210A Week 8 Notes

Last week we did comparison between two groups. We were mostly interested in comparing two groups on a continuous dimension, but it can generalize to proportions by two groups and (by use of dummy sets) to non-binary categorical variables. This time will focus more narrowly on the association between categorical variables and use simpler techniques for handling this.

Cross-tabs (Contingency Tables)

Start with a table of one categorical variable by another. The book calls these "contingency tables" but sociologists usually call them "cross-tabs" because they tabulate one variable crossed by another. And show how many fall in each possible combination of the categories. For example we can compare "married" and "occupation" in the NLSW. (For simplicity I'm only keeping the first four occupations).

```
. sysuse nlsw88.dta, clear (NLSW, 1988 extract)
. keep if occ>=1 & occ<=4 (837 observations deleted)
. tab occ married
                      |        married
           occupation |    single    married |     Total
----------------------+----------------------+----------
Professional/technica |       113        204 |       317
       Managers/admin |       106        158 |       264
                Sales |       227        499 |       726
   Clerical/unskilled |        36         66 |       102
----------------------+----------------------+----------
                Total |       482        927 |     1,409
```

In this case married is the column variable and occupation is the row variable, but this is arbitrary. The intersections of row and column in the center of the table are called "cells" and so, for instance we see that there are 113 women who are both single and professionals. On the outside of the table are the "marginals," which simply tell us the distribution of each variable, regardless of association. For instance, the row marginal tells us that there are 482 single women (regardless of job) and 317 professional women (regardless of marital status).

Another way to view the data is not to look at the raw numbers in the sample, but the percentage. But percentage of what? What is the denominator? When the whole sample is the denominator, we speak of the "cell" percentage. In contrast, if you want to look at the percentages on one variable within categories of the variable, you look at the probabilities "conditional" on the value of the other variable. Stata calls conditional probabilities "row percentages" or "column percentages." In Stata these are considered options so it comes after the comma. (The options are cell, row, col, and nofreq where nofreq suppresses the raw numbers). Here are the occupation probabilities conditional on marital status and the cell percentages.

```
. tab occ married, col nofreq
                      |        married
           occupation |    single    married |     Total
----------------------+----------------------+----------
Professional/technica |     23.44      22.01 |     22.50
       Managers/admin |     21.99      17.04 |     18.74
                Sales |     47.10      53.83 |     51.53
   Clerical/unskilled |      7.47       7.12 |      7.24
----------------------+----------------------+----------
                Total |    100.00     100.00 |    100.00
 . tab occ married, cell nofreq
                      |        married
           occupation |    single    married |     Total
----------------------+----------------------+----------
Professional/technica |      8.02      14.48 |     22.50
       Managers/admin |      7.52      11.21 |     18.74
                Sales |     16.11      35.42 |     51.53
   Clerical/unskilled |      2.56       4.68 |      7.24
----------------------+----------------------+----------
                Total |     34.21      65.79 |    100.00
```

χ^2 Test of Statistical Independence

Now that we have the cross-tab, what we'd like to know is if the two variables are associated. (Only continuous variables can be "correlated," categorical variables are "associated"). This takes the form of a statistical test and like all statistical tests is based on contrasting the data with a null hypothesis. In this context, the null hypothesis is that you could guess the cells based on knowing the marginals. To take a simple example, if you have two binary variables each of which is .50, .50; the null hypothesis is that the cells should be about .25, .25, .25, .25. These are the expected frequencies, meaning the frequencies you would expect if you know the marginals. Formally stated, the null is that the difference between the observed frequencies and the expected frequencies are zero.

The first step to the test is calculating the expected frequencies. The subscript "e" means expected (similar to "hat" or "^") and the subscripts "i" and "j" refer to particular rows and columns.

$f_{eij} = \left(\frac{f_i}{f}\right) f_j$

Stata can generate these expected frequencies for us (for pedagogical purposes, there's not much reason to do this in practice).

```
 . tab occ married, expected nofreq
                      |        married
           occupation |    single    married |     Total
```

```
----------------------+---------------------+----------
Professional/technica |     108.4      208.6 |     317.0
       Managers/admin |      90.3      173.7 |     264.0
                Sales |     248.4      477.6 |     726.0
   Clerical/unskilled |      34.9       67.1 |     102.0
----------------------+---------------------+----------
                Total |     482.0      927.0 |   1,409.0
```

You then compare the expected frequency for the cell (f_{eij}) to the actual frequency for the cell (f_{ij}). When you compare the prediction to reality this is called a "residual" and this will become important when you do regression. For instance, our actual number of single professionals is 113, but our prediction is 108.4, so our residual for the cell is is 4.6. You calculate the squared residual as a proportion of the expectation (in our example, 0.19) for each cell in the table and sum them to show how far off the null is from the observed data.

$\chi^2 = \sum \frac{(f_{ij} - f_{eij})^2}{f_{eij}}$

This sum is the test statistic and the inrepretation is analogous to "*t*." It's written χ^2 and is pronounced chi-squared. If you're not using LaTeX or other typesetting that lets you show a Greek χ character, it's usually better to spell it out as "Chi" instead of using the Latin alphabet letter "X" because in math "X" usually means a matrix.

The interpretation is very similar to "*t*" in that large figures mean more statistical significance and the interpretation is sensitive to degrees of freedom. However degrees of freedom has a different definition, instead of being based on the number of cases, it's based on the number of categories in the variables such that df=(row-1)(col-1)

```
. tab occ married, cell nofreq chi2
                      |        married
           occupation |    single    married |     Total
----------------------+---------------------+----------
Professional/technica |      8.02      14.48 |     22.50
       Managers/admin |      7.52      11.21 |     18.74
                Sales |     16.11      35.42 |     51.53
   Clerical/unskilled |      2.56       4.68 |      7.24
----------------------+---------------------+----------
                Total |     34.21      65.79 |    100.00
          Pearson chi2(3) =   7.2784   Pr = 0.064
```

Stata treats the χ^2 test as an option to the tab command. Here there are three degrees of freedom because we have two marital statuses and four occupations so df=(4-1)(2-1)=3. The χ^2 test statistic is 7.28 because this is what it works out to if you sum the squared differences as proportion of expectation for each cell in the expected versus observed tables. Remember this basically tells you whether your marginals tell you everything you need to know (the null, indicated by a low χ^2) or there is some noteworthy association in the table (the alternate hypothesis, indicated by a high χ^2). As with a *t*, there's no simple way to interpret it so you look up in the table how to interpret χ^2at *df*. The table in our book indicates that at 3 degrees of freedom, α=.05 corresponds to χ^2 of 7.81. Since our figure of 7.28 is a bit below this, we can not claim to have shown a statistically significant association between occupation and marital status (though since it's close to .05 we can't rule it out either). Stata can calculate an exact *p* value for us and it shows .064, which as with the table in the book is a bit shy of statistical significance.

There are a few important points of contrast between t and χ^2. First, there's no such thing as a two-tailed χ^2test so you don't need to double *p*. Second, whereas *t* values of two or higher are usually significant, you can often have fairly large χ^2 values that are not. The 1.96 standard is *not* meaningful with a χ^2 test. Third, χ^2 is even more sensitive to degrees of freedom. Fourth, unlike *t* where degrees of freedom is approximately the number of cases in the dataset, with χ^2, *df* is approximately the number of cells in the table.

Note that the number of cases in the underlying dataset does not *directly* enter into it anywhere. We have yet to see "*n*" in our formulas but this does not mean that *n* is not important. The accuracy of a χ^2 increases with *n* and generally speaking you need at least 5-10 cases in each cell for χ^2 to be trustworthy. If this condition is not met you need to use more exotic metrics like Fisher's exact test. In practice this is an issue of choosing a different option in the software (Stata offers five different alternatives to χ^2 in its basic "tab" syntax and even more options are available in other commands.)

Frequency Weights

A technical issue that is especially important for cross-tabs is that Stata and other stats software assume that each record in the dataset describes one case. This is true for the NLSW data we've mostly been playing with but it isn't true for everything. Many datasets describe a type of case and then list how many cases are of that type. The listing of how many cases match the description is called a "frequency weight." For instance, if you read a journal article and it described there being 110 black women, 90 black men, 200 white women, and 190 white men in the paper's sample you could enter this as

```
black  female  fweight
1      1       110
1      0       90
0      1       210
0      0       190
```

Although this dataset only has four records it's describing 600 cases and is equivalent to having 110 rows each of which has black=1 and female=1, then 90 rows each of which had black=1 and female=0, and so on. This is much more compact and so simple datasets are often distributed like this. Many US government datasets use frequency weights and most of the Stata example datasets use them. However they're not well-suited for datasets with continuous variables, which is why NLSW has a separate row for each case. The only issue is that you have to let Stata know that you want to use the weights. For instance, the Stata example dataset "voter" describes how different income groups broke between the three candidates in the 1992 election and uses the variable "pop" as a frequency weight. Note that for some reason the fweight syntax comes in brackets before the comma even though you'd expect it to be a post-comma option.

```
. sysuse voter, clear
. tab inc candidat [fweight=pop],  chi2
    Family |    Candidate voted for, 1992
    Income |   Clinton       Bush      Perot |     Total
-----------+---------------------------------+----------
     <$15k |   127,947     49,878     39,035 |   216,860
   $15-30k |   167,292    130,116     74,352 |   371,760
   $30-50k |   190,527    176,586     97,587 |   464,700
   $50-75k |   123,920    130,116     55,764 |   309,800
     $75k+ |    72,493     96,658     32,219 |   201,370
-----------+---------------------------------+----------
     Total |   682,179    583,354    298,957 | 1,564,490
          Pearson chi2(8) =  3.9e+04   Pr = 0.000
```

If we leave out the weights, the results are silly.

```
. tab inc candidat,  chi2
    Family |    Candidate voted for, 1992
    Income |   Clinton       Bush      Perot |     Total
-----------+---------------------------------+----------
     <$15k |         1          1          1 |         3
   $15-30k |         1          1          1 |         3
   $30-50k |         1          1          1 |         3
   $50-75k |         1          1          1 |         3
     $75k+ |         1          1          1 |         3
-----------+---------------------------------+----------
     Total |         5          5          5 |        15
           Pearson chi2(8) =   0.0000   Pr = 1.000
```

Note that there are actually two different kinds of weights. Frequency weights (fweight) describe how many cases match the description. If you have a dataset (like nlsw88.dta) you can flatten it into an fweight dataset (like voter.dta) by using the "contract" command. Sampling weights (pweight) are used to correct for weird sampling design and get it to match some known description of the population. For instance if you intentionally oversample a small subpopulation you need to use pweight to tell Stata that you have too many cases from that group (compared to a random sample).

Residuals

The χ^2 test only tells you whether the table as a whole has something unusual about it, but it doesn't get any more specific than that. To identify which cells are most responsible for the weirdness you need to calculate residuals.

The default Stata syntax doesn't help you calculate residuals for crosstabs. Fortunately one of the many Stata ado files written by the statistician Nick Cox, is tabchi, which you can install by typing

```
net describe tab_chi, from(http://fmwww.bc.edu/RePEc/bocode/t)
net install tab_chi
```

From now on you can treat tabchi just like any other command.

A basic residual is just $f_{ij} - f_{eij}$ but this is hard to interpret as it doesn't scale for the marginals. The book suggests using standardized residuals but I think it makes more sense to use Pearson residuals which are defined as ${f_{ij}-f_{eij}}/{\sqrt{f_{eij}}}$. The basic interpretation of a Pearson residual is that zero means the cell fits the marginals, a large positive number means it's bigger than expected, and a large negative number means it's smaller than expected. Here's the Pearson residuals with both observed and expected frequencies suppressed.

```
. tabchi occ married, pearson noo noe
          Pearson residual
----------------------------------------
                       |     married
            occupation |  single  married
-----------------------+-----------------
Professional/technical |   0.438   -0.316
        Managers/admin |   1.651   -1.190
                 Sales |  -1.355    0.977
    Clerical/unskilled |   0.187   -0.135
----------------------------------------
          Pearson chi2(3) =   7.2784  Pr = 0.064
 likelihood-ratio chi2(3) =   7.2166  Pr = 0.065
```

As can be seen, professionals and clericals fit the marginals pretty well but managers are unusually likely to be single whereas saleswomen are unusually likely to be married. So to the extent that we think there's an association between marital status and occupation, it's driven by managers and sales.

Odds Ratios

Another way to compare categorical variables is with odds ratios. If you don't gamble, odds are harder to get a feel for than probabilities but they have some nice mathematical properties. The one we'll deal with now is that odds-ratios are good for comparing the association of two binary variables.

Odds are simply how often something happens (or is expected to happen) over how often something does not. Odds are always a ratio and if no denominator is stated the implicit denominator is "1." Odds have a range of zero (not gonna happen) to one (even money) to infinity (sure thing). Note that the terms "success" and "failure" are derived from gambling and are used loosely in statistics. In some examples "success" can be downright weird if you take it literally. For instance, on page 236 of the book, everyone in the data is a murderer and for these murderers "success" is killing a white person and "failure" is killing a black person.

$odds = \frac{\pi}{1-\pi} = \frac{successes}{failures}$

In contrast, probability is how often something happens out of the total. Probability ranges from zero (not gonna happen) to half (even money) to one (sure thing).

$\pi = \frac{odds}{odds+1} = \frac{successes}{successes+failures}$

You won't deal with them until 210B, but log-odds or logits are just the logarithm of odds. They range from negative infinity (not gonna happen) to zero (even money) to infinity (sure thing). This is a nice property for regression.

One thing you can do with odds is use them to compare how likely something is for two groups. It's natural to want to compare probabilities by subtracting or dividing them, but doing so isn't very meaningful, which is why we use odds ratios. To do this you just divide the odds for one by those of the other. There are several ways to do this. Note that the Greek letter theta ("θ") can stand for "odds ratio." If you have two groups called 1 and 2, the odds-ratio tells you how much more common some event is for group 1 vs group 2. It tells you how much more likely "success" is for group one versus group two and thus you can view it as a measure of association.

	group1	group2
success	success_1	success_2
failure	failure_1	failure_2

$\theta = \frac{odds_1}{odds_2} = \frac{\frac{successes_1}{failures_1}}{\frac{successes_2}{failures_2}} = \frac{successes_1 * failures_2}{successes_2 * failures_1}$

As an example we can take our earlier example of marriage by occupation and look at the odds of marriage for managers vs sales.

$\theta = \frac{158*227}{499*106} = \frac{35866}{52894} = 0.678$

So we can see that women in management are two-thirds as likely to be married as women in sales. Note that $\theta < 1$ means success is more likely for group 2, whereas $\theta > 1$ means success is more likely for group 1, and right at 1 it's even money. When you get to logistic regression in 210B, odds-ratios will be important to interpreting the output of logit regression in a similar fashion to how *diff* (as used in a t-test of means) is to interpreting OLS regression.

Philosophy of Science

Sociologists tend to divide themselves into "positivists" and "social constructions" (with "enlightened positivist" sometimes being a middle ground), but these terms don't do justice to the philosophy of science and if taken seriously neither model is very appealing. Likewise, many scientists will tell you that they follow Popper's falsifiable hypothesis logic but neither does this reflect the way science is actually done (or ought to be done). We'll go over several approaches to science, all of which agree that science is possible and desirable, but differ as to exactly what this means. The central problem that all of them are directly or indirectly attempting to grasp with is that of rigorous induction – how do we translate observations about the universe into understandings of natural law without being blinded by our preconceptions.

Popper

Karl Popper was originally very excited by the social theories of Marx and the psychology of Freud and Adler, at one point working in Adler's lab. He grew frustrated though when he saw how absolutely any evidence could be made to fit within their theories with even facially disconfirming evidence being interpreted as the result of a previously unstated contingency or of the system's ability to sublimate contradiction. In contrast, when Einstein stated his theory of general relativity he extrapolated from it a very specific prediction about how during a solar eclipse it would be apparent that the gravity of the sun bends starlight. Sir Arthur Eddington observed an eclipse and found that Einstein's predictions were correct, but the point is not that Einstein was right but that he gave a specific prediction that could have been wrong. Popper found this contrast fascinating and set out to build an intellectual career out of contrasting science (exemplified by Einstein) with pseudo-science (exemplified by the Marxists and Freudians). Note that Popper thought that there was nothing inherently pseudo-scientific about social or behavioral inquiries, he just wasn't fond of these particular examples.

Popper's essential insight from this contrast was that confirmation is cheap. He gave the example of a theory that all swans are white. It would be fairly easy to make a long list of white swans in much the same way that Freud made a long list of people whose neuroses derived from sublimated sexuality. Popper said a much better thing to do would be to search for black swans and fail to find any. In fact (as seen in the photo) there are black swans so we can reject the "all swans are white" hypothesis. Freud never looked for his black swans, which would be neurotics without sublimated sexuality. Worse yet, Freud didn't really have a non-tautological measure of sublimation so his theory is literally not falsifiable. In contrast, Einstein made a very specific prediction about how the stars would appear during a solar eclipse such that any astronomer could examine a photo of an eclipse and see whether it matched Einstein's prediction.

Popper can be summed up as emphasizing the importance of "falsifiable hypotheses" as the definitive characteristic of science. Such a definition worked for him as he was far less interested in how science works than in defining what is *not* science. This is why one of the worst things you can say about a scientist is that his work is "not even wrong" as it implies that the scientist has lapsed into metaphysics. Philosophers call this agenda "demarcation" and sociologists call it

Black swans at the Franklin Park Zoo in Boston.

"boundary work." In our time the principle demarcation problem is creationism[1] whereas for Popper himself it was mostly about Marxism and psychoanalysis. Positivists also use obviously ridiculously things like astrology, pyramidology, and parapsychology for calibrating the gun sights, with the assumption being that any good demarcation criteria should be able to explain why astrology is bullshit. Popper went so far as to say that the theory of natural selection is not scientific because in practice "fitness" is defined tautologically as "that which is associated with survival." However in a famous lecture he eventually recanted and argued that even if it is tautological to label any given common allele as promoting fitness, we can make a falsifiable hypothesis that selective advantage is more important for explaining complex organs than such descent processes as genetic drift.

While the idea of the "falsifiable hypothesis" was Popper's key contribution, it's worth also reviewing the logical positivist school with which he was loosely affiliated. The positivists drew a very strong distinction between synthetic (empirical data) and analytic (math and logic). Any statement that could not be

[1] Creationism comes in two major forms, a hardcore "young Earth" version and a more squishy "intelligent design." Young Earth creationism argues that Genesis is literally true and about 6000 years ago God created the heavens and the Earth in 144 hours and a few generations later created a massive flood that essentially rebooted the world. Intelligent design accepts the broad outlines of the conventional scientific view of the age of the Earth and the procession of natural history but argues that divine intervention routinely adjusts natural history, chiefly through being responsible for speciation. One bizarre consequence of this is that intelligent design is too vague to test, whereas young Earth creationism gives very concrete predictions (all of which are demonstrably false). Thus in strict Popperian terms intelligent design is more pseudo-scientific than young Earth creationism as the latter gives testable (albeit false) hypotheses whereas the former does not.

described as either synthetic or analytic they derided as metaphysics (or more whimsically, "music" or "poetry"). Popper's work fit within the positivist framework as it assumed a sort of deduction-induction cycle where the scientist would use logic to *derive* falsifiable hypotheses from theory, then collect data to *test* these hypotheses. Sociologists usually use the term "positivist" casually as the opposite of "deconstructionist" or "postmodernist" to mean someone who believes that science is possible without being hopelessly mired by subjectivity. Our usage completely loses any philosophic notions about strong distinctions between analytic, synthetic, and metaphysical.

Quine and Underdeterminism

W. V. Quine published "Two Dogmas of Empiricism" in the *Philosophic Review* in 1951 and basically destroyed positivism. The two dogmas in question are that 1) there is a meaningful boundary between synthetic and analytical and 2) that a discrete synthetic statement can be evaluated. Quine feels that really these two dogmas have the same conceptual flaw but he treats them as relatively distinct so as to make his critique isomorphic to positivism itself. On the synthetic/analytic dichotomy, Quine's critique is basically that it gets very messy distinguishing between a definition and a finding as they take the same grammatical form. Even more radically, he claimed that the profound weirdness of quantum physics demonstrates that even abstract logic is an empirical question.

The second critique is more directly of interest to practicing scientists. This "underdeterminism" problem shares a lot with an earlier argument by Pierre Duhem and so is sometimes known as the Duhem-Quine thesis. The positivists understood a version of this, called the "auxiliary hypotheses" problem but they underestimated how problematic it was.[2] When we state a hypothesis, it is implied that the hypothesis is expected to hold *ceteris paribus*. The assumption is that the evidence will test a hypothesis if (and only if) the auxiliary hypotheses are well behaved. This raises the problem that when we encounter evidence that is facially contrary to a hypothesis we cannot be sure if this is really evidence against the hypothesis, or is it only one of the *ceteri* making mischief by failing to remain *parili*s.

One of the best known problems of this sort was Marx's various failed historical predictions, most notably that the revolution would occur in an industrial nation (but also other expectations such as the emisseration of the proletariat). Many people, including both Popper and Gramsci, took this to mean that Marx was simply wrong. However many Marxists argued that there was nothing wrong with the theory of dialectical materialism as such, it merely hadn't explicitly anticipated the skill and charisma of Lenin or the agency that the bourgeois state showed in creating the welfare state to defuse class struggle in the industrial world. Thus in this imagining Marx's hypothesis (the Germans or English will

[2]Ethnomethodologists have similar concepts and refer to "secondary elaborations" used to salvage the truth of "incorrigible proposition" against facial challenges.
Mehan, Hugh and Houston Wood. 1975. *The Reality of Ethnomethodology.* New York: Wiley Interscience.

have a socialist revolution) was suppressed by the auxiliary hypothesis that unusually capable leaders would not show up in backwaters that had only recently abandoned serfdom. The positivists didn't see this case as especially problematic because they thought that the Marxist apologia of auxiliary hypotheses were embarrassingly ad hoc.

The Duhem-Quine underdeterminism thesis is that auxiliary hypotheses are literally infinite. Some of the examples of these infinite auxiliary hypotheses philosophers give are kind of silly, like "elves do *not* cause equipment to give inaccurate readings on Tuesdays" or "it does *not* matter what color shirt the lab worker is wearing" but it's hard to say whether blaming elves is really any sillier than claiming that men make history but not in the circumstances of their choosing *except* for V I Lenin because he's just so awesome. However positivists would have no problem saying that when a scientist makes such an interpretation it is clearly the fault of the scientist and not of either elves or Lenin. Unfortunately, Duhem and Quine argued that there are a very large number of very plausible auxiliary hypotheses. Prime among these are the innumerable ways in which data collection can suffer measurement error. Furthermore there is the fact that many of our measurement tools are themselves based in theories which conceivably could be wrong. For example, say your hypothesis is that if you heat a gas in an airtight and rigid chamber the pressure will rise and your barometer finds that the pressure does not rise. This could be interpreted as evidence against Boyle's law or it could be that Boyle was right and you have one of the following problems:

- your chamber is not airtight and/or rigid
- your heater and/or thermometer are broken
- your barometer is broken
- barometers don't actually measure pressure
- an infinite number of other more or less plausible auxiliary hypotheses

If we ignore the infinite number of unstated auxiliary hypotheses and focus on the specific ones, you can imagine testing each of them in turn. For instance, you could measure that your chamber is indeed airtight by putting a rat in it and seeing that it suffocates. But these verifications are themselves beset with problems such as that maybe the rat had a heart attack despite an abundance of oxygen, or perhaps it takes more ventilation to sustain a rat than to relieve a slowly expanding gas. The problem is recursive so that ultimately you can always spin a (progressively more convoluted) story that your original hypothesis was correct. In some cases this is actually a good thing to do since things like sloppy lab work are pretty common and if we never blamed anomalous results on auxiliary hypothesis we'd soon run out of theories.

While it's easiest to illustrate with the hard sciences, the issue of theory dependent tools is also a problem for social science. For instance, before we can get to the *real* problems, social scientists have to implicitly or explicitly decide issues like whether income is an adequate proxy for total consumption, how to reduce millions of jobs into hundreds of occupations and ultimately into something as manageable as the EGP class schema or the SEI index, the magnitude of social desirability bias, and how long (if ever) it takes informants to relax and act normally in front of an ethnographer. In my own subfield of radio studies, everyone agrees what the key hypothesis is (broadcasting monopolies create diverse content) and what the evidence shows facially (yes they do) but there is a big debate over an essentially auxiliary hypothesis about the quality of the evidence (whether "format" is a meaningful proxy for content).

Despite his rejection of positivism, Quine was no nihilist or skeptic. Indeed, he was explicit about offering a post-positivist way to recover empiricism. Quine felt that *ultimately* we cannot test any discrete hypothesis but only the entire system of science. However even in limited and narrow cases we must accommodate the evidence so that if we wish to salvage a particular hypothesis against contradictory evidence we must displace the doubt to some more or less specific auxiliary hypotheses. Quine speaks of belief as a web, fabric, or force field and treats surprising observations as not necessarily discrediting any particular belief but prodding the field as a whole and deforming it. There is a loose coupling between observations and beliefs so the main hypothesis may withstand the contrary observation but the anomaly's evidentiary weight has to go *somewhere*. Implicit in Quine's positive agenda is that parsimony is a worthy goal (lest evidence be diffused through the web indefinitely) but it is debatable whether parsimony is a distinctly scientific value or merely an aesthetic principle. This post-positivist empiricist agenda (usually called "holism") is a bit fuzzy and its ambiguities would not be resolved until Kuhn.

Kuhn and Paradigms

Thomas Kuhn was not trained as a philosopher but as a physicist when he started teaching a course on the history of science at Harvard and as such his work is mostly descriptive whereas Popper and Quine are prescriptive. In teaching this course he developed the ideas that he eventually published as *The Structure of Scientific Revolutions*. While the term "paradigm shift" is what most people remember about this book – and which became fashionable among business types in the 1990s – the real interest is what happens *within* the paradigms.

A paradigm is a more or less cohesive agenda and set of guiding principles. The paradigm thus shows scientists what sort of questions to ask and what are reasonable ways to go about answering them. Kuhn refers to work within a paradigm as "normal science" which mostly consists of "puzzle-solving." Contra Popper, normal science consists of working out minor puzzles about exactly *how* the paradigm works, but not *if* the paradigm works as it is taken for granted. This is holism in practice in both a positive and negative sense. In the positive sense, the paradigm provides coherence to the universe and presents manageable chunks of reality for the scientist to chew on.[3] As anticipated by Duhem, one of the things the paradigm does is tell the scientist what a meaningful problem looks like. In the negative sense, the paradigm can blind the scientist to evidence against it for when observations contradict the theory they are impossible to interpret and must be shunted off to some auxiliary hypothesis.

Whenever evidence contradicts predictions of the paradigm this does not cause the rejection of theory but merely presents "anomalies." These anomalies can be temporarily accommodated by speculation about measurement error or epiphenomenal forces. Eventually though such perturbations in the web of belief make it so convoluted as to be less than useful. At this point a scientist or small circle of scientists creates a new paradigm which can accommodate the anomalies. Unlike the cruft-besotten old paradigm, the new paradigm is internally consistent, but it is also fairly vague and may lack details as to exactly how the paradigm work. Fleshing out these details thus becomes puzzle-solving for a new generation of normal science and we come full circle.

Thus, Kuhn essentially took holism and ran with it, but added a somewhat Popperian element during unsettled periods. It is especially worth noting that Victorian scientists had identified numerous anomalies in the Newtonian paradigm which Einsten was able to solve with the general relativity paradigm. Thus Kuhn's model of science is a very good description of the positivists' favorite case of good science.

Combining holism and positivism is something of a Goldilocks problem. If Popper wants us to be ready to discount theory and Quine wants us to be ready to discount data, Kuhn finds a way that they can both be right. Most of the time his model resembles holism, but when the web becomes too knotted by accomodating anomalies it shatters and we get a paradigm shift somewhat like Popper's falsification.

[3] These virtues of paradigm-dependence imply that Duhem, Quine, and Kuhn would all find grounded theory to be Quixotic at best.