Chapter 3

Computing Effects

Measures of frequency

Consider a hypothetical trial that showed that drug A was superior to placebo: twenty percent of the patients who took the drug have recovered within one year as compared with ten percent of their counterparts who took placebo pills. Table 3–1 shows detailed results for that causal contrast.

Table 3–1. Results of a hypothetical trial

Treatment Group	Recovery	No recovery	Total
Drug A	200	800	1,000
Placebo	100	900	1,000

How will we estimate the effect of taking the drug as compared with taking placebo?

To compute a *measure of association* (which is assumed to be a measure of effect) we may either subtract or divide the frequency of recovery in the two treatment groups. There are, however, three possible measures of frequency we may subtract or divide: proportion, odds, and rate. All three measures make up a ratio—the number of events (here, the number of patients who recovered) divided by some reference count.

When the reference count in the denominator is the total number of people, including those who will eventually count as events, the frequency is called *proportion*. When the reference count includes only people who remained event-free, the frequency is called *odds*. When the reference count is person-time at risk of the event, the frequency is called *rate*. Notice again that all three measures share one numerator—the number of people who had an event—but each measure relies on a unique denominator.

I will elaborate below on each measure of frequency and then compute and discuss the derived measures of effect.

Proportion, proportion difference, and proportion ratio

Everyone knows what a proportion is. If 1,000 patients received a drug and 200 recovered within one year, then the proportion that recovered is 0.2 (or 20 percent). But from here on we enter a territory of dense fog and heated debates about the meaning and origin of that number. For me, and for many others I think, 0.2 is somehow related to the probability of recovery—a one-year summary of causal propensities to recover that were generated by the drug and by other causes of recovery. For others, called subjective Bayesians, the word probability belongs to a different school of thought and should only reflect "degrees of beliefs" that reside in the human mind: the higher the probability of an assertion, the greater the belief in its truth. In-between these extremes, there are gradations and equivocations. Not surprisingly perhaps, the two extreme views of the meaning of the number 0.2 are tightly linked to one's preference for a model of causation. While indeterminism embraces the probabilistic interpretation of some

proportions (though not all), determinism prefers to keep the physical phenomenon of an empirical proportion clearly separated the abstract concept of probability. We have already seen seeds of the disagreement in the first chapter when a determinist and an indeterminist voiced diverging opinions about the reading of an imaginary *risk-o-meter*. In the next chapter you will see a full explanation for their disagreement.

Regardless of labels, two measures of effect may be computed from proportions (here, percentages): If we subtract the percentage of patients who recovered while taking placebo from the percentage of patients who recovered while taking the drug (20%-10%), the resulting effect size is 10 percentage *points* (not 10 percent!) in favor of the drug. If we divide the numbers (0.2/0.1), the effect size is 2 in favor of the drug; no units attached.

In the first computation, we estimated the effect on an *additive scale*: taking the drug, rather than taking placebo, will add 10 percentage points to the percentage that will recover. In the second computation, we estimated the effect on a *multiplicative scale*: taking the drug will double the percentage that will recover with placebo. Naturally, those 10 percentage points are called a difference measure of effect (*proportion difference* or *probability difference*), whereas the number 2 is called a ratio measure of effect (*proportion ratio* or *probability ratio*). Since a ratio describes the size of a numerator relative to the size of a denominator, *relative proportion* and *relative probability* frequently substitute for proportion ratio and probability ratio. For now, don't be tempted to replace proportion or probability with the word "risk."

Although we estimated the effect of our causal contrast on recovery, we may also estimate its effect on the complementary outcome namely, on "no recovery", which might have included death. Among patients who took the drug, 80 percent have remained sick or died as compared with 90 percent of their counterparts who took placebo. Following this computation, the proportion difference remains unchanged, 10 percentage points in favor of the drug, but the proportion ratio changes to 0.89 (0.8/0.9=0.89). Therefore, we have to infer that taking that drug, rather than taking placebo, will double the proportion that will recover but not halve the proportion that will not recover. For some reason the effect on recovery when measured on a multiplicative scale is very different from the effect on remaining sick or dying which is the complementary condition. Which is the true number?

Both numbers are true. This small surprise is a mathematical property of proportions that has not received much attention. In a 2x2 table there are two proportion ratios for a binary outcome, each reflecting a category of interest. One way around the difficulty is to turn our back to the proportion ratio in favor of the proportion difference, a quantity that does not depend on the chosen category of the outcome. And indeed, many writers have argued that a difference measure of effect is superior to a ratio measure. The arguments they supply, however, rest on entirely different reasons. For example, a difference measure of effect fits well with deterministic causation and not so well with indeterministic causation.

From an indeterministic viewpoint, a ratio measure is undoubtedly superior because it quantifies unambiguously the strength of one force relative to another—whatever those pairs of forces may be. When we say that one causal propensity toward recovery is twice as strong as another (say, taking drug A relative to taking placebo), we have claimed all that needs to be claimed about reality. No units are needed. To me, this simple argument is powerful and convincing. Moreover, indeterminism allows for the possibility that the effect on recovery would be different from the effect on non-recovery, even though the two conditions are complementary.

Before turning to the odds, notice an important shortcoming of proportion-based measures of effect, whether a difference or a ratio. The upper limit of a proportion (1.0) imposes an undesirable constraint on the effect size. For example: if the proportion that recovered with placebo were 0.8, the effect size for any drug relative to placebo could not exceed 20 percentage points (100%-80%) on an additive scale and 1.25 (1.0/0.8) on a multiplicative scale. Yet from the viewpoint of indeterministic causation any constraint on a ratio measure of effect is unacceptable. Just as there is no limit to the relative strength of two magnetic fields, there should be no limit to the relative strength of two causal propensities. As we'll see later, other measures of effect do not share that shortcoming of a proportion.

Odds, odds difference, and odds ratio

Instead of asking how many patients have recovered per 1,000 recipients of drug A (which was 200 per 1,000 = 0.2), we may ask how many patients have recovered per 1,000 recipients of that drug *who remained sick*. (Again, the category "remained sick" includes deaths, if any.) That ratio, called the *odds of recovery*, is computed below for each treatment group using the data in Table 3–1.

Odds of recovery while taking the drug: 200 recovered / 800 remained sick = 250/1,000 = 0.25
Odds of recovery while taking placebo: 100 recovered / 900 remained sick = 111/1,000 = 0.11

A little arithmetic will show that the odds are a simple function of a proportion (or a probability). For example:

$$\text{Odds of recovery while taking the drug: } \frac{200}{800} = \frac{200/1{,}000}{800/1{,}000} = \frac{\text{Proportion recovered}}{\text{Proportion remained sick}}$$

$$\text{Or in notation:} \quad \text{Odds (Z)} = \frac{\text{Pr (Z)}}{\text{Pr (not Z)}} = \frac{\text{Pr (Z)}}{1 - \text{Pr (Z)}}$$

where odds (Z) reads "odds of Z" and Pr denotes proportion or probability, depending on personal preference and context. (Again, even in indeterminism not every proportion may be called probability. We'll see an example in a later chapter.)

Finally, notice that the odds of not recovering, 800/200, are just the inverse of the odds of recovering, 200/800. And in general: Odds (not Z) = 1 / Odds (Z)

The odds are a favorite measure of frequency among gamblers, but have not been treated with similar respect among epidemiologists, often portrayed as a weak measure that survives on its ability to estimate the proportion of a rare event. (When the event is rare— say a proportion of 0.1— the proportion and the odds are similar. Compare, for example, 100/1,000 to 100/900.) The odds, unlike proportion, have no upper limit, ranging from zero to infinity.

To find the range of the odds, we will use the definition Odds (Z) = Pr (Z) / Pr (not Z) and compute the odds for two extreme values of Pr (Z):

When Pr (Z) = 0, then Pr (not Z) = 1, and their ratio—the odds—is equal to zero. When Pr (Z) = 1, then Pr (not Z) = 0 and their ratio is infinite. Therefore, $0 \leq \Pr(Z) \leq 1$ implies $\text{Odds}(Z) \geq 0$.

In principle, both the *odds difference* and the *odds ratio* (also called *relative odds*) may serve as measures of effect, analogous to the proportion difference and the proportion ratio. Nonetheless, nobody seems to compute the odds difference, perhaps because the scale is unknown to the layperson. The odds ratio, in contrast, is widely used by epidemiologists, but mostly in case-control studies and cross-sectional studies and often reluctantly. The same prejudice against the odds has been carried to the odds ratio whose claim to fame, in most minds, rests on its ability to estimate the proportion ratio of a rare event. Can the odds ratio claim an independent status as a measure of effect? The answer is not simple at all.

On the one hand, since the odds are superior to the proportion for their lacking of an upper limit, the odds ratio is superior to the proportion ratio for the very same reason. Recalling a previous example: If 80% of 1,000 patients have recovered while taking placebo (odds of 800/200), the proportion ratio for any causal contrast with placebo cannot exceed 1.25 (1.0/0.8). The odds ratio, however, can take larger values because the odds of recovery have no upper limit. For instance, if 999 of 1,000 patients who took drug A have recovered, their odds of recovery is 999/1 and the corresponding odds ratio for recovery is (999/1) / (800/200) ≈ 250. And if all 1,000 recipients of that drug have recovered, both their odds of recovery and the odds ratio for recovery are infinite: (1,000/0)/(800/200).

A second reason to prefer the odds ratio as a measure of effect has to do with our earlier bafflement about which proportion ratio to compute—for recovery or for "no recovery". The odds ratio for "not Z" is always the inverse of the odds ratio for "Z": if taking a drug doubles the odds of recovery relative to taking placebo, it also halves the odds of "no recovery". To convince ourselves, let's consider again the causal contrast in Table 3–1: drug A relative to placebo.

$$\text{Odds ratio for recovery:} \quad \frac{200 / 800}{100 / 900} = 9/4$$

$$\text{Odds ratio for "no recovery":} \quad \frac{800 / 200}{900 / 100} = 4/9$$

This property of the odds ratio follows a property of the odds that was mentioned earlier: Odds (not Z) = 1 / Odds (Z). The proof is shown below for the causal contrast of a drug with placebo.

$$\text{Odds ratio for “not Z”} = \frac{\text{Odds (not Z) given DRUG}}{\text{Odds (not Z) given PLACEBO}} = \frac{1 \,/\, \text{Odds (Z) given DRUG}}{1 \,/\, \text{Odds (Z) given PLACEBO}} =$$

$$= \frac{\text{Odds (Z) given PLACEBO}}{\text{Odds (Z) given DRUG}} = \frac{1}{\dfrac{\text{Odds (Z) given DRUG}}{\text{Odds (Z) given PLACEBO}}} =$$

$$= \frac{1}{\text{Odds ratio for Z}}$$

Unfortunately, these two advantages of the odds ratio succumb to one, possibly fatal, mathematical deficiency, which is called "non-collapsibility". Even if the causal contrast of a drug with placebo has a constant effect on recovery for each person, such as an odds ratio of 10, the *observed* odds ratio in the sample might not yield that number. To illustrate this puzzling mathematical property, I chose a hypothetical sample of 1,000 that is composed of two groups (Table). One group (100 people) has much higher chances of recovery than the other (900 people), yet both groups share the same odds ratio for recovery: in both groups the odds of recovery if taking drug A are 10-times the odds of recovering if taking placebo.

		Odds of recovery	Probability of recovery*	Number expected to recover	Number expected not to recover
100 people	If taking drug A	0.90	0.47	100 x 0.47 = 47	53
	If taking placebo	0.09	0.08	100 x 0.08 = 8	92
		OR=10	PR = 5.9		
900 people	If taking drug A	0.10	0.09	900 x 0.09 = 81	819
	If taking placebo	0.01	0.01	900 x 0.01 = 9	891
		OR=10	PR = 9		

* Odds(Z)=Pr(Z)/(1-Pr(Z)). If we isolate "Pr(Z)", we get Pr(Z)=Odds(Z)/(1+Odds(Z))
OR denotes odds ratio; PR denotes probability ratio

Based on these data, you would probably expect that the observed odds ratio for recovery in the whole sample of 1,000 people would be 10 as well. Well, it is not. After displaying the count of people who are expected to recover and the count of people who are expected to not recover (Table), the computed odds ratio turns out to be 8.5.

	Recovered	Not recovered	Total
If taking drug A	128(=47+81)	872(=53+819)	1000
If taking placebo	17(=8+9)	983(=92+891)	1000

$$OR = \frac{128/872}{17/983} = 8.5 \qquad PR = \frac{128/1000}{17/1000} = 7.5$$

Notice that the probability ratio in the whole sample (7.5) resides between the probabilities we computed for the two groups (5.9 and 9), which seems reasonable. It is their average. This fact has inspired many writers to belittle the odds ratio and claim that the proportion ratio is the "right" measure of effect. Well, neither measure is deficiency-free so I am not sure how to reach a verdict, especially when a better measure is waiting just around the corner, in the next section (rate ratio). Eventually, both the proportion ratio and the odds ratio may reserve a seat in causal inquiry only to the extent that they estimate the rate ratio. Or maybe the issue is much deeper than I can explain here.

I will conclude this section with two short notes: one is mathematical; the other, semantic.

The odds ratio and the proportion ratio always behave according to the following rules of inequality (proof omitted):

If the odds ratio > 1, then the odds ratio > the larger of the two proportion ratios.
If the odds ratio < 1, then the odds ratio < the smaller of the two proportion ratios.

For example: In our hypothetical trial the odds ratio for recovery (drug A relative to placebo) was 2.25—larger than the corresponding proportion ratio, which was 2 (and obviously larger than the proportion ratio for "no recovery".) Likewise, the odds ratio for "no recovery" is 0.44—smaller than the corresponding proportion ratio, which is 0.89. This fact has inspired many writers to criticize the odds ratio for exaggerating the effect size, implicitly assuming that the proportion ratio is the "right" measure of effect. Well, neither is deficiency-free so I am not sure how to reach a verdict, especially when a better measure is waiting just around the corner, in the next section.

On a semantic note: You may have noticed that "the odds of recovery are 200 to 800" or "the odds ratio is 9/4" sound much like probabilistic expressions. Nonetheless, opponents of empirical probabilities use these expressions freely for lack of a substitute. No one in the odds department has yet invented a counterpart to the word proportion.

Rate, rate difference, and rate ratio

In every group of people, both the proportion of an event and the odds of an event, say death, depend on follow-up time. Bizarre exceptions aside, the proportion of deaths within one minute is close to zero (as are the odds), whereas the proportion of deaths within 100 years is about 1 (and the odds are large or infinite.) In-between, the value of each measure of frequency almost always increases and never decreases.

This dependency of proportion and odds on follow up time detracts from the merit of derived measures of effect such as proportion ratio and odds ratio. Imagine, for example,

a 100-year long trial of the effect of some drug on death as compared with placebo. Toward the end of that trial, both the proportion ratio and the odds ratio converge toward 1—no effect of the drug—even if the drug prolongs life as compared with placebo. And if we dismiss this trial as unrealistically long, another embarrassing difficulty arises in a trial that is too short, say, one-hour long. In either example it is hard to explain exactly what went wrong. Why does the length of a trial strikingly distort the estimated effect? Moreover, how and where to draw a theoretical range for permissible trial length?

Trials may indeed be too long or too short, but it only means that they are economically inefficient—either consuming extra resources for no extra return or costing too little to produce useful results. It seems awkward, however, to call a trial flawed just because it was allowed to continue for many years. From a theoretical viewpoint, the fault is not rooted in the length of any study but in those measures of frequency from which we compute measures of effect. Rather than quantifying the *strength* of causal forces, proportion and odds quantify the *cumulative effect* of those forces up to arbitrary time. And a cumulative effect is produced from a mixture of two ingredients: the strength of causal forces, which may be constant over time, and follow up time, which continually increases.

To measure the strength of causal forces alone, we should look for a measure of frequency that does not obligatorily change over time—a measure that may produce a single number instead of a series of numbers that depend on follow-up time. Such a measure shall describe the *velocity* at which an event happens: the higher the velocity, the stronger the underlying causal forces. This measure is called the *rate* of an event—the number of events divided by the person-time at risk of that event.

When "person-time at risk" is presented to students for the first time, many of them are puzzled about the logic behind this idea. (I was one.) Drawing an analogy from persons to cars and from person-time to driving-time might help. If you were asked to compare the causal forces toward a car crash in two towns, you would not be satisfied with knowing the number of crashes relative to the number of cars in each town. You would also want to know how much time drivers in each town spend on the road (and perhaps how many miles they drive.) Suppose, for example, that the number of cars were identical in the two towns, the proportion of crashes in one town twice the proportion in the other, yet drivers in the crash-prone town also spend twice as much time on the road. If so, the number of crashes per car-on-the-road-time—that is, the *rate* of crashes—would have been the same in these towns.

We may gain deeper understanding of "person-time at risk" from thinking some more about the dimensions in which causes operate and their effects occur. One obvious dimension is "head count"—how many people are at risk of death or how many cars are at risk of a crash. Another equally important dimension is the axis of time because no event will happen if time were to freeze. "Person-time at risk" fully incorporates the time dimension, by tallying the time-at-risk for each person, whereas proportion and odds treat time naively, as if it were uniform for all people at risk. For instance: when we say that the odds of one-year death in a group are 0.25 (1 death per 4 survivors), we implicitly assume that every group member was known to be at risk for one full year, which means that no one was lost to follow up and that all deaths occurred on the 365th day.

Though almost always false, the assumption of uniform time-at-risk may be tolerated under the following condition: Both events and losses to follow up occur infrequently. If a 1,000-patient trial lasted one year, 2 percent of the patients have died during that year, and 2 percent have been lost to follow-up, then we don't grossly err by assuming that all

1,000 patients were at risk of death for one full year. On the other hand, if half the patients have died in the first six months, then at least half the patients were not at risk of death for a full year. (After having an event, one is no longer at risk of precisely the same event even when recurrence is possible.)

Must the rate be constant over time?

Since the rate measures the velocity at which events occur and since that velocity mirrors underlying causal forces, any change in these forces will affect event velocity—that is, the rate will change. The recovery rate among recipients of a drug may be slow during the first month of treatment, fast in the next five months, and slow again thereafter. Such fluctuation may be explained by accentuation or attenuation of causal forces toward recovery, including the force generated by the drug (if any). Now, stretching our scientific-mathematical imagination, we may picture an extreme situation where the rate *continually* changes during some period such that every moment—every time *t*—has a unique rate. We'll call that number the *instantaneous rate*. The simplest analogy again comes from the driving world. The speed of a car, as recorded at each moment on the speedometer, need not be constant, and may change from moment to moment by applying force to the accelerator or the brake.

To define the instantaneous rate at a time point *t*, we need to follow a mental exercise in which we compute the rate repeatedly for a shrinking time interval $[t, t+\Delta t]$, where Δt is positive and tends to zero. The resulting series of numbers will converge toward some value—the instantaneous rate at time *t*, denoted R (t). (Read "R of t" or "R as a function of t.")

Using notation: $R(t) =$ limit of R in the interval $[t, t+\Delta t]$ when $\Delta t \rightarrow 0$

The instantaneous rate is often called the *hazard rate* or just the *hazard*, a historical misnomer. The term "hazard" was coined in industrial studies in which events were always bad—like the premature failure of a light bulb. And indeed, in statistical textbooks you may find expressions such as "the hazard of failure" and "failure-time analysis" where "failure" stands for "event". This jargon resonates well in some human studies, say, failure of a drug to prevent death, but sounds silly in others. Cancer patients would be delighted to hear that "their hazard of failure is high", if the speaker is talking about the instantaneous rate of cancer remission.

I have criticized proportion and odds for their time-dependency, so you may wonder why I am not condemning the rate for *its* time-dependency. Though time affects all three measures of frequency, it affects them for fundamentally different reasons. The rate may change over time because causal reality could change—causal propensities are not obliged to remain constant—whereas proportion and odds change over time, in part because *we* change the follow-up time. *We* determine how many days or months or years a study will last, and thereby influence the values of proportion and odds. Of course, arbitrary follow-up time also enters into every rate we compute (by setting the maximum person-time at risk) but that does not diminish one bit from the theoretical standing of the instantaneous rate. When the instantaneous rate is constant over some follow-up time, our computed rate for that period estimates that number directly. When it is not constant, our computed rate is some average of several, or possibly infinite, instantaneous

rates. Keep in mind that the instantaneous rate is a *parameter*—a number that exists out there, so we think, but forever remains unknown.

Now, I have a small pedagogical confession to make. When I criticized proportion (and odds) for their time-dependency, I deliberately avoided using the word probability because it's not true that probability always mixes follow-up time with causal propensities. We can define the *instantaneous probability* (and, therefore, the *instantaneous odds*) in complete analogy to the instantaneous rate, and it is even possible to derive one from the other. Nonetheless, there is a key difference between probability and rate: the leap from instantaneous probability to computed proportion does not resemble the leap from instantaneous rate to computed rate. Computed proportion does not estimate—nor is it the average of—instantaneous probabilities. Follow-up time does mix in.

After grasping the meaning of a rate, the rest should be simple. Like proportion and odds, the rate provides us with two derived measures of effect: rate difference, with units attached, and rate ratio, which is unit free. From an indeterministic viewpoint, the instantaneous rate ratio has a complete set of desired properties: it quantifies the relative strength of two causal propensities without mixing in follow-up time; it is unit free; and it has no upper limit. What else could we ask for?

Rate ratio and rate difference can be computed only when we obtain data from a trial or a cohort study. In other designs the time at risk remains unknown, so we must turn to proportion-based measures of effect or to odds-based measures of effect. Fortunately, however, when the event happens infrequently, both the proportion ratio and the odds ratio estimate the rate ratio reasonably well: all three ratio-measures of effect yield similar results (proofs omitted). And as we'll see in chapter 14, in one version of the case-control study the odds ratio directly estimates the rate ratio, regardless of how frequently the event happens.

To conclude this section, we'll compute the rate difference and the rate ratio for the example that opened this chapter, using hypothetical data on person-time at "risk" of recovery (another example of a misnomer).

Suppose that no patient was lost to follow up during that one-year trial and that recovery of patients in both treatment groups has spread *uniformly* over time. When an even number of events spread uniformly over one year, we can pair all of the patients who had the event such that each pair contributes exactly one person-year at risk. For example, a person who had the event after one month will be paired with a person who had the event after 11 months, for a combined person-time of one year. If so, the 1,000 recipients of drug A have contributed 900 person-years: 100 person-years by the 200 patients who have recovered (200/2) and 800 person-years by those 800 patients who remained sick. Similar calculation yields 950 person-years of 1,000 placebo recipients, recalling that 100 of them have recovered (Table 3–1).

The rate difference is therefore,
$200/900 - 100/950 = 222/1{,}000 - 105/1{,}000 = 117$ per 1,000 person-years.

In words: If 1,000 patients were to take the drug rather than placebo for one year, 117 additional patients will recover during that year.

And the rate ratio is, $\dfrac{200 / 900}{100 / 950} = 2.1$

In words: the causal propensity to recover while taking the drug is 2.1 times as strong as the causal propensity to recover while taking placebo.

What is "risk"? (three paragraphs on fuzzy language)

Terms such as "risk", "rate", "confounding", and "bias" are found in the vocabulary of the layperson who fires them easily—sometimes appropriately and more often not. Rate, in particular, is often misused or ambiguously used by laypersons and scientists alike in phrases such as "prevalence rate" (should be prevalence proportion) and "one-year event rate" (rate or proportion?)

But the first prize for ambiguity undoubtedly goes to the word "risk" and its main derivative the "relative risk". For some writers, risk is synonymous with probability. For others it is a synonym for both probability and rate but not for odds. For others yet, risk is a generic descriptor of all three measures of frequency. I have even encountered writers who felt a need to tack "risk" in front of odds (risk-odds), perhaps to help the poor odds gain some credentials for causal inquiry.

A remedy may be found when precision replaces ambiguity and some conventions emerge from within the guilds of epidemiologists and statisticians. Until then, I suggest that you try to avoid the word "risk" whenever possible. Most often, though not always, there is a more precise word you can use: proportion, probability, odds, rate, hazard. And if you must make a statement such as "The odds ratio estimates the relative risk", be sure to explain that risk means proportion to you (if it does) and that ratio and relative are interchangeable words. Or better, just write: "The odds ratio estimates the proportion ratio" or "The relative odds estimates the relative proportion".

Effects on continuous variables

If the world were made up only of binary outcomes—crash or no crash, dead or alive, recovered or not—this chapter could have ended right after the last section. Many effects, however, do not come from binary outcomes. Some dwell in discrete, related categories such as full recovery, partial recovery and no recovery, whereas others make up a continuous variable: blood pressure, hemoglobin concentration, lung function—to name a few examples.

When the effect takes the form of discrete categories or when we artificially transform a continuous variable into a categorical one, we can still rely on proportion and odds and rate to estimate effects, although it's not always that simple. For instance: does it make sense to compute the odds of partial recovery counting in the denominator patients who remained sick *and* patients who have fully recovered? Perhaps not. Is it appropriate to count in that denominator only patients who remained sick? Certainly not. If you discard one of three categories of a variable, the ratio of the remaining two categories is not odds. Any gambler will tell you.

We will leave these technical difficulties unresolved and turn our attention to a far more common challenge, which would require fresh methods: how to estimate the effect of an exposure on a continuous variable? But first, one paragraph on a semantic matter.

What a cause does to a continuous variable is to determine deterministically, or to help determine through a causal propensity, a value—blood pressure, hemoglobin concentration, lung function. Since the phrase "smoking is a cause of the value of lung

function" sounds awkward, a cause of a continuous variable takes on another name—*determinant*, a semantic twist on the word "cause" that unfortunately sounds as if it was derived from the word determinism. Keep in mind, therefore, that there is no theoretical difference between the expressions "smoking is a cause of lung cancer" and "smoking is a determinant of a measure of lung function called FEV_1" (short for forced expiratory volume in one second.) Both expressions state a causal theory; neither implies a particular model of causation.

Arithmetic mean difference

Because a determinant of a continuous variable does not bring about an event—an FEV_1 of 2.1 liters is not "an event"—we cannot estimate effects from measures of frequency. Another measure of effect in needed. The simplest substitute is *the difference in the arithmetic mean between two causal assignments*, for example, the difference in mean FEV_1 between smoking and former smoking. I will illustrate using real data.

Table 3–2 shows the first few observations (or rows or records) from a dataset that contained the FEV_1 values of 3,968 smokers and 4,931 former smokers. The larger one's FEV_1, the better is one's lung function.

Table 3–2. First 8 observations of smoking status and FEV_1

Observation	Smoking Status	FEV_1 (liters)
1	Smoker	3.15
2	Smoker	2.58
3	Smoker	2.02
4	Former smoker	2.64
5	Smoker	3.11
6	Smoker	1.14
7	Former smoker	2.36
8	Former smoker	3.55

Source: ARIC

In that dataset, the arithmetic mean of FEV_1 among former smokers was 3 liters, whereas the arithmetic mean among smokers was 2.6 liters. Therefore, the estimated effect of that causal contrast on FEV_1 is a difference of 0.4 liter in favor of former smoking.

But before building up too much faith in that estimate, think for a moment about the variable SMOKING STATUS and its causal assignments. There is no doubt that this variable is located giant steps away from any theoretical exposure of interest. Both continued smoking and former smoking are composed of many different causal assignments—continued smoking of 2 packs a day for 20 years, continued smoking of 2 cigarettes a day for 5 years, quitting smoking 20 years ago after having smoked 5 cigarettes a day for 5 years, recent quitting after having smoked 2 packs a day for 20 years, and many more. Do all nested causal contrasts between smoking and former smoking produce a difference in FEV_1 of 0.4 liter? Of course not. But for pedagogical purposes let's ignore that difficult reality and pretend that both smokers and former smokers were

homogenous groups—say, all smokers had smoked a pack a day for 20 years and all former smokers had quit 10 years before their FEV_1 was measured, after having smoked a pack a day for 10 years. To alleviate fear of confounding (chapter 6), we'll also pretend that smoking status was randomly assigned. With these generous assumptions in mind, 0.4 liter is the estimated effect of not smoking for 10 years after having smoked a pack a day for 10 years.

What are the properties of the mean difference?

As its name implies, the mean difference belongs to the class of difference measures of effect, quantifying effects on an additive scale. And like its counterparts in that class the mean difference carries the units along, a property that makes it dependent on the measurement scale. But the main drawback of this measure has to do with its claim about causal reality: After 10 years of smoking, continued smoking for another 10 years, rather than abstaining, should lead to a difference in FEV_1 of 0.4 liter regardless of whether the value of FEV_1 at the start point was 3 liters or 2 liters or 1 liter. The mean difference tells us that a causal contrast works with a constant absolute effect, no matter what the start point may be (Figure 3–1.) Thinking back for a moment, the same message was delivered by difference measures of frequency. Is this how Nature works? Always? Never?

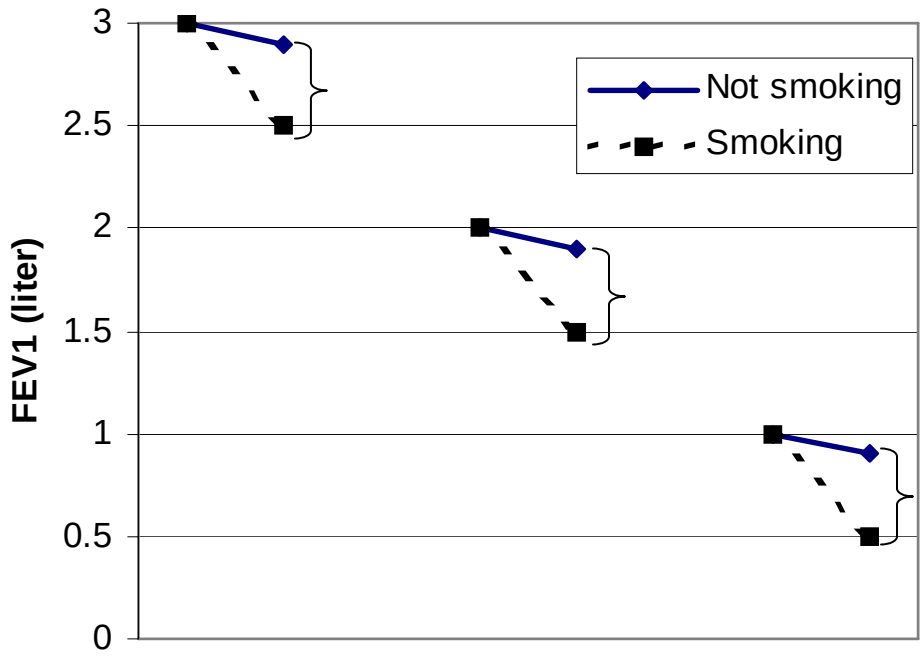

Figure 3–1. Estimated effects of smoking status on the FEV_1 of three people based on a mean difference of 0.4 liter

Nobody knows the answer. No more than anyone knows whether a rate difference describes causal reality better than a rate ratio or, at a deeper level, which model of causation operates in this ultra-complex world. Some writers provide examples where a constant difference seems empirically appropriate and other examples where it does not,

but there is a whiff of circular reasoning here. We rely on data to compute a difference measure of effect, not knowing that we chose the appropriate measure, yet argue that data-driven results could somehow retroactively certify or discredit our choice.

The mean difference unquestionably fails in one situation and rests on shaky grounds in several others. If the effect variable takes only positive values and is bound at zero, the mean difference cannot do the job of describing causal reality for some causal contrasts. Suppose, for example, that the difference in mean FEV_1 between *smoking* and *never smoking* were 1 liter in favor of never smoking. Can that number describe the effect of taking up smoking for people whose initial FEV_1 is 0.9 liter? It cannot. Smoking or not smoking, their FEV_1 will never decline to −0.1 liter (not to mention that a range of above zero values is incompatible with life).

We should also question the performance of the mean difference whenever the distribution of the effect variable is skewed rather than symmetrical. The arithmetic mean of a severely skewed distribution does not point to the center of the distribution because one tail strongly pulls the value in that direction. And if the meaning of the mean is questionable, so is the meaning of any mean-derived measure.

Two other shortcomings of the arithmetic mean are often cited by statisticians: Similar to skewness, outlying values of an otherwise symmetrical distribution will pull the mean toward their end, and unequal variance of two means will undermine the variance of the mean difference, a crucial statistic that will show up in later chapters.

Geometric mean ratio

Regardless of whether a mean difference makes sense, we should look for an alternative—for a unit-free ratio measure that can quantify causal propensities for effects that reside in a continuous variable. A natural candidate is the *geometric mean ratio*, the ratio of two geometric means. Table 3–3 shows the analogy between the geometric mean and the arithmetic mean.

Table 3–3. A comparison of the geometric mean and the arithmetic mean

	Arithmetic mean	Geometric mean
Measurement scale	Additive scale	Multiplicative scale
Definition	$\frac{a_1 + a_2 + ... + a_n}{n}$	$(a_1 \cdot a_2 \cdot ... \cdot a_n)^{1/n}$ or, written differently: $\sqrt[n]{a_1 \cdot a_2 \cdot ... \cdot a_n}$
Shorthand symbols*	$\frac{\sum_{i=1}^{n} a_i}{n}$	$\left(\prod_{i=1}^{n} a_i\right)^{1/n}$
Measure of effect	Arithmetic mean difference	Geometric mean ratio

* Σ is shorthand for repeated summation; Π is shorthand for repeated multiplication.

Despite crowded notation, the computation is simple: to calculate the arithmetic mean, we add up n numbers and divide the sum by n; to calculate the geometric mean, we multiply n numbers and take the n-root of the product. To gain intuitive understanding

of why the geometric mean describes the center of a distribution, think first about a group whose members have the same value of a ($a_1=a_2=\ldots=a_n$.) If so, the geometric mean is the n root of n self-multiplications of a, which is equal to a—as expected. Now think about a group of n members whose values are not exactly the same but are rather scattered around some center. The n root of the product of all these values should inform us where that center is located.

We can compute the ratio of two geometric means, say group A relative to group B, using the formula $\sqrt[n]{a_1 \cdot a_2 \cdot \ldots \cdot a_n} \,/\, \sqrt[n]{b_1 \cdot b_2 \cdot \ldots \cdot b_n}$), but that would require many multiplications and working with extremely large products. Just think about multiplying 3,968 FEV_1 values of smokers and 4,931 values of former smokers…

Which brings us to logarithmic transformation—a handy tool for numerous statistical methods of causal inquiry. (This may be a good time to refresh your memory of the algebra of logarithms and exponents.)

It is not too difficult to show that the following equality holds for any logarithmic base (say, base 10):

$$\log(\text{geometric mean}) = \frac{\log a_1 + \log a_2 + \ldots + \log a_n}{n}$$

Or in words: the logarithm of the geometric mean is equal to the arithmetic mean of log-transformed values.

Proof (if interested):

$$\log\left(\sqrt[n]{a_1 \cdot a_2 \cdot \ldots \cdot a_n}\right) = \log\left[\left(a_1 \cdot a_2 \cdot \ldots \cdot a_n\right)^{1/n}\right] = \log\left(a_1^{\,1/n} \cdot a_2^{\,1/n} \cdot \ldots \cdot a_n^{\,1/n}\right) =$$

$$\log\left(a_1^{\,1/n}\right) + \log\left(a_2^{\,1/n}\right) + \ldots + \log\left(a_n^{\,1/n}\right) = \frac{1}{n}\log(a_1) + \frac{1}{n}\log(a_2) + \ldots + \frac{1}{n}\log(a_n) =$$

$$\frac{\log(a_1) + \log(a_2) + \ldots + \log(a_n)}{n}$$

This identity will help us to compute the *log of the geometric mean ratio* for group A relative to group B. (Then, one quick exponentiation at the end will give us the geometric mean ratio.)

$$\log\left(\frac{\text{geometric mean}_{\,A}}{\text{geometric mean}_{\,B}}\right) = \log\,(\text{geometric mean}_{\,A}) - \log\,(\text{geometric mean}_{\,B}) =$$

$$= \frac{\log a_1 + \log a_2 + \ldots + \log a_n}{n_A} \;-\; \frac{\log b_1 + \log b_2 + \ldots + \log b_n}{n_B}$$

Exponentiation of the last expression (taking the antilog) will yield the geometric mean ratio, which was our goal. But check again what the last mathematical expression means. It is simply the arithmetic mean difference of a new variable, a variable whose values were computed from the original variable by taking the logarithm.

The geometric mean ratio will fail when the effect variable takes negative values, but that doesn't happen often in biomedical science. A far more common problem is a cluster of zero values. In such cases, we add a small positive constant to every value, whether zero or not, thereby slightly shifting the distribution to the right. A word of caution, though: I have seen results that were sensitive to the choice of that small constant—adding 0.1 and adding 0.01 have led to substantially different geometric mean ratios even though the original values ranged from 0 to 50.

Back to our example. To compute the geometric mean ratio of FEV_1 for smokers relative to former smokers, we'll first take the log of every FEV_1 value in the dataset, compute the mean of the new variable among smokers (0.393) and among former smokers (0.462), compute the mean difference (–0.069), and transform back the result by exponentiation ($10^{-0.069} = 0.85$). And if we choose to subtract in reverse order, the mean difference will be positive (0.069) and the geometric mean ratio greater than one ($10^{0.069}$=1.17), which is just the inverse value ($10^{0.069} = 1 / 10^{-0.069}$). Any base will work as long as it is used throughout.

What does the number 0.85 say about causal reality?

In the language of indeterminism this number compares, on a ratio scale, two causal propensities to determine, or set, or influence the value of FEV_1. As compared with quitting, continued smoking for another 10 years is trying to set the value of FEV_1 at a lower value—at 0.85 of whatever value the other causal assignment is aiming at. If abstaining from smoking is aiming at an FEV_1 of 2 liters after 10 years, continued smoking is aiming at 0.85 that value, which is 1.7 liter (15% lower). And if abstaining is aiming at 1 liter, continued smoking is aiming as 0.85 liter (again, 15% lower.) A causal contrast works to produce an effect whose value is a constant ratio, which implies that the absolute effect will not be constant. Unlike the arithmetic mean difference, the absolute effect *will* depend on one's FEV_1 at the start point (Figure 3–2.)

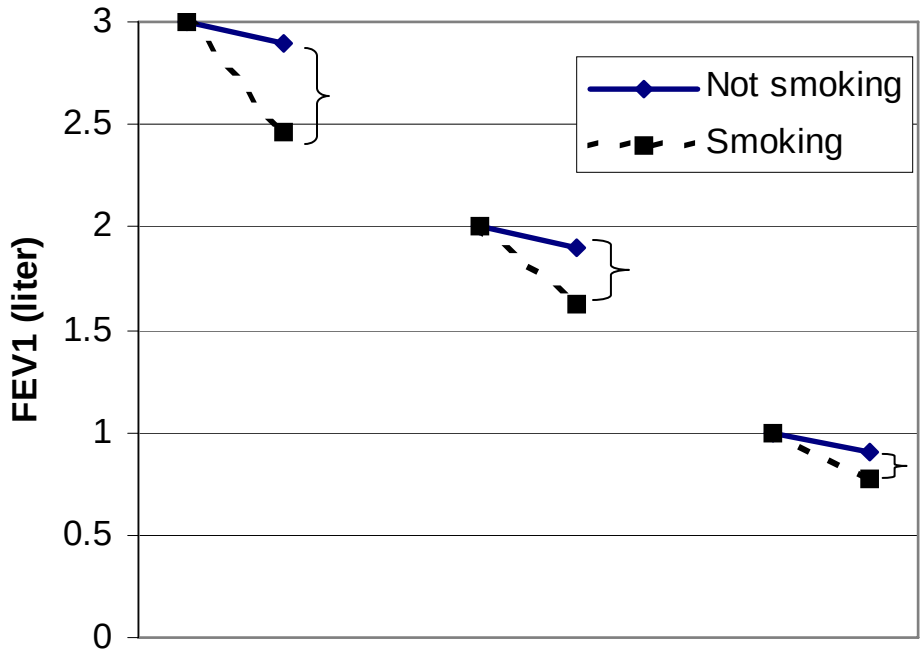

Figure 3–2. Estimated effects of smoking status on the FEV_1 of three people based on a geometric mean ratio of 0.85 (or 1.17)

Arithmetic mean difference or geometric mean ratio?

Comparing Figure 3–2 with Figure 3–1, we realize that the geometric mean ratio and the arithmetic mean difference are not just two ways to estimate an effect. They are alternative models of causal reality that sharply disagree about the magnitude of effects beyond any reconciliation. One model says that a causal contrast tends to bring about a constant difference whereas the other says that a causal contrast tends to bring about a constant ratio (which means a constant percentage change.)

To see how acute the disagreement is, consider another example—the effect of two alternative treatments on the length of hospital stay. Suppose we computed for that causal contrast an arithmetic mean difference of 2 days and a geometric mean ratio of 0.5. What inference will we draw from each number?

The arithmetic mean difference tells us that one treatment will shorten the hospital stay by 2 days as compared with the other: from 3 days to 1 day, from 6 to 4, from 10 to 8. In contrast, the geometric mean ratio tells us that one treatment will shorten the hospital stay by 50%: from 3 days to 1.5 days, from 6 to 3, from 10 to 5. These are very different claims about causal reality no study can adjudicate: you have to choose one or the other. If you subscribe to indeterminism the choice is easy—compute a ratio. If you subscribe to determinism, you may be able to rationalize your way to the mean difference. And if you don't want to commit to either model of causation, you will have to choose by other means: statistical assumptions and arguments, or weak theoretical arguments (for example, I *think* that the causal contrast works by cutting a fixed number of hospital days.)

Many statisticians advocate the use of logarithmic transformation—and therefore, the geometric mean ratio—when some statistical assumptions justify transformation. Which brings up interesting questions about the interplay between mathematical reasoning and causal reasoning. Should scientific inquiry comply with statistical preference or should statistical procedures comply with scientific preference? Is statistics only a tool in the hands of scientific inquiry or is it also a prescription for how scientific inquiry should proceed? Which comes first and how much influence should one have on the other?

We will return to these fascinating questions in later chapters in connection with other crucial decisions in causal inquiry. For now I will offer my general answer and let others offer dissenting views. I don't like the idea that I should succumb my philosophy of science to statistical dictates, if there are any. (Statistics speaks in many voices—hypothesis testing, estimation, frequentist, likelihood, Bayes.) In my opinion, statistics is a toolkit to choose from to one's scientific liking. Some statistical tools serve well my views about causal inquiry, others do not serve so well, and others do not fit at all. No amount of statistical-philosophical ink, for example, will convince me to use anything in statistics that reads like a Bayesian thought (with all due respect to my Bayesian colleagues.)