Journal of Multidisciplinary Research, Vol. 5, No. 2, Summer 2013, 23-33.
ISSN 1947-2900 (print) • ISSN 1947-2919 (online)
Compilation Copyright © 2013 by St. Thomas University. All rights reserved.

Some Statistical Illusions and the Debate on Discrimination

Helgi Tomasson
University of Iceland

Abstract

Discrimination is a charged issue in many parts of the world. Frequently scientists assess the extent and nature of discrimination by means of data analysis. All data analysis rests on a bridge between a scientific model and a statistical model. In this article, the author reviews some simple issues that can give misleading results. The methodological literature on statistics explains the nature of these issues, but they might slip past the attention of scientists and policy makers with less formal training in statistics. In this article, the author illustrates this by means of some simple examples.

Keywords: statistical model, omitted variable bias, Simpson paradox, discrimination, affirmative action

Introduction

Discrimination is a highly debated issue in Western countries. Legislators and other authorities have been keen on passing laws and regulations on quotas and other types of affirmative action. The underlying motive is that some forces in society treat some groups unfairly. The underlying forces and reasons causing this unfairness are not always clear. There simply is a statement of unfairness, and that some action is necessary. A group of multidisciplinary specialists, lawyers, economists, sociologists, and others in administration and politics work together in forming this policy. The policy stands on data of uneven quality, and statistical modeling of the underlying process is prone to be fuzzy. Due to the heterogeneity of the group of policy makers, it is inevitable that the greatest common denominator for statistical expertise is likely to be low. Reporting of common statistical measures can easily generate some misconception as I show by illustrating some simple statistical facts. Although an experiment on possible gender discrimination in wages is the main line of reasoning behind this article, in principle, the examples apply to all applied statistical work. For readers with formal training in econometrics and statistics, the examples are trivial, but hopefully, eye opening for some.

In recent years, the fact that women as a group receive lower pay than men in most industrial countries has caused a debate concerning whether this is somehow unjust – that is, whether the labor market systematically discriminates against women. Many supporters of this statement base their views on quantitative data, surveys, tax-data, official statistics, etc. When making statements on the association of, say, gender and wages, one is doing statistical inference in a statistical model. An important property of the statistical model is that it should have some potential in replicating the observed data. In this article, I illustrate some plausible errors in inference by means of a few examples.

The first example illustrates that inference based on 2x2 tables that might arise in surveys of official statistics (e.g., U.S. Census Bureau) can be misleading. The point of the example is that it is necessary to take into account all-important variables simultaneously. Textbooks on econometrics and statistics state this clearly and unambiguously. The second example illustrates the importance of homogeneity of a group in statistical analysis. That is, that aggregation of heterogeneous groups can give misleading results. The third example shows the impact of measurement error in covariates, i.e., the presence of a measurement error usually results in biased estimates of the importance of the covariates. Finally, I mention the economic motivation for discrimination.

It is now more than 50 years since the publication of *How to Lie with Statistics* (Huff, 1954). This is a classic reference, and one issue of the 2005 volume of the journal *Statistical Science* discussed various types of statistical traps to celebrate its 50th anniversary (Best, 2005). The view that the difference in average wages between men and women confirms discrimination is perhaps one of the most widespread statistical illusions in modern times.

Some Statistical Issues

Two by Two Tables are Misleading

This author created the artificial data in Table A (in the Appendix) for the purpose of illustrating statistical issues that arise in the analysis of wage data. Here, the understanding is that gender=1 is male and gender=0 is female. There are two age groups: young (age=0) and old (age=1), two types of occupations (0/1), and two steps of seniority (not senior=0, senior=1). If one presents the average wages in a sequence of all possible two by two tables, one gets tables 1, 2, and 3. In all three tables, the males have higher wages in all groups, both occupation groups, both age groups, and both levels of seniority. For many people, it is tempting to infer that the labor market discriminates against women. This, however, is not necessarily the case.

Table 1
Average Wages by Gender and Occupation

	Occupation=0	Occupation=1
Males	77.000	122.142
Females	63.333	103.333

Table 2
Average Wages by Gender and Age

	Age=0	Age=1
Males	72.500	118.750
Females	61.250	97.500

Table 3
Average Wages by Gender and Seniority

	Seniority=0	Seniority=1
Males	78.750	152.500
Females	69.090	120.000

It is Necessary to take Account of all Important Covariates Simultaneously

The tables above are textbook material for illustration of statistical fallacies. All statistical analysis stands on some kind of statistical modeling. A statistical model is a simplified model of some reality, and the scientist uses data to infer about that reality. Statistical research consists of analyzing the properties of statistical methods, i.e., a pseudo real-world. The reference is *the true statistical model,* and the features of statistical methods stand on their properties for that model. If a method gives a biased or somehow distorted view of the true model, it shows a characteristic of the method. Some people who work with data do not realize this and claim they are not working with a model; they are just listening to data and letting data speak for themselves. A statistician does not accept this type of argument and claims they are working with some kind of model; they just do not know it. Someone who wishes to make a strong inference based on Table 1 is essentially basing this inference on a model in which the only covariates of interest for explaining average wages are gender and occupation. In this kind of model, one ignores other variables that might affect the wage formation. If occupation is important for average wages then it is necessary to include the occupation variable in all models concerning wages.

A fundamental tool in statistical analysis is the linear regression analysis. A simple version, with dummy (0/1) covariates, is of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

Here y_i can, for example, represent wages of individual i in a given period of time, x_{i1} represent the gender of individual i, x_{2i} represent the occupation type of the individual, and ε_i

represent the individual deviation from the expected wages of individual i. If this is the true model, one can try to guess the importance of the covariates by estimating the β_i's based on data. If, say, x_{i1} is the gender of an individual then one might interpret the coefficient β_1 as the impact of gender on wages, corrected for the factors $x_{i2}, \cdots, x_{ik}$. In the current example, one can do this. If x_{i2} is the occupation of an individual and x_{i3} is the seniority of an individual then one obtains the following:

$$\text{average wages} = 64.520 + 10.340x_1 + 16.270x_2 + 56.930x_3.$$

The interpretation is, that in the given period of time, a senior individual earns on average $56.930 more than a non-senior, occupation 1 earns $16.270 more than occupation 0, and males $10.340 more than females. Is it then correct to state that, when corrected for important covariates, the gender-pay difference is $10.340? No! There are more important covariates that affect wages, and it is necessary to include them in the model. If the age variable, x_4, is added to the model, one obtains the following:

$$\text{average wages} = 48.130 - 5.810x_1 + 24.640x_2 + 51.300x_3 + 44.330x_4.$$

This reversed the sign of the coefficient of the gender variable, so in this model, the females seem to earn more than males. Actually, this example is a textbook case for illustrating the impact of omitting important variables. The following formula generated the true data:

$$\text{wages} = 50.000 - 5.000x_1 + 20.000x_2 + 50.000x_3 + 40.000x_4 + 10.000x_3x_4,$$

In reality, females earned more than males, due to the realistic feature that the aging-effect is not constant over occupation groups (an occupation-age interaction term).

This example is, of course, a simplification of reality. The covariates take only two values (0/1). But, as a textbook example, it makes a point. It is a laboratory model with artificially generated data, in which favors females over males. A superficial statistical data analysis based on 2x2 tables or simple regression models easily can yield misleading results. The model builder always has to be concerned whether she has included all-important variables, and specified them correctly in the model. It is easy to construct examples in which all 2x2 tables will give misleading results.

Oaxaca (1973) describes a regression model for labor market data. His data set consisted of thousands of records. At that time (1973), working with so much data was a big achievement. The amount of data might lead some people to believe the results must be correct just because of the huge amount of data. A statistically-minded individual knows that perhaps some variables were missing from their model, and, therefore, the inference on male-female differentials might be biased. Even if there are many observations, measurements on important variables may be lacking. Indeed, some of the missing variables may be latent and impossible to measure directly. Therefore, methods for dealing with unobserved and unobservable variables are sometimes necessary. An important model design is the panel-data, or repeated measures design. Panel-data methods consist of statistical methods for models in which there are many observations of the same individual. Hausman and Taylor (1981) derive an improved methodology for panel-data analysis. They apply their method for filtering out unmeasured individual variables to a data set on men's wages from PSID (Panel-Study-of-Income-Dynamics). Greene (2003) shows their method and part of their numerical results in a textbook on econometrics. In Table 13.3 on page 306 in Greene's (2003) text, one can read the ordinary-least-squares method of estimating the model suggests the racial wage-gap is around 8.5%. The table also shows output of their new method, which suggests the racial wage-gap is around 1.8%. The message of this is that much of the reported wage discrimination might be a result of an improper statistical method.

Becker (1993) reacts to a report from the Boston Federal Reserve concerning the HMDA (Home Mortgage Disclosure Act) for monitoring minority access to the mortgage market. The authors of that report later published it in the *American Economic Review* (Munnell, Tootell, Browne, & McEneaney, 1996). The report suggests that, even after controlling for various external variables, results suggest discrimination against minorities. Becker's answer is to turn the problem around: If the banks discriminate against minorities then the banks should profit more from lending to minorities, which does not seem to be the case. The economic motive seems to be missing.

Another Nobel Prize winner analyzes the concept of discrimination. In his article, Detecting Discrimination, Heckman (1998) gives statements like the following.

> [C]areful reading of the entire body of available evidence confirms that most of the disparity in earnings between [B]lacks and [W]hites in the labor market of the 1990s is due to the differences in the skill they bring to the market, and not to discrimination in the market, and (p. 101)...the evidence from the current U.S. labor market is that discrimination by employers alone does not generate large economic disparities... (p. 112)

In the years before 1970, there was a certain tendency in econometrics to search for a large global model that explained everything. One of the skeptics of this approach was C. W. J. Granger, the Nobel Prize winner in economics for 2003. In the early 1970s, he lectured on the concept of spurious regression (Granger & Newbold, 1974). This article was basically a reinvention and improvement on the argument made by Yule (1926). The general idea is that unrealistic assumptions in the statistical model reduce the validity of the outcome. In the time-series models Granger discussed, the assumptions concerned issues like stationarity. Textbooks of econometrics and statistics, e.g., (Greene, 2003), typically assume a correctly specified model, homogenous underlying groups and covariates without measurment error. It took the economic profession a long time (10 to 20 years) to appreciate Granger's skeptical thoughts. Even now, people in economics, and perhaps in other disciplines, are practicing spurious-regression—regressing time-series data without specifying a particular time-series model, obtaining high R^2, and wrongly interpreting this as a highly significant (and important) result.

Aggregation of Heterogeneous Groups can Give Misleading Results

It is easy to construct an example in which all firms discriminate against men, but the aggregate seems to discriminate against women. The following example (with these numbers) occurs in many textbooks (Poirier, 1995; Lancaster, 2004), although the accompanying stories are different. Tables 4 and 5 show that the A- and B-type firms discriminate against men. However, when aggregated, the A+B-firm discriminates against women.

A teacher in biostatistics reviewed this example in her lecture. The story was that in a certain population the carriers of a certain gene were more likely to acquire a disease. However, the population was heterogeneous and consisted partly of Native Americans (Indians), and partly of Americans of European origin. Among the Native Americans, the gene carrier was not more likely to acquire the disease. Likewise, among European Americans, the gene carrier was not more likely to acquire the disease. It just so happened that both the gene and the disease were more common in the European American population. It would have been a wrong policy to use the aggregated data to take some action based on whether an individual carried that gene or not.

Table 4

Wages in Firm A. 70% of the Females have High Wages, 60% of the Males have High Wages

	High wages	Low wages
Males	18	12
Females	7	3

Table 5

Wages in Firm B. 30% of the Females have High Wages, 20% of Males have High Wages

	High wages	Low wages
Males	2	8
Females	9	21

Table 6

Wages in Firm A+B. 40% of Females have High Wages, 50% of Males have High Wages

	High wages	Low wages
Males	20	20
Females	16	24

The Impact of Measurement Error

It is wrong to ignore the presence of measurement error just because every individual is equally likely to be subject to it. Many textbooks in econometrics (e.g., Greene, 2003) show that a measurement error in a covariate will lead to biased and inconsistent estimates, i.e., the bias will not disappear by increasing sample size. A measurement error in one covariate will in general affect the estimates of the impact of all other covariates in a regression model.

If we assume 30% of males have a senior job, and 5% of females have a senior job, and further assume the probability of a wrong classification of a senior worker is 30% and the probability of a wrong classification of a non-senior worker is 5%. Further, assume that Table 7 shows the true wage distribution. Then one can use Bayes rule to derive the observed distribution in Table 8. Here, the truth for both males and females is that the senior worker earns twice as much as a non-senior worker. The truth is gender neutral. The measurement-error (classification error) is completely gender neutral, yet the observed values are likely to report a difference in gender among both seniors and non-seniors. Someone might conclude the reward for seniority is much smaller for females than for males.

This type of statistical calculation is well known in the medical literature. The medical literature uses the terms "sensitivity," the proportion of positives the test correctly classifies, and "specificity," the proportion of negatives the test correctly classifies. The properties of the predictive value of the test, i.e., the probability of having a disease, when the test is positive, is something completely different (Altman, 1991).

Measurement errors are a natural phenomenon in all practical data observations. For example, in the 1960 U.S. Census, there are 62 women age 15-19 who have more than 12 children, and widows below the age of 14 are common (De Veaux & Hand, 2005). Textbooks in econometrics and statistics suggest various solutions (e.g., the instrumental technique, such as Greene, 2003). The treatment of measurement error in general is a complicated issue and requires sophistication on behalf of the scientists. A text on measurement error models is Fuller (1987).

Table 7
True Wage Relations between Senior and Non-Senior Workers

	Seniors	Non-seniors
Males	2	1
Females	2	1

Table 8
The Relations between Senior and Non-Senior Workers of Table 7 Observed with Measurement Error

	Seniors	Non-seniors
Males	1.86	1.12
Females	1.42	1.02

Conclusion

When debating discrimination, one needs a clear definition. The economic intuition of, say, Becker (1971) tells us that, even if the intention to discriminate is there, the discrimination cannot be very widespread. The scope for employers to practice discrimination is particularly limited. In a free market with easy access to information, employers maximizing their profit will try to minimize labor cost; therefore, the demand for workers who are willing to work for low pay will rise. These workers, in turn, will be able to raise their wages. Marriage is an important institution. Partners will decide how to allocate their participation in the labor market. Historically, the nature of this allocation has been such that the male partner has much higher income. Some authors have sought explanations for this (e.g., Korenman & Neumark, 1991). O'Neill and O'Neill (2005) state the following.

> There is no gender gap in wages among women and men with similar family roles. Comparing the wage gap between men and women age 25-43 who have never married

> and never had a child, we find a small observed gap in favor of women, which becomes insignificant after accounting for differences in skills and job and workplace characteristics. (p. 34)

The marriage is important. Farrell (2005) cites facts from the U.S. Census Bureau in the 1950s to show there was very little difference in the earnings of unmarried women and unmarried men between 45 and 54 (p. xxi). This period took place well before the introduction of *affirmative action* and the Equal Pay Act of 1963. Farrell (2005) gives a follow-up of these results in modern times, and the pattern seems to be the same. Of course, one might claim that unmarried women are somehow different from unmarried men, i.e., that one needs to model the marriage selection process. This kind of modeling is likely to be a complicated statistical exercise and beyond the scope of this article. The professional literature on econometrics and statistics is typically of a high mathematical level. The research in that field is highly focused on mathematically proving asymptotic properties of estimators. This makes the field inaccessible to many professionals educated in the social sciences; perhaps, therefore, policy makers with varying educational backgrounds might fall victim to simple statistical traps. At least, it seems to be a reasonable doubt about the benefit of affirmative action policy. Becker (1971) also suggests that it is of dubious social value to reward or punish employers for discrimination (beyond other criminal conduct). Discrimination is not in their economic control. The lawyer, Farrell (2005), asks the compelling question whether the main impact of the Equal Pay Act of 1963 was to earn a living for a group of lawyers (he claims to be one of them). Still legislators carry on, and O'Neill (2010) bursts out: "Washington's Equal Pay Obsession. There's no epidemic of gender discrimination. So why is Congress proposing another law?" (http://online.wsj.com/article/SB1000 1424052748703326204575616450950657916.html). A probable cure for the debate might be more cooperation across academic disciplines. Professionals in econometrics and statistics have to give more respect and reward to ordinary statistical work, and other groups have to be more open to accepting statistical advice.

References

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall.

Becker, G. S. (1971). *The economics of discrimination* (2nd ed.). Chicago: University of Chicago Press.

Becker, G. S. (1993). Evidence against banks does not prove bias. *Bloomberg Businessweek: Businessweek Archives*, http://www.businessweek.com/stories/1993-04-18/the-evidence-against-banks-doesnt-prove-bias, May 31st, 2013.

Best, J. (2005). Lies, calculations, and constructions: Beyond how to lie with statistics. *Statistical Science*, *20*(3), 210-214.

De Veaux, R. D., & Hand, D. (2005). How to lie with bad data. *Statistical Science*, *20*(3), 231-238.

Farrell, W. (2005). *Why men earn more*. New York: Amacom.

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley & Sons.

Granger, C., & Newbold, P. (1974). Spurious regression in econometrics. *Journal of Econometrics*, *2*, 111-120.

Greene, W. (2003). *Econometric analysis* (5th ed.). New York: Prentice Hall.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, 49(6), 1377-1398.

Heckman, J. J. (1998). Detecting discrimination. *Journal of Economic Perspectives*, *12*(2), 101-116.

Huff, D. (1954). *How to lie with statistics*. London: Penguin Books.

Korenman, S., & Neumark, D. (1991). Does marriage really make men more productive? *The Journal of Human Resources*, 26(2), 282-307.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Oxford: Blackwell Publishing.

Munnell, A. H., Tootell, G. M. B., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, 86(1), 25-53.

Oaxaca, R. (1973). Male-female wage differentials in urban labour markets. *International Economic Review*, *14*(3), 693-709.

O'Neill, J. E. (2010, November 16). Washington's equal pay obsession: There's no epidemic of gender discrimination. So why is Congress proposing another law? *Wall Street Journal*, European Edition, http://online.wsj.com/article/SB10001424052748703326204575616450950657916.html

O'Neill, J. E., & O'Neill, D. (2005, April). What do wage differentials tell us about labour market discrimination. *National Bureau of Economic Research Working Paper Series.* Technical Report 11240, http://www.nber.org/papers/w11240

Poirier, D. J. (1995). *Intermediate statistics and econometrics: A comparative approach*. Cambridge: The MIT Press.

Yule, G. (1926). Why do we sometimes get nonsense-correlation between time-series? – A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89, 1-63.

Appendix

Table A1
Data for Illustrative Examples

Gender	Occupation	Age	Seniority	Wages
1	1	1	1	165
1	1	1	1	165
1	1	1	1	165
1	1	1	0	115
1	0	1	0	85
1	0	1	0	85
1	0	1	0	85
1	1	0	1	115
1	1	0	0	65
1	1	0	0	65
1	0	1	0	85
1	0	0	0	45
0	0	0	0	50
0	0	0	0	50
0	0	0	0	50
0	0	0	0	50
0	0	1	0	90
0	0	1	0	90
0	0	0	0	50
0	0	0	0	50
0	0	1	0	90
0	1	1	0	120
0	1	0	1	120
0	1	0	0	70

About the Author

Helgi Tómasson (helgito@hi.is) has a B.S. degree in applied mathematics from the University of Iceland and a Fil.Dr. (Ph.D.) in statistics from the University of Gothenburg, Sweden. He has been a fellow at the department of Biostatistics at the International Agency for Research on Cancer in Lyon, France. He has been a director of an institute of labor market research in Reykjavik and now holds a tenured position as a professor of econometrics and statistics at the Faculty of Economics at the University of Iceland, Reykjavik, Iceland. He has been a visiting scholar in Denmark, England, Sweden, and the USA. He has authored articles in scientific journals. He is interested in statistical computing, and is the author and maintainer of the statistical package ctarma=continous-time-auto-regressive-moving-average.

Discussion Questions

1. It should be clear that affirmative action is not free. Is it possible to do a cost-benefit analysis of affirmative action?

2. How widespread are the statistical fallacies described in the text in scientific work?

3. The equal-pay equal-rights industry employs how many?

4. Do politicians use statistical illusions to buy votes?

To Cite this Article

Tomasson, H. (2013, Summer). Some statistical illusions and the debate on discrimination. *Journal of Multidisciplinary Research, 5*(2), 23-33.