241
Column Total	5948	3122	749	181	10000

$\frac{(O_{u,v}-\hat{E}_{u,v})^2}{\hat{E}_{u,v}}$	$0 \leq V_i < 1$	$1 \leq V_i < 2$	$2 \leq V_i < 3$	$3 \leq V_i \leq 6$	Row Total
$-4 < M_i \leq -2$	0.08	0.54	0.41	0.02	1.05
$-2 < M_i \leq -1$	0.01	0.24	0.10	0.89	1.24
$-1 < M_i \leq 0$	0.16	0.03	0.97	0.22	1.39
$0 < M_i \leq 1$	0.08	0.36	0.02	1.32	1.79
$1 < M_i \leq 2$	0.19	0.14	0.12	0.04	0.49
$2 < M_i < 6$	0.01	0.30	0.91	0.09	1.32
Column Total	0.53	1.62	2.53	2.59	7.26

Table 1: Chi-Square Test of Independence

The simulated data described in the caption to Figure 1 are dropped into a two-way contingency table. There are $s = 6$ rows, $t = 4$ columns, and $(s-1)\cdot(t-1) = 15$ degrees of freedom in the test statistic calculated (keeping the row and column sums, strike out any row and any column in the body of the first panel, and the remaining 15 cells allow you to deduce what you struck out). I find $\sum_{u=1}^{s}\sum_{v=1}^{t}\frac{(O_{u,v}-\hat{E}_{u,v})^2}{\hat{E}_{u,v}} = 7.26$, with a p-value of 0.95. Note, for example, that 73.7 in bold font in the (1,2) position of the second panel is calculated using the first row and second column totals (taken from the first panel, and also in bold font) as

$$73.7 = \frac{3122}{10000}\cdot\frac{236}{10000}\cdot 10000.$$

We cannot reject the null hypothesis that the two attributes are independent. Note that a couple of cells have $\hat{E}_{u,v} < 5$, which is, technically, a violation of the assumptions required to run the test. In this case, however, the violation is minor.

That is, the *multinomial theorem* holds:

$$(p_1 + \cdots + p_r)^n = \sum_{\substack{\nu_1,\ldots,\nu_r \\ s.t.\ \nu_1+\cdots+\nu_r=n}} \binom{n}{\nu_1,\nu_2,\ldots,\nu_r} p_1^{\nu_1}\cdots p_r^{\nu_r}. \tag{27}$$

For example, in the simple case with $r = 2$ possible outcomes and n trials, the *binomial theorem* holds:

$$(p_1 + p_2)^n = \sum_{\substack{\nu_1,\nu_2 \\ s.t.\ \nu_1+\nu_2=n}} \binom{n}{\nu_1,\nu_2} p_1^{\nu_1} p_2^{\nu_2} = \sum_{\nu_1=0}^{n} \binom{n}{\nu_1} p_1^{\nu_1}(1-p_1)^{n-\nu_1}, \tag{28}$$

because $\binom{n}{\nu_1,\nu_2} = \binom{n}{\nu_1} = \frac{n}{\nu_1!(n-\nu_1)!}$ and $\nu_2 = n - \nu_1$. We can see that the general term in Equation 28 is just the binomial probability $P(\nu_1; n, p_1) = \binom{n}{\nu_1} p_1^{\nu_1}(1-p_1)^{n-\nu_1}$.

For example, suppose you toss a fair coin (with $r = 2$ outcomes) $n = 3$ times. Then the sample space is $\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$, which is $\binom{3}{3,0} = \binom{3}{3} = 1$ way of getting three heads, $\binom{3}{2,1} = \binom{3}{2} = 3$ ways of getting two heads, $\binom{3}{1,2} = \binom{3}{1} = 3$ ways of getting one head, and $\binom{3}{0,3} = \binom{3}{0} = 1$ ways of getting no heads.

More generally, $\binom{n}{\nu_1,\nu_2,\ldots,\nu_r}$ counts the number of *permutations* of ways in which you can get ν_1 occurrences of outcome 1, ν_2 occurrences of outcome 2, ..., ν_r occurrences of outcome r, each occurring with probability $p_1^{\nu_1} p_2^{\nu_2} \cdots p_r^{\nu_r}$.

How do we confirm the asymptotic distribution in Equation 14? Unlike the direct manipulation and decomposition of terms in Crack et al. (2018) in the case of the exact result for the sample variance, we need to rely upon an asymptotic approximation argument (using *characteristic functions* here, and a central limit theorem in Section 6).

A characteristic function is a unique function associated with a probability distribution. The characteristic function of random variable X is given by $\phi(t) = E[e^{itX}]$, where $i = \sqrt{-1}$ and t is an auxiliary variable. If the series is expanded, the coefficient of $(it)^k/k!$ is given by $E(X^k)$ (Evans et al., 1993). So, expansions of characteristic functions can be used to generate moments. Note that X is integrated out of the characteristic function. So, $\phi(\cdot)$ is not a function of X at all, but only of the auxiliary variable t.

If we can show that our multinomial sample, appropriately scaled, has, asymptotically, the characteristic function of a multivariate normal distribution with $r - 1$ standardized variates, then squaring and adding these up will yield a chi-squared distribution with $r - 1$ degrees of freedom. In what follows, we will look at the characteristic functions of Bernoulli, binomial, multinomial, and multivariate singular normal distributions.

Suppose that X_j is distributed Bernoulli with parameter p. Then X_j has characteristic function

$$\begin{aligned}\phi_{Bernoulli}(t) = E\left[e^{itX_j}\right] &= e^{it1} \cdot p + e^{it0} \cdot (1-p) \\ &= \left[pe^{it} + (1-p)\right]. \qquad (29)\end{aligned}$$

The characteristic function of the sum of IID random variables is easily seen to be the product of the characteristic functions of the underlying IID variables.[5] For example, suppose that $X \sim Bin(n,p)$ is given by $X = X_1 + \cdots + X_n$, where the X_j are distributed IID Bernoulli for $j = 1, \ldots, n$. Then

$$\begin{aligned}\phi_{binomial}(t) &= E\left[e^{itX}\right] \qquad (30) \\ &= E\left[e^{it\sum_{j=1}^{n} X_j}\right] \\ &\overset{*}{=} E\left[\prod_{j=1}^{n} e^{itX_j}\right] \\ &\overset{**}{=} \prod_{j=1}^{n} E\left[e^{itX_j}\right] \\ &= \left[pe^{it} + (1-p)\right]^n, \qquad (31)\end{aligned}$$

where step-($*$) follows from the properties of the exponential function, step-($**$) follows from independence of the X_j, and the final step follows from Equation 29.

[5] The converse is not necessarily true. That is, if $\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$, then X and Y are not necessarily independent. Hamedani and Volkmer (2009) present a counter example involving "subindependence." Another counter example is if $X = Y$ is Cauchy (see https://en.wikipedia.org/wiki/Subindependence).

Alternatively, let j range over all possible outcomes for $X \sim Bin(n,p)$, then

$$\begin{aligned}
\phi_{binomial}(t) &= E\left[e^{itX}\right] && (32)\\
&= \sum_{j=0}^{n} e^{itj} \cdot P(X=j)\\
&= \sum_{j=0}^{n} e^{itj} \binom{n}{j} p^j (1-p)^{n-j}\\
&= \sum_{j=0}^{n} \binom{n}{j} \left(pe^{it}\right)^j (1-p)^{n-j}\\
&= \left[pe^{it} + (1-p)\right]^n, && (33)
\end{aligned}$$

where the last step follows from the binomial theorem in Equation 28.

Now let us consider the multinomial case, as a generalization of the derivation leading to the binomial case in Equation 33. Note that the auxiliary variable is now an $(r \times 1)$ vector $\vec{t} = [t_1, \ldots, t_r]'$, and we obtain the joint characteristic function:

$$\begin{aligned}
\phi_{multinomial}(\vec{t}\,) &= E\left[e^{i\vec{t}\,'\vec{\nu}}\right]\\
&= E\left[e^{i(t_1\nu_1 + t_2\nu_2 + \cdots + t_r\nu_r)}\right]\\
&= E\left[e^{it_1\nu_1} e^{it_2\nu_2} \cdots e^{it_r\nu_r}\right]\\
&= \sum_{\substack{\nu_1,\nu_2,\ldots,\nu_r\\ s.t.\ \nu_1+\nu_2+\cdots+\nu_r=n}} P(\vec{\nu}; n, p_1, p_2 \ldots, p_r) e^{it_1\nu_1} e^{it_2\nu_2} \cdots e^{it_r\nu_r}\\
&= \sum_{\substack{\nu_1,\nu_2,\ldots,\nu_r\\ s.t.\ \nu_1+\nu_2+\cdots+\nu_r=n}} \left[\binom{n}{\nu_1, \nu_2, \ldots, \nu_r} p_1^{\nu_1} p_2^{\nu_2} \cdots p_r^{\nu_r}\right] e^{it_1\nu_1} e^{it_2\nu_2} \cdots e^{it_r\nu_r}\\
&= \sum_{\substack{\nu_1,\nu_2,\ldots,\nu_r\\ s.t.\ \nu_1+\nu_2+\cdots+\nu_r=n}} \binom{n}{\nu_1, \nu_2, \ldots, \nu_r} (p_1 e^{it_1})^{\nu_1} (p_2 e^{it_2})^{\nu_2} \cdots (p_r e^{it_r})^{\nu_r}\\
&= \left(p_1 e^{it_1} + p_2 e^{it_2} + \cdots + p_r e^{it_r}\right)^n, && (34)
\end{aligned}$$

by the multinomial theorem in Equation 27.

Now let

$$X_j = \frac{\nu_j - np_j}{\sqrt{np_j}}, \qquad (35)$$

for $j = 1, \ldots, r$. Then, from Equation 12, we see that

$$C = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} = \sum_{j=1}^{r} X_j^2. \tag{36}$$

Note that Equation 35 is not a standardizing transformation. That is, although $E(X_j) = 0$, $V(X_j) = (1 - p_j) \neq 1$, which was pointed out already in Equation 16.

Let us now derive the joint characteristic function of $\vec{X} = [X_1, \ldots, X_r]'$. Note that $\vec{X}$ is not distributed multinomial, but rather it is a transformed version of a multinomial random variable. Drawing upon our previous analysis, we have that

$$\begin{aligned}
\phi_{\vec{X}}(\vec{t}) &= E\left[e^{i\vec{t}'\vec{X}}\right] \\
&= E\left[e^{i(t_1X_1 + \cdots + t_rX_r)}\right] \\
&= E\left[e^{i\sum_{j=1}^{r} t_j\left(\frac{\nu_j - np_j}{\sqrt{np_j}}\right)}\right] \\
&= \left(e^{-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}}\right) E\left[e^{i\sum_{j=1}^{r}\frac{t_j\nu_j}{\sqrt{np_j}}}\right] \\
&= \left(e^{-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}}\right) E\left[\prod_{j=1}^{r} e^{i\frac{t_j\nu_j}{\sqrt{np_j}}}\right] \\
&\overset{*}{=} \left(e^{-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}}\right) \sum_{\substack{\nu_1,\ldots,\nu_r \\ s.t.\ \nu_1+\cdots+\nu_r=n}} \left[P(\vec{\nu}; n, p_1, \ldots, p_r)\prod_{j=1}^{r} e^{i\frac{t_j\nu_j}{\sqrt{np_j}}}\right] \\
&= \left(e^{-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}}\right) \sum_{\substack{\nu_1,\ldots,\nu_r \\ s.t.\ \nu_1+\cdots+\nu_r=n}} \left[\binom{n}{\nu_1, \ldots, \nu_r} p_1^{\nu_1}\cdots p_r^{\nu_r}\prod_{j=1}^{r}\left(e^{i\frac{t_j}{\sqrt{np_j}}}\right)^{\nu_j}\right] \\
&= \left(e^{-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}}\right) \sum_{\substack{\nu_1,\ldots,\nu_r \\ s.t.\ \nu_1+\cdots+\nu_r=n}} \left[\binom{n}{\nu_1, \ldots, \nu_r}\prod_{j=1}^{r}\left(p_je^{i\frac{t_j}{\sqrt{np_j}}}\right)^{\nu_j}\right] \\
&= \left(e^{-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}}\right)\left(p_1e^{i\frac{t_1}{\sqrt{np_1}}} + \cdots + p_re^{i\frac{t_r}{\sqrt{np_r}}}\right)^n, \qquad (37)
\end{aligned}$$

by the multinomial theorem. Note that the expected value of the product does not resolve to the product of the expected values at step-$(*)$ because the ν_j are dependent, not independent, as shown in Equation 9.

Now, to get at the heart of the matter, let us take natural logarithms of Equation 37. After this, we will expand the result using a vector MacLaurin expansion involving a score vector and a hessian matrix. We get that

$$\begin{aligned}\ln\left[\phi_{\vec{X}}(\vec{t}\,)\right] &= \left(-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}\right) + n\cdot ln\left(p_1 e^{i\frac{t_1}{\sqrt{np_1}}} + \cdots + p_r e^{i\frac{t_r}{\sqrt{np_r}}}\right)\\ \ln\left[\phi_{\vec{X}}(\vec{t}\,)\right] &= \left(-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}\right) + n\cdot ln\left(\mathcal{P}(\vec{t}\,)\right), \qquad (38)\end{aligned}$$

where $\mathcal{P}(\vec{t}\,) = \left(p_1 e^{i\frac{t_1}{\sqrt{np_1}}} + \cdots + p_r e^{i\frac{t_r}{\sqrt{np_r}}}\right)$.

Let $\vec{S}(\vec{t}\,) = \frac{\partial\mathcal{P}(\vec{t}\,)}{\partial\vec{t}}$ be the score vector, and let $H(\vec{t}\,) = \frac{\partial^2\mathcal{P}(\vec{t}\,)}{\partial\vec{t}'\partial\vec{t}}$ be the Hessian matrix. Direct evaluation yields

$$\begin{aligned}\left[\vec{S}(\vec{t}\,)\right]_k &= \frac{\partial\mathcal{P}(\vec{t}\,)}{\partial t_k} = i\sqrt{\frac{p_k}{n}}e^{i\frac{t_k}{\sqrt{np_k}}}\\ \left[H(\vec{t}\,)\right]_{kk} &= \frac{\partial^2\mathcal{P}(\vec{t}\,)}{\partial t_k^2} = \frac{i^2}{n}e^{i\frac{t_k}{\sqrt{np_k}}} = -\frac{1}{n}e^{i\frac{t_k}{\sqrt{np_k}}}\\ \left[H(\vec{t}\,)\right]_{kl} &= 0 \text{ for } k\neq l\end{aligned}$$

So, it follows that

$$\mathcal{P}(\vec{0}\,) = 1,\ \vec{S}(\vec{0}) = \frac{i}{\sqrt{n}}\begin{pmatrix}\sqrt{p_1}\\ \sqrt{p_2}\\ \vdots\\ \sqrt{p_r}\end{pmatrix},\ \text{and } H(\vec{0}) = -\frac{1}{n}\begin{pmatrix}1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1\end{pmatrix} = -\frac{1}{n}I_r, \qquad (39)$$

where I_r is the $(r\times r)$ identity matrix.

Then, plugging the above into a vector MacLaurin expansion yields

$$\begin{aligned}\mathcal{P}(\vec{t}\,) &= \mathcal{P}(\vec{0}\,) + \vec{S}(\vec{0})'\vec{t} + \frac{1}{2}\vec{t}\,'H(\vec{0})\vec{t} + \text{ higher-order terms} \\ &= 1 + \frac{i}{\sqrt{n}}\sum_{j=1}^{r} t_j\sqrt{p_j} - \frac{1}{2n}\sum_{j=1}^{r} t_j^2 + O(n^{-3/2}). \qquad (40)\end{aligned}$$

Recall the Taylor series expansion of the natural logarithm, $ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4}\cdots$, for $|x| < 1$. Plugging Equation 40 into the Taylor series expansion of the logarithm yields

$$ln\left[\mathcal{P}(\vec{t}\,)\right] = \frac{i}{\sqrt{n}}\sum_{j=1}^{r} t_j\sqrt{p_j} - \frac{1}{2n}\sum_{j=1}^{r} t_j^2 - \frac{1}{2}\left(\frac{i}{\sqrt{n}}\sum_{j=1}^{r} t_j\sqrt{p_j}\right)^2 + O(n^{-1/2}). \qquad (41)$$

Now we plug Equation 41 back into Equation 38, note the cancellation of the $\frac{i}{\sqrt{n}}\sum_{j=1}^{r} t_j\sqrt{p_j}$ term, the pre-multiplication by n, and that $i^2 = -1$, to obtain

$$\ln\left[\phi_{\vec{X}}(\vec{t}\,)\right] = -\frac{1}{2}\sum_{j=1}^{r} t_j^2 + \frac{1}{2}\left(\sum_{j=1}^{r} t_j\sqrt{p_j}\right)^2 + O(n^{-1/2}). \qquad (42)$$

Thus, we get that

$$lim_{n\to\infty}\phi_{\vec{X}}(\vec{t}\,) = e^{-\frac{1}{2}\left[\sum_{j=1}^{r} t_j^2 - \left(\sum_{j=1}^{r} t_j\sqrt{p_j}\right)^2\right]} = e^{-\frac{1}{2}Q(\vec{t}\,)}, \qquad (43)$$

where

$$Q(\vec{t}) = \left[\sum_{j=1}^{r} t_j^2 - \left(\sum_{j=1}^{r} t_j\sqrt{p_j}\right)^2\right] = \vec{t}\,'\Lambda\vec{t} \qquad (44)$$

is the quadratic form built using moment matrix $\Lambda = I_r - \vec{\theta}\vec{\theta}'$, for $\vec{\theta} = [\sqrt{p_1}, \ldots, \sqrt{p_r}]'$.

Now let $\vec{u} = \Theta\vec{t}$, for $(r \times 1)$ vector $\vec{u}$, and $(r \times r)$ orthogonal matrix Θ, but with the special property of Θ that the last element of $\vec{u}$ satisfies $u_r = \sum_{j=1}^{r} t_j\sqrt{p_j}$. A "spectral theorem" is a theorem that shows when a linear operator, or matrix, can be diagonalized via some transformation. In our case, consider $\Theta\Lambda\Theta' = \Theta(I_r - \vec{\theta}\vec{\theta}')\Theta'$, for $\vec{\theta} = [\sqrt{p_1}, \ldots, \sqrt{p_r}]'$.

It might not be immediately obvious, but Λ is idempotent (that is, $\Lambda \cdot \Lambda = \Lambda$; Note that the only idempotent matrix that is invertible is the identity matrix.) This means that Λ's eigenvalues are either 0 or 1. The number of eigenvalues of an idempotent matrix that are equal to 1 is equal to the trace of the matrix (i.e., the sum of the diagonal entries), and in this case, $rank(\Lambda) = trace(\Lambda) = r - 1$ (Rao, 1973, p. 28). For our Λ, the matrix Θ just mentioned diagonalizes Λ, and $\Theta\Lambda\Theta'$ has only ones and zeros on the diagonal (corresponding to its eigenvalues).[6] The transformation $\vec{u} = \Theta\vec{t}$ yields $\vec{t} = \Theta'\vec{u}$, and thus

$$\begin{aligned}
\vec{t}'\Lambda\vec{t} &= \vec{u}'(\Theta\Lambda\Theta')\vec{u} \\
&= \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{r-1} \\ u_r \end{pmatrix}' \cdot \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{r-1} \\ u_r \end{pmatrix} \\
&= \left(\sum_{j=1}^{r-1} 1 \cdot u_j^2 \right) + \left(0 \cdot u_r^2 \right) = \sum_{j=1}^{r-1} u_j^2. \qquad (45)
\end{aligned}$$

In other words,

$$Q(\vec{t}) = \left[\sum_{j=1}^{r} t_j^2 - \left(\sum_{j=1}^{r} t_j \sqrt{p_j} \right)^2 \right] = \sum_{j=1}^{r-1} u_j^2. \qquad (46)$$

It follows that for large n, the the joint characteristic function of $\vec{X}$ converges to $e^{-\frac{1}{2}Q(\vec{t})}$, where the quadratic form $Q(\vec{t})$ is non-negative and of rank $r-1$. The limiting function $e^{-\frac{1}{2}Q(\vec{t})}$ is the characteristic function of a singular normal distribution (Cramér, 1946, Section 24.3).[7] What this means is that $\sum_{j=1}^{r} X_j^2$ has, asymptotically, the same distribution as $\sum_{j=1}^{r} \eta_j^2$, where $\eta_1, \dots, \eta_{r-1}$ are IID N(0,1), and $\eta_r \sim N(0,0)$ (i.e., with probability 1, η_r is equal to zero) (Cramér, 1946, Section 24.3). It then follows that $C = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} = \sum_{j=1}^{r} X_j^2 \overset{A}{\sim} \sum_{j=1}^{r} \eta_j^2 = \sum_{j=1}^{r-1} \eta_j^2 \sim \chi_{r-1}^2$ (Cramér, 1946, p. 314).

Note that $\vec{X}$ is subject to the constraint $\vec{X}'\vec{\theta} = 0$ (i.e., the probability mass of $\vec{X}$ lies in th $(r-1)$-dimensional hyperplane $\vec{X}'\vec{\theta} = 0$ in $\mathbb{R}^r$). So, it should not be surprising that the end result has $r-1$ degrees of freedom, the same as the dimension of the hyperplane that

[6] For any square symmetric matrix M, there exists an orthogonal matrix Θ such that $\Theta M\Theta'$ is diagonal, with the diagonal composed only of the eigenvalues of M. Also, $trace(\Theta M\Theta') = trace(\Theta M\Theta') = trace(\Theta'\Theta M) = trace(IM) = trace(M)$ (Ayers, 1962, p. 163; Rao, 1973, p. 39).

[7] More generally, $\phi(\vec{t}) = e^{-\frac{1}{2}Q(\vec{t})}$, where $Q(\vec{t}) = \vec{t}'C\vec{t}$ is of full rank, is the characteristic function of the r-variate normal distribution with zero mean vector and variance-covariance matrix C (Feller, 1971, p. 522).

the probability mass sits in.

How do we build Θ in practice? There is no unique answer, but begin by noting that Λ annihilates the vector $\vec{\theta} = [\sqrt{p_1}, \ldots, \sqrt{p_r}]'$:

$$\begin{aligned} \Lambda\vec{\theta} &= (I_r - \vec{\theta}\vec{\theta}')\vec{\theta} \\ &= \vec{\theta} - \vec{\theta}(\vec{\theta}'\vec{\theta}) \\ &= \vec{\theta} - \vec{\theta} \cdot (1) \\ &= \vec{0}. \end{aligned} \tag{47}$$

So, given that $\vec{\theta}$ is non-trivial, Λ cannot be invertible (else run Λ^{-1} through Equation 47 to deduce $\vec{\theta} = \vec{0}$, a contradiction). (Of course, we already knew that Λ was non-invertible because one of its eigenvalues is zero.) Equation 47 also implies that $\vec{\theta}$ is an eigenvector of Λ corresponding to eigenvalue 0. If we take all r eigenvectors of Λ, and arrange them as rows of the matrix Θ, with θ' as the last row, then Θ is an orthogonal matrix that meets our needs.

6 Asymptotic Distribution II: Via Multivariate CLT

The derivation in this section is half as long as the derivation in Section 5, but this one builds upon much of the work appearing already in several of the previous sections.

We use a multivariate central limit theorem (CLT). It exploits the independence across trials (i.e., for $i = 1, \ldots, n$) while also recognizing the dependence in the outcomes (i.e., for $j = 1, \ldots, r$) in the functional form of the off-diagonal terms in the variance-covariance matrix.

Let us give two helpful results. First, Rao (1973, p. 128) presents a multivariate version of the Lindeberg-Lévy CLT. Suppose that $\vec{U}_i = [U_{i1}, U_{i2}, \ldots, U_{i,r}]'$ has finite mean $E(\vec{U}_i) = \vec{\mu}$ and finite variance $V(\vec{U}_i) = \Sigma$, and the sequence $\vec{U}_i$ for $i = 1, 2, \ldots$ is IID, then

$$\sqrt{n}\left(\bar{U}_n - \vec{\mu}\right) \xrightarrow{d} N_r(0, \Sigma), \tag{48}$$

where $\bar{U}_n = \frac{1}{n}\sum_{i=1}^{n}\vec{U}_i$ is understood to be a vector, even though it has no vec symbol above it, and $N_r(0, \Sigma)$ denotes an r-variate normal distribution with mean zero vector and invertible variance-covariance matrix Σ, having r-variate normal density function

$$N_r(\vec{u}-\vec{\mu}|\vec{0},\Sigma) = (2\pi)^{-r/2}\,|\Sigma|^{-1/2}\,e^{-\frac{1}{2}(\vec{u}-\vec{\mu})'\Sigma^{-1}(\vec{u}-\vec{\mu})}, \tag{49}$$

where $|\cdot|$ denotes the determinant.

Second, suppose that Σ is a symmetric matrix. Then, as mentioned already, there exists orthogonal Θ such that $\Sigma = \Theta D\Theta'$, where D is diagonal. If we let $D^{1/2}$ denote the diagonal matrix whose diagonal is composed of the square root of the diagonal elements of D, in the same order, then $\Sigma = \Theta(D^{1/2}D^{1/2})\Theta' = (\Theta D^{1/2}\Theta')(\Theta D^{1/2}\Theta')$, because Θ is orthogonal. If we let $\Sigma^{1/2} = \Theta D^{1/2}\Theta'$, then $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$, and we may call $\Sigma^{1/2}$ the *square root* of Σ. In the case where Σ is invertible, $\Sigma^{-1/2}$ is just $(\Sigma^{-1})^{1/2}$.

Recall from Equation 3 that $\vec{X}_i$ is an $(r \times 1)$ vector, distributed multivariate Bernoulli. That is, $\vec{X}_i$ follows the simplest multinomial distribution, with $n = 1$ and with probability vector $\vec{p}$.

Then, following our earlier analysis, we know that

$$E(\vec{X}_i) = \vec{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_r \end{pmatrix}, \text{ and } V(\vec{X}_i) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_r \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_r \\ \vdots & & \ddots & \vdots \\ -p_rp_1 & -p_rp_2 & \cdots & p_r(1-p_r) \end{pmatrix} = \Sigma, \tag{50}$$

say, where

$$\begin{aligned} V(\vec{X}_i) = \Sigma &= \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_r \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_r \\ \vdots & & \ddots & \vdots \\ -p_rp_1 & -p_rp_2 & \cdots & p_r(1-p_r) \end{pmatrix} \\ &= \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & p_r \end{pmatrix} - \vec{p}\vec{p}' \\ &= diag(\vec{p}) - \vec{p}\vec{p}', \end{aligned} \tag{51}$$

where $diag(\cdot)$ is understood to diagonalize its argument.[8]

If we look closely at Σ, we see that the sum of any row is zero. Similarly, the sum of any column is zero. This means that the rows, and the columns, are linearly dependent, and thus $|\Sigma| = 0$, and Σ is non-invertible. With $|\Sigma| = 0$ and Σ^{-1} undefined, the density function in Equation 49 is not defined. So, we cannot apply the CLT in Equation 48 to $\vec{X}_i$.

The dependence in Σ is a signal that we can drop one row in each of $\vec{X}_i$ and $\vec{p}$, and strike out one row and one column in Σ, while still retaining all the information we need. Given the symmetry, it does not matter which row we drop; let us choose to drop the r^{th} row, as follows.

Let $\vec{X}_i^\dagger = [(\vec{X}_i)_1, (\vec{X}_i)_2, \ldots, (\vec{X}_i)_{r-1}]'$, let $\vec{p}^{\,\dagger} = [p_1, p_2, \ldots, p_{r-1}]'$, and let $\Sigma^\dagger$ denote the $((r-1) \times (r-1))$ upper left corner of Σ (i.e., Σ with the r^{th} row and r^{th} column struck out). Then

$$V(\vec{X}_i^\dagger) = \Sigma^\dagger = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{r-1} \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_{r-1} \\ \vdots & & \ddots & \vdots \\ -p_{r-1}p_1 & -p_{r-1}p_2 & \cdots & p_{r-1}(1-p_{r-1}) \end{pmatrix} = diag(\vec{p}^{\,\dagger}) - \vec{p}^{\,\dagger}(\vec{p}^{\,\dagger})' \quad (52)$$

Having removed one element from each of $\vec{X}_i$ and $\vec{p}$, and having struck out one row and one column in Σ, we now have that $\Sigma^\dagger$ is invertible. After some algebra, it can be shown that

$$\left[V(\vec{X}_i^\dagger)\right]^{-1} = \left(\Sigma^\dagger\right)^{-1} = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_r} & \frac{1}{p_r} & \cdots & \frac{1}{p_r} \\ \frac{1}{p_r} & \frac{1}{p_2} + \frac{1}{p_r} & \cdots & \frac{1}{p_r} \\ \vdots & & \ddots & \vdots \\ \frac{1}{p_r} & \frac{1}{p_r} & \cdots & \frac{1}{p_{r-1}} + \frac{1}{p_r} \end{pmatrix} = diag(1./\vec{p}^{\,\dagger}) + \frac{1}{p_r}\vec{\iota}_{r-1}\vec{\iota}_{r-1}{}', \quad (53)$$

where $1./\vec{p}^{\,\dagger}$ denotes the vector built from the reciprocal of each element of $\vec{p}^{\,\dagger}$, and $\vec{\iota}_{r-1}$ is the $((r-1) \times 1)$ vector of ones. (See Footnote 8 regarding the function $diag(\cdot)$.)

Exercise: *Prove that $\Sigma^\dagger(\Sigma^\dagger)^{-1} = I_{r-1}$, the $(r-1) \times (r-1)$ identity matrix. It took me only one page of handwritten algebra. Hint: note that $\sum_{j=1}^{r-1} p_j = 1 - p_r$.*

Now let $\bar{X}_n^\dagger = \frac{1}{n}\sum_{i=1}^n \vec{X}_i^\dagger$ be the $((r-1) \times 1)$ vector average of the first n of the $\vec{X}_i^\dagger$

[8] Note that if $\vec{e}^{\,(j)}$ is the j^{th} $(r \times 1)$ basis vector in $\mathbb{R}^r$, then $diag(\vec{p}) = \sum_{j=1}^r \vec{e}^{\,(j)}\vec{p}\,'\vec{e}^{\,(j)}\left(\vec{e}^{\,(j)}\right)'$.

variables, then, using our multivariate CLT in Equation 48, we get that

$$\sqrt{n}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right) \xrightarrow{d} N_{r-1}(0, \Sigma^\dagger). \tag{54}$$

If we now scale Equation 54 by $\left(\Sigma^\dagger\right)^{-1/2}$, as defined just after Equation 49, we get that

$$\sqrt{n}\left(\Sigma^\dagger\right)^{-1/2}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right) \xrightarrow{d} N_{r-1}(0, I_{r-1}). \tag{55}$$

Given that the RHS of Equation 55 is multivariate IID standard normal, let me write

$$Z_n^\dagger = \sqrt{n}\left(\Sigma^\dagger\right)^{-1/2}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right). \tag{56}$$

Then Equation 55 says that $Z_n^\dagger$ is asymptotically multivariate IID standard normal:

$$Z_n^\dagger \xrightarrow{d} N_{r-1}(0, I_{r-1}). \tag{57}$$

It follows, by the definition of the χ^2 distribution, that

$$\left(Z_n^\dagger\right)' Z_n^\dagger \xrightarrow{d} \chi^2_{r-1}. \tag{58}$$

Let me now sketch out a proof that $\left(Z_n^\dagger\right)' Z_n^\dagger$ is, in fact, Pearson's chi-square statistic. I have omitted a few lines in the middle of the proof.

$$\begin{aligned}
\left(Z_n^\dagger\right)' Z_n^\dagger &= \left[\sqrt{n}\left(\Sigma^\dagger\right)^{-1/2}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right)\right]' \sqrt{n}\left(\Sigma^\dagger\right)^{-1/2}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right) \\
&= n\left[\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right)'\left(\Sigma^\dagger\right)^{-1/2}\right]\left(\Sigma^\dagger\right)^{-1/2}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right) \\
&= n\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right)'\left(\Sigma^\dagger\right)^{-1}\left(\bar{X}_n^\dagger - \vec{p}^{\,\dagger}\right) \\
&= n\left[\frac{1}{n}\begin{pmatrix}\nu_1\\ \vdots \\ \nu_{r-1}\end{pmatrix} - \begin{pmatrix}p_1\\ \vdots \\ p_{r-1}\end{pmatrix}\right]'\left(\Sigma^\dagger\right)^{-1}\left[\frac{1}{n}\begin{pmatrix}\nu_1\\ \vdots \\ \nu_{r-1}\end{pmatrix} - \begin{pmatrix}p_1\\ \vdots \\ p_{r-1}\end{pmatrix}\right] \\
&= n\left[\frac{1}{n}\begin{pmatrix}\nu_1 - np_1\\ \vdots \\ \nu_{r-1} - np_{r-1}\end{pmatrix}\right]'\left(diag(1./\vec{p}^{\,\dagger}) + \frac{1}{p_r}\vec{\iota}_{r-1}\vec{\iota}_{r-1}{}'\right)\left[\frac{1}{n}\begin{pmatrix}\nu_1 - np_1\\ \vdots \\ \nu_{r-1} - np_{r-1}\end{pmatrix}\right] \\
&\vdots \\
&= \frac{1}{n}\left\{\sum_{j=1}^{r-1}\frac{(\nu_j - np_j)^2}{p_j} + \frac{1}{p_r}\left[\sum_{j=1}^{r-1}(\nu_j - np_j)\right]^2\right\} \\
&= \sum_{j=1}^{r}\frac{(\nu_j - np_j)^2}{np_j} = C,
\end{aligned} \tag{59}$$

where, at the last step, I used the fact that $\sum_{j=1}^{r-1}(\nu_j - np_j) = -(\nu_r - np_r)$, from which it follows that $\frac{1}{p_r}\left[\sum_{j=1}^{r-1}(\nu_j - np_j)\right]^2 = \frac{(\nu_r - np_r)^2}{p_r}$. Thus, the desired result follows immediately:

$$C = \sum_{j=1}^{r}\frac{(\nu_j - np_j)^2}{np_j} \xrightarrow{d} \chi^2_{r-1}, \tag{60}$$

completing the proof.

7 References

Ayers, Frank Jr., 1962, *Matrices*, Schaum's Outline Series; McGraw-Hill: New York, NY.

Cochrane, W.G., 1934, "The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance," *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 20 No. 2, (April), pp. 178–191.

Crack, Timothy Falcon, Michael J. Osborne, Malcolm A. Crack, Mark J. Osborne, 2008, "A Constructive Demonstration of the Distribution of the Student-t Test Statistic for the Mean," Working Paper, 22pp.

Cramér, Harald, 1946, *Mathematical Methods of Statistics*, Princeton University Press: Princeton, NJ.

DeGroot, Morris H., 1989, *Probability and Statistics*, Addison-Wesley: Reading, MA.

Evans, Merran, Nicholas Hastings, and Brian Peacock, 1993, *Statistical Distributions*, Second Edition, John Wiley and Sons: New York, NY.

Feller, William, 1971, *An Introduction to Probability Theory and its Applications*, Volume II, Second Edition, John Wiley and Sons: New York, NY.

Fisher, Ronald A., 1915, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, Vol. 10 No. 4, (May), pp. 507–521.

Fisher, Ronald A., 1922, "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, Vol. 85 No. 1, (January), pp. 87–94.

Fisher, Ronald A., 1925, "Applications of "Student's" Distribution," *Metron*, Vol. 5 No. 3, (December), pp. 90–104.

Geary, R. C., 1936, "The Distribution of "Student's" Ratio for Non-Normal Samples," *Supplement to the Journal of the Royal Statistical Society*, Vol. 3 No. 2, (no month), pp. 178–184.

Hamedani, G.G. and H.W. Volkmer, 2009, "Letter to the Editor (regarding De Paula, A. (2008) "Conditional Moments and Independence," The American Statistician, 62, 219221)," *The American Statistician*, Vol. 63 No. 3, (August), p. 295.

Johnson, Norman L., Samuel Kotz, and Adrienne W. Kemp, 1993, *Univariate Discrete Distributions*, Second Edition, John Wiley and Sons: New York, NY.

Pearson, Karl, 1900, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philosophical Magazine Series 5*, Vol. 50 No. 302, pp. 157–175.

Pearson, Karl, 1904, *On the Theory Of Contingency and Its Relation To Association and Normal Correlation*, appearing in Drapers’ Company Research Memoirs, Biometric Series I, Mathematical Contributions to the Theory of Evolution, part XIII, Dulau and Co., London.

Plackett, R. L., 1983, “Karl Pearson and the Chi-Squared Test,” *International Statistical Review*, Vol. 51 No. 1, (April), pp. 59–72.

Rao, Calyampudi Radhakrishna, 1973, *Linear Statistical Inference and its Applications*, John Wiley & Sons, Inc., NY.

Student (Gosset, William Sealy), 1908, “The Probable Error of a Mean,” *Biometrika*, Vol. 6 No. 1 (March), pp. 1–25.

Zwillinger, Daniel, and Stephen Kokoska, 2000, *Standard Probability and Statistics Tables and Formulae*, Chapman & Hall/CRC, New York, NY.