Explorations in statistics: the analysis of ratios and normalized data

Douglas Curran-Everett
Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado; and Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado

Submitted 31 May 2013; accepted in final form 11 July 2013

Curran-Everett D. Explorations in statistics: the analysis of ratios and normalized data. Adv Physiol Educ 37: 213–219, 2013; doi:10.1152/advan.00053.2013.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This ninth installment of Explorations in Statistics explores the analysis of ratios and normalized—or standardized—data. As researchers, we compute a ratio—a numerator divided by a denominator—to compute a proportion for some biological response or to derive some standardized variable. In each situation, we want to control for differences in the denominator when the thing we really care about is the numerator. But there is peril lurking in a ratio: only if the relationship between numerator and denominator is a straight line through the origin will the ratio be meaningful. If not, the ratio will misrepresent the true relationship between numerator and denominator. In contrast, regression techniques—these include analysis of covariance—are versatile: they can accommodate an analysis of the relationship between numerator and denominator when a ratio is useless.

Analysis of covariance; model II regression; ordinary least-squares regression

This ninth paper in Explorations in Statistics (see Refs. 13–20) explores the analysis of ratios. As researchers, we compute a ratio—a numerator divided by a denominator—to compute a proportion for some biological response or to derive some standardized variable.1 In each situation, we want to control for differences in the denominator when the thing we really care about is the numerator: that is, we compute a ratio to normalize—or standardize—the numerator to the denominator. Often, we want to know if a ratio differs between groups: for example, between old and young or between treated and control.

Physiology is rife with ratios (Table 1). No wonder. A ratio is seductively simple. There is just one problem: the analysis of a ratio is complex. In this exploration, we will see why. First, we need to review the software we will use to investigate ratios.

R: Basic Operations

The first paper in this series (13) summarized R (46) and outlined its installation. For this exploration there are two additional steps: download Advances_Statistics_Code_Ratios.R2 to your Advances folder and install the extra packages lmodel2 and smatr.3

To install these packages, open R and then click Packages | Install package(s) . . . . 4 Select a CRAN mirror close to your location and then click OK. Select lmodel2 and smatr and then click OK. When you have installed these packages you will see package ‘lmodel2’ successfully unpacked and MD5 sums checked
package ‘smatr’ successfully unpacked and MD5 sums checked
in the R Console.

To run R commands. If you use a Mac, highlight the commands you want to submit and then press ⌘- during key (command key + enter). If you use a PC, highlight the commands you want to submit, right-click, and then click Run line or selection. Or, highlight the commands you want to submit and then press Ctrl+R.

The Trouble with Ratios: an Overview

Unless there is reason to believe that the regression passes through the origin, the ratio is of dubious value. It may do for rough work, but careful experimentation deserves [a] more efficient statistical method.

George Snedecor (1946)

If a ratio controls for the impact of the denominator on the numerator, then there will be no relationship between the ratio and the denominator: the magnitude of the ratio will remain constant across all observed values of the denominator. But there is a paradox: if there is no relationship between the numerator and the denominator, then the mere calculation of the ratio will create a relationship between the ratio and the denominator (Fig. 1). If there is a relationship between the numerator and the denominator, a ratio will be effective only when the relationship between the numerator and the denominator is a straight line that intersects the origin (Fig. 2). This inescapable phenomenon is the foundation of Snedecor’s understated warning.

In a 1949 paper published in the Journal of Applied Physiology, Tanner (57) illustrated the perils associated with ratios using physiological metrics like stroke volume and glomerular filtration rate. People in other disciplines have followed suit (1–4, 6, 35, 42, 43, 48).

1 When we compute a proportion, the numerator and denominator represent the same biological response; we measure them at different times but in the same units. When we derive a standardized variable, the numerator and denominator represent different biological responses; we measure them at the same time but in different units.

2 This file is available through the Supplemental Material link for this article at the Advances in Physiology Education website.

3 The lmodel2 package executes regression in situations where the predictor variable X includes measurement error, and the smatr package executes standardised major axis estimation and testing routines. This exploration requires also the beeswarm and coin packages from our preceding exploration of permutation methods (20).

4 The notation click A | B means click A, then click B.
A recent paper (5) in the Journal of Applied Physiology reinforced the notion that ratios can be troublesome, and it advocated transformation as a remedy. When we explored the bootstrap (16), we discovered that transformation can indeed be useful. If we transform a ratio, however, we simply rescale it: that fails to address its fundamental mathematical flaws. In other words, if we transform a ratio, we try to finesse a square peg into a round hole.

Now that we have a sense there is trouble lurking in the shadows of a ratio, let us explore the havoc a ratio can wreak.

### The Trouble with Ratios: an Example

In order to see the trouble ratios can cause, we need some data. Unlike most of our earlier explorations (13–19), however, this time we want to choose our data. Because the pig has played a conspicuous role within physiology,5 suppose we want to estimate—in pigs—the relationship between weight gain and the amount of feed consumed (Table 2). It turns out this relationship may provide critical information about some metabolic pathway, or it may provide essential economic information for pig farmers.

If we compute the ratio $y/x$ to standardize the numerator $y$, weight gain, to the denominator $x$, the amount of feed consumed, then we discover that our sample of 10 pigs gained, on average, 10 units of weight, while they consumed 4 units of feed.

<table>
<thead>
<tr>
<th>Numerator $Y$: Main Variable</th>
<th>Denominator $X$: Reference Variable</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any physiological response</td>
<td>Initial level of that response</td>
<td>53</td>
</tr>
<tr>
<td>Blood flow or ventilation</td>
<td>Organ volume, weight, oxygen uptake</td>
<td>11, 22, 53</td>
</tr>
<tr>
<td>Flow-mediated dilation</td>
<td>Shear rate</td>
<td>41, 45, 47, 58</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Housekeeping gene expression</td>
<td>24, 29</td>
</tr>
<tr>
<td>One metabolite: IgG₄, GSH</td>
<td>Another metabolite: IgE, GSSG</td>
<td>28, 36</td>
</tr>
<tr>
<td>Muscle mass</td>
<td>Body weight</td>
<td>33</td>
</tr>
<tr>
<td>Cerebral oxygen uptake</td>
<td>Cerebral uptake of glucose and lactate</td>
<td>52</td>
</tr>
<tr>
<td>Maximal oxygen uptake</td>
<td>Lean body mass</td>
<td>23, 28, 33, 34</td>
</tr>
<tr>
<td>RNA</td>
<td>Protein concentration or tissue mass</td>
<td>29, 31</td>
</tr>
<tr>
<td>Smooth muscle tension</td>
<td>Minimum or maximum tension</td>
<td>44</td>
</tr>
<tr>
<td>Surface area</td>
<td>Mass</td>
<td>34</td>
</tr>
<tr>
<td>Actual volume: FEV₁, TLC</td>
<td>Predicted volume</td>
<td>9, 50, 56</td>
</tr>
<tr>
<td>One volume: FEV₁, Vd, RV</td>
<td>Another volume: FVC, Vr, TLC</td>
<td>9, 39, 50</td>
</tr>
</tbody>
</table>

The numerator $Y$ is divided by the denominator $X$ to create the ratio $Y/X$. FEV$_1$, forced expiratory volume in 1 s; FVC, forced vital capacity; TLC, total lung capacity; Vd, dead space; RV, residual volume; Vr, tidal volume. In 2009, the Journal of Applied Physiology published 409 original research articles. Of these, 80% (327 articles) reported some kind of ratio or normalized response.

---

### Example Data

![Fig. 1](http://advan.physiology.org/...)

Fig. 1. Scatterplots of the numerator $y$ (left) and the ratio $y/x$ (right) against the denominator $x$ for 1000 observations. Despite the absence of a relationship between the numerator and the denominator ($r = 0.00$), there is an obvious relationship between the ratio and the denominator. This phenomenon of mathematical coupling (17) occurs regardless of whether the observations are drawn at random from a normal (top) or a uniform (bottom) distribution. The commands in lines 50–51 of Advances_Statistics_Code_Ratios.R return the correlations between the numerator and the denominator; your values will differ slightly. The commands in lines 27–82 of Advances_Statistics_Code_Ratios.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 1: first line to Figure 1: last line.

---

![Fig. 2](http://advan.physiology.org/...)

Fig. 2. The numerator $y$ (left) and the ratio $Y/X$ (right) against the denominator $X$ for four models (A–D). Only when a straight-line relationship between $Y$ and $X$ goes through the origin (model A) does the ratio control for the relationship between $Y$ and $X$.

5 For example, see Ref. 32.
average, 0.20 lb for every pound of feed consumed (Fig. 3). In Fig. 3, the line that depicts this relationship intersects the origin. On the other hand, if we estimate the relationship between weight gain and the amount of feed consumed using ordinary least-squares regression (see Ref. 19), then we obtain

$$\hat{y} = b_0 + b_1x = -22.1 + 0.26x \ . \ \ (1)$$

In Eq. 1, $\hat{y}$ represents the predicted weight gain, $b_0$ estimates the weight gain when no feed is eaten, and $b_1$ estimates the weight gain when 1 lb of feed is eaten. How do we interpret this result? Our pigs gained 0.26 lb for every pound of feed consumed. This approach assumes there is no appreciable measurement error in $x$, the amount of feed consumed.

When we explored regression, we saw that if $x$ does include measurement error, then our estimate of the slope between $y$ and $x$ will be depressed (19). If we estimate the relationship between weight gain and the amount of feed consumed using model II

$$y = \frac{y}{x}$$

regression, a form of regression that can account for measurement error in $x$ (38, 55, 61), then we obtain

$$\hat{y} = b_0 + b_1x = -72.6 + 0.41x \ . \ \ (2)$$

Our pigs gained 0.41 lb for every pound of feed consumed. The commands in lines 135–145 of Advances_Statistics_Code_Ratios.R return these values.

The theoretical flaws of a ratio have now shown themselves. In most cases, a ratio perverts the relationship between the numerator, the thing we really care about, and the denominator, the thing for which we want to adjust. Because many denominators in physiology are subject to measurement error, we will continue our exploration using model II regression. We will revisit ordinary least-squares regression in Practical Considerations.
If we plunge ahead and calculate a ratio to standardize glomerular filtration rate to kidney weight, we find that glomerular filtration rate per gram was 1.20 units in the treated group and 1.15 units in the control group (Fig. 4). If we assess our null hypothesis using either a two-sample t test or a permutation method (20), we reject the null hypothesis ($P < 0.001$) and conclude that glomerular filtration rate per gram differs between the two groups. The commands in lines 301–309 of Advances_Statistics_Code_Ratios.R return these values and results.

Let us now step back, however, and consider model II regression as a tool to study the relationship between glomerular filtration rate and kidney weight in these two groups of pigs. First, we need to see how we generated the observations in Table 3.

Staying Current

**Fig. 4.** Ratios calculated from the simulated data in Table 3. The horizontal lines depict the group means of 1.15 and 1.20. The 99% confidence interval for the difference in ratios, estimated using normal-theory methods (14) and confirmed by the bootstrap (16), is [0.03, 0.06]. The commands in lines 279–293 of Advances_Statistics_Code_Ratios.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figures 4–7: first line to Figures 4–7: last line. To replay this data graphic, highlight and submit lines 485.

If we plunge ahead and calculate a ratio to standardize glomerular filtration rate to kidney weight, we find that glomerular filtration rate per gram was 1.20 units in the treated group and 1.15 units in the control group (Fig. 4). If we assess our null hypothesis using either a two-sample t test or a permutation method (20), we reject the null hypothesis ($P < 0.001$) and conclude that glomerular filtration rate per gram differs between the two groups. The commands in lines 301–309 of Advances_Statistics_Code_Ratios.R return these values and results.

Let us now step back, however, and consider model II regression as a tool to study the relationship between glomerular filtration rate and kidney weight in these two groups of pigs. First, we need to see how we generated the observations in Table 3.

**Fig. 5.** The relationship between glomerular filtration rate, the numerator $y$, and kidney weight, the denominator $x$, in a control group (group 0) and in a treated group (group 1) when the true slopes of the relationship are identical. For each group the line depicts the relationship estimated using model II regression: these lines share a common slope ($P = 0.62$). Inset: the ratio $y/x$, plotted against the denominator $x$ on a condensed scale of 85–115, fails to normalize the numerator $y$ to the denominator $x$. The commands in lines 324–404 of Advances_Statistics_Code_Ratios.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figures 4–7: first line to Figures 4–7: last line. To replay this data graphic, highlight and submit lines 486–487.

**Fig. 6.** The relationship between glomerular filtration rate, the numerator $y$, and kidney weight, the denominator $x$, in a control group (group 0) and in a treated group (group 1) when the true slopes of the relationship differ. For each group the line depicts the relationship estimated using model II regression: these lines differ in slope ($P < 0.001$). Inset: the ratio $y/x$, plotted against the denominator $x$ on a condensed scale of 85–115, fails to normalize the numerator $y$ to the denominator $x$. The commands in lines 324–404 of Advances_Statistics_Code_Ratios.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figures 4–7: first line to Figures 4–7: last line. To replay this data graphic, highlight and submit lines 494–495.

**Fig. 7.** Scatterplots of residuals from the model II regressions depicted in Figs. 5 and 6 against the denominator $x$. These regression models are appropriate: there is no discernible pattern to the residuals. The commands in lines 429–455 of Advances_Statistics_Code_Ratios.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figures 4–7: first line to Figures 4–7: last line. To replay this data graphic, highlight and submit lines 488 and 496.
also that the random error of 0 and a standard deviation \( \varepsilon \) represents random error in the numerator of covariance) was used. When the standard deviation of the random error associated with the measurement of \( Y \) and the measured values of \( X \) listed in Table 3. Your values will differ slightly.

If we use model II regression (60, 61) to estimate the true relationship between glomerular filtration rate and kidney weight is similar in the two groups (\( P = 0.62 \)) but that the elevation of the relationship differs (\( P < 0.001 \)); see Fig. 5. The commands in lines 337–468 of Advances_Statistics_Code_Ratios.R execute the model II regressions and return these values. Your values will differ slightly.

Now we have trouble. If we analyze the ratio \( y/x \), then we conclude that glomerular filtration rate per gram of kidney weight is greater in the treated group. But if we analyze the relationship between glomerular filtration rate and kidney weight using regression (see Fig. 5), then we conclude that the relationship between changes in glomerular filtration rate and changes in kidney weight—the slope \( \Delta y/\Delta x \)—is identical in the two groups. We conclude also that, at some kidney weight, glomerular filtration rate in the treated group is about 3 units greater than in the control group.

To simplify our lives, suppose that the first-order model

\[
Y = \beta_0 + \beta_1 X + \varepsilon
\]

defines the true relationship between glomerular filtration rate \( Y \) and kidney weight \( X \). In Eq. 3, \( \beta_0 \) represents the \( Y \) intercept, \( \beta_1 \) represents the slope of the relationship between \( Y \) and \( X \), and \( \varepsilon \) represents random error in \( Y \) at each value of \( X \) (19). Suppose also that the random error \( \varepsilon \) is distributed normally with a mean of 0 and a standard deviation \( \sigma_\varepsilon = 1 \).

It turns out that we defined the coefficients \( \beta_0 \) and \( \beta_1 \) as

<table>
<thead>
<tr>
<th>Group</th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

in lines 166–173 of Advances_Statistics_Code_Ratios.R. By using these values in Eq. 3, we generated the observed values of glomerular filtration rate \( Y \) for the control (group 0) and treated (group 1) groups listed in Table 3.

We generated the measured values of kidney weight by adding random measurement error to each known value of kidney weight \( X \):

\[
X + \xi
\]

where \( \xi \), the random error associated with the measurement of \( X \), is distributed normally with a mean of 0 and a standard deviation \( \sigma_\xi = 1 \). The commands in lines 193–218 of Advances_Statistics_Code_Ratios.R generate the observed values of \( y \) and the measured values of \( x \) listed in Table 3. Your values will differ slightly.

Table 4. Impact of measurement error in \( x \) on estimated slopes

<table>
<thead>
<tr>
<th>( \sigma_\xi )</th>
<th>( y_{\text{OLS}} )</th>
<th>( y_{\text{model II}} )</th>
<th>( \Delta y )</th>
<th>99% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_\xi = 1 )</td>
<td>0.245</td>
<td>0.250</td>
<td>0.005</td>
<td>( -0.008 ) to ( -0.002 )</td>
</tr>
<tr>
<td>( \sigma_\xi = 5 )</td>
<td>0.222</td>
<td>0.209</td>
<td>0.022</td>
<td>0.007 to 0.040</td>
</tr>
</tbody>
</table>

Values represent means and 99% confidence intervals obtained from 100,000 bootstrap replications of 100 simulations (see Ref. 16). In each simulation we drew at random 2 samples, each with 30 observations, to estimate the true group difference in slopes of 0.25 (see Fig. 8). In each set of 100 simulations, the standard deviation of \( X \) was \( \sigma_X = 8.8 \), and the standard deviation of the random error associated with the measurement of \( X \) was identical in the two groups: \( \sigma_\xi = 1 \) or \( \sigma_\xi = 5 \). When \( \sigma_\xi = 1 \), the estimated difference in slopes was similar regardless of whether model II or ordinary least-squares (OLS) regression (analysis of covariance) was used. When \( \sigma_\xi = 5 \), the blunting of slopes was more pronounced with OLS regression (see Eq. 4), but the estimated difference in slopes (0.200) still resembled the difference estimated using model II regression (0.222).

Fig. 8. The impact of measurement error in the denominator \( x \) on the slopes estimated using ordinary least-squares (OLS) and model II regression. Here, the standard deviation of the random error associated with the measurement of \( x \) was 1 for both groups: \( \sigma_\xi = 1 \). Because \( \sigma_\xi > \sigma_X = 8.8 \), the slopes estimated using ordinary least-squares regression resemble the slopes estimated using model II regression. Note that each approach returns about the same group difference in slopes: 0.20 (model II) and 0.23 (OLS). The true group difference in slopes, defined in the simulation, was 0.25 (see Table 4).
rate per gram differs between the two groups \((P < 0.001)\). If we use model II regression, we learn that

\[
\begin{align*}
\text{group } 0: & \quad \hat{y} = 14.1 + 1.0x \\
\text{group } 1: & \quad \hat{y} = 16.1 + 1.24x \quad .
\end{align*}
\]

These slopes also differ convincingly \((P < 0.001)\). Once again, however, the ratio \(y/x\) distorts the magnitude of the relationship between the numerator \(Y\) and the denominator \(X\).

When we explored regression (19), we discovered that residual plots help us decide if our statistical model of the relationship between \(Y\) and \(X\) is appropriate. Residual plots confirm that our model II regression models are appropriate (Fig. 7).

**Practical Considerations**

Let us pretend we have some real data that others express typically as ratios (see Table 1). How do we proceed to analyze these data? If we are wedded to the notion of a ratio, then the first thing we must do is plot the data (see Ref. 19, p. 351, Rule 1).\(^8\) If the relationship between the numerator \(y\) and the denominator \(x\) is a straight line through the origin, which means the magnitude of the ratio \(y/x\) remains constant across observed values of the denominator \(y\), then the ratio can be a meaningful quantity (see Fig. 2). What happens if this condition is not satisfied? The ratio will misrepresent the true relationship between \(Y\) and \(X\) (see Figs. 3 and 5) and will render meaningless between-group comparisons of the ratio.

In contrast, regression techniques are versatile: they can accommodate an analysis of the relationship between a numerator \(Y\) and a denominator \(X\) when the ratio \(Y/X\) is useless. When we used ordinary least-squares regression to estimate the relationship between weight gain and the amount of feed consumed by pigs (see Fig. 3), our approach was tantamount to analysis of covariance, a technique that blends regression with analysis of variance (12, 25–27, 40, 51) and one that others (1–8, 30, 42, 43, 48, 49, 57, 59) have advocated for the analysis of ratios. Like ordinary least-squares regression, analysis of covariance assumes there is no measurement error in the denominator \(X\).

If we use ordinary least-squares regression to estimate the relationship between a numerator \(Y\) and a denominator \(X\) when there is measurement error in \(X\), then our estimate of the slope of the relationship between \(Y\) and \(X\) will be smaller by a factor of

\[
\frac{\sigma_Y^2}{\sigma_X^2 + \sigma_\epsilon^2} \quad .
\]

In Eq. 4, \(\sigma_X\) is the standard deviation of \(X\) and \(\sigma_\epsilon\) is the standard deviation of the random error associated with the measurement of \(X\) (10, 19). If \(\sigma_X\) swamps \(\sigma_\epsilon\), then the slope estimated using ordinary least-squares regression will resemble the slope estimated using model II regression.

If we want to compare in two groups the relationship between \(Y\) and \(X\), the question is, does analysis of covariance (ordinary least-squares regression) provide a reasonably accurate estimate of the group difference in slopes even though it may underestimate the actual magnitudes of the individual group slopes? The answer appears to be yes (Fig. 8). For different values of \(\sigma_\epsilon\), the standard deviation of the random error associated with the measurement of \(X\), analysis of covariance returns about the same group difference in slopes as does model II regression (Table 4). This is useful: analysis of covariance is more common in statistical packages than is model II regression. Of course, you can always collaborate with a statistician (Ref. 21, guideline 1).

**Summary**

*Any criticism of the use of ratios by biologists [can elicit] a Pavlovian response . . . similar to that resulting from [a criticism] of motherhood, America, and apple pie.*

William Atchley and Dwane Anderson (1978)

*Researchers love ratios. Statisticians loathe them.*

Michal Jasieniński and Fakhri Bazzaz (1999)

As this exploration has demonstrated, the computation of a ratio, a quantity designed to normalize a numerator to some denominator, either as a proportion or as a standardized variable, is deceptively simple. The value of that ratio, however, is often suspect. Why? A ratio accounts for differences in the denominator only if the relationship between numerator and denominator is a straight line through the origin. If this condition is not satisfied, then all bets are off: the ratio will misrepresent the true relationship between numerator and denominator. As a result, we may fail to find a group difference that does exist, or we may find a group difference that does not exist. Who wants to gamble like this?

In contrast, regression techniques\(^9\) that estimate the numerator as a function of the denominator can accommodate an analysis of the relationship between numerator and denominator when the ratio is useless.

In the next installment of this series, we will explore the analysis of a change in some physiological response. Often, we express actual change as percent change so we can account for different initial values. But this creates a problem: percent change is just another ratio. In the next exploration, we will investigate how we can analyze the change in some physiological response.

**ACKNOWLEDGMENTS**

I thank Gerald DiBona (University of Iowa College of Medicine, Iowa City, IA), Gordon Drummond (University of Edinburgh, Edinburgh, UK), John Ludbrook (Department of Surgery, The University of Melbourne, Victoria, Australia), and Matthew Strand (National Jewish Health, Denver, CO) for their helpful comments and suggestions.

**REFERENCES**


\(^8\) No matter how we analyze the data, it is essential to plot the data.

\(^9\) These regression techniques include analysis of covariance.