Supplemental 1

Ratios

This lecture provides supplemental material concerning the handling of ratios in statistics. When you analyze, say with a t-test, a collection of ratios, the raw data are the ratios, and you are asking and answering questions about the individual ratios. When you are interested, however, in the population ratio, this is a mathematically incorrect way of asking and answering the question. We discuss a better way below.

Population Ratios

For the moment, let's consider traffic fatalities in 1995 and in 1996 in a country with two states. The data below have been stratified by the two states. Percentage increase, or decrease, in fatalities is computed as the difference in fatalities over the number of fatalities in 1995. This is the ratio minus one, so it is a quantity that is very closely related to the ratio. For starters, we are simply going to consider ratios for now.

	1995	1996	ratio
State 1	185	200	1.081
State 2	589	556	0.944
Totals	774	756	0.974

If you wanted to know the ratio for the country, however, you would not average the ratios 1.081 and 0.944 to report an average of 1.0125. Instead, you would total the fatalities for the two years and compute the ratio of the totals which gives 0.974. Even that statistic is slightly off. It is biased and, on average, is slightly away from where it should be. There are formulas, however, that, while complicated, are reasonable to compute.

Let

$$\boldsymbol{R = \frac{\mu_y}{\mu_x}}$$

be a population ratio of interest. Let $\boldsymbol{\bar{X}}$ and $\boldsymbol{\bar{Y}}$ be the sample means, $\boldsymbol{s_X^2}$ and $\boldsymbol{s_Y^2}$ be the sample variances, $\boldsymbol{\hat{\rho}}$ be the sample correlation, and $\boldsymbol{n}$ be the sample size. A bias-corrected estimate of $\boldsymbol{R}$ is given by the following formula which uses all of the symbols just defined and can be calculated in Excel:

$$\boldsymbol{\hat{R} = \frac{\bar{Y}}{\bar{X}} - \left(\frac{1}{n}\right)\left(\frac{1}{(\bar{X})^2}\right)\left(\left(\frac{\bar{Y}}{\bar{X}}\right)s_X^2 - \hat{\rho}s_X s_Y\right)}$$

Its estimated variance is

$$\text{var}\left(\hat{\boldsymbol{R}}\right) = \left(\frac{\mathbf{1}}{\boldsymbol{n}}\right)\left(\frac{\mathbf{1}}{(\bar{\boldsymbol{X}})^{\mathbf{2}}}\right)\left(\left(\frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}}\right)^{\mathbf{2}} \boldsymbol{s}_{\boldsymbol{X}}^{\mathbf{2}} + \boldsymbol{s}_{\boldsymbol{Y}}^{\mathbf{2}} - \mathbf{2}\left(\frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}}\right)\hat{\boldsymbol{\rho}}\boldsymbol{s}_{\boldsymbol{X}}\boldsymbol{s}_{\boldsymbol{Y}}\right)$$

so that the standard error of the estimator $\hat{\boldsymbol{R}}$ is

$$\boldsymbol{S.E.}(\hat{\boldsymbol{R}}) = \sqrt{\text{var}\left(\hat{\boldsymbol{R}}\right)}$$

All of that may look complicated, but the formulas only involve quantities that are easy to calculate from the data.

Example: Did raising the speed limit increase fatalities?

Did the law allowing stated to raise the speed limit on American highways increase fatalities in car crashes? You did a paired analysis on these data treating the ratios in each state as independent data of interest. But what if we wanted an estimate of the percentage increase in fatalities in the collection of states that chose to increase the speed limit and we wanted to compare that estimate to the estimate from the collection of states that chose to retain the speed limit. Then we need the collection of estimates in the formula for estimating the collective ratio above. Let $\boldsymbol{X}$ denote the 1995 fatalities and $\boldsymbol{Y}$ denote the 1996 fatalities.

States that increased their speed limits	States that retained their speed limits
$\bar{\boldsymbol{X}} = \mathbf{941.5}$ $\quad \bar{\boldsymbol{Y}} = \mathbf{950.8}$ $\boldsymbol{s}_{\boldsymbol{X}}^{\mathbf{2}} = \mathbf{893,057}$ $\quad \boldsymbol{s}_{\boldsymbol{Y}}^{\mathbf{2}} = \mathbf{934,842}$ $\boldsymbol{n} = \mathbf{32}$ $\hat{\boldsymbol{\rho}} = \mathbf{0.993}$ $\hat{\boldsymbol{R}} = \mathbf{1.010}$ $\quad \text{S.E.}(\hat{\boldsymbol{R}}) = \mathbf{0.021}$	$\bar{\boldsymbol{X}} = \mathbf{614.8}$ $\quad \bar{\boldsymbol{Y}} = \mathbf{612.6}$ $\boldsymbol{s}_{\boldsymbol{X}}^{\mathbf{2}} = \mathbf{218,486}$ $\quad \boldsymbol{s}_{\boldsymbol{Y}}^{\mathbf{2}} = \mathbf{217,704}$ $\boldsymbol{n} = \mathbf{19}$ $\hat{\boldsymbol{\rho}} = \mathbf{0.997}$ $\hat{\boldsymbol{R}} = \mathbf{0.996}$ $\quad \text{S.E.}(\hat{\boldsymbol{R}}) = \mathbf{0.012}$

Based on these estimates and their standard errors, it does not seem like there is any difference between the states who chose to increase their speed limits and those that chose to retain them at 55 mph. We can do a formal, but approximate, test:

$$\boldsymbol{Z} \approx \frac{\hat{\boldsymbol{R}}_{\mathbf{1}} - \hat{\boldsymbol{R}}_{\mathbf{2}}}{\sqrt{\left(\text{S.E.}(\hat{\boldsymbol{R}}_{\mathbf{1}})\right)^{\mathbf{2}} + \left(\text{S.E.}(\hat{\boldsymbol{R}}_{\mathbf{2}})\right)^{\mathbf{2}}}} = \frac{\hat{\boldsymbol{R}}_{\mathbf{1}} - \hat{\boldsymbol{R}}_{\mathbf{2}}}{\sqrt{\left(\text{var}(\hat{\boldsymbol{R}}_{\mathbf{1}})\right) + \left(\text{var}(\hat{\boldsymbol{R}}_{\mathbf{2}})\right)}}$$

is approximately normally distributed. It is only approximate, but we have no formal correction like when we use a t-distribution instead of a standard normal distribution, and so we simply use the normal approximation. Note that the variance of the sum or difference of two independent random variables is the sum of the variances - variances add, standard deviations and standard errors do not.

In our case,

$$\boldsymbol{Z \approx \frac{1.010 - 0.996}{\sqrt{(0.021)^2 + (0.012)^2}} = 0.55,}$$

which is clearly an insignificant score.

The data follow. After that there is a technical explanation about why these formulas work. The technical explanation requires an understanding of calculus.

Increased speed limit

State	1995 deaths	1996 deaths	percent change
AL	1114	1146	2.9
AZ	1025	994	-3
AR	631	615	-2.5
CA	4192	3989	-4.8
CO	645	617	-4.3
DE	121	116	-4.1
FL	2805	2753	-1.9
GA	1488	1573	5.7
ID	262	258	-1.5
IL	1586	1477	-6.9
IS	527	465	-11.8
KS	442	490	10.9
MD	671	608	-9.4
MA	444	417	-6.1
MI	1530	1505	-1.6
MS	868	811	-6.6
MO	1109	1148	3.5
MT	215	200	-7
NE	254	293	15.4
NV	313	348	11.2
NM	485	485	0
ND	74	85	14.9
OH	1360	1391	2.3
OK	669	772	15.4
PA	1480	1469	-0.7
RI	69	69	0
SD	158	175	10.8
TN	1259	1239	-1.6
TX	3183	3742	17.6
UT	325	321	-1.2
WA	653	712	9
WY	170	143	-15.9

Retained speed limit

State	1995 deaths	1996 deaths	percent change
AK	87	81	-6.9
CT	317	310	-2.2
D.C.	58	62	6.9
HI	130	148	13.8
IN	960	984	2.5
KY	849	842	-0.8
LA	894	902	0.9
ME	187	169	-9.6
MN	597	576	-3.5
NH	118	134	13.6
NJ	774	814	5.2
NY	1679	1593	-5.1
NC	1448	1494	3.2
OR	574	526	-8.4
SC	881	930	5.6
VT	106	88	-17
VA	900	877	-2.6
WV	377	348	-7.7
WI	745	761	2.1

The Delta Method: Taylor Series Expansions and Approximation Methods

The reason why the formulas work is that we have expanded a function of one or more random variables in a Taylor series about the mean(s). Recall that the Taylor series expansion of a function of one variable about a specific point $\boldsymbol{x_0}$ is:

$$\boldsymbol{f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + \cdots}$$

For our purposes, we will only consider out to the linear or the quadratic terms. There is a more complicated, but similar, formula for a function of two variables. This section just describes functions of a single variable, so we will not derive the formula for ratios here.

If $\boldsymbol{X}$ is our random variable and $\boldsymbol{\mu = EX}$ is its mean, then the first degree (linear) Taylor series approximation is

$$\boldsymbol{f(X) \approx f(\mu) + (X - \mu)f'(\mu)}$$

The mean is a good place to expand about because it is usually a good guess as to what a random data point will be and one generally expects a lot of data to be around the mean value. We use $\boldsymbol{Ef(X)}$ to denote the mean of this function of the random variable $\boldsymbol{X}$. An initial approximation to this mean is

$$\boldsymbol{Ef(X) \approx f(\mu) + E(X - \mu)f'(\mu) = f(\mu) + 0 = f(\mu)}$$

since the expected value of a constant is just the constant and the expected value of $\boldsymbol{X}$ is $\boldsymbol{\mu}$. Our approximation to the variance of $\boldsymbol{f(X)}$ is

$$\text{var}(\boldsymbol{f(X)}) \approx \text{var}\left(\boldsymbol{f(\mu) + (X - \mu)f'(\mu)}\right) = \text{var}(\boldsymbol{X})(\boldsymbol{f'(\mu)})^2$$

A better approximation to the mean uses the second order (quadratic) Taylor series expansion of $\boldsymbol{f(X)}$ and corrects for bias:

$$\boldsymbol{Ef(X) \approx f(\mu) + E(X - \mu)f'(\mu) + \frac{E(X - \mu)^2}{2}f''(\mu) = f(\mu) + 0 + f''(\mu)}\text{var}(\boldsymbol{X})\boldsymbol{/2}$$

Note that all of those formulas involve things that we do not know, such as the mean and the variance of the population. However, we use the "plug" in principle in statistics that says we should plug in our best guesses, namely the mean and variance of the data. That gives the following formulas:

$$\boldsymbol{Ef(X) \approx f(\bar{X}) + f''(\bar{X})s_X^2/2}$$

and

$$\boldsymbol{(var)(f(X)) \approx s_X^2(f'(\bar{X}))^2}$$

Example 1 An exponential random variable, $\boldsymbol{X}$, with mean $\mathbf{1}$ has variance $\mathbf{1}$. Use this information to approximate the mean and variance of $\boldsymbol{\ln(X)}$.

Let $\boldsymbol{f(x) = \ln(x)}$. Then $\boldsymbol{f'(x) = 1/x}$ and $\boldsymbol{f''(x) = -1/x^2}$. So that $\boldsymbol{E\ln(X) \approx \ln(1) - (1/2) = -1/2}$ and $\boldsymbol{(var)(\ln(X)) \approx 1}$. These are not great approximations because the exponential distribution and the log of it are both very spread out. Simulations suggest that the mean is about $\boldsymbol{-0.57}$ and the variance is about $\mathbf{1.7}$. However, if instead of a single member of the population, we were dealing with population averages, the approximation would improve. See the next example. ■

Example 2 Suppose we have a sample of size 100 from any distribution and the sample mean is $\mathbf{1}$ and the sample variance is $\mathbf{1}$. Approximate the mean and variance of $\boldsymbol{\ln(\bar{X})}$. The difference here is that the sample average, $\boldsymbol{\bar{X}}$ is approximately normal and is close to its mean.

Note that $\boldsymbol{\ln(1) = 0}$ but the $\boldsymbol{\text{var}(\bar{X}) = 1/100}$ since the variance of the mean decreases proportional to the sample size. Thus, $\boldsymbol{E\ln(X) \approx \ln(1) - (1/200) = -0.005}$ and $\boldsymbol{(var)(\ln(X)) \approx 1/100 = 0.01}$. Simulations suggest that these approximations are quite accurate. ■

For a derivation of the formula for ratios, see Rice: Mathematical Statistics and Data Analysis.

References and Readings

[1] John Rice. *Mathematical Statistics and Data Analysis.* Brooks/Cole, 3^{rd} edition, 2006.

Exercises for Lecture 1

1. –

2. –