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

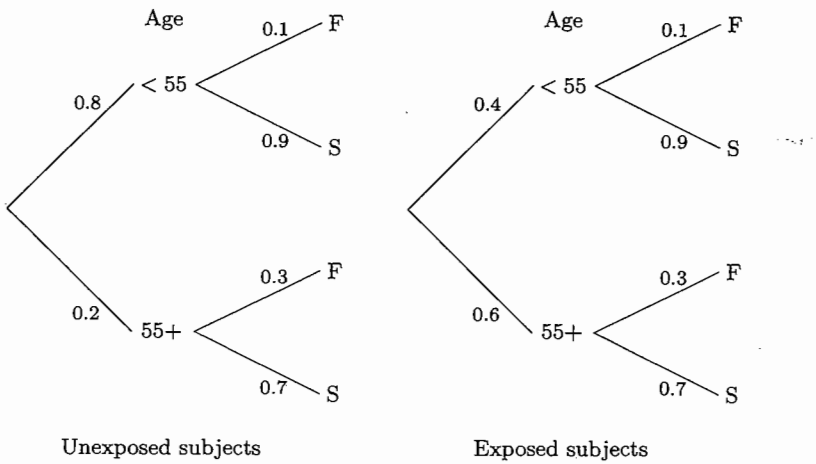


Fig. 14.1. Confounding by age.

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.

Extreme examples of confounding

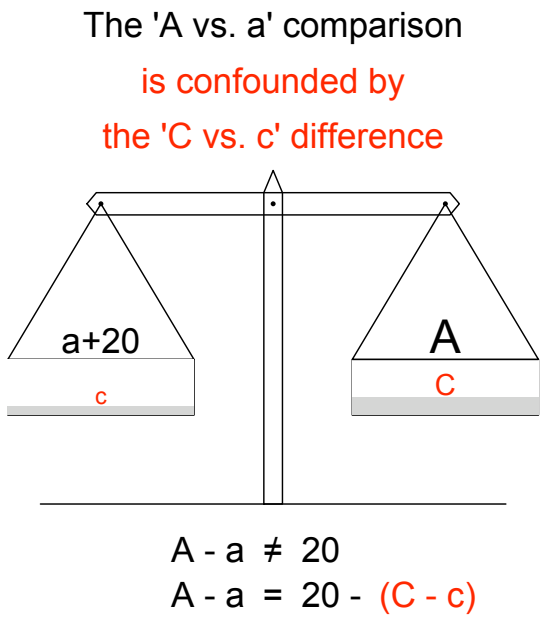
Rather that rely on made-up examples, it is also good to have real ones, and even extreme ones, to make the point. JH likes the extreme one Does Smoking Improve Survival? in the Expansion Modules in the website for the Moore and McCabe Statistics text [also given in the 1st chapter of Rothman's 2002 introductory text, with finer age-categories] Twenty-year survival status for 1314 women categorized by age and smoking habits at the time of the original survey. `http://www.whfreeman.com/statistics/ips/eesee4/eesees4.htm`.

JH's other favourite (again of *extreme* confounding) is the apparent gender-bias in admissions to the graduate schools at Berkeley (cf. "Sex Bias in Graduate Admissions: Data from Berkeley", an article by P. J. Bickel et al. in Science 7 February 1975: Vol. 187. no. 4175, pp. 398 - 404. [faculty-specific data are also in worked example of M-H technique in JH's notes for 'Chapter 9 epi' of course 607.] Most confounding is less extreme than in these two examples.

JH has a third example (a story told as a joke), which involves the taboo subjects of sex, religion and politics – topics that we are told we should not bring up in polite conversation, but which he is willing to tell anyway.

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.



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). Oner 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.

Supplementary Exercise 14.1 Sharper and **Fairer** Comparisons:

.

Effect of sexual activity on the longevity of male fruitflies

[Limit analysis to 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 category) relative to sexually inactive flies (reference category), 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 up the thorax range into 3 (roughly equal-sized) strata: small, medium and large. 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 same strata, compute the mean longevity for each of the 6 cells. Then, again using the overall proportions of flies in each stratum as the 3 weights, compute a standardized 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 the last few pages of JH's notes appended to the end of the Science article) 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 mother's-milk and IQ 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 variate). 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?