INDEPENDENCE & PROBABILITY

A. Before beginning this section, it is important that a common misunderstanding is prevented. If this section is not carefully followed, it is easy to mistakenly conclude that \( \Pr( A \& B ) = \Pr( A ) \times \Pr( B ) \) (i.e., that the probability that the events, A and B, jointly occur equals the product of the probability that event A occurs times the probability that event B occurs). In fact, for any given sample estimates of these probabilities are unlikely to exactly yield this equivalence. So, for the record, please keep the following in mind:

1. It is always true that \( \Pr( A \& B ) = \Pr( A \mid B ) \times \Pr( B ) \) (i.e., that the probability that the events, A and B, jointly occur equals the product of the probability that event A occurs given that the event B has occurred times the probability that event B occurs).

2. In the special case when A and B are independent events, however, \( \Pr( A \& B ) = \Pr( A ) \times \Pr( B ) \).

B. That said, let us now begin with an imaginary population. We draw two units (call them Joe and Mary) from the population at random and evaluate them according to some characteristic (e.g., age). (Note that we are assigning the names, Joe and Mary, irrespective of the units' actual names or genders.)

1. At this point we might ask ourselves, "What is the probability that Joe is older than the mean age for the population?" On the one hand, the age distribution might be skewed to the right (i.e., there may be many more younger people than older people). If couples are having fewer children, the distribution might be more symmetric (i.e., fewer babies
may be "balanced" by the number of older people from large birth cohorts who have died off). On the other hand, if the population is of people who abstain from sex altogether (e.g., a religious community like that of the Shakers), the age distribution might be skewed to the left. In the first case, the probability would be less than 0.5 (because the population's mean age, \( \mu_A \), would be larger than the population's median). In the last case, it would be greater than 0.5 (because it would be smaller than the population's median). Because the median is the value of a variable such that 0.5 of the distribution has higher values (and 0.5 has lower values) on the variable, it is only in the second case (when the distribution is symmetric and mean=median) that we may conclude the following:

\[
Pr( J ) = 0.5 , \text{ where } J \text{ is the event that Joe's age } > \mu_A .
\]

a. The probability that a single event occurs is sometimes referred to as the MARGINAL PROBABILITY of the event. It is the probability of an event irrespective of whether or not any other events may have occurred.

b. Note that this use of the \( Pr(*) \) notation is somewhat different from our earlier usage. The use of \( J \) here (and \( M, A, B, \) etc. later on) corresponds to a specific event, not unlike the discrete attributes of nominal-level variables. Prior references to \( Pr( X > k ) \) allow \( X \) to take values along a continuum, not unlike the attributes of ratio-level variables.

2. O.K. Then let's assume that Joe and Mary were randomly sampled from a population with a symmetric age distribution. A JOINT PROBABILITY is the probability that two or more events occur. Accordingly, we might
ask, "What is the joint probability that both Joe’s and Mary’s ages are greater than the population’s mean age?" To help in answering this question, we can use a 2×2 table to depict the four relevant combinations of events regarding Joe’s and Mary’s ages:

<table>
<thead>
<tr>
<th></th>
<th>above</th>
<th>below</th>
</tr>
</thead>
<tbody>
<tr>
<td>above</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>below</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If Joe and Mary were sampled at random, then the probability that Mary’s age is greater than the population mean is independent of (i.e., uneffected by) whether or not Joe’s age is greater than the mean. Note that whenever two events are independent (as are units of analysis’s values on variables are when they are sampled at random), the joint probabilities of the events equals the product of the marginal probabilities of the events. Using probability notation,

\[
\Pr( J \& M ) = \Pr( J ) \times \Pr( M ) = .5 \times .5 = .25 .
\]

Accordingly, the joint probability represented by each cell in the table equals .25. Thus "the probability that EITHER Joe’s OR Mary’s (but NOT BOTH’s) ages are greater than the mean" can be calculated as follows:

\[
\Pr( A ) + \Pr( B ) = \Pr( J \& \text{not-}M ) + \Pr( \text{not-}J \& M ) = .25 + .25 = .5
\]

NOTICE that this ONLY HOLDS WHEN Joe’s and Mary’s ages are RANDOM events, as they would not be were Mary and Joe married to each other, for example.
3. DEFINITION: Two random variables are statistically independent if the "conditional distribution" of one is the same within each level of the other variable.

Thus, if Joe has one level of the age variable (e.g., being older than the mean age), the probabilities are .5 that Mary (when selected) will be younger than the mean age and .5 that Mary will be older than this. If Joe’s level on the age variable is "younger than the mean age," these probabilities remain the same (i.e., the conditional distribution of Mary’s age is the same no matter what Joe’s age level is). Got it?

C. Knowing a STATISTICAL DEPENDENCE when you see one

Let’s now consider an example that shifts our thinking from probability distributions for two units of analysis, to empirical distributions for two groups (or clusters) of these units. In examining people’s general anxiety about life, you might hypothesize that the more religious people are, the less anxious they are. This can be shown graphically as follows:

Since the distribution of anxiety is different for different religiosity levels, anxiety and religiosity are statistically dependent.

NOTE: If you assume that each of these superimposed distributions are normal, you might wish to test whether the mean anxiety scores of their
corresponding subpopulations are significantly different. This is what is
done when you use a t-test, which we shall discuss at length later in the
semester.

D. Why random sampling is so important

1. Random sampling is a data-collection strategy for ensuring statistical
independence of observations among units of analysis (like Mary and
Joe). As a consequence, any dependence found in one’s data matrix can
only have originated in one of two ways:

a. On the one hand, the dependence may be due to nonrepresentative
peculiarities among the units of analysis randomly selected into the
sample. This is what statisticians mean when they speak of results’
being due to sampling error. Always recognizing this possibility,
but armed with the central limit theorem, the statistician estimates
the probability that this has occurred. If this probability is
sufficiently small (commonly, less than .05), the dependence may have
an alternative, "nonerroneous" source.

b. On the other hand, if one’s units of analysis are representative of
the population from which they were sampled (i.e., if sampling error
is not the culprit), then any statistical dependencies detected in
one’s sample must reflect dependencies among variables in this larger
population. Accordingly, given that their random sampling allows
them to assume observations among subjects to be statistically
independent, statisticians generally speak of independence and
dependence among variables (not subjects).

2. Statisticians tend to think of statistical independence and dependence
in terms of the "random variables" of a data matrix. For example, consider the following matrix with data on "n" units of analysis and "k" variables:

```
Var1 Var2 Var3 . . . Vark
person #1 2 4 2 . . . . 3
person #2 1 5 7 . . . . 2
person #3 3 2 4 . . . . 3
. . . . . . . . .
. . . . . . . . .
. . . . . . . . .
person #n 3 4 5 . . . . 1
```

Each number in this matrix is the value taken by a RANDOM VARIABLE. Imagine that you have not yet collected your data and that you have just made plans to identify the attributes of "n" persons on "k" variables. Before collecting your data you can think of yourself as having an empty data matrix with n-times-k place-holders for values that will be set only once you have randomly selected the persons to be included in your sample. These n-times-k place-holders are the random variables that you have at your disposal. Note how the values taken by a "variable" (i.e., the values down a column of the matrix) are the values taken by the "random variables" associated with it.

Because random sampling ensures that data on persons (or, more generally, on units of analysis) are statistically independent, any dependencies found among one's data are most likely due to statistical DEPENDENCE among one's variables and not among one's subjects! This is important because researchers are generally NOT interested in particular units of analysis (except in that they are representative of their populations-of-interest). Instead, the researcher's objective is to understand associations (or dependencies) among variables. When units of analysis have been sampled at random, not only does the central limit
theorem hold but you have ensured that associations found in your data are ones among variables. I cannot overemphasize the importance of this conclusion.

E. Thus far it has been claimed that when two variables are statistically independent,
\[ \Pr(A \& B) = \Pr(A) \times \Pr(B) . \]
If \( A \& B \) are two persons' values on the same variable (e.g., age) then this equality holds whenever these persons have been RANDOMLY sampled.

F. Now let us imagine that we also know that Joe (the first person sampled) is Catholic and that the Catholics in our population are (on the average) younger than other subjects. (Say, they have lots of babies.) Then given that Joe is Catholic, the probability that Mary is older than Joe is greater than the probability that she is not—assuming that Mary is selected at random from our population. Let's take this example more seriously:

Each of the following is a CONDITIONAL PROBABILITY (i.e., a probability of an event's occurrence given your knowledge that another event[s] has occurred). Taken together, they both comprise the conditional distribution of respondents' above- or below-mean ages given that they have Roman Catholic religious affiliation:

\[ \Pr(X > \mu_A | X = \text{Catholic}) = .4 \]
\[ \Pr(X < \mu_A | X = \text{Catholic}) = .6 \]

ASIDE: The "|" in these expressions is read as "given." For example, the first equality is read, "Given (|) that \( X \) is Catholic, the probability is .4 that \( X \) has an age greater than (>) the average age in the population (\( \mu_A \))."
We can illustrate this using a VENN DIAGRAM:

If you know that \( X = \text{Catholic} \), then you are only looking at a subpopulation of the respondents. The conditional probability refers to how many of the Catholics are old vs. young. To find the joint probability that someone is BOTH old and Catholic requires that we know something more than is depicted in the diagram. If you are drawing a sample from a population of Catholics, then the probability of sampling a Catholic is one. In this case you would be "given" that each unit of analysis is Catholic, leaving the joint probability of being both old and Catholic equal to the conditional probability of being old given Catholic affiliation.

Instead, let’s assume that we are studying residents of Antwerp and that 70% of the people in this Belgian city are Catholic. I.e., we have the following marginal distribution of religion:

\[
\Pr( X = \text{Catholic} ) = .7 \\
\Pr( X \neq \text{Catholic} ) = .3
\]

Now, let us consider the probability that someone is BOTH old and Catholic:

1. If we make the unlikely assumption that age is symmetrically distributed among Antwerp’s residents, we can conclude (given the discussion at the outset of this section) that the marginal distribution of age is as follows:
2. Blindly applying the formula for statistically independent variables, we calculate that

\[ \Pr( X > \mu_A ) = .5 \]
\[ \Pr( X < \mu_A ) = .5 \]

But this is NOT the probability that a respondent is an old Catholic, since we know that if \( X = \text{Catholic} \), then the probability that \( X > \mu_A \) equals .4! I.e., it is NOT .5 as is \( \Pr( X > \mu_A ) \) in the population.

So we need to change our formula slightly:

\[ \Pr( X = \text{old and Catholic} ) = \Pr( X = \text{Catholic} ) \times \Pr( X = \text{old} | X = \text{Catholic}) = .7 \times .4 = .28 \text{ (which is clearly less than .35)} \]

We can illustrate these probabilities in a table:

Table 1: Table of Joint and Marginal Probabilities of Religion and Age of Residents of Antwerp, Belgium.

<table>
<thead>
<tr>
<th>Age</th>
<th>Catholic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>.28</td>
<td>(.22)</td>
</tr>
<tr>
<td>Young</td>
<td>(.42)</td>
<td>(.08)</td>
</tr>
<tr>
<td>.7</td>
<td>.3</td>
<td>.5</td>
</tr>
</tbody>
</table>

COMMENTS:

a. CONDITIONAL PROBABILITIES are calculated as follows:

\[ \Pr( X > \mu_A | X = \text{Catholic}) = .4 = \frac{\Pr( \text{Old} \ & \ \text{Catholic})}{\Pr( \text{Catholic})} \]
(Recall that if you know that \( X = \text{Catholic} \), then you are only looking at a subpopulation of 70% of the total. To find the conditional probability of being old given being Catholic, you need only ask, "How many of these 70% are old?")

Thus, the relation among conditional, marginal, and joint probabilities is

\[
\text{Conditional} = \frac{\text{Joint}}{\text{Marginal}}. 
\]

NOTE: Be sure you can distinguish among "joint," "conditional," and "marginal distributions"!!

b. \( \Pr(O \& C) = \Pr(O) \times \Pr(C | O) = .5 \times (.28)/.5 = .28 \)
\[
= \Pr(C) \times \Pr(O | C) = .7 \times (.28)/.7 = .28 
\]

BUT, note that \( \Pr(C | O) \neq \Pr(O | C) \) !!!

c. When \( \Pr(O \& C) = \Pr(C) \times \Pr(O | C) \times \Pr(C) \times \Pr(O) \), then the two events, \( O \& C \), are not statistically independent:

RECALL that statistical independence requires that the CONDITIONAL DISTRIBUTION of one variable (e.g., \( \Pr(O | C) \)) equals its MARGINAL DISTRIBUTION (e.g., \( \Pr(O) \)) for all levels of a second variable (e.g., \( C \)).

d. Finally, when one variable is independent of the other, the other is statistically independent of it. That is,
\( \Pr(C | O) = \Pr(C) \) implies \( \Pr(O | C) = \Pr(O) \).

This follows, since by definition,
\( \Pr(O \& C) = \Pr(O) \times \Pr(C | O) = \Pr(C) \times \Pr(O | C) \).

And if \( \Pr(C | O) = \Pr(C) \), then \( \Pr(O | C) = \Pr(O) \) !!!
G. NOW, THE BIG QUESTION: What is a knowledge of marginal, joint, and conditional distributions good for?

1. Imagine that you wish to draw a multistage cluster sample of residents from a city of 60,000. The sample is to be drawn in two stages and with **probability proportional to size** (i.e., in a manner ensuring that each resident has the same probability of being included in the sample). A sample of 300 is to be obtained by randomly sampling 10 residents within each of 30 randomly sampled blocks.

2. In the first stage of your sampling, you wish to sample each block (B) with a probability that ensures that each of its residents (R) has the same chance of being included in the final sample. Assuming that these probabilities can be correctly assigned (e.g., based on census records) and given your motivation to ensure that each of the city's 60,000 residents has the same probability of being included in a sample of size 300, this joint probability of sampling a resident's block and a resident within that block would be

\[
\Pr( B \& R ) = \frac{300}{60,000} = .005 .
\]

3. Now, imagine that a block with 100 residents is selected. Given that 10 residents are to be sampled from each sampled block, the probability of the event that one of these 100 residents is included in the final sample is

\[
\frac{10}{100} = .1 .
\]

We can generalize this result to refer to any block with an arbitrary number (say, k) of residents, to yield the conditional probability of the event that one of a block's "k" residents is included in the final
4. Now let's consider the "probability proportionate to size" question:
Given the probability at which you wish each resident to be selected and the probability at which a resident is randomly sampled from a block of size, $k$, at what probability must a block of this size be sampled? Here we are asking for the marginal probability at which the 30 blocks are to be sampled. The question's answer follows directly from our knowledge of the relation among marginal, joint, and conditional distributions:

$$\Pr(R \mid B) = \frac{10}{k}$$

5. Final comments regarding multistage cluster sampling: Once you have found the probability at which blocks are to be sampled, there is sampling software to help you in sampling blocks "weighted" according to these probabilities. Also note that before doing this, you must ensure that each block has at least 10 residents (can't sample 10 if there are only 6) and no more than 2000 residents (doesn't make sense to sample a block at a probability greater than one). If you are unclear on the latter point, try calculating $\Pr(B)$ for any $k > 2000$, and you'll see what I mean.

H. Marginal, joint, and conditional probabilities are also fundamental to understanding the chi-square statistic. Let's assume that we have obtained a random sample of 800 Antwerp residents, and that our data are as depicted in Table 2. Based on these data we should be able to address research questions such as, "Are Catholics younger (or older) than other residents?"
Table 2: Catholic and Non-Catholic Residents by Age.*

<table>
<thead>
<tr>
<th></th>
<th>Catholic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>224</td>
<td>176</td>
</tr>
<tr>
<td>Young</td>
<td>336</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>560</td>
<td>240</td>
</tr>
</tbody>
</table>

* Hypothetical data.

1. We can begin answering this question by determining whether Catholics’ ages are different from what one would EXPECT ON THE BASIS OF THE MARGINAL AGE DISTRIBUTION OF ALL RESIDENTS. We can estimate the marginal probabilities associated with being old and Catholic as follows:

\[
\hat{Pr}(O) = \frac{400}{800} = .5, \quad \text{where } \hat{Pr}(O) \text{ is an estimator of } Pr(O) - \text{the probability that an Antwerp resident is old, and}
\]

\[
\hat{Pr}(C) = \frac{560}{800} = .7, \quad \text{where } \hat{Pr}(C) \text{ is an estimator of } Pr(C) - \text{the probability that an Antwerp resident is Catholic.}
\]

2. Joint probabilities are obtained using numbers within the cells of the table. For example,

\[
\hat{Pr}(O \& C) = \frac{224}{800} = .28.
\]

3. Now to the question at hand: "Is this joint probability different from what one would expect to find by chance alone (i.e., from what one would expect if age and religion were in fact unrelated among Antwerp..."
residents)?" We can answer this in parts:

a. We know that if age and religion are unrelated (a.k.a., independent),
\[ \Pr(O \& C) = \Pr(O) \times \Pr(C). \]

b. Given this, we can now estimate how many old Catholics would one expect to find in this sample (of size, \( n = 800 \)) if age and religion were unrelated among all Antwerp residents:
\[ f_e = \hat{\Pr}(O) \times \hat{\Pr}(C) \times n = .5 \times .7 \times 800 = 280. \]

1) Note that the joint probability assumed here (i.e., given the assumption of independence) equals .35 (\( .5 \times .7 \)), whereas the joint probability estimated from Table 2 is considerably smaller than this (namely, .28 as calculated above).

2) As a direct consequence, this expected frequency (\( f_e \)), 280, is 56 larger than the observed frequency (\( f_o \)) of 224 in Table 2.

3) Given one expected frequency (e.g., \( f_e = 280 \)), we can immediately determine all the other EXPECTED CELL FREQUENCIES by ensuring that cell frequencies add up to the table's marginal frequencies:

Table 3: Expected Frequencies of Catholic and Non-Catholic Residents by Age.

<table>
<thead>
<tr>
<th>Religion</th>
<th>Catholic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>280</td>
<td>120</td>
</tr>
<tr>
<td>Young</td>
<td>280</td>
<td>120</td>
</tr>
</tbody>
</table>

560 240 800
4. O.K. Now we know both the expected and observed cell frequencies associated with our sample. BUT how different do these frequencies have to be to be SIGNIFICANTLY DIFFERENT? This question is answered with the CHI-SQUARE statistic.

Calculating chi-square requires that for each cell of your table, you first subtract the expected from the observed cell size, then square the difference, and divide this by the expected cell size. The sum of all these CONTRIBUTIONS TO CHI-SQUARE has a specific probability distribution. In particular, the sum is distributed as chi-square with the number of degrees of freedom in your table. The formula for chi-square is as follows:

\[
\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}
\]

a. There is only one degree of freedom in a 2x2 table. As mentioned previously, this is because if you know the frequency in 1 cell, you can determine the frequencies of all other cells based on both this frequency and the marginal frequencies of the table. In general, you can decide how many degrees of freedom you have by using the formula, 

\[
(r - 1) \times (c - 1),
\]

where \( r \) = the number of rows in the table and \( c \) = the number of columns in the table. If you added a third variable with "d" attributes to the table, the resulting 3-dimensional table would have 

\[
(r - 1) \times (c - 1) \times (d - 1)
\]

degrees of freedom. This can be generalized further.

b. O.K., so we calculate it and discover that chi-square = 74.67. Now
what? Well, we want to know if Catholics are younger than we would expect on the basis of the marginals alone. To find out, we look at Table C from among the tables handed out in class. There it indicates that in a table with 1 degree of freedom,

$$\Pr( \chi^2_1 > 10.827 ) = .001 .$$

1) Whereas in the standard normal table (Table A) the body of the table contains probabilities and values of the $z$-statistic head the rows and columns, in your chi-square table (Table C) the body of the table contains values of the chi-square-statistic, columns are headed by probabilities, and rows are headed by "degrees of freedom" for your table.

2) Because our chi-square of 74.67 is (considerably) larger than this, the probability is less than .001 that sampling error accounts for why our observed frequencies differ so much from the frequencies we would have expected if age and religion were statistically independent. This is because it is very unlikely that the chi-square of 74.67 reflects a chance occurrence due to peculiarities of our sample.

I. STATISTICAL SIGNIFICANCE versus THEORETICAL CONFIRMATION

1. Once you know that a table contains a statistically significant association, you still do not know if the association is in the hypothesized direction. For example, chi-square could equal 74.67 because Catholics are significantly older than non-Catholics. So it is only by returning to the table that we can definitively establish whether Catholics are significantly younger than non-Catholics. In the above case we can conclude that they are, because

54
2. One of the most common among mistakes made by social scientists is to conclude that a statistically significant finding supports their theory despite the fact that (without their having noticed it) the direction of the significant association is opposite to that suggested in their theory. Please be careful not to mistake statistical significance for theoretical confirmation.

J. STATISTICAL SIGNIFICANCE versus SUBSTANTIVE IMPORTANCE

IMAGINE that our data on the Belgian city are as follows:

Table 4: Catholic and Non-Catholic Residents by Age.*

<table>
<thead>
<tr>
<th>Religion</th>
<th>Catholic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>48%</td>
<td>55%</td>
</tr>
<tr>
<td></td>
<td>267 (exp.=280)</td>
<td>133 (exp.=120)</td>
</tr>
<tr>
<td>Young</td>
<td>52%</td>
<td>45%</td>
</tr>
<tr>
<td></td>
<td>293 (exp.=280)</td>
<td>107 (exp.=120)</td>
</tr>
</tbody>
</table>

* Hypothetical data.

The table shows Catholic residents seven percent (7%) more likely to be young than residents with other religious affiliations.

1. Is this statistically significant? Yes, it is (at \( \alpha = .05 \)). The calculated value of chi-square equals \( \chi^2 = 4.02 \), which (since it is larger than \( \chi^2_{1,.05} = 3.841 \) [see Table C]) indicates that a result this strong or stronger would occur by chance in only one same-sized random sample in twenty.
2. Is the result SUBSTANTIALLY IMPORTANT? What do you think?

a. THERE IS NO STATISTICAL BASIS FOR DECIDING how much of a difference is substantively important. You must consult your theory, your colleagues, and your common sense.

b. Using the .05 significance level and considering a 10% difference to be substantively important, the findings in Table 4 are statistically significant, but not substantively important. Because statisticians (like you) must decide what is substantively important, their statistics are unable to "speak for themselves."

K. RELATING CHI-SQUARE BACK TO THE CONCEPT OF "STATISTICAL INDEPENDENCE":
Let's see just what we have done here. We have two events, O = the event of being "old" and C = the event of being Catholic. When we collect our data, they are in this form:
Table 5: Dummy Table of Catholics and Old Residents.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>( \bar{C} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>( \bar{0} )</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

\begin{align*}
ap + b & \quad a + c \\
\text{a + b} & \quad n = a + b + c + d
\end{align*}

The question is, then, whether or not \( 0 \) and \( C \) are statistically independent. To decide this, we need only find out if

\[
\Pr(0 \text{ and } C) = \Pr(0) \times \Pr(C)
\]

(i.e., if the joint probability of \( 0 \) and \( C \) equals the product of the marginal probabilities of the events \( 0 \) and \( C \)).

1. The marginal probabilities of \( 0 \) and \( C \) are estimated as follows:

\[
\hat{\Pr}(0) = \frac{a + b}{n} \quad \hat{\Pr}(C) = \frac{a + c}{n}
\]

2. The joint probability of \( 0 \) and \( C \) is estimated as follows:

\[
\hat{\Pr}(0 \text{ and } C) = \frac{a}{n}
\]

3. Thus, a good test of statistical independence should evaluate whether

\[
\Pr(0 \text{ and } C) = \Pr(0) \times \Pr(C).
\]

In terms of our dummy table, it should evaluate the extent that

\[
\hat{\Pr}(0 \text{ and } C) = \hat{\Pr}(0) \times \hat{\Pr}(C)
\]

or that

\[
\frac{a}{n} = \frac{a + b}{n} \times \frac{a + c}{n}
\]

RECALL that in calculating chi-square, we sum up
For the a-cell in Table 4, the observed frequency is \( f_o = a \) and the expected frequency is \( f_e = \frac{a + b}{n} \times \frac{a + c}{n} \times n \).

So, the contribution to chi-square of the a-cell is

\[
\frac{(f_o - f_e)^2}{f_e} = \left[ \frac{a - \left[ \frac{(a + b)}{n} \times \frac{(a + c)}{n} \times n \right]}{\frac{(a + b)}{n} \times \frac{(a + c)}{n} \times \frac{1}{n}} \right]^2
\]

Multiplying by \( \frac{1/n^2}{1/n^2} \) we get

\[
= \left[ \frac{a}{n} - \left[ \frac{(a + b)}{n} \times \frac{(a + c)}{n} \times \frac{1}{n} \right] \right]^2
\]

\[
= \left[ \frac{\hat{Pr}(O \& C) - \hat{Pr}(O) \times \hat{Pr}(C)}{\hat{Pr}(O) \times \hat{Pr}(C)} \right]^2 \times n
\]

This is the contribution of cell "a" to chi-square in a two-variable table. A few comments about this "contribution" are in order:

a. The magnitude of the contribution meets one important criterion of a good test of statistical independence, since it equals zero when

\[
\hat{Pr}(O \& C) = \frac{a}{n} = \frac{a + b}{n} \times \frac{a + c}{n} = \hat{Pr}(O) \times \hat{Pr}(C).
\]

b. In a three-variable table this contribution would look like this:

\[
n \times \left[ \frac{\hat{Pr}(A \& B \& C) - \hat{Pr}(A) \times \hat{Pr}(B) \times \hat{Pr}(C)}{\hat{Pr}(A) \times \hat{Pr}(B) \times \hat{Pr}(C)} \right]^2.
\]
This can be generalized further.

c. NOTICE how the 'n' in the formula for the contribution suggests that the larger one's sample size, the more likely a statistically significant difference between $\hat{P}(0) \ast \hat{P}(C)$ and $\hat{P}(0 \& C)$ will be detected.

L. About the chi-square distribution

1. Chi-square and sample size

a. We have a 2 x 2 table:

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>$\bar{C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6 (7)</td>
<td>4 (3)</td>
</tr>
<tr>
<td>$\bar{0}$</td>
<td>8 (7)</td>
<td>2 (3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>6</td>
<td>20</td>
</tr>
</tbody>
</table>

$$\chi^2 = \frac{(6-7)^2}{7} + \frac{(8-7)^2}{7} + \frac{(4-3)^2}{3} + \frac{(2-3)^2}{3}$$

$$= \frac{6}{21} + \frac{14}{21} = \frac{20}{21} = .95$$

We now can estimate the probability this happened by chance. From Table C, we find that $\Pr(\chi^2 > 1.074) = .30$. (Note that 1.074 is as close to our chi-square of .95 as we can find in Table C.) Thus (using a bit of mental interpolation), the probability of getting a chi-square as large as .95 by chance is about .34. That is, one would expect a chi-square this large or larger in about one out of every three tables this size. (I.e., it is VERY probable.)

b. Now, all other things remaining equal, imagine that we have a sample

59
ten times as large:

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>\bar{C}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60 (70)</td>
<td>40 (30)</td>
</tr>
<tr>
<td>\bar{0}</td>
<td>80 (70)</td>
<td>20 (30)</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>60</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \frac{(60-70)^2}{70} + \frac{(80-70)^2}{70} + \frac{(40-30)^2}{30} + \frac{(20-30)^2}{30} \]

\[ = \frac{60}{21} + 140 \frac{21}{21} = 200 \frac{21}{21} = 9.5 \]

From Table C we find that

\[ \Pr( \chi_1^2 > 10.827 ) = .001 \text{ and } \Pr( \chi_1^2 > 6.635 ) = .01 . \]

Again using a bit of mental interpolation, we can conclude that

\[ \Pr( \chi_1^2 > 9.5 ) \approx .004 \text{ (or so). That is, the probability of getting a } \chi_1^2 \text{ this large by chance is about 4 in 1,000 samples (i.e., NOT probable at all).} \]

NOTE that this chi-square is exactly 10 times as large as the first one and is based on a sample exactly 10 times larger. (This is no coincidence.) If two tables have the same relative cell sizes, but one is based on a sample k times as large, the chi-square for the larger table will be k times that of the smaller. (An algebraic proof of this statement is given on page 58.)

c. We can make use of this insight by addressing a new question: How large a sample would we need for the same relative cell sizes to be detected as statistically significant at the .05 level?
We know the following:

20 is the original sample size.

.95 is the chi-square for this sample.

3.841 is the size of chi-square we need to detect.

Since $20 \cdot k = n$ and $.95 \cdot k = 3.841$, then $k = 4.04$ and $n = 81$.

2. Chi-square should only be used when $f_e \geq 5$ in all cells of a 2x2 table. When one's table is larger than 2x2, 75% of the table's cells should have $f_e \geq 5$ and all cells should have $f_e \geq 1$. If $f_e < 5$ for too many cells, then Fisher's exact test (for 2x2 tables) or an extension of this test (neither covered in this course) should be used. Moreover, larger expected cell sizes can be ensured by dropping or collapsing the table's categories, or by collecting more data.

3. Trivia

a. The shapes of chi-square distributions change with different degrees of freedom. When the degrees of freedom get larger than 20 or so, chi-square takes on the shape of a normal distribution.

b. The MEAN of a chi-square distribution equals its degrees of freedom; its VARIANCE equals twice its degrees of freedom (i.e., $\text{Var}(\chi^2_{df}) = 2\cdot df$).

c. If $Z \sim N(0,1)$, then $Z^2$ - chi-square with one degree of freedom. You can verify this by comparing Table A with Table C. For example, $\Pr( |Z| > 1.96 ) = .05 = \Pr( \chi^2_1 = Z^2 > 3.84 = [1.96]^2 )$.

---

(NOTE: "=" means "equals by definition." ) ALSO the sum of two squared standard normal random variables is distributed as a chi-square random variable with TWO degrees of freedom. This can be generalized further to sums of more squared standard normal random variables.

M. One final example:

You have a contingency table with four variables:

- Religious affiliation: Catholic, Protestant, Jewish, Buddhist.
- Gender: Male, Female.
- Political affiliation: Republican, Democrat, Independent, Other.
- Ethnicity: Black, White, Other.

We calculate a chi-square for the table and it equals 26. This value of chi-square does NOT provide significant evidence (at $\alpha = .05$) that the data in this table vary significantly from what you would expect by chance. Drawing this conclusion begins by determining that the appropriate degrees of freedom for this table equal 18 ($=\left[\begin{array}{c}4-1 \\ 2-1 \\ 4-1 \\ 3-1 \end{array}\right]^2$). Then consulting Table C, we find that $\chi^2_{18,.10} = 25.989$ and $\chi^2_{18,.05} = 28.869$. Accordingly, one would expect a $\chi^2_{18}$ this large in about 1 in 10 samples.
N. GENERAL CONCLUSIONS

1. Normal distribution

   a. Indicates the probabilities of random fluctuations around a population parameter.

   b. The larger one's sample (n), the more closely the sampling distribution of each unbiased point-estimate-statistic will approximate a normal distribution with a mean of the population parameter—the one that the statistic estimates—and with a variance that is inversely proportional to the sample size.

2. Chi-square distribution

   a. Indicates the probabilities of SQUARED random fluctuations around a population parameter. (Recall that \( Z^2 \) — chi-square with one degree of freedom, that the sum of two squared standard normal random variables is distributed as a chi-square random variable with two degrees of freedom, etc.) In this sense, the chi-square distribution allows you to determine whether you have a "normal" amount of variation. Differently put, chi-square measures the degree to which your actual data vary from what you would expect knowing only your marginal distributions.

   b. The larger one's sample (n), the larger the value of chi-square for any given set of joint probabilities. In fact all things equal, increasing one's sample size by a factor of "k" will increase chi-square by exactly this amount. That is, whereas point-estimate-statistics approach population parameters' values for increasingly larger samples, chi-square-statistics approach infinity.
These two distributions are the most important in all of statistical theory. As do all probability distributions, they allow us to make probabilistic statements about interrelations among variables when these variables measure attributes of randomly sampled subjects (or other units of analysis).