# CORRELATION, DETERMINATION AND CAUSALITY IN INTRODUCTORY STATISTICS

**Milo Schield, Augsburg College**
**Dept. of Business & MIS. 2211 Riverside Drive. Mpls., MN 55454**

**KEY WORDS**: Critical Thinking, Prediction, Explanation, Teaching

**ABSTRACT:** *Most students study statistics to learn about causality: how to measure it and how to discover it. This is most difficult in the social sciences where experiments are often impossible and observational studies are the norm. Yet in introductory statistics correlation and determination are discussed openly while causality, if it is mentioned, is only mentioned briefly and negatively: "correlation is not causality." By excising causality from statistics we have disavowed the historical roots of our discipline. We should not let our early history with unwarranted generalizations about causality leave us in denial about the proper relation between correlation and causality in statistics. To better satisfy the interests of our students and to be inclusive of our own history, we must emphasize causality more in teaching statistics. This paper argues that there are many things that can be taught about causality that are not discipline specific. Students must be able to detect whether a given statement is assertive of causality or -- as is more common -- is ambiguous about causality. Students should be taught how to detect the causal connotations of words and phrases. Students must be taught to be proactive in seeking alternative explanations for differences, ratios and correlations in observational studies. Students must be taught the causal differences between description, prediction and explanation. Statistics should be expanded to include causality in ways that are discipline independent and professionally appropriate.*

**INTRODUCTION:** This paper examines the relation between correlation, determination and causality in introductory statistics. This paper holds that our goal is not merely variance reduction in modeling a given set of data; our goal is to use variance reduction as a means for discovering confounding variables and thus making better predictions and explanations. Our ultimate goal is to help our users understand causality under uncertainty. This paper assumes the following: correlation is an observable phenomenon whereas causality is always inferred; correlation may be a sign of causality, but correlation is never sufficient to infer causality.

## I. CORRELATION AND CAUSALITY

The only time causality is mentioned in most courses is in relation to correlation: (A) "Correlation is not causality". This statement is made by both textbooks and instructors. The purpose of this statement is to warn students against presuming that a large correlation always signifies causality. This statements should certainly be made, since students often believe that correlation is sufficient for causality. But we should do much more than simply reiterate this true statement. First, this statement is ambiguous since the quantity (none, some or all) is unstated. Ambiguity often arises from critical omissions or from words whose meanings are vague and not very specific.

1. *We should teach students to detect ambiguities due to omitted quantifiers.* Because of this ambiguity, some students may even conclude that "Correlation is never causal". These students will either deny any causal associations or will see this formulation as meaningless since there are times when it is obviously false.

2. *We should teach students to add quantifiers to make ambiguous statements more precise:* (B) "Some correlation is not causal" or "Correlation is not always causality". We might speak of necessity and say that "Correlation is necessary but not sufficient for causality", but now the statement is equivocal on 'correlation'. Equivocal words or phrases have multiple meanings but the meanings are distinct and the intended meaning is generally obvious from the sentence (Connell, 1973). The word 'pen' is equivocal but its' meaning in "The farmer cleaned the pen" is obvious. Ambiguous phrases have multiple meanings but the meanings overlap and the intended meaning is not generally obvious from the sentence.

3. *We should teach students to identify causally related words that are equivocal.* 'Correlation' is equivocal. Generally speaking, 'correlation' is a common noun synonymous with 'association'. In this non-technical sense, correlation is necessary for causality. But in statistics, 'correlation' signifies a proper noun -- the Pearson linear product-moment correlation. In this technical sense, correlation is not necessary for causality since a causal relationship may be non-linear.

Consider again the statement "Correlation is not always causality". This statement is still ambiguous. Does the statement signify all forms of causality or just some forms of causality?

4. *We should teach students to classify causal relationships using some schema.* We can certainly distinguish natural causality from artificial (man-made) causality. Within natural causality, we can and should classify the relationship between a predictor and a predicted variable as follows:

- direct causality by the predictor on the predicted,
- reverse causality by the predicted on the predictor, or
- common causality by a common cause acting on both the predictor and the predicted.

More complex forms of causality (reciprocal causality) can be built from combinations of these basic forms. Thus we could use these distinctions and say that (C) "Correlation is not always due to direct causality by the predictor on the predicted".

Note that these three statements (A, B, and C) are vastly different; the first is very ambiguous in relation to the third. If the first is stated but the third is intended, students may not understand what was intended. More fundamentally, these statements are all negative. They deny a universal by saying "Some correlation is not direct causality". But this is not very disputable and they all sidestep the fundamental question: "If the correlation is not pure coincidence then what kind of causality is involved?".

5. *Given some categories of causation, we should teach students to create and assess arguments for and against each type of causality.* What evidence would they need to try to settle the issue. In this way, we would be teaching students to think creatively and critically.

## II. CORRELATION & MAN-MADE CAUSALITY

High correlations are generally the exception in the social sciences. When students encounter a high correlation, they often presume the large magnitude is strong evidence in support of direct causality. Students are not aware that a high correlation may result from an association that is man-made. Students are unaware that some types of human causality don't fit under the classification of natural causality mentioned previously. Consider two common situations:

- For families, there may be a high correlation between the size of the family and the poverty-level income or between the income of a family and the taxes they pay on that income.
- For businesses, there may be a high correlation between fixed assets and total assets, between revenues and profits, between quantity sold and revenues, or between current liabilities and quick ratios.

In both situations, the causality involves human action; the causality is artificial rather than natural (non-artificial).

In the first situation, poverty level and income tax are completely man-made (artificial). These two artificial variables are associated with (mathematically dependent on) two independent variables: size of family and income of family. More precisely, there is some method or definition that relates the artificial dependent variable to the independent predictor variable. Thus, the values of these two artificial dependent variables are largely 'determined by' these two independent predictor variables. In one sense, a change in the predictor variable (family size or family income) "causes" a change in the artificial variable (poverty income or income tax). In another sense, it is human causality that defines the dependent variable in terms of the independent variable so that their correlation is extremely high.

In the second situation, the concepts are related by means of a formula involving a sum (addition of parts into a whole), a difference (subtraction), a product (multiplication) or a ratio (division). There is a mathematical or formal relation between the two variables.

Since the percentages in the various parts of a whole sum to 100%, the inverse correlation between their respective shares is sometimes presented as a formal correlation (Sachs, 1989). This formal relationship seems more natural than man-made. But it was a human choice to focus on a particular whole and to measure the parts in relation to that whole. So, in that sense, this association is better classified as being man-made.

By their design or nature, correlations involving an intentional or man-made relationship often have extremely high values. While high values are not typical of coincidence, high values do not necessarily indicate natural causality. High correlations may

indicate human causality. Artificial correlations may still be worth modeling using regression. In other situations, one may want to exclude all such relationships from a model to focus on empirical causality instead of man-made relationships.

6. W*e should teach students to distinguish between natural and artificial (man-made) correlations.*

7. W*e should teach students to infer causality using appropriate procedures.* In experimental studies, Mills methods are applicable (Kelly, 1994). But in observational studies we lack the necessary control. In both kinds of studies, causality is not proven so much as other possibilities are rejected. Lothar Sachs (Sachs, 1984) presents a procedure involving elimination. First, eliminate pure chance, then eliminate man-made (or formal) causality, and finally eliminate common-cause natural causality. What is left over is direct natural causality between two things. Obviously, this is easier said than done.

## III. AGENCY AND CAUSALITY

The simplest examples of causal relations involve things which "act": human beings, living things or the forces of nature. In each case there is some form of agency that acts. In forming statements of determination, one can focus on the factor or on the model. Focusing on the factor as the agency of determination may unwittingly support the conclusion that the relationship is causal.

- "Sex explains 55% of the variability in wage."

Focusing on the model might withhold such unintended support.

- "Controlling on sex decreases the variability in wage 55%"

8. *We should teach students to focus on the act of modeling as the active agency in observational studies.* The emphasis should not be on the predictor variables as causal factors.

## IV. DIRECTIONALITY AND CAUSALITY

The form of the sentence involving determination can be suggestive of causality. Since causality is normally one-way, directional forms of sentences lend support to the idea of causality whereas bi-directional forms of statements are lacking in such support.

Directional statements of determination focus on one factor as appearing more potent. This apparent priority may be stated using either the active or passive voice as follows

- "Shoe size' explains' 60% of variability in height"
- "60% of the variation in height is 'explained by' shoe size"

Bi-directional statements of determination give no priority to either factor since they explicitly mention the symmetry involved. Examples of bi-directional statements include:

- "Given two factors, height and shoe size, 60% of the variability in either is explained by knowing the value of the other"
- 60% of the variability in either variable is 'associated with' (correlated with) the value of the other variable".

These bi-directional forms are beneficial because they do not imply that the correlation is directional; since causality is generally directional, this silence gives no support for presuming direct causality.

9. *We should teach students to be aware of how directional statements of determination may implicitly support the notion of causality*. We should give students alternative ways of making such statements that have a lesser implication of causality.

## V. EXPLANATION AND CAUSALITY

In most statistics texts, explanation and prediction are simply different aspects of the same relationship. If two variables have a correlation of .7, then we can form several statements:

- Knowing the value of the predictor <u>decreases</u> the variability in the predicted variable by about 50%.
- Knowing the value of the predictor <u>explains</u> about 50% of the variability in the predicted.

In the first case, the emphasis is on improving the prediction; in the second case, the emphasis is on the explanation. Mathematically prediction and explanation are perfectly symmetric. Once you have one, you automatically have the other. Of course there is some arbitrariness in how one allocates the quality of an association among more than one predictor. But in human affairs, prediction and explanation are not symmetric.

10. *We must teach students to recognize that less latitude is allowed in forming explanations than in forming predictions.* Since most students know this implicitly, perhaps teachers of statistics need to remind themselves of this fact.

For example, consider predicting the weather based on the height of a column of mercury -- a perfectly acceptable approach. But would we say that the height of a column of mercury explains the weather? Not really. Both the height of the mercury and the forthcoming weather are affected by a common cause -- the atmospheric pressure. In summary, we accept and use predictions based on common causes, but we require that explanations be as close to direct causality as is practical.

Thus, to speak of explanations as mirror images of predictions is to redefine a very fundamental term in our scientific language.

11. *We should teach students to identify a statistical 'explanation' as a mathematical or non-causal explanation.* This will help distinguish it from a causal explanation.

## VI. CORRELATION: CAUSE AND EFFECT.

In trying to understand causality, students often fail to distinguish between cause and effect in building models. Suppose one wants to model years of schooling using either parental education or current income. One might view parental education as a causal factor (psychologically) in predicting the amount of schooling completed by the child. One might view current income as a consequence of years of schooling such that more years of schooling are a cause of a higher income. It may be that current income has a higher correlation with years of schooling, but this does not imply that current income causes one's years of schooling. If students want to model a particular variable (schooling) as the effect, then there are some predictor variables (current income) that must be excluded. Until students recognize this, their models may have an indigestible mixture of causes and effects among the predictor variables.

12. *In cases of natural direct causality, we should teach students to consider whether one variable is a cause or effect in relation to another.* If the relation is natural, are we modeling a relation reflecting a common cause or is there a direct causality between the variables being measured? If there is a direct causality between the measured variables, what kind of predictive model is involved? Are we using effects to predict a cause, or are we using causes to predict an effect. Obviously a given model can be mixture of all of these.

## VII. PREDICTION AND CAUSALITY.

In discussing relationships, instructors reiterate that "correlation is not causality". In discussing regression, instructors use correlation as the basis for prediction and for explanation. This leaves students in a mental quandary. Correlation is often sufficient for a prediction or an explanation, yet causality is not necessary for a correlation. Does this mean that causality is not necessary for a prediction or an explanation? This issue is seldom addressed.

In one sense, causality is not necessary for successful prediction. In predicting the movements of masses in a gravitational fields, scientists can generate very accurate predictions. But they have almost no idea of how gravity works. Actuaries predict the frequencies of various events. Neither the scientists nor the actuaries know the causality involved. Nevertheless, in making predictions, both assume that the relationship will persist through time -- that the process is "under control". In this sense, correlation without some kind of causality is not a sound basis for modeling relations, for making predictions or for identifying explanations. Causal stability is absolutely necessary if our models, predictions and explanations are to be accurate in the future. We depend upon the ability of the observed relationships to persist through time to make accurate predictions. We depend upon observed relationships to persist through time to have a meaningful explanation of how things come to be. But in order for regularities to persist through time, they must involve things that have an identity or nature that persists through time. Past correlations without some kind of underlying causal identity are almost irrelevant in predicting the future.

Consider a case in the social sciences of a large correlation without much underlying causal stability. Suppose we had a 100% correlation. Suppose that in every election for President of the U.S., the taller (tallest) of the candidates had won the election. Without some underlying model of what causes this observed regularity, without some reason to envision this relationship as persisting, even though the correlation were 100%, we would have good reason to doubt its predictive power. We have little reason to think that height is so strongly associated with any of

the qualifications of being elected President. Thus our prediction about the next election might completely ignore height. Lacking some underlying causality, predicting the future based on a past correlation is just like concluding that all swans are white simply because we had never seen a black swan. Knowing that color is a very peripheral characteristic of biological organisms we have reason to doubt that our generalization is true. The underlying causal stability is lacking.

13. *We should teach students to reflect on the mechanism or nature involved in supporting a correlation.* What evidence do we have to say it will continue? The nature of the process may be unknown, but if the human mind is to have real knowledge, it is not enough to know that something is so; we need to know, to the extent possible, why it is so.

## VIII. DETERMINATION AND CAUSALITY

Common words used technically in statements of determination can be very suggestive of causality. These words can be either verbs or nouns. Consider the verb 'explains' in the sentence "Sex explains 60% of the variability in height". When asked how strongly various verbs implied causality, a group of statistics teachers[1] indicated that 'cause' was definitely causal, while other words such as 'decreases', 'indicates', 'predicts', 'implies', 'influences, 'affects', 'accounts for' and 'explains' are very ambiguous concerning causality. Although 'correlates' was considered the least likely to imply causality, there was still considerable variation among the respondents. In a statistical context, many of these words have technical meanings that are silent about causality. To indicate these technical usages of common words, authors of statistics texts often put the verbs 'accounts for' and 'explains'[2] in quotes to indicate their special technical status. But generally they do not indicate why the quotes are used. Students see most verbs of determination as implying causality.

The point is that the two groups (amateurs and professionals) view the same words differently. Students commonly use most of these verbs as causal; statisticians use most of these verbs technically as indeterminate toward causality. This difference between common usage and technical usage creates a great opportunity for mistakes by students, for deception by those who are unethical opportunists and for silence by professionals who don't see themselves as responsible for correcting the mistakes of others.

Nouns of determination can be very suggestive of causality. Nouns of determination that are ambiguous about causality include 'a factor', 'an influence', 'an explanatory variable' and 'a predictor'. Ambiguous nouns of determination have both a common and a technical meaning -- just like their counterparts in the verbs of determination. Technically, these ambiguous nouns of determination assert nothing about causality even though in common speech they generally imply causality.

Even within technical usage, 'factor' has some additional ambiguity. In experiments, 'factor' often refers to a causal relationship (c.f., ANOVA). But in observational studies in the social sciences, 'factor' has no causal implication.

14. *We should teach students to note when an ambiguous verb or noun has a causal connotation.* We should not eschew ambiguous verbs or nouns as unscientific; we should include such terms and teach students to be aware of their latent power to suggest causality.

## IX. ORTHOGONALITY AND CAUSALITY

Suppose that in a large random sample "55% of the variability in wage is accounted for (or explained by) by sex after controlling on age". And suppose that this relationship is indeed one of natural causality. Does it follow that sex must be the primary cause? Students often draw this conclusion. Their reasoning might be correct if all causes were in some sense "pure".

The simplest case of purity is where things are exclusive. Suppose that 40% of college students are juniors and seniors. Since these classes are exclusive, we are certain that most undergraduates are freshman and sophomores. One reason for presuming exclusivity in the statement of determination is the presumption that all causes should explain all of the variability. When things are classified into categories that are mutually exclusive and jointly exhaustive, then these categories will account for and contain 100% of the things being classified. Thus it appears that the statement of determination implies a causal model which is exclusive. But students quickly realize that

[1] Exploratory survey of statistics teachers at MAA STATS conference in Oshkosh Wisconsin, June 1995.

[2] 'Explanation' is not usually indexed in statistics books. Cf. Moore and McCabe (2nd Edition), Freedman et al (2nd Edition) and Iman (1st Edition).

causes are seldom exclusive. They quickly realize that purity between two variables involves something more complex than exclusivity.

Another form of purity between two variables is independence. If two variables are independent, then knowing the value of the first gives no additional information about the value of the second. Thus, it would seem that students are presuming that all other causes are independent of wages. Students usually recognize that dependency is the norm -- not the exception. Independence is the exception -- not the norm. Students may wonder why we use percentages which often identify the fraction of a whole to designate this relationship when the parts are not necessarily independent. Is the number really useful in identifying something as a cause?

A much more complex form of purity between two variables involves orthogonality -- the lack of correlation. When two variables are uncorrelated, they may be (and often are) independent. But orthogonality is not sufficient for independence. Since two orthogonal variables have no correlation, it is impossible for one to account for any part of the variability in the other. Thus, it seems that students are assuming orthogonality as a hidden premise. But if told, most students would have no better understanding of their error.

If told they were committing the fallacy of independence, students might grasp the idea of their error.

15. *We should teach students to realize that factors of determination are often correlated -- they are not generally independent. To assume orthogonality between factors is like assuming independence or exclusivity. Such assumptions must be justified. High wages and extensive experience are not exclusive; wages and years of experience are not independent and they certainly are not uncorrelated.*

Instead of focusing on orthogonality (independence or exclusivity), one could focus on non-orthogonality or correlation (dependence or inclusivity). One could mention explicitly that there are many other variables that can 'account for' or 'explain' a given reduction in variation.

16. *We should teach students to realize that controlling on a variable does not exclude the effects of any other variables that are correlated but uncontrolled.* Thus we should say that "55% of the variability in wage is determined by sex (or by any other uncontrolled variable (or cluster of variables) that is more correlated with wage)".

Consumers of statistics often conclude that a statistically significant correlation implies direct causality. Thus it seems unethical to hide behind our technical use of common terms. It would seem more ethical to state explicitly what the phrase "determined by" really means.

**SUMMARY**

Statistics is a tool for detecting causality. We must expand statistics to include causality as a central topic. And since causality is deeply embedded in our language, we should first teach our students to see when statistical words, phrases and sentences have causal connotations without necessarily asserting a causal relation. We must recognize that statistics is at least as much a problem of language as it is of mathematics. We must teach students in the social sciences to be aware of the extreme difficulties in concluding a relation is one of direct causality. Students are interested in causality. By including this topic we can link our statistical techniques to the issues our students want to address. Statistics was born at the boundary between mathematical probability and natural causality. Statistics began as a technique for identifying causal laws in the social sciences. Causality is the missing link in teaching statistics today. It is time to return to the task that motivated our founders to invent statistics in the first place, to study those aspects of causality that are discipline independent and to thereby help our students use statistics properly in the search for knowledge.

**REFERENCES:**

Connell, Richard J. (1973). Logical Analysis: A New Approach.

Kelley, David (1994). The Art of Reasoning. Second Edition.

Sachs, Lothar (1984). Applied Statistics A Handbook of Techniques, 2nd Edition. Springer-Verlag. P.393.

Schield, Milo (1994), "Random Sampling versus Representative Samples," ASA 1994 Proceedings of the Section on Statistical Education, P.107-110

**ACKNOWLEDGMENT:**

Dr. Bruce Reichenbach, Thomas V.V. Burnham and Gerald Kaminski made comments on earlier drafts.

Dr. Schield may be reached at schield@augsburg.edu.