ACADEMIC PRESS, INC.
1250 Sixth Avenue, San Diego, CA 92101

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
24-28 Oval Road, London NW1 7DX

Library of Congress Cataloging-in-Publication Data

Gastwirth, Joseph L.
Statistical reasoning in law and public policy.

(Statistical modeling and decision science)
Includes bibliographies and indexes.
Contents: v. 1. Statistical concepts and issues of fairness—v. 2. Tort law, evidence, and health.
1. Forensic statistics—United States. 2. Statistics.
I. Title. II. Series.
KF8968.75.G37 1988 349.73'021 87–33367
ISBN 0–12–277160–5 (v.1) 347.30021
ISBN 0–12–277161–3 (v.2)

Printed in the United States of America
88 89 90 91 9 8 7 6 5 4 3 2 1

Contents

TABLE 6.24. Wiring Configurations at Homes of all Cases and Controls Stratified by Traffic Level

	HCC	LCC	All	Fraction Exposed
		Heavy Traffic Areas		
Cases	39	35	74	.527
Control	14	34	48	.298
Total	53	69	132	.434
		Other Areas		
Cases	146	271	417	.350
Control	87	335	424	.210
Total	235	606	841	.279
		Simple Aggregation of the Data Ignoring Traffic Level		
Case	185	306	491	.377
Control	103	359	462	.223
Total	288	665	953	.302

Source: Table 9 of Wertheimer, N. and Leeper, E. (1979), the source of Table 6.23.

traffic is that the odds ratio 2.16 is a more accurate estimate of the effect of exposure to HCC. It should be mentioned that the association found in the study is in dispute. A study[75] in Rhode Island did not find a similar association, while two other case-control studies[76] and several other health studies[77] did.

This particular study illustrates several basic issues in case control studies:

(a) Finding all cases or a representative sample of them.
(b) Finding a suitable (comparable) control group.
(c) Assessing the exposure status (for this study exposure to ELF was not directly measured, proximity to HCC wiring was used).
(d) Accounting for possible confounders or alternative explanation of the observed elevated risk of exposed individuals.

As a general principle it is best to rely on more than one epidemiologic study (of either or both types) before concluding that exposure to a substance is related to getting a particular disease. In chapter 10 we will discuss statistical techniques which are useful in analyzing data counting the number of events, each of which has a small probability of occurring, that happen over a period of time and will apply them to cohort or prospective studies. We close this chapter with a fundamental result,[78] due to Cornfield, concerning the potential effect of a possible confounder that is not considered in a study.

Suppose we compare the disease rate of an exposed group (II) with a control group (I) and find a relative risk of R (e.g., 3). A skeptic asserts that a new factor (confounder) X really caused the observed relative risk and that really there is no association. If the new agent X acts on the disease producing mechanism independently of exposure, then in order for it to fully explain the observed relative risk of R (which is greater than 1) two conditions must hold

(a) The relative risk (RX) associated with factor X must exceed R (the risk found in the study), and

(b) The fraction (F) of persons in the exposed group (II) on whom factor X operates (e.g., who were exposed to X or inherently possess it if it is age, sex) must be *at least R times* the corresponding fraction of the control group.

Suppose we observe a relative risk of 5 when studying the association of a chemical exposure and a worker subsequently getting cancer of the renal pelvis after stratifying the data to account for smoking behavior of the study groups. A trade association challenges this conclusion asserting that a greater proportion of the exposed population drank alcoholic beverages, which caused the disease, rather than the chemical. The previous result implies that in order for this assertion to fully explain the result of the study:

(a) Drinking alcoholic beverages increases one's chance of getting the cancer by 5 (which is not the case), and

(b) The fraction of persons who drink in the exposed group is at least 5 times that of the control group.

Notice that if more than 20% of the control group drinks alcoholic beverages then (b) implies that alcohol consumption *cannot* explain the observed risk of 5, since the fraction of exposed persons who drink would have to exceed 1.0, which is impossible.

Therefore, large and statistically significant values of odds ratios estimated from studies which account for the major potential confounding variables (e.g., age, smoking) usually indicate a real effect of exposure on one's chances of getting the disease studied. The magnitude of the estimated association (relative risk) may change somewhat when other factors which have a small influence are included.

On the other hand, had our hypothetical study on cancer of the renal pelvis estimated a relative risk of 2.0 but failed to adjust for smoking,

which has a relative risk of about 7.6 for males and 5.8 for females,[79] then condition (a) is satisfied and it is conceivable that 20–25% of the control group smoked while 45–50% of the exposed group smoked. Thus, the observed relative risk of 2.0 might have been due to the different smoking patterns of the cases and controls rather than exposure to the chemical.

Problems

1.* When we stratified the case control data to account for the effect of living near heavy traffic, we noticed the risk of cancer was higher for persons exposed to HCC and living near heavy traffic than for persons only exposed to HCC, as the odds ratios were 2.72 vs. 2.03. On the other hand, the p-values of the test for the equality of the exposure fractions gave a smaller and hence more significant p-value for the data referring to persons not living near heavy traffic. Is there an error in our calculations? If not, explain our results.

2.* Suppose one used the data in Table 6.24 to estimate the relative risk of exposure to HCC for persons in high traffic areas by computing

$$\frac{\text{fraction of cases exposed}}{\text{fraction of controls exposed}} = \frac{.527}{.298} = 1.77.$$

Why is this computation incorrect?

3. Suppose that probability of a child getting cancer before 19 is 1 in a 1000 and that the relative risk due to living near a high current power line is 2.0.

(a) What other information would you need to decide whether these power lines pose a serious public health problem?

(b) Find some relevant data and/or make some reasonable assumptions to enable you to estimate the number of childhood cancers that might be due to exposure to high current power lines.

(c) Suppose it were possible to reroute the major power lines in the nation which would lead to half as many children being exposed as are today. How many lives would be saved each year? What factors might enter into the ultimate decision?

4.* It is usually possible to suggest an extra factor or variable that might be related to success which was not used in an analysis. Show that Cornfield's observation on how strong and prevalent this factor must be in the exposed group relative to the control group applies in the following EEO settings to support the findings made:

(a) In *Boykin*, plaintiff's showed (Table 6.6) that blacks had half the probability of whites of receiving the better initial assignments after consideration of prior experience. Suppose the defendant asserted that a higher proportion of whites had a college education and this meant they had more promise as an employee. Is it plausible that this omitted factor could explain why whites had *twice* the chance of blacks of receiving the better job?

(b) In *Agarwal*, we noted that minorities had 70% of the odds of promotion as majority employees. How plausible is it that differences in the occupations held by the two groups could explain this difference?

Answers to Selected Problems

1. Remember that statistical significance depends on the sample size. Only 132 subjects in the study lived near high traffic routes, while 841 lived in other locations. Therefore, the statistical fluctuation or standard deviation of the difference between the exposure fractions is much smaller for the second set of data, enabling us to detect a smaller difference as significant.

2. The relative risk, R, we are interested in is the

$$R = \frac{\text{probability an exposed person gets cancer}}{\text{probability an unexposed person gets cancer}},$$

not the ratio of the exposure probabilities. Because the probabilities of getting a particular disease are small, the R is well approximated by the odds ratio and the odds ratio (and only the odds ratio) is the same for the prospective and retrospective organization of the data in a follow-up study.

4. The jobs at issue in *Boykin* were relatively low-skilled blue collar jobs which usually are not sought after by college graduates. Thus, the fraction of applicants who completed college was probably quite small. Moreover, college education is not likely to double one's chances of being a capable employee, especially among persons who had prior job experience. Thus, this explanation lacks credibility.

APPENDIX. Cornfield's Condition for a New Factor to Explain an Observed Association

This appendix presents a formal proof of the result of Cornfield (1959), which stated that a new agent or factor, X, which was not considered in a study that found a relative risk, $R_0 > 1$, between persons exposed to the agent studied relative to an unexposed or control group would need to possess a relative risk,

$$R_X \geq R_0 \tag{6A.1}$$

and prevalence ratio, $\theta > R_0$, in order to have caused the observed relative risk, R_0. Recall that the prevalence ratio is the ratio of the fraction, f_2, of persons exposed to the original study agent who were also exposed to agent X to the fraction, f_1, of persons in the control (unexposed) group who were exposed to agent X.

Let p denote the small probability a person not exposed to the study agent or new agent X has of contracting the disease. Now, the $100f_1\%$ of the control group who were exposed to agent X have probability pR_X of getting the disease, while the remaining fraction $(1 - f_1)$ have probability p.

By the total probability theorem (2.11), the probability a member of the control group contracts the disease is

$$P_I = f_1R_Xp + (1 - f_1)p. \tag{6A.2}$$

Similarly, the probability a member of group II contracts the disease is

$$P_{II} = f_2R_Xp + (1 - f_2)p. \tag{6A.3}$$

Notice that in deriving expressions (6A.2) and (6A.3), we assume that the study agent has no effect, so persons who are not exposed to agent X have the natural probability, p, of contracting the disease.

Now the observed relative risk, R_0, is

$$R_0 = \frac{P_{II}}{P_I} = \frac{f_2pR_X + (1 - f_2)p}{f_1pR_X + (1 - f_1)p} = \frac{R_Xf_2 + (1 - f_2)}{R_Xf_1 + (1 - f_1)}. \tag{6A.4}$$

The *largest* possible value of R_0 occurs when the numerator $R_Xf_2 + (1 - f_1)$ is large and the denominator $R_Xf_1 + (1 - f_1)$ is small. Since $R_X > 1$, this happens when $f_2 = 1$ and $f_1 = 0$. Thus, the *largest* possible value of R_0 is R_X, i.e., $R_0 \leq R_X$.

Next we need to determine the ratio θ of the fraction, f_2, of persons exposed to the study agent (group II) who were also exposed to agent X to the corresponding fraction, f_1, of persons not exposed to the study agent but were exposed to agent X, that would be required for agent X to have caused the observed relative risk. Letting $f_2 = \theta f_1$, expression (6A.4) becomes

$$R_0 = \frac{\theta f_1R_X + 1 - \theta f_1}{f_1R_X + (1 - f_1)} = \frac{\theta f_1(R_X - 1) + 1}{f_1(R_X - 1) + 1}. \tag{6A.5}$$

Multiplying both sides of (6A.5) by $[f_1(R_X - 1) + 1]$ yields

$$R_0f_1(R_X - 1) + R_0 = \theta f_1(R_X - 1) + 1$$

or

$$R_0 - 1 = (\theta - R_0)(R_X - 1)f_1,$$

and, finally,

$$\theta = R_0 + \frac{R_0 - 1}{(R_X - 1)f_1}. \tag{6A.6}$$

Since both R_0 and R_X are greater than 1, as we found an association between the original agent studied and the disease, the second term

$$\frac{(R_0 - 1)}{(R_X - 1)f_1}$$

is positive, so the prevalence ratio, θ, must exceed R_0.

Formula (6A.6) then tells us more than the statement of Cornfield, as it enables us to calculate the prevalence of agent X in the original exposed group that is required to explain the observed association from its prevalence, f_1, in the control group.

Note that in formula (6A.6) the prevalence ratio, θ, decreases as R_X increases. This makes sense because the greater the association between the new factor X and the disease, the less prevalent it needs to be in order to cause the observed disease rate.

Problems

1. Show that in order for a new agent to explain an observed relative risk, R_0, its prevalence ratio lies between R_0 and $R_0 + 1/f_1$.

2. Show that if $R_0 = 3$ and f_1 is at least 1/3, then *no* new agent X can explain the observed association.
Hint: Show that f_2 would have to exceed 1.0, which is impossible.