14
Confounding and standardization

14.1 Confounding

Epidemiological studies generally involve comparing the outcome over a period of time for groups of subjects experiencing different levels of exposure. Such studies are usually not controlled experiments but 'experiments of nature' of which the epidemiologist is a passive observer. In such investigations, there is always the possibility that an important influence on the outcome, which would have been fixed in a controlled experiment, differs systematically between the comparison groups. It is then possible that part of an apparent effect of exposure is due to these differences, and the comparison of the exposure groups is said to be *confounded*. Statistical approaches to dealing with the problem of confounding aim to correct, during analysis, for such deficiencies in the design of experiments of nature.

A particularly important potential confounding variable (or *confounder* in many epidemiological studies is the age of subjects. We shall consider an example in which subjects in a follow-up study are classified according to whether their age at the start of follow-up was less than 55 years or 55 years or more. Suppose that the breakdown between the two age groups is $0.8 : 0.2$ and that the conditional probability of failure is 0.1 in the first age group and 0.3 in the second. When age is ignored the overall or *marginal* probability of failure is

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14.$$

Now suppose that the age distribution differs between the two exposure groups, being $0.8 : 0.2$ in the not exposed group but $0.4 : 0.6$ in the exposed group (see Fig. 14.1). The marginal probability of failure for the unexposed group is still

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14,$$

but for the exposed group it is now

$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22.$$

The marginal probabilities of failure now suggest an apparent effect of

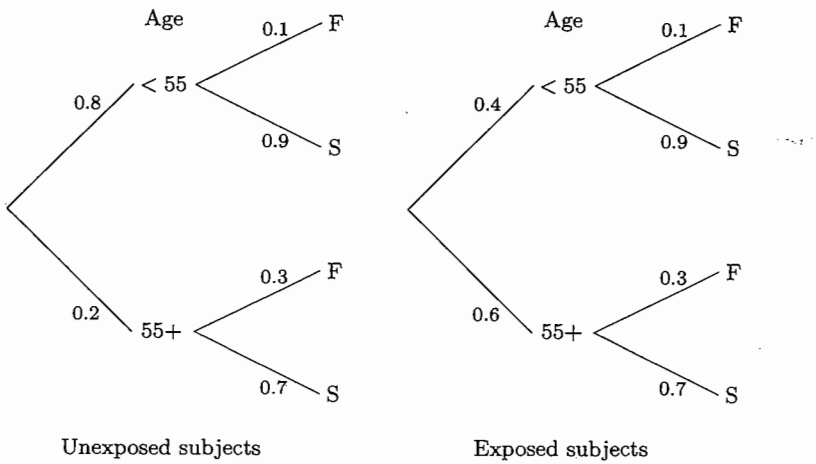


Fig. 14.1. Confounding by age.

exposure, but this is entirely due to the difference in age distributions between the exposed and unexposed subjects.

In this example the apparent effect of exposure is entirely due to age differences but confounding may also be partial, acting either to exaggerate or to dilute a real relationship. As an example of this, suppose the effect of exposure is to raise the probability of failure from 0.1 to 0.2 in the younger age group and from 0.3 to 0.5 for older subjects. When the age distribution is 0.8 : 0.2 in both exposure groups the overall effect of exposure is to increase the marginal probability of failure from

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14$$

in the unexposed group to

$$(0.8 \times 0.2) + (0.2 \times 0.5) = 0.26$$

in the exposed group. When the age distribution is 0.8 : 0.2 in the unexposed group and 0.4 : 0.6 in the exposed group the overall effect of exposure is to increase the marginal failure probability of failure from

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14$$

in the unexposed group to

$$(0.4 \times 0.2) + (0.6 \times 0.5) = 0.38$$

in the exposed group. Thus the overall effect of exposure appears greater

when the age distributions differ than when they are the same.

These examples demonstrate that a third variable, such as age, can distort the relationship between an exposure and failure provided it is related to both exposure and failure. This dual relationship is often taken as the definition of a confounder. However, although it is a necessary condition for a variable to be a confounder, it is not sufficient: a confounder must also be a variable which would have been held constant in a controlled experiment. For example, in perinatal epidemiology, we might ask whether birthweight could be regarded as confounding the relationship between the receipt of proper antenatal care and the risk of perinatal death. Although birthweight is related to both antenatal care and perinatal risk, it cannot be regarded as a confounder since one of the *results* of successful antenatal care should be adequate birthweights. Since it would not make sense to envisage an experiment in which we varied the provision of antenatal care while maintaining the distribution of birthweight constant, differences in birthweight distribution cannot be regarded as a deficiency in the design of the experiment of nature. It is not, therefore, a confounder.

14.2 Correction for confounding

The linking of confounding to an imaginary experiment helps to clarify the ideas which lie behind statistical methods for dealing with the problem. There are two rather different approaches, and these closely mimic the ways in which extraneous influences are dealt with in experimental science.

The classical approach to experimentation is to hold constant all influences other than the experimental variable(s) of interest. For example, to avoid confounding by age, we would simply compare failure risks in exposed and unexposed subjects *of a fixed age* or, at least, falling within a narrow range of ages. The statistical comparison would then be of failure probabilities conditional upon age. The same comparison can be made in an non-experimental study by the analytical strategy called *stratification*. By dividing (or stratifying) the data according to age, the single experiment of nature in which age has not been adequately controlled is transformed into a series of smaller experiments within which age is closely controlled. The analysis then compares probabilities of failure between exposure groups within age bands. However, a consequence of this strategy is that individual strata may contain too little data to be informative on their own. The more finely we stratify the data, the more closely we control for confounding, but the sparser our data becomes within strata. This impasse may only be broken by making the further assumption that the comparisons estimate the same quantity within each stratum, and then combining the information from the separate strata. We shall defer further discussion of this approach to Chapter 15.

Holding extraneous variables constant is not the only model for good ex-

perimentation, although it is certainly the most familiar. In the twentieth century, experimentation has become a valuable tool in fields of study such as biology, in which such close control of experimental material and conditions is not possible. The idea of *randomization* has been central to this development; if we cannot ensure that experimental groups are identical in all important respects, then by assigning subjects to groups *at random*, we ensure that the probability distributions for extraneous variables do not differ between exposure groups. Comparisons between the groups can then be safely made.

Returning to the comparison of failure probabilities between exposure groups, it is rarely possible, in epidemiology, to use randomization to ensure that extraneous variables have equal distributions in the different exposure groups. However, it is possible to take account of differences in the distribution of a specific variable, such as age, by predicting the outcome for exposure groups which have the same age distribution. This is done by first estimating the age-specific probabilities of failure for each exposure group, and then using these to predict the marginal probabilities of failure for exposure groups which have a standard age distribution. This forms the basis of the second statistical approach to dealing with confounding, known in epidemiology as *direct standardization.*

14.3 Standardized rates

The remainder of this chapter concerns the use of direct standardization to compare *rates.* Since rates are probabilities per unit time they can be compared in the same way as failure probabilities. Age-specific failure rates are estimated for each of the groups being compared, and these are used to predict the marginal rates which would have been observed if the age distributions in the comparison groups had been the same as the standard age distribution. These estimates are called *standardized rates.*

The choice of the age distribution to use for standardization depends on the purpose of the analysis. It is quite common for the overall distribution of age, added over exposure groups, to be used as the standard, thus simulating the results of an experiment in which the total study group was randomly allocated between exposure categories. However, if one of our aims is to facilitate comparisons with other published studies, it is more useful to use an age distribution which is in general use. Several distributions are commonly used for this purpose. One is the age distribution of the world population, another is the age distribution for developed countries. Since there is no 'correct' standard there is much to be said in favour of using a *uniform* age distribution where the percentage falling in each age group is the same. One advantage of using a uniform age distribution is that the standardized rate is then directly proportional to the *cumulative rate* for a subject experiencing the age-specific rates from the study

Table 14.1. IHD incidence rates per 1000 person-years

	Exposed (< 2750 kcal)			Unexposed (≥ 2750 kcal)		
Age	Cases	P-yrs	Rate	Cases	P-yrs	Rate
40–49	2	311.9	6.41	4	607.9	6.58
50–59	12	878.1	13.67	5	1272.1	3.93
60–69	14	667.5	20.97	8	888.9	9.00
Total	28	1857.5	15.07	17	2768.9	6.14

throughout life.

Direct standardization is most commonly used when comparing quite large groups, such as the populations of different countries or regions. When used with less extensive data it will yield statistically unreliable estimates if some of the age-specific rates, although based on very few cases, receive appreciable weight in the analysis.

To illustrate the technique of direct standardization we shall return to study of ischaemic heart disease and energy intake, discussed in Chapter 13. The incidence of ischaemic heart disease in the exposed group (low energy-intake) is 15.1 per 1000 person-years while the rate in the unexposed group is 6.1 per 1000 person-years. These rates, which take no account of any possible confounding effect of age, are often referred to as *crude* rates to distinguish them from standardized rates.

Table 14.1 shows the data stratified by 10-year age bands. The age distribution is different in the two exposure groups; this may be seen by converting the person-years to a proportion of the total person-years in each group giving 0.168, 0.472, and 0.359 in the three age bands for the exposed (low energy-intake) group and 0.210, 0.459, and 0.321 for the unexposed (high energy-intake) group. These age differences might explain some of the difference in the crude IHD incidence rates.

Using the uniform age distribution as standard, our estimate of the marginal rate for a group of exposed subjects with a uniform age distribution is

$$(0.333 \times 6.41) + (0.333 \times 13.67) + (0.333 \times 20.97) = 13.67$$

per 1000 person years and, for a group of unexposed subjects with a uniform age distribution, it is

$$(0.333 \times 6.58) + (0.333 \times 3.93) + (0.333 \times 9.00) = 6.50$$

per 1000 person-years. The standardized rates for the two groups are therefore 13.7 and 6.5 per 1000 person-years. These do not differ greatly from the crude rates of 15.1 and 6.1 per 1000 person-years, showing that the

confounding effect of age is small in this case.

Exercise 14.1. Find the standardized rates for the exposed and not exposed groups using as standard the age distribution with probabilities of 0.2, 0.5, and 0.3 in the three age bands.

⋆ 14.4 Approximating the log likelihood

When there are three age bands, as in the IHD and energy example, the standardized rate parameter takes the form of a weighted sum of the age-specific rate parameters,

$$W^1\lambda^1 + W^2\lambda^2 + W^3\lambda^3,$$

where

$$\lambda^1, \lambda^2, \lambda^3$$

are the rate parameters for the age bands and

$$W^1, W^2, W^3$$

are the probabilities of the standard age distribution. Since λ^1, λ^2 and λ^3 have independent log likelihoods, we can use the ideas introduced in section 13.4 and Appendix C to derive a Gaussian approximation to the profile log likelihood for the standardized rate. The most likely value is

$$W^1M^1 + W^2M^2 + W^3M^3$$

where $M^1 = D^1/Y^1$ is the most likely value of the age-specific rate parameter in band 1, and similarly expressions hold for bands 2 and 3. The standard deviation of the Gaussian approximation is

$$\sqrt{(W^1S^1)^2 + (W^2S^2)^2 + (W^3S^3)^2}$$

where $S^1 = \sqrt{D^1}/Y^1$ is the standard deviation of the Gaussian approximation to the log likelihood for λ^1, again with similar expressions for bands 2 and 3.

For the IHD and energy example the proability weights are

$$W^1 = W^2 = W^3 = 0.333.$$

The age-specific rate for the first age band of the exposed group is 6.41 and the corresponding standard deviation is

$$\sqrt{2}/311.9 = 0.00453,$$

or 4.53 per 1000 person-years. The most likely values for the rates in the other two age bands are 13.67 and 20.97 with standard deviations 3.94 and 5.61 per 1000 person-years. The standard deviation of the standardized rate is therefore

$$\sqrt{(0.333 \times 4.53)^2 + (0.333 \times 3.94)^2 + (0.333 \times 5.61)^2} = 2.74$$

per 1000 person-years.

Exercise 14.2. Show that the standard deviation of the standardized rate for the unexposed group is 1.63 per 1000 person-years.

LOG TRANSFORMATION OF STANDARDIZED RATES

Just as for any other rate, Gaussian approximations to the log likelihood are more accurate when related to the *log* of the standardized rate. The most likely value on the log scale is, of course, just the log of the standardized rate, and the corresponding standard deviation can be calculated by using the rule described in Chapter 9. There we saw that the standard deviation of the Gaussian approximation to the likelihood for $\log(\lambda)$ is obtained from the standard deviation of the Gaussian approximation to the likelihood for λ by multiplying by $1/M$, where M is most likely value of λ. It follows that for the example of energy intake and IHD incidence, the standard deviations of the standardized rates on a log scale are $2.74/13.67 = 0.200$ and $1.63/6.50 = 0.251$.

A simple extension of the same ideas allows us to calculate estimates and confidence intervals for the ratio of two standardized rates. The log of this ratio is equal to the difference between the logarithms of the two standardized rates, and from section 13.4 and Appendix C the standard deviation of the log of the ratio of the standardized rates is

$$\sqrt{(0.200)^2 + (0.251)^2} = 0.321.$$

This can be used to obtain a confidence interval for the ratio of the standardized rates by using the error factor

$$\exp(1.645 \times 0.321) = 1.696.$$

Exercise 14.3. Use this error factor to find an approximate 90% confidence interval for the ratio of the two standardized rate parameters.

Solutions to the exercises

14.1 The estimated standardized rates are

$$(0.2 \times 6.41) + (0.5 \times 13.67) + (0.3 \times 20.97) = 14.41$$

for the exposed group, and

$$(0.2 \times 6.58) + (0.5 \times 3.93) + (0.3 \times 9.00) = 5.98$$

for the unexposed group.

14.2 The standard deviations of the age-specific rates are 3.29, 1.76, and 3.18 respectively. The standard deviation of the standardized rate is

$$\sqrt{(0.333 \times 3.29)^2 + (0.333 \times 1.76)^2 + (0.333 \times 3.18)^2} = 1.63.$$

14.3 The ratio of standardized rates is $13.67/6.50 = 2.10$ and the 90% range for this is from $2.10/1.696 = 1.24$ to $2.10 \times 1.696 = 3.56$.

14 Confounding and Standardization

14.1 Confounding

Experimental vs. non-experimental

JH prefers this implied distinction to the 'experimental' vs. 'observational' that many authors use. After all, all studies (even randomized trials) make observations. The word 'observational' might also be confused with the term 'observed **only**' for those in the 'no treatment' arm of a treated vs. not treated comparison – even if that comparison is formed experimentally. ***The word experiment*** (check any dictionary) refers to 'a distortion deliberately introduced in order to learn about its effects'

Miettinen glossary: ***experiment***: "a study in which a determinant is intentionally perturbed for reasons none other than the goals of the study itself."

C&H's depiction of the epidemiologist as a 'passive observer' also focuses on this key 'intentional vs not' distinction.

In 2021, a *new* and helpful distinction came to the fore: *experimental* (RCT) versus '***real-world***'. If you Google 'real-world vaccine efficacy' or other terms involving these two words, you will get several hits. In the real world, those who get vaccinated (or get to get vaccinated) are different in many relevant aspects from those who don't. As soon as the COVID-19 vaccines were rolled out, we had to be on the lookout for, and deal with, these differences.

EXTREME EXAMPLES OF CONFOUNDING – FROM 'BC' [1]

Rather that rely on made-up examples, it is also good to have real ones, and even extreme ones, to make the point. JH likes the two given in the very 1st chapter of Rothman's 2002 introductory text[2]

Rothman's **first** example ... [verbatim]

> Common sense tells us that residents of Sweden, where the standard of living is generally high, should have lower death rates than residents of Panama, where poverty and more limited health care take their toll. Surprisingly, however, a greater proportion of Swedish residents than Panamanian residents die each year. This fact belies common sense. The explanation lies in the age distributions of the populations of Sweden and Panama. Figure 1-1 shows the population pyramids of the two countries. A population pyramid displays the age distribution of a population graphically. The population pyramid for Panama tapers dramatically from younger to older age groups, reflecting the fact that most Panamanians are in the younger age categories. In contrast, the population pyramid of Sweden is more rectangular, with roughly the same number of people in each of the age categories up to about age 60 and some tapering above that age. As these graphs make clear, Swedes tend to be older than Panamanians. For people of the same age in the two countries, the death rate among Swedes is indeed lower than that of Panamanians, but in both places older people die at a greater rate than younger people. Because Sweden has a population that is on the average older than that of Panama, a greater proportion of all Swedes die in a given year, despite the lower death rates within age categories in Sweden compared with Panama.

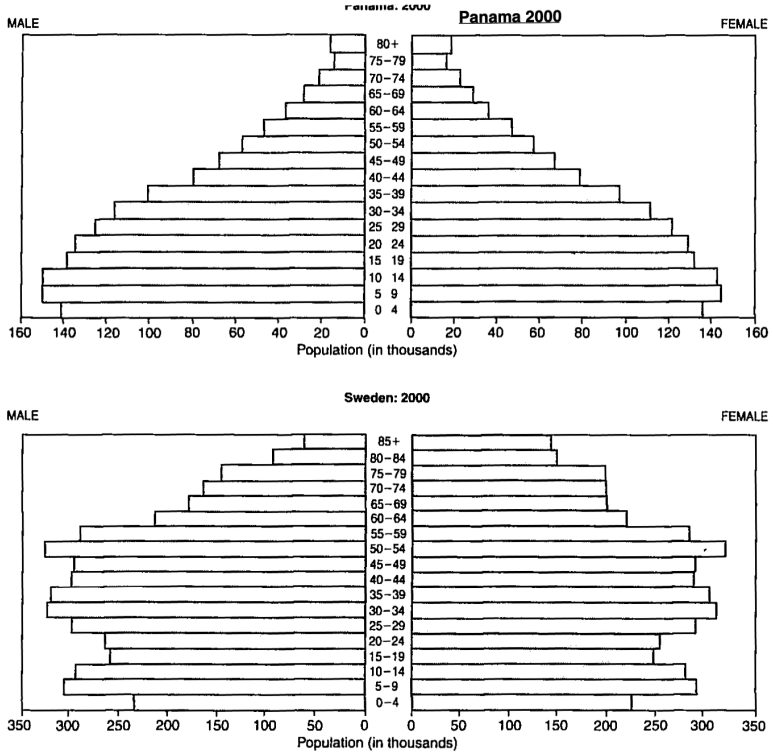


Figure 1–1. Age distribution of the populations of Panama and Sweden (population pyramids). Source: U.S. Census Bureau, International Data Base.

This situation illustrates what epidemiologists call *confounding*. In this example, age differences between the countries are confounding the differences in death rates.

Confounding occurs commonly in epidemiologic comparisons.

[1] Before COVID

[2] Epidemiology: An introduction. Kenneth J Rothman. Oxford University Press.

Rothman's **second** example ... [verbatim]

Consider the following mortality data, summarized from a study that looked at smoking habits of residents of Whickham, England, in the period 1972-1974 and then tracked the survival over the next 20 years of those who were interviewed? Among 1314 women in the survey, nearly half were smokers. Oddly, proportionately fewer of the smokers died during the ensuing 20 years than nonsmokers. The data are reproduced in Table 1-1.

Table 1-1. Risk of death in a 20-year period among women in Whickham, England, according to their smoking status at the beginning of the period*

Vital Status	Smoker	Nonsmoker	Total
Dead	139	230	369
Alive	443	502	945
Total	582	732	1314
Risk (dead/total)	0.24	0.31	0.28

*Data from Vanderpump et al.

Only 24% of the women who were smokers at the time of the initial survey died during the 20-year follow-up period. In contrast, 31% of those who were nonsmokers died during the follow-up period. Does this difference indicate that women who were smokers fared better than women who were not smokers?

Not necessarily. One difficulty that many readers quickly spot is that the smoking information was obtained only once, at the start of the follow-up period. Smoking habits for some women will have changed during the follow-up. Could those changes explain the results that appear to confer an advantage on the smokers? It is theoretically possible that all or many of the smokers quit soon after the survey and that many of the nonsmokers started smoking. While possible, this scenario is implausible, and without evidence for these changes in smoking behavior, this implausible scenario is not a reason- able criticism of the study findings.

A more realistic explanation for the unusual finding becomes clear if we examine the data within age categories, as shown in Table 1-2 (the risks for each age group were calculated by dividing the number who died in each smoking group by the total of those dead or alive).

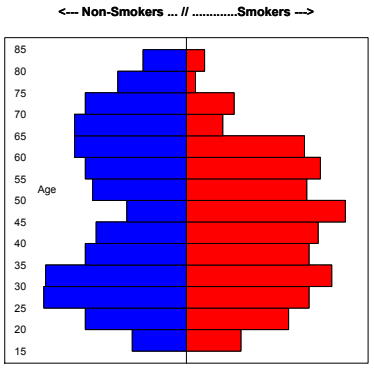
Table 1-1 combines all of the age categories listed in Table 1-2 into a single table, which is called the *crude* data. The more detailed display of the same data in Table 1-2 is called an *age-specific* display, or a display stratified by age. The age-specific data show that in the youngest and oldest age categories there was little difference between smokers and nonsmokers in risk of death. Few died among those in the younger age categories, regardless of whether they were smokers or not, whereas among the oldest women, nearly everyone died during the 20 years of follow-up. For women in the middle age categories, however, there was a consistently greater risk of death among smokers than nonsmokers, a pattern contrary to the impression gained from the crude data in Table 1-1.

Why did the nonsmokers have a higher risk of death in the study population as a whole? The reason is evident in Table 1-2: a much greater proportion of the nonsmoking women were in the highest age categories, the age categories that contributed a proportionately greater number of deaths. The difference in the age distributions between smokers and nonsmokers reflects the fact that, for most people, lifelong smoking habits are determined early in life. During the decades preceding the study in Whickham, there was a trend for increasing proportions of young women to become smokers. The oldest women in the Whickham study grew up during a period when few women became smokers, and they tended to remain nonsmokers for the duration of their lives. As time went by, a greater proportion of women who were passing through their teenage or young adult years became smokers. The result is a strikingly different age distribution for the female smokers and non- smokers of Whickham. Were this difference in the age distribution ignored, one might conclude erroneously that smoking was not related to a higher risk of death. In fact, smoking is related to a higher risk of death, but confounding by age has obscured this relation in the crude data of Table 1-1. In Chapter 8, we return to these data and show how to calculate the effect of smoking on the risk of death after removing the age confounding.

Table 1-2. Risk of death in a 20-year period among women in Whickham, England, according to their smoking status at the beginning of the period, by age*

Age (years)	Vital Status	Smoker	Nonsmoker	Total
18–24	Dead	2	1	3
	Alive	53	61	114
	Risk	0.04	0.02	0.03
25–34	Dead	3	5	8
	Alive	121	152	273
	Risk	0.02	0.03	0.03
35–44	Dead	14	7	21
	Alive	95	114	209
	Risk	0.13	0.06	0.09
45–54	Dead	27	12	39
	Alive	103	66	169
	Risk	0.21	0.15	0.19
55–64	Dead	51	40	91
	Alive	64	81	145
	Risk	0.44	0.33	0.39
65–74	Dead	29	101	130
	Alive	7	28	35
	Risk	0.81	0.78	0.79
75+	Dead	13	64	77
	Alive	0	0	0
	Risk	1.00	1.00	1.00

*Data from Vanderpump et al.[5]



Made by JH using `histbackback` function in `Hmisc` package, with data in `mosaicData` package in `R`.

The Whickham story is also told in DR Appleton, JM French, MPJ Vanderpump. *Ignoring a covariate: an example of* ***Simpson's paradox****.* (1996) American Statistician, 50(4):340-341. Indeed, the individualized `R` data are 'synthesized' from the table given in that article.

A memorable [not involving age] example of [extreme] confounding

Apparent **gender-bias in admissions to the Berkeley graduate schools**[3]
(Most confounding is less extreme than in these 'classic' examples.)

Berkeley Data: M:F Comparative parameters Odds Ratio (OR), Risk Ratio (RR) and Risk Difference (RΔ)

(Using KKM table 17.16 notation)

	D	$\bar{D}$	Men: E = a	Women: $\bar{E}$ = b		m_1
			c	d		m_0
			n_1	n_0		n

Faculty	Admitted?	Men	Women	All	for RΔ: a/n_1	b/n_0	RΔ	for OR: a•d / b•c	a•d / n	b•c / n	for RR: a•n_0 / b•n_1	a•n_0 / n	b•n_1 / n	for RΔ: var(RΔ)*	w = 1/var	w•RΔ
A	Y	512	89	601	0.62	0.82	−0.20	0.35	10.4	29.9	0.75	59.3	78.7	1.63E−3	614	−125
	N	313	19	332												
	All	825	108	933												
B	Y	353	17	370	0.63	0.68	−0.05	0.80	4.8	6.0	0.93	15.1	16.3	9.12E−3	110	−5
	N	207	8	215												
	All	560	25	585												
C	Y	120	202	322	0.37	0.34	+0.03	1.13	51.1	45.1	1.08	77.5	71.5	1.10E−3	913	26
	N	205	391	596												
	All	325	593	918												
D	Y	138	131	269	0.33	0.35	−0.02	0.92	42.5	46.1	0.95	65.3	69.0	1.14E−3	879	−16
	N	279	244	523												
	All	417	375	792												
E	Y	53	94	147	0.28	0.24	+0.04	1.22	27.1	22.2	1.16	35.7	30.7	1.51E−3	661	25
	N	138	299	437												
	All	191	393	584												
F	Y	22	24	101	0.06	0.07	−0.01	0.83	9.8	11.8	0.84	10.5	12.5	3.41E−4	2935	−33
	N	351	317	668												
	All	373	341	769												
All	*Y*	*1198*	*557*	*1755*	*0.44*	*0.30*	*+0.14*	*1.84*			*1.47*					
	N	*1493*	*1278*	*2771*												
	All	*373*	*341*	*4526*												
								∑: 145.8	161.1			263.4	278.7		6113	−129

$OR_{MH} = \frac{145.8}{161.1} = \mathbf{0.91}$ $RR_{MH} = \frac{263.4}{278.7} = \mathbf{0.94}$ $R\Delta_w = \frac{\sum w \bullet R\Delta}{\sum w} = \frac{-129}{6113} = \mathbf{-0.02}$

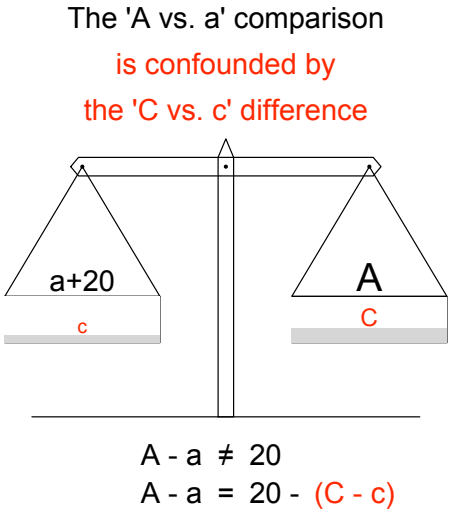
* var(RΔ) = Sum of 2 binomial variances

Confounding by age (Fig 14.1)

The key is that the crude comparison is distorted by age: the 'exposed vs. unexposed' comparison is really a comparison of 'somewhat younger exposed' vs. 'somewhat older exposed'. The diagram below explains confounding with fewer numbers: the comparison of the more- ('A') vs. less- ('a') exposed is distorted or confounded: the 'pan' that supports A is – by itself – heavier (by an amount $C - c$) than the one that supports a.

Although this dual relationship ["a third variable can distort a relationship between an exposure and failure provided it is related to both exposure and failure"] is often taken as a necessary condition for a variable to be a confounder, it is not sufficient.

C&H give an example involving the variable birthweight as the third variable. Because it is one of the factors that is improved by proper antenatal care and that reduces risk, it is often referred to as an '**intermediate**' variable, or the



variable that ***mediates*** [is a mechanism for] the effect of proper antenatal care on perinatal risk.

A CLOSER TO HOME EXAMPLE: Montreal vs. Toronto

Jean Drapeau was Mayor of Montreal from 1954 to 1957, and again from 1960-1986. He believed that **water fluoridation** was an infringement on personal liberties, and refused to implement it in Montreal.

When Jean Doré succeeded him, hopes were hight that it would be adopted, but several groups opposed it, some on the grounds that it might cause **cancer**. City Hall organized a public meeting on the topic, and invited Dr Sandy McPherson (Public Health, Toronto) to address these people's concerns. JH was not in attendance, but was later told that McPherson addressed the meeting this way

> I understand you have concerns that adding fluoride to Montreal's drinking water might increase the number of cancers. So, before this meeting, I had a careful look at the cancer incidence rates in Toronto, which has been adding fluoride to its drinking water for decades, and Montreal, which had never put fluoride it its drinking water. And, I came up with very puzzling numbers, and I am hoping you can help me understand them: Toronto, *with* fluoridation, has *fewer* new cancers per capita than Montreal, which is *without* fluoridation. [In my comparisons, I adjusted for the different age-sex compositions of the two cities.] **Why might this be?**

[3] Sex Bias in Graduate Admissions: Data from Berkeley. Bickel PJ et al., Science 7 February 1975: Vol. 187. no. 4175, pp. 398 - 404.

2021 EXAMPLES: EXTREME CONFOUNDING

2021 has brought a large number of confounded contrasts of the efficacy of COVID-19 vaccines, naive contrasts which give the appearance that the vaccines are not as good as they were in the RCTs; some really extreme ones can even make the vaccinated look like they have worse outcomes than the unvaccinated.

By Googling "Simpson's paradox vaccinations" you will find several examples of such distorted (unfair) comparisons.

This Washington Post article points us to a helpful blog: "The University of Pennsylvania biostatistician Jeffrey Morris wrote an especially thorough and widely shared blog post making this point".

Nor surprisingly, the earliest examples are found in the 'real-world' Israeli data, but examples of Simpsons' paradox [4] also showed up in data from the UK, the home of 'The Simpson the paradox is named for'. [He is not to be confused with 'OJ' Simpson, whose name will forever be connected withe type of crime that was the topic here.]

As Olli Saarela (the one who alerted JH to these data) noted "The vaccine efficacy has two components, against infection, and against hospitalization or death if infected. The former has been waning in Israel for the early vaccinated, but the latter component is still very much there. In the media the numbers are often misinterpreted by ignoring the denominators and focusing on the numerators only (e.g. x% of the hospitalized are vaccinated/unvaccinated)"

Here are links to the UK data from a report on 6 August 2021. The Simpson's paradox Olli was talking about is in comparing hospital admissions or deaths between vaccinated (2 doses) and unvaccinated — among positive delta cases – in the North-East and South-East corners of Table 5, page 18.

Important: since COVID, it has been easy to find real examples of Simpson's paradox, you might think that confounding only refers to contexts where the true direction (slope, ratio, ...) is reversed when you fail to dis-aggregate the data by the confounding variable. This is not true. The term confounding also applies to contexts where failure to dis-aggregate just weakens – or exaggerates – the association measure, but maintains the same direction seem in the confounder-specific strata. A good example is the recent data from Scotland, where the fully-Pfizer-vaccinated vs. unvaccinated case-fatality comparison suggests a VE of just 40%, even though it is above 80% in each age-stratum. *In other words, Simpson's paradox is just a very extreme case of confounding.*

[4] Other links re. Simpson's paradox: here, here and here.

2022 EXAMPLE: CONFOUNDING

This blog tried to correct the message contained in an email sent in June 2022 morning to 6 million Americans from the New York Times *The Morning*. The email, entitled COVID and Race, reported that the "death rate for White Americans (W) has recently exceeded the rates for Black (B), Latino and Asian Americans. The disparities seem to have flipped."

While the blogger (Katelyn Jetelina, who calls herself "your local epidemiologist") begins by addressing the more complex 'time × race interaction' using the full (continuous) time scales in her first 2 graphs, in the exercise below, we will examine the yearly B:W mortality rate ratios. Even in this simpler context, however, it has the same features that she considers, but maybe not extreme enough to produce a full-blown 'Simpson's Paradox'

> The context missing here is Simpson's Paradox— a "statistical phenomenon where an association between two variables in a population *emerges, disappears or reverses* when the population is divided into subpopulations."
>
> In other words, if we just slap data on a graph, it looks like one very clear story. However, when we take into account confounders—or other variables that could also explain this phenomenon—it tells another story.

and we may merely find that the association between two variables in a population *changes* when the population is divided into subpopulations. [In fact, the label 'Simpson's Paradox' is reserved for those contexts where the pattern is *reversed*, so that the (say) age-specific rate ratios are all on one side of the null, but the overall rate ratio (the one that ignores age) is on the other side of the null.

Just like Dr Jetelina did, JH extracted the following data from a database called CDC WONDER. [As she says the 2022 death data are provisional (which means it's not the official count because death certificates take a long time to process), but it's the best we have. JH made a dataset of all COVID deaths for the full 2020-2022 period, but just for White and Black Americans.

Dr Jetelina says that when she looked at just the data for 2022, the COVID mortality rate was 43 per 100,000 White Americans compared with 37 per 100,000 for Black Americans. After she adjusted for age, "the story changes: Whites account for 31 per 100,000 while Blacks account for 40 per 100,000. A complete switch."

You are asked to look further into this below (see exercise 14.6).

14.2 Correction for confounding

C&H offer two options for minimizing confounding. The first is the 'classical' one of holding constant all factors except the one of interest. If one has the option, one can do this by 'blocking', or matching, on these extraneous factors ahead of time (if one has that option; in the analysis one then combines the results of the within-statum (within-block) contrasts, under the assumption that each of these is an estimate of the same (common) parameter value. The second is the use – when possible – of randomization to make the compared groups more equal from the outset, and not just on measured, but also on unmeasured confounders.

C&H present direct standardization as though it were an alternative way of combining the results of the within-statum (within-block) contrasts. But in fact, as is described in the next section of these notes, it can sometimes be regarded as a weighted average of these stratum-specific contrasts.

14.3 Standardized Rates

The key is the use of the *same* set of weights $W_1, \ldots, W_K$ to form the weighted average (w.a.) $\hat{\lambda}_{0,w.a.} = \sum_k W_k \hat{\lambda}_{0,k}$ of the K stratum-specific rates observed in the unexposed (0), and $\hat{\lambda}_{1,w.a.} = \sum_k W_k \hat{\lambda}_{1,k}$ of the stratum-specific rates observed in the exposed(1).

One can also see the *difference* of these two standardized (weighted averages of the stratum-specific) rates as a weighted average of the stratum-specific rate differences, since

$$\hat{\lambda}_{1,w.a.} - \hat{\lambda}_{0,w.a.} = \sum_k W_k \{\hat{\lambda}_{1,k} - \hat{\lambda}_{0,k}\}.$$

Although JH does not advocate calculating a weighted average of ratios (preferring, as Mantel does to take a single ratio of sums), one can – provided all of the ratios are finite – also write the *ratio* of these two standardized (weighted average of the) rates as a (different) weighted average of the K stratum-specific *rate ratios* $[\hat{\lambda}_{1,k}/\hat{\lambda}_{0,k}]$:

$$\frac{\hat{\lambda}_{1,w.a.}}{\hat{\lambda}_{0,w.a.}} = \frac{\sum_k W_k \hat{\lambda}_{1,k}}{\sum_k W_k \hat{\lambda}_{0,k}} = \frac{\sum_k [W_k \hat{\lambda}_{0,k}] \times [\hat{\lambda}_{1,k}/\hat{\lambda}_{0,k}]}{\sum_k W_k \hat{\lambda}_{0,k}} = \frac{\sum_k W'_k \times [\hat{\lambda}_{1,k}/\hat{\lambda}_{0,k}]}{\sum_k W'_k}.$$

In this re-expression, the ratio of the two standardized rates is a weighted average of the observed stratum-specific rate ratios, with weights $W'_k = W_k \hat{\lambda}_{0,k}$.

Correction via 'regression-models' vs. 'standardization' (JH)

Increasingly, corrections for confounding are carried out using generalized linear model versions of what in the simplest case is classically called 'analysis of covariance'. These `glm's` (and others such as Cox regression) are described in C&H chapters 22 and beyond. However, before we get there, it is good to appreciate the basic difference between the type of standardization described in section 14.3, and these regression models.

One way to think of the difference is via an example where we would like to create an unbiased (i.e., a fair) comparison between two groups of students, one that had experienced experimental condition "1" (e.g., distance learning) and the other under experimental condition "0' (e.g., face-to-face in class contact with the teacher on-site). Let's denote the two conditions by the subscripts 1 and 0. Suppose that it was unavoidable that one of the classes was on average older than (and thus at an advantage relative to) the other.

Correction by standardization

We could think of two ways to reduce (eliminate) the age-difference, and arrive at an unbiased estimate of the true difference (Δ) in the means – assumed to be constant across ages. The first is to stratify the students into K age-bands and take (the same) weighed average of the within-age-band mean scores for each group, to arrive at $\bar{y}_{1,w.a.} = \sum_k W_k \bar{y}_{1,k}$ and $\bar{y}_{0,w.a.} = \sum_k W_k \bar{y}_{0,k}$ respectively. As discussed above, the difference of these two standardized means is also a weighed average of the within-age-band differences in the mean scores, i.e.,

$$\sum_k W_k \{\bar{y}_{1,k} - \bar{y}_{0,k}\}.$$

One can think of this as the numerical equivalent of artificially 'evening up' the two teams/classes: it is as though one forced some of the distance students to take the face-to-face version, and vice versa, so that the two classes had the same age-composition $(W_1, \ldots, W_K)$.

Say that the age distributions in those who had intended to take the course were:

age-band:	20-25	25-30	30-35
no. who applied to be 'distance' students:	20	33	46
no. who applied to be 'on-site' students:	50	35	14

Then one possibility would be to – if it were possible – 'transfer some students from one to the other format' so that the age distributions in the classes were:

age-band:	20-25	25-30	30-35
no. of 'distance' students:	35	34	30
no. of 'on-site' students:	35	34	30

If actual transfers were not possible, one could still '*mathematically*' move some students from one to the other format. In other words, one would leave the students in the class they applied for, and use the observed results to create results for *two synthetic classes with the same age-distribution in each*. Suppose the actual results in the 20, 33 and 46 who took the distance class, and the 50, 35 and 14 who took the on-site class were:

age-band:	20-25	25-30	30-35
means for actual 'distance' students:	$\bar{y}_{d,1}$	$\bar{y}_{d,2}$	$\bar{y}_{d,3}$
means for actual 'on-site' students:	$\bar{y}_{o,1}$	$\bar{y}_{o,2}$	$\bar{y}_{o,3}$

From these we could create results for two synthetic or hypothetical classes, with the same age-distribution, say $\{35, 34, 30\}$ in each, just as above:

	mean for 'synthetic' class
'distance':	$(35 \times \bar{y}_{d,1} + 34 \times \bar{y}_{d,2} + 30 \times \bar{y}_{d,3})/99$
'on-site':	$(35 \times \bar{y}_{o,1} + 34 \times \bar{y}_{o,2} + 30 \times \bar{y}_{o,3})/99,$

and compare these two weighted averages.

Since these 2 'classes' are synthetic or hypothetical, the choice of weights is not restricted by the same constraints we had in the situation we we actually transferred students from one to the other class. Thus, we could just as well have, say $\{33, 33, 33\}$ – or $\{43, 33, 23\}$ – in each of the two synthetic classes.

Correction by a regression model

The other way out of this confounding by age is via a regression model. It requires a somewhat stronger assumption than a 'constant (or common) across ages Δ': its also requires that we use a model that links the mean response at each age to *age*. The most commonly used model is a basic analysis-of-covariance model, with parallel lines for the distance (d=1) and on-site (d=0) classes:

$$E[y|age, d] = \mu_{y|age,d} = \beta_0 + \beta_{age} \times age + \beta_d \times d.$$

In our example, the average ages in the distance and on-site classes are 28.8 and 25.7 respectively, a difference of 3.1 years, and so we can obtain an adjusted difference by subtracting a correction factor from the crude difference. This correction is the product of the $\widehat{\beta_{age}}$ and the 3.1 years. The crude and adjusted difference are therefore:

mean of:	y	age
actual 'distance' students:	$\overline{y}_d$	$\overline{age}_d$
actual 'on-site' students:	$\overline{y}_o$	$\overline{age}_o$
(crude) difference:	$\overline{y}_d$ - $\overline{y}_o$	3.1 years
adjusted difference:	$(\overline{y}_d$ - $\overline{y}_o) - \widehat{\beta_{age}} \times 3.1$	

One can see from this that the magnitude of the correction is a function of how strong the effect of age is and how different the average age is in the compared groups.

In the *(synthetic) standardization* approach, conceptually one alters the *composition* of the two compared groups – it is as though one adds distance subjects to, or takes away some distance subjects from, the 3 age-strata of the distance arm, and likewise adds on-site subjects to, or takes away some on-site subjects from, the age-strata of the on-site arm. This way one creates two '*pseudo-samples*', to use a term used by Robins in causal inference to describe the samples formed by inverse probability of treatment weighting

(IPTW). One can also think of the adding and taking away of students as giving different weights to the contributions of students in different age-bands. For example, in the distance class, the result of each student in the youngest age-band is up-weighted and given a weight of 35/20; likewise the results of each student in the middle age-band is slightly up-weighted and given a weight of 34/33, while the result of those in the oldest age-band is down-weighted and given a weight of 30/46. the corresponding up/down-weightings for the results of each student in the on-site class are 35/50, 35/34 and 30/14 in the youngest, middle and oldest age-bands respectively.

To see why Robins calls it IPTW, consider the first age-band, where of the 70 students, 20 took the distance course and 50 the on-line one. So the probability that a student in this band took the distance course is 20/70 and that (s)he took the on-line one is 50/70. The inverses of these probabilities are 70/20 and 70/50, double the 35/20 and 35/50 used above, and the same if we scale the IPTW's so that our pseudo-sample is the same size as our actual sample.

In the *regression* approach, conceptually one takes the group means of the two entire samples of subjects and then adjusts their scores to those of persons of the mean age.

Exposure to Scientific Theories Affects Women's Math Performance

Ilan Dar-Nimrod and Steven J. Heine*

On 14 January 2005, Lawrence Summers, then president of Harvard University, speculated that one reason why women are underrepresented in science and engineering professions is because of a "different availability of aptitude at the high end" (*1*). These remarks were met with much outcry by some critics of President Summers, and social scientists were divided in their reaction to his comments. The question of sex differences in math in the context of the nature-versus-nurture debate is not new and remains contentious. For this paper, we did not explore whether such innate sex differences exist. Instead, we investigated how women's math performance is affected by whether they are considering genetic or experiential accounts for the stereotype of women's underachievement in math. Such a question is relevant to how people respond to scientific arguments and science education more generally.

Stereotype threat is a phenomenon in which the activation of a self-relevant stereotype leads people to show stereotype-consistent behavior, thereby perpetuating the stereotypes (*2*). For example, African Americans perform worse on intelligence tests when their race is highlighted (*2*), and women's math performances decrease when their gender is made salient (*3*). Stereotype threat can be reduced when people focus on the malleability of the traits at hand (*4*).

Past research reveals that people respond differently to genetic and experiential accounts of behaviors. Undesirable behaviors with experiential causes are seen as more voluntary and blameworthy than behaviors with genetic causes (*5*). Experiential causes, in contrast to genetic ones, appear to be viewed as less impactful and more controllable. We reasoned that stereotypes about one's groups are often perceived as inescapable, because many stereotypes are viewed in essentialized terms (*6*). That is, people may view the origin of some stereotypes as resting on the perceived genetic basis that distinguishes these groups. If individuals share the same genetic foundation at the base of the stereotype, they might feel that the stereotype applies to them and hence are vulnerable to stereotype threat. In contrast, we propose that people might react differently if the origins of the group differences were perceived to rest on the specific experiences that people's groups have had. People may reason that their own experiences are different or that they can resist the effects of their experiences.

Our studies manipulated participants' beliefs regarding the source of gender differences in math and measured their subsequent math performance (Fig. 1). In study 1 (*7*), women undertook a Graduate Record Exam–like test in which they completed two math sections separated by a verbal section. The verbal section contained the manipulation in the form of reading comprehension essays. Each test condition used a different essay. Two of the essays argued that math-related sex differences were due to either genetic (G) or experiential causes (E). Both essays claimed that there are sex differences in math performance of the same magnitude. Two additional essays served as a traditional test of stereotype threat. One essay, designed to eliminate underperformance, argued that there are no math-related gender differences (ND). The other essay, designed as a standard stereotype-threat manipulation (S), primed sex without addressing the math stereotype. Controlling for performance on the first math section, we used analyses of covariance to demonstrate that women in the G and the S conditions exhibited similar performances on the second math test ($F < 1$). Women in the E and the ND conditions, although not different from each other ($F < 1$), significantly outperformed women in G and S conditions (all P values ≤ 0.01).

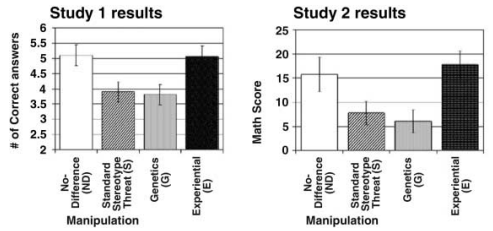


Fig. 1. (Left) Study 1 results. Scores on second math test (controlling for scores on first test) after reading essays. **(Right)** Study 2 results. Scores on math test after hearing manipulation.

These findings were replicated in a second study (*7*) that used a different experimental design. An analysis of variance identified significant performance differences between the conditions [$F(3,88) = 4.15$, $P < 0.01$]. Fisher probable least-squares difference (PLSD) comparisons revealed that women in G and S conditions performed comparably ($P > 0.50$) but significantly worse than women in E and ND conditions (all P values < 0.02), which did not differ ($P > 0.50$).

These studies demonstrate that stereotype threat in women's math performance can be reduced, if not eliminated, when women are presented with experiential accounts of the origins of stereotypes. People appear to habitually think of some sex differences in genetic terms unless they are explicitly provided with experiential arguments. It remains to be seen whether the results generalize to stereotypes about other groups and abilities.

Whether there are innate sex differences in math performance remains a contentious question. However, merely considering the role of genes in math performance can have some deleterious consequences. These findings raise discomforting questions regarding the effects that scientific theories can have on those who learn about them and the obligation that scientists have to be mindful of how their work is interpreted. What President Summers perhaps intended to be a provocative call for more empirical research on biological bases of achievement may inadvertently exacerbate the gender gap in science through stereotype threat.

References and Notes

1. L. H. Summers, "Remarks at NBER conference on diversifying the science and engineering workforce," 14 January 2005, www.president.harvard.edu/speeches/2005/nber.html.
2. C. M. Steele, *Am. Psychol.* **52**, 613 (1997).
3. S. J. Spencer, C. M. Steele, D. Quinn, *J. Exp. Soc. Psychol.* **35**, 4 (1999).
4. J. Aronson, C. Fried, C. Good, *J. Exp. Soc. Psychol.* **38**, 113 (2002).
5. J. Monterosso, E. B. Royzman, B. Schwartz, *Ethics Behav.* **15**, 139 (2005).
6. D. A. Prentice, D. T. Miller, *Psychol. Sci.* **17**, 129 (2006).
7. Materials and methods are available on *Science* Online.
8. J. Lau, J. Sim, and R. Vella-Zarb provided assistance. E. Buchtel, E. Dunn, and A. Norenzayan commented on drafts. S.J.H. acknowledges funding from National Institute of Mental Health in the USA (R01 MH60155-01A2) and from Social Sciences and Humanities Research Council of Canada (410-2004-0795).

Supporting Online Material
www.sciencemag.org/cgi/content/full/314/5798/435/DC1
Materials and Methods

9 June 2006; accepted 15 August 2006
10.1126/science.1131100

2136 West Mall, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

*To whom correspondence should be addressed. E-mail: heine@psych.ubc.ca

math_output.txt **Pag**
Printed: Thursday, November 2, 2006 3:46:46 PM

```
Math / Gender Data..
(Ilan Dar-Nimrod and Steven Heine, Science 4314 20 Oct 2006, p 435)

proc format ;
value codes 1="G" 2="E" 3="ND" 4="S";
run;
data a;   * 1=G  2=E 3=ND 4=S ;
array ic(4) G E ND S;
*infile "unix:mathdata.txt";
infile "Macintosh HD:Users:jameshanley:Documents:Courses:626:MathGender:mathdata.txt";
input c math1 math2;
do i =  1 to 4; ic(i)=(c=i); end;
math1c = math1 - 4.9099099;
run;


proc means n min mean max; format c codes. ;  var math1 math1c math2; run;

The SAS System 11:20 Tuesday, October 24, 2006

          Variable    N       Minimum          Mean       Maximum
          ------------------------------------------------------
          MATH1     111     1.0000000     4.9099099    13.0000000
          MATH1C    111    -3.9099099  9.9099106E-9     8.0900901
          MATH2     111             0     4.4414414    10.0000000
          ------------------------------------------------------

proc means n min mean max; format c codes.; class c; var math1c math2;

      C  N Obs  Variable    N       Minimum          Mean       Maximum
     ---------------------------------------------------------------------
     G      28  MATH1C     28    -2.9099099    -0.4456242     5.0900901
                MATH2      28     1.0000000     3.5714286     9.0000000

     E      27  MATH1C     27    -2.9099099     0.3123123     3.0900901
                MATH2      27     2.0000000     5.2222222     9.0000000

     ND     27  MATH1C     27    -3.9099099    -0.3913914     4.0900901
                MATH2      27     2.0000000     4.8888889    10.0000000

     S      29  MATH1C     29    -3.9099099     0.5038832     8.0900901
                MATH2      29             0     4.1379310     8.0000000
     ---------------------------------------------------------------------

proc reg data=a; model math2 = S G E ;

Dependent Variable: MATH2

                          Analysis of Variance

                          Sum of         Mean
   Source           DF    Squares       Square      F Value       Prob>F

   Model             3     45.73062     15.24354        3.473       0.0187
   Error           107    469.63875      4.38915
   C Total         110    515.36937

    Root MSE       2.09503     R-square       0.0887
    Dep Mean       4.44144     Adj R-sq       0.0632
    C.V.          47.17003

                          Parameter Estimates

                        Parameter      Standard    T for H0:
   Variable  DF         Estimate         Error   Parameter=0    Prob > |T|

   INTERCEP   1        4.888889    0.40318855        12.126        0.0001
   S          1       -0.750958*   0.56027753        -1.340        0.1830
   G          1       -1.317460    0.56508076        -2.331        0.0216
   E          1        0.333333    0.57019472         0.585        0.5601


proc reg data=a; model math2 = S G E math1c;

Dependent Variable: MATH2

                          Analysis of Variance

                          Sum of         Mean
   Source           DF    Squares       Square      F Value       Prob>F

   Model             4    185.10374     46.27594       14.852       0.0001
   Error           106    330.26563      3.11571
   C Total         110    515.36937

    Root MSE       1.76514     R-square       0.3592
    Dep Mean       4.44144     Adj R-sq       0.3350
    C.V.          39.74247

                          Parameter Estimates

                        Parameter      Standard    T for H0:
   Variable  DF         Estimate         Error   Parameter=0    Prob > |T|

   INTERCEP   1        5.103533    0.34121364        14.957        0.0001
   S          1       -1.241939** 0.47772816        -2.600        0.0107
   G          1       -1.287718    0.47612189        -2.705        0.0080
   E          1       -0.052588    0.48386265        -0.109        0.9137
   MATH1C     1        0.548414    0.08199698         6.688        0.0001

**Example of adjusted difference... (see class notes on confounding by jh)

S vs ND(ref) ..

unadjusted difference  = ( 4.1379310 -  4.8888889 ) = -0.750979*

  adjusted difference  =  -0.750979 - 0.548414 * (0.5038832 - (-0.3913914) )

                       =  -0.750979 - 0.548414 * 0.8952746

                       =  -0.750979 - 0.490981

                       =   1.2419**
```

```
===== raw data (courtesy of 1st author) ========
```

i	group	math1	math2	G	E	ND	S
1	2	8	6	0	1	0	0
2	3	9	8	0	0	1	0
3	3	4	4	0	0	1	0
4	1	3	1	1	0	0	0
5	1	5	2	1	0	0	0
6	1	5	4	1	0	0	0
7	1	6	5	1	0	0	0
...							
...							
...							
103	3	6	3	0	0	1	0
104	4	5	4	0	0	0	1
105	4	3	3	0	0	0	1
106	4	3	6	0	0	0	1
107	3	5	3	0	0	1	0
108	4	4	2	0	0	0	1
109	3	6	6	0	0	1	0
110	4	5	4	0	0	0	1
111	3	3	3	0	0	1	0

```
Full dataset available on website
```

Confounding: Reducing it by Regression
(page 1)

Preamble

- Don't overlook classical, "non-regression" methods
- Regression methods are more "synthetic" (i.e. "artificial")
- Cf chapter 3 by Anderson et al. (c622; readings from aahovw)

Definitions ... / synonyms

Original (statistical, in design of experiments)

- inability to estimate higher order interactions (so typically assume they are zero)

- "mixed up with other effects" or "inextricable"

Epidemiological

- (osm)

Other terms

- "Lurking" (i.e. "hidden") variable

- "Simpson's Paradox" is the most extreme form

*(see collection of Simpson's paradox examples under **Other Resources** on **c626**)*

Examples...

- Does using a Macintosh lead to sloppier writing? a
- Better Service from Canada Post after "Major Restructuring"a
- Salaries of Master's and PhD's a
- Outcomes of Pregnancy during Residency for women and wives of their male classmates• Admissions of Males & Females to Berkeley Graduate Schools b
- Percentage of White & Black Convicts Receiving Death Penalty a
- Intelligence Quotient (IQ) - Mother's Milk; Other Variables a
- Lung Function of Vanadium Factory Workers Other resources, c697
  - vs. reference group (matched for smoking and age) that was 3.4 cm different in ave. height
- Blood Pressure and Altitude - age; height; weight; country b
- Longevity - Sexual Activity; thorax size c622
- Fatalities & Speed Limit Change - Time a
- NEURODEVELOPMENT OF CHILDREN EXPOSED IN UTERO TO ANTIDEPRESSANT DRUGS b
- What Does It Take to Heat a New Room? dataset, c697

a notes on Ch 2, c607 b resources this course (678), session 5

Confounding: Reducing it by Regression
(page 2)

Adjustment via regression ...

- "Outcome" **Y**
- Contrast with respect to **X** ("Exposure" variable)
  (for now, say X is binary **X=1** and **X=0**)
- Confounder **C**

CRUDE CONTRAST:

via $E[Y \mid X] = b_0 + b_X X$

b_X = crude difference = $\bar{Y}_{X=1} - \bar{Y}_{X=0}$

ADJUSTED CONTRAST:

$E[Y \mid X, \underline{C}] = b_0^* + b_X^* X + b_C C$

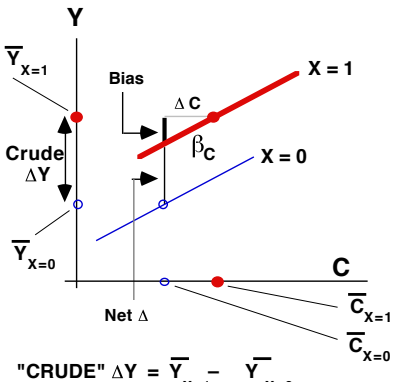
b_X^* = adjusted difference

$= \bar{Y}_{X=1} - \bar{Y}_{X=0}$ ***(CRUDE Δ)***

minus

$b_C (\bar{C}_{X=1} - \bar{C}_{X=0})$ ***(ADJUSTMENT)***

In Pictures... (cf Anderson et al. chapter)



"CRUDE" $\Delta Y = \bar{Y}_{X=1} - \bar{Y}_{X=0}$

$\Delta C = \bar{C}_{X=1} - \bar{C}_{X=0}$

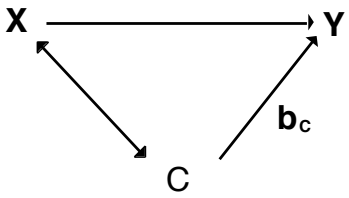
Bias = $\beta_C \times \Delta C$

"Net" $\Delta Y = \bar{Y}_{X=1} - \bar{Y}_{X=0} - \beta_C \times \Delta C$

Confounding: Reducing it by Regression
(page 3)

Anatomy of the "Adjustment"

b_C ($\bar{C}_{X=1}$ – $\bar{C}_{X=0}$)



- for a NON-ZERO ADJUSTMENT...

b_C NON ZERO

AND

($\bar{C}_{X=1}$ – $\bar{C}_{X=0}$) NON ZERO

Special issues

1.
- Adjustment uses a LINEAR relation Y <--> C

If Y <--> C relationship not linear, using a linear relation will not produce correct adjustment

e.g. Y= birthweight and C = Age in residents' study

2.
- If Y <--> C relationship not same at different levels of X

(ie if C is a modifier of X<->Y rel'n, or X is a modifier of C<->Y rel'n i.e. if X<-->C "interaction")

then cannot make a unique "adjustment" (adjustment different at different levels of C)

e.g. gender D's in salary (C = # years experience)
c.f. Miettinen diagram (covariate as a modifier, confounder, or both)

3.
- **Inappropriate Adjustment...**

X ---> C ---> Y

X ---> Y ---> C

Supplementary Exercise 14.1a

Refer to the data on the Berkeley graduate school admissions shown on p. 3 and to the article by Bickel, Hammel, and O'Connell in Science in 1975 Sex Bias in Graduate Admissions: Data from Berkeley.

i. In 1 paragraph, summarize the article in words that would be understood by a professional (e.g. a lawyer) who has little knowledge of statistics or epidemiology.

ii. Imagine that 933, 585, ... 769 applicants applied to the six Faculties A-F respectively but that all 4,526 were women. If they had the same success rates as the women actually achieved, what proportion of the 4,526 women would have been admitted? Calculate the variance of this proportion.
Imagine that 933, 585, ... 769 applicants applied to the six Faculties A-F (i.e., 4,526 in all) but that all 4,526 were men. If they had the same success rates as the men actually achieved, what proportion of them would have been admitted? Calculate the variance of this proportion.

iii. What would C&H call these two proportions?

iv. Calculate (a) the female-male *difference* between these two proportions, and a CI for it (b) the *ratio* of the proportions, and a CI for it (Hint: do your calculations in the log(ratio) scale, and convert back. (c) the *ratio* of the proportions that were *not* admitted, and a CI for it. (d) the ratio of the *odds* of being admitted, and a CI for it. Comment on your findings, and give reasons for which of the four metrics you prefer.

Supplementary Exercise 14.1b

Refer again to the data on the Berkeley graduate school admissions shown on p. 3 ('MH' stands for 'Mantel-Haenszel').

i. The three summary measures (OR, RR and RD) at the bottom of the Table lack accompanying confidence intervals. Find and cite (but do not implement) the appropriate formulae you could use to calculate them.

Supplementary Exercise 14.1c

Refer again to the Berkeley graduate school admissions data shown on p. 3.

i. Fit the four measures via binomial regression (`glm`) with the logit, log and identity links, using 'men' as the reference category, and 'women' as the index category, and using 'faculty' as a categorical variable. Use the relevant fitted coefficient and its SE to obtain a CI.

Supplementary Exercise 14.2a

Refer to Exposure to Scientific Theories Affects Women's Math Performance. and the appended supplementary material, along with the data the authors provided to JH.

i. In 1 paragraph, summarize the article in words that would be understood by a professional (e.g. an educator) who has little knowledge of statistics or epidemiology.

ii. For now, limit your analysis to the S (index) vs. ND (reference) contrast, involving 56 women.

How '(im)balanced' were these groups with respect to `math1`? What are the implications of this?

iii. Categorize the `Math1` scores into 4 bins, so that 13 obtained a score of 1-3, 13 obtained a 4, 20 obtained a score of 5-6 and 10 had a score of 7-13.

Imagine that these 13, 13, 20 and 10 women (56 in all) were all in the ND group. If they had the same mean `math2` scores as the ND women in these 4 `math1`-bins actually achieved, what would the overall `math2` mean of the 56 women have been? Calculate the variance of this weighted mean.

Imagine that these 13, 13, 20 and 10 women (56 in all) were all in the S group. If they had the same mean `math2` scores as the S women in these 4 `math1`-bins actually achieved, what would the overall `math2` mean of the 56 women have been? Calculate the variance of this weighted mean.

iv. Calculate the difference of these two weighted means, along with the variance of this difference. Compare them with the results from the `proc reg data=a; model math2 = S G E math1c;` fitting given on the right hand side of p7. Comment on any differences.

Supplementary Exercise 14.2b

i. Estimate the between-group differences in `math2` using a linear model with an intercept (for a suitable reference group) and 3 indicator variables. [For interest, run it as a traditional 'anova' as well].

ii. How '(im)balanced' were the groups with respect to `math1`[5] and how serious is this in terms of the 'fairness' of the comparison you have made in i. ?

[5]To make the group differences easier to interpret, use a centered version of it – i.e. derive a version that has an overall mean of 0.

iii. Use the (centered) `math1` variable as a covariate in the linear model, and report the (adjusted) estimates of the between group differences in `math2`.

iv. Some investigators would have adjusted for baseline scores by subtracting `math1` from `math2` and using this difference in the linear model. How different is this from the approach in iv. ? *Hint*: rewrite the fitted model in iii. so that the left hand side of the regression equation involves both `math2` and `math1`. Why approach do you prefer?

Supplementary Exercise 14.3

Sharper and **Fairer** Comparisons:

See the article "Sexual activity reduces lifespan of male fruitflies" and accompanying material in this website.

Limit your analysis to the 50 fruitflies with 1 partner/2 days .. the effect is obvious in those with 8.

Aside: When we first analyzed this dataset, student PE, now on McGill faculty, argued that thorax size cannot be used as a predictor or explanatory variable since fruitflies who die young may not be fully grown, i.e., it is also an "intermediate" variable. Later, student NK (now on faculty elsewhere) had studied entomology and assured us that fruitflies do not grow longer after birth; i.e., thorax length is not time- (age)-dependent!

i. Use `lm` in `R` to calculate the difference in mean longevity (mean days lived) of sexually active flies (index cat.) relative to sexually inactive flies (reference cat.), ignoring other covariates. Is this difference (i) substantial? (ii) statistically significant at the conventional $\alpha = 0.05$ level?

ii. Again ignoring other covariates, calculate the overall *mortality rate* (no. deaths / 100 fruitfly-days lived – effectively, apart from the scaling by 100, the reciprocal of mean longevity) for each of the two compared categories.

iii. How different are the mean thorax lengths of the active and inactive flies? Is this difference "statistically" significant? Is it substantial? Is statistical significance a non-issue here anyway? Explain.

iv. (Independently of which flies were subsequently assigned to an active/inactive partner) divide the thorax range into 3 (roughly equal-sized) strata: S, M and L. Compute the mortality rates (no. deaths / fruitfly-days) for the resulting 6 cells. Then, using the overall proportions of flies in each stratum as the same 3 weights for both, compute standardized mortality rates for the active and inactive groups.

v. Using these strata, compute the mean longevity for each of the 6 cells. Then, using the overall proportions of flies in each stratum as the 3 weights, compute a mean longevity for each of the two compared groups.

vi. If – other things being equal – flies 0.01 mm larger live on average 1 day longer, how much of a longevity "advantage" would the active flies have from the outset as a result of their larger average thorax size? On this basis, how much lower would the mean longevity of active than inactive flies be if it were "adjusted" for the difference in thorax size?

vii. Instead of using the "out of the air" value of 1day/0.01mm, use multiple regression to simultaneously estimate the additional mean days/mm and the decrease in days associated with (due to) activity i.e., fit the model:

$$E[longevity \mid thorax,\ activity] = \beta_0 + \beta_{thorax} \times thorax + \beta_{active} \times\ active.$$

viii. Verify that if you correct/adjust the comparison as in (vi) but using the fitted β_{thorax} from (vii) instead of the 'out of the air' 0.01, and using the the thorax difference in (iii), you arrive at the β_{active} obtained in (vii). Hint: cf schematic diagram in JH notes on confounding.

ix. Use the correction for confounding in the **Women and Math** study (see above) to explain – in just a few sentence, and in English rather than in 'Statistical-ese' – to your father-in-law how 'adjustment by regression' works.

x. In the Breast milk and subsequent IQ in children born preterm study, Lucas et al use multiple regression to correct for *several* IQ determinants that are 'imbalanced' between the 'Mother's milk' and 'No-mothers-milk' groups. To understand how it works, extend the 'Adjusted Contrast' equation on page 2 of JH's Notes on Confounding: Reducing it by Regression (the same ones at the end of the Women and Math article) so that it accommodates imbalances in several variables (hint: think of X as a vector rather than a scalar). This time, using Tables I, II and IV, explain the (now multivariable) correction/adjustment to your grandparents – who strongly believe that the mother's milk - IQ link is causal. Use Tables I, II and IV.

xi. {A '*sharper*' comparison} The p-value for the activity contrast in (vii) is smaller (and the associated CI narrower) than the corresponding one in (i). One reason is that the larger adjusted estimate of the effect (the numerator of the t-test on adjusted difference); another is the smaller SE of the estimated effect (the denominator of t-test).
Why is the SE of the estimated longevity difference from analysis (vii) smaller?

Notes: JH introduced the 'shaper and fairer' terminology in an 1983 article entitled Appropriate uses of multivariate analysis in the Annual Review of Public Health (also available under REPRINTS/TALKS on his home page). The same issues are illustrated in notes he appended to the article 'Exposure to Scientific Theories Affects **Women's Math Performance**' (see above) and in excerpts from '**Breast Milk and Subsequent IQ** in Children Born Preterm' article, under the ETIO-gnosis heading on the website Regression and Multivariable Analysis Sept. 26, 2013.

Supplementary Exercise 14.4

Table 1. Vaccine Effectiveness in Preventing Death from Covid-19, Stratified According to Age Group, Vaccination Status, and Vaccine (All Community Cases from April 1 to August 16, 2021, with Follow-up Conducted until September 27, 2021).*

Age Group, Vaccination Status, and Vaccine	Person-Years of Follow-up	No. of Persons	No. of Deaths	Rate per 100,000 Person-Years	Adjusted Hazard Ratio (95% CI)†
16 to 39 Years of Age					
Unvaccinated	8669.5	35,449	17	0.20	—
One vaccine dose 0–27 days before test					
ChAdOx1 nCoV-19	56.6	150	0	0.00	—
BNT162b2	2338.4	10,535	1	0.04	—
One vaccine dose ≥28 days before test or two doses with second dose 0–13 days before test					
ChAdOx1 nCoV-19	463.0	1,793	0	0.00	—
BNT162b2	1706.3	10,167	1	0.06	—
Two vaccine doses with second dose ≥14 days before test					
ChAdOx1 nCoV-19	767.7	4,140	0	0.00	—
BNT162b2	567.3	3,040	0	0.00	—
40 to 59 Years of Age					
Unvaccinated	1230.3	4,803	33	2.68	Reference
One vaccine dose 0–27 days before test					
ChAdOx1 nCoV-19	453.8	1,497	2	0.44	0.24 (0.06–1.01)
BNT162b2	86.9	286	0	0.00	0.00 (0.00–∞)
One vaccine dose ≥28 days before test or two doses with second dose 0–13 days before test					
ChAdOx1 nCoV-19	1865.2	7,945	2	0.11	0.04 (0.01–0.15)
BNT162b2	477.9	2,022	0	0.00	0.00 (0.00–∞)
Two vaccine doses with second dose ≥14 days before test					
ChAdOx1 nCoV-19	1707.4	9,587	16	0.94	0.12 (0.07–0.24)
BNT162b2	629.8	3,318	2	0.32	0.05 (0.01–0.21)
≥60 Years of Age					
Unvaccinated	81.4	380	24	29.49	Reference
One vaccine dose 0–27 days before test					
ChAdOx1 nCoV-19	19.1	46	0	0.00	0.00 (0.00–∞)
BNT162b2	0.2	1	0	0.00	0.00 (0.00–∞)
One vaccine dose ≥28 days before test or two doses with second dose 0–13 days before test					
ChAdOx1 nCoV-19	213.9	692	2	0.93	0.03 (0.01–0.14)
BNT162b2	69.8	190	4	5.73	0.25 (0.09–0.74)
Two vaccine doses with second dose ≥14 days before test					
ChAdOx1 nCoV-19	973.8	5,262	73	7.50	0.10 (0.06–0.16)
BNT162b2	351.0	1,952	24	6.84	0.13 (0.07–0.23)

* Vaccine effectiveness was estimated as 1 minus the hazard ratio. Some adults had received the mRNA-1273 vaccine (Moderna) at the time of their positive test (4135 persons, contributing 379 person-years of follow-up). No deaths from coronavirus disease 2019 (Covid-19) occurred among the persons who received the mRNA-1273 vaccine, and estimates and numbers are not provided in the table.
† Hazard ratios are not provided for the 16-to-39-year age group because only two deaths occurred among vaccinated persons in this group and no deaths occurred among those who were fully vaccinated (i.e., those who had received two doses with the second dose received ≥14 days before testing).

The following data were extracted from Table 1 [overleaf] of '*BNT162b2 and ChAdOx1 nCoV-19 Vaccine Effectiveness against Death from the Delta Variant*' based on a Scotland-wide surveillance platform (Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 [EAVE II] that includes individual-level linked data on vaccination, testing, viral sequencing, primary care, hospital admissions, and mortality among 5.4 million people (approximately 99% of the Scottish population). The New England Journal of Medicine, Oct 20, 2021

Age		P-Y	P	D	D/100,000PY†	VE
16-39						
	Unvaccinated	8669.5	35449	17	0.20	
	Vaccinated*	567.3	3040	0	0.00	vv.v
		9236.8				
40-59						
	Unvaccinated	1230.3	4803	33	2.68	
	Vaccinated*	629.8	3318	2	0.32	vv.v
		1860.1				
≥ 60						
	Unvaccinated	81.4	380	24	29.49	
	Vaccinated*	351.0	1952	24	6.84	vv.v
		432.4				
===		===	===	==	===	
ALL						
	Unvaccinated	9981.2	40632	74	xx.xx	
	Vaccinated*	1548.1	8310	26	yy.yy	VV.V
		11529.3				

P-Y: Person-Years of Follow-up; P: No. of Persons ; D: No. of Deaths: VE: Vaccine Efficacy; *With BNT162b2; Second dose ≥ 14 days before test.

i. Correct the Column Heading† (copied from the NEJM table).

ii. Using the data in the 2 'ALL' rows at the bottom, calculate the missing rates, xx.xx and yy.yy – and from them a point estimate of the VE against death, and associated CI.

iii. Explain to a lay person why these rates, and the resulting VE, are misleading, and why you should have refused to calculate the CI!

iv. Consider a total of 11529.3 person years of follow-up, with 9236.8, 1860.1 and 432.4 of them contributed by the 3 age groups shown. If the mortality rates in these 3 segments of follow-up time were the same as in the 3 *unvaccinated* segments, how many deaths would you expect? [6] Convert this number into a death rate, and compute its variance.

Consider the same total of 11529.3 person years of follow-up, with again the same 9236.8, 1860.1 and 432.4 of them contributed by the 3 age groups shown. If the mortality rates in these 3 segments of follow-up time were the same as in the *vaccinated* segments, how many deaths would you expect? Convert this number into a death rate, and compute its variance. Mention any reservations you have about your variance calculation.

v. Calculate the difference of these two weighted rates, along with the SE of this difference. Compare them with the results from a GLM fit of a Poisson model with the identity link.[7] Comment on any differences/difficulties.

You can do this by putting the D's, P-Y's and 'Vaccinated' indicator into vectors of length 6,

```
D   = c(   17,     0,       33,     2,    24,    24)
PY = c(8669.5, 567.3 ,  1230.3, 629.8,  81.4, 351.0)
Vaccinated = rep( c(0,1),3)
Stratum    = rep( 1:3,each=2)

# additive (rate difference)

V.PY = Vaccinated * PY

#crude
summary(glm(D~ -1+PY+V.PY,
            family=poisson(link="identity") ) )

# as fn. of age band

S.1 = (Stratum==1); S.1.PY = S.1 * PY
S.2 = (Stratum==2); S.2.PY = S.2 * PY
S.3 = (Stratum==3); S.3.PY = S.3 * PY

summary(glm(D~ -1+ S.1.PY + S.2.PY + S.3.PY ,
            family=poisson(link="identity") ) )
```

[6] You can follow the same calculations as C&H, but do not use the 1/3, 1/3, 1/3 weights that they did; instead, use the observed distribution of the person-years of follow-up.

[7] See here and here.

```
# not easy to fit!
summary(glm(D~ -1+ S.1.PY + S.2.PY + S.3.PY + V.PY  ,
            family=poisson(link="identity") ) )
```

vi. Calculate the ratio of these two weighted rates, and an associated CI.

vii. Compare them with the results from a GLM Poisson model with (canonical) log link (if need be, see previous link). Comment on any differences.

```
# multiplicative (rate ratio)

#crude
fit = glm(D~ Vaccinated + offset(log(PY)),
    family=poisson)
summary(fit)
round(exp(fit$coefficients),3)

# incl. age
fit = glm(D~ as.factor(Stratum) + Vaccinated +
                offset(log(PY)),  family=poisson)
summary(fit)
round(exp(fit$coefficients),3)
```

viii. Summarize the new elements learned during this exercise.

ix. Indicate which approaches were not quite as satisfactory as you might like. (This might be a commercial for Chapter 15!)

Supplementary Exercise 14.5

Have a quick look at the UK and Israeli studies, and tell us

i. whether the focus in each one is on efficacy against infection, or against hospitalization or death if infected.

ii. whether the measures they used were rates (with PT denominators) or risks (with persons as denominators),

iii. whether there is a good case, when the target is case-hospitalization or case-fatality rates, to go with the 'risk' measure. [Hint: if, in the Scotland study, the follow-up were extended to 6, 12, 24 months post-Dx, what would happen to the rates? the risks? and their differences and ratios]

iv. which of the three studies has the most data/information as for the fully vaccinated vs. unvaccinated contrasts.

Supplementary Exercise 14.6 – COVID-19 Mortality rates in Black vs White Americans

See 2022 EXAMPLE on page 4 for context. The dataset JH was able to extract from CDC WONDER had 65 rows [3 years × 11 age groups × 2 races, minus one cell (2020, age band 1-4 years, Black) where (presumably) no COVID deaths were reported.] Since, for any age-race stratum, CDC WONDER reports the same population-size for all 3 years, JH added in a row with 0 deaths, to form this dataset of 66 rows.
Presumably, the data from 2020 cover just the portion of the year where there was the (new) cause of death code for COVID-19, 2021 covers a full 12-months, and 2022 covers what has been received up to the time the data were downloaded, in late October 2022. Moreover, there is a possibility, especially in the 2022 data, that there may be different (ie. race-specific) time lags in the notifications of death. Nevertheless, the data can help illustrate what has been going on, and why we need to pay attention to differences in age distributions when comparing death rates.

For each year separately,

i. Using different symbols or colours for the two races, plot the logs of the death rates against age, and try to indicate how 'stable' each datapoint is. How close do the rates follow Gompertz law? And how 'parallel' are the 2 sets of logRates?

ii. Plot the corresponding B:W RateRatios against age, again taking care to not over-emphasize the least precise ones.

iii. Using different symbols or colours for the two races to overlay both on the same plot, plot the proportions in each age-category against age, and comment on the difference. Also, for each race, calculate and show the mean age and the crude mortality rate.

iv. From the latter, calculate the crude mortality rate ratio and rate difference. In a soundbite/sentence, explain to a lay audience why the crude differences (and, presumably the differences for all cause mortality as well) do not align with the well-established 'vitality' differences between US Whites and Blacks.

v. Suggest ways to make a fairer comparison. You don't need to carry out the calculations, but do illustrate them using formulae that your research assistant could code up in `R` or in a spreadsheet.

Supplementary Exercise 14.7 – Association of SARS-CoV-2 Infection during Early Weeks of Gestation with Situs Inversus

This reminds JH of another disturbing observation – in 1941, following a different viral epidemic!