Statistical Artifacts in the Ratio of Discrete Quantities

Roger G. Johnston (*)
Shayla D. Schroder
Rajika Mallawaaratchy

Los Alamos National Laboratory

(*) Since October 2007:
Argonne National Laboratory;
Fax: 630-252-7323; email: rogerj@anl.gov

LAUR-02-1782

Common Problems in Using the Ratio

- Choosing units poorly
- Keeping wrong number of digits
- Ignoring covariance in error analysis
- Undefined when denominator = 0

Common Problems in Using the Ratio (con't)

These problems are fairly well recognized (except by students),

but two others aren't...

Other Problems with the Ratio (Less Widely Recognized)

- Artifacts in the ratio when the numerator & denominator are discrete

- Lexicon

The Statistical Artifact

- Weird fine-structure (sometimes not so fine) shows up in the histogram of the ratio of two discrete variables.

- This can, and often has, been misinterpreted as instrumentation problems, or as potentially interesting science or engineering.

- But it really is an artifact of ratioing discrete numbers.

- Yet the artifact is not a binning error!

Batting Average: an instructive example

$$BA = \frac{\text{number of hits}}{\text{number of at bats}}$$

$$.000 \leq BA \leq 1.000$$

(1001 possible batting averages)

Batting Average (con't)

Batting .333 is "easy".

I can go: 1 for 3, 2 for 6, 3 for 9, ...

but

Batting .334 is difficult!

I must go: 96 for 287, 97 for 290, 98 for 293, ...

Batting Average (con't)

Many players don't get 287 official at bats
in an entire season, so they never
even get a shot at batting .334!

(Thus, .334 is nearly unobtainable.)

Just to be Specific...

Consider the ratio R=A/B

where A & B are:

- integers in the range 0-255
- uncorrelated
- given by a Gaussian probability distribution with
  mean=127.5, s=32

and R values are:

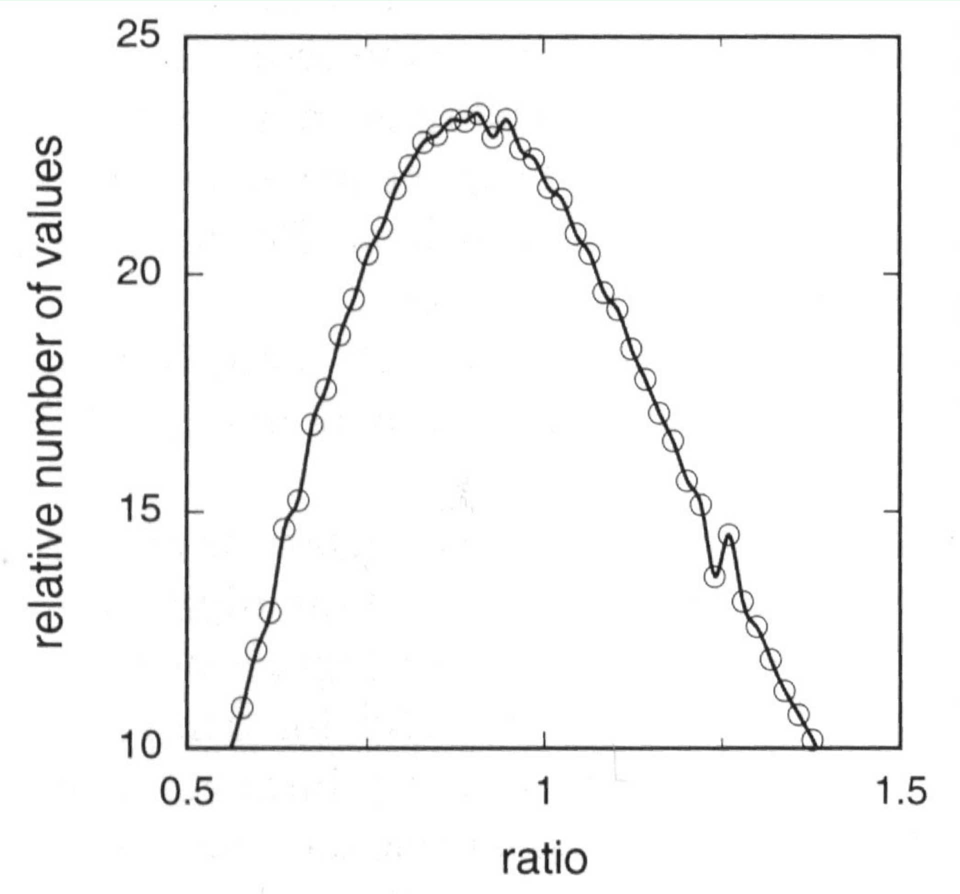
- in the range 0-5
- digitized (quantized) over 256 values (channels or bins)

What Does the Ratio Histogram Look Like?

A: integers, 0-255
(256 bins = 8 bit resolution)

B: integers, 0-255
(256 bins = 8 bit resolution)

R=ratio=A/B: 0-5
in 256 bins
(8-bit resolution)

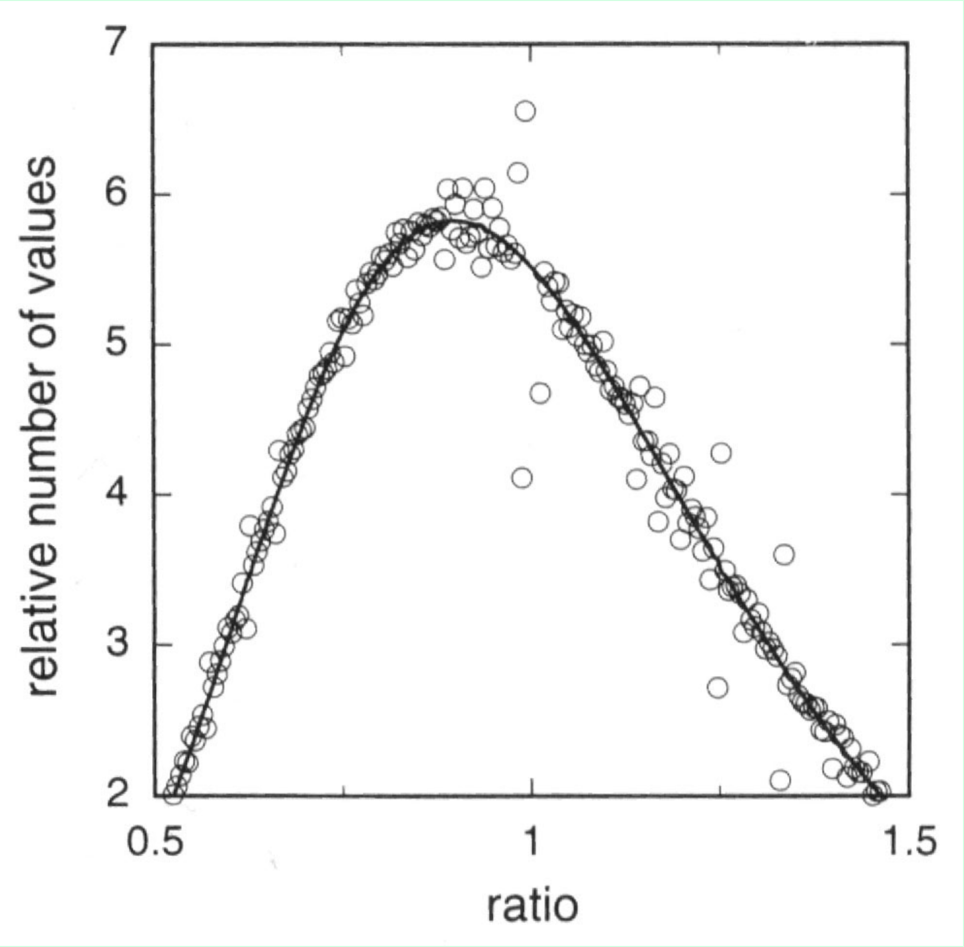


The Artifact Gets Worse With Higher Histogram Resolution!

A: integers, 0-255
 (256 bins = 8 bit resolution)

B: integers, 0-255
 (256 bins = 8 bit resolution)

R=ratio=A/B: 0-5
    in 1024 bins
  (10-bit resolution)



Thus, the Artifact is Not Due to Binning Errors!

With higher resolution for the ratio, the histogram artifact gets worse, not better.

Why?

Because there are more nearly unobtainable "batting averages"

Artifacts in the Ratio Histogram

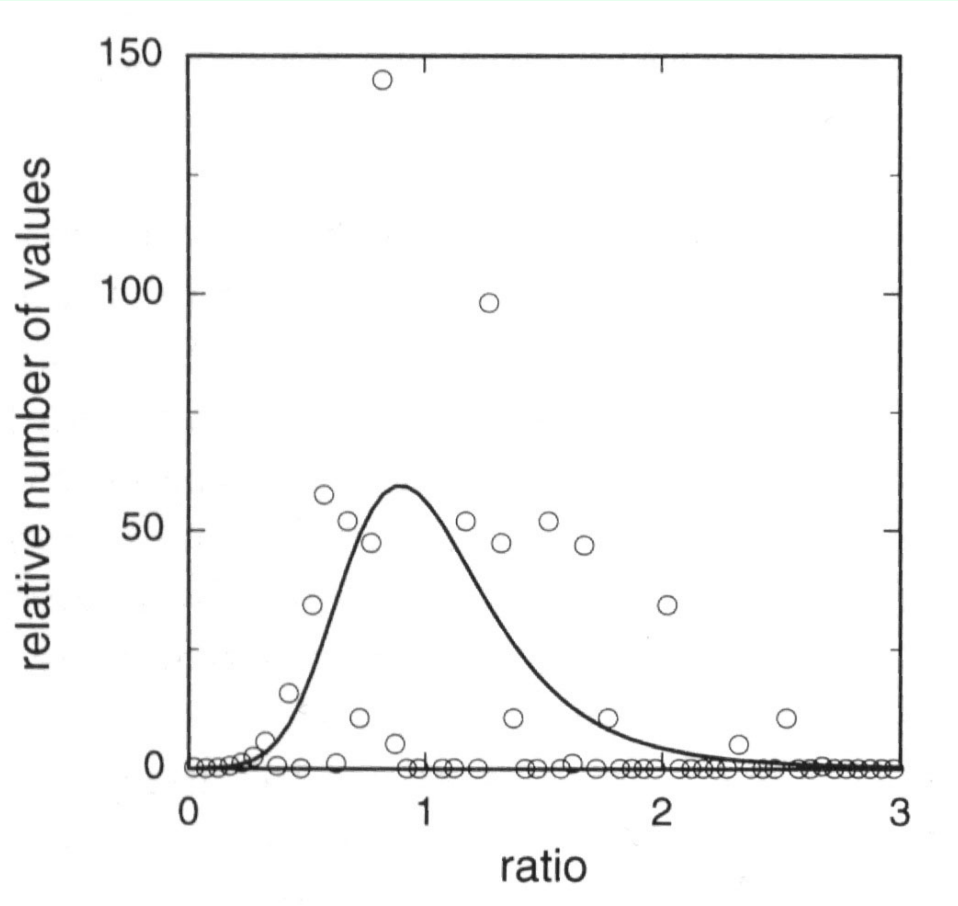
How bad can it get?

Artifacts in the Ratio Histogram (con't)

A: integers, 0-9

B: integers, 0-9

R=ratio=A/B: 0-5
in 100 bins



Artifacts in the Ratio Histogram (con't)

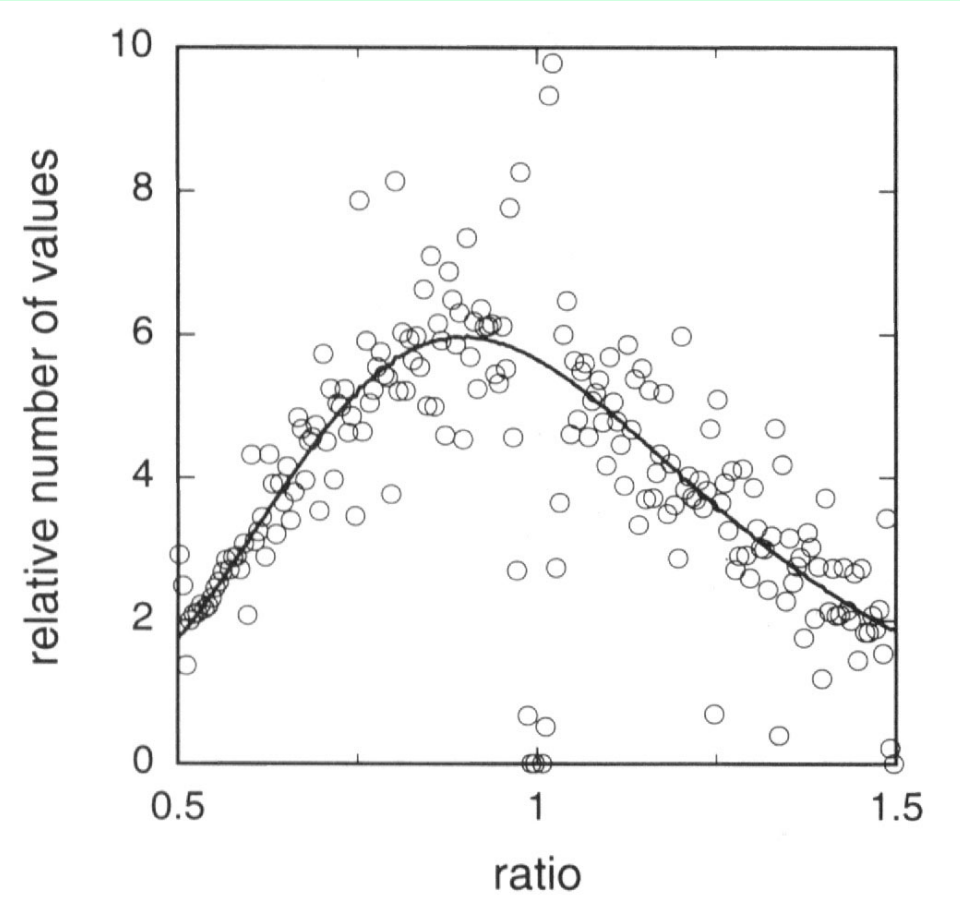
So how CAN we reduce the artifact?

Artifacts in the Ratio Histogram (con't)

A: integers, 0-99
(100 bins)

B: integers, 0-99
(100 bins)

R=ratio=A/B: 0-5 in 1000 bins

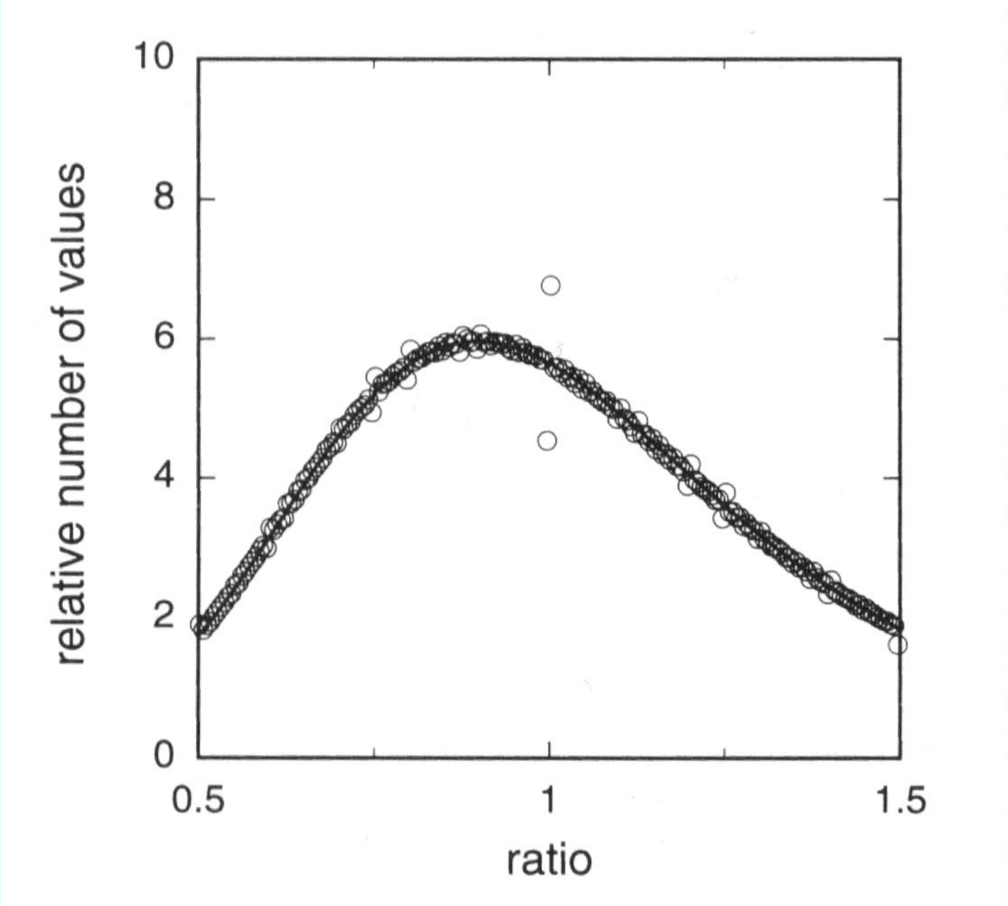


More Bins for A & B Reduces the Artifactual Fine Structure!

A: integers, 0-999
(1000 bins)

B: integers, 0-999
(1000 bins)

R=ratio=A/B: 0-5 in 1000 bins



Artifacts in the Ratio Histogram (con't)

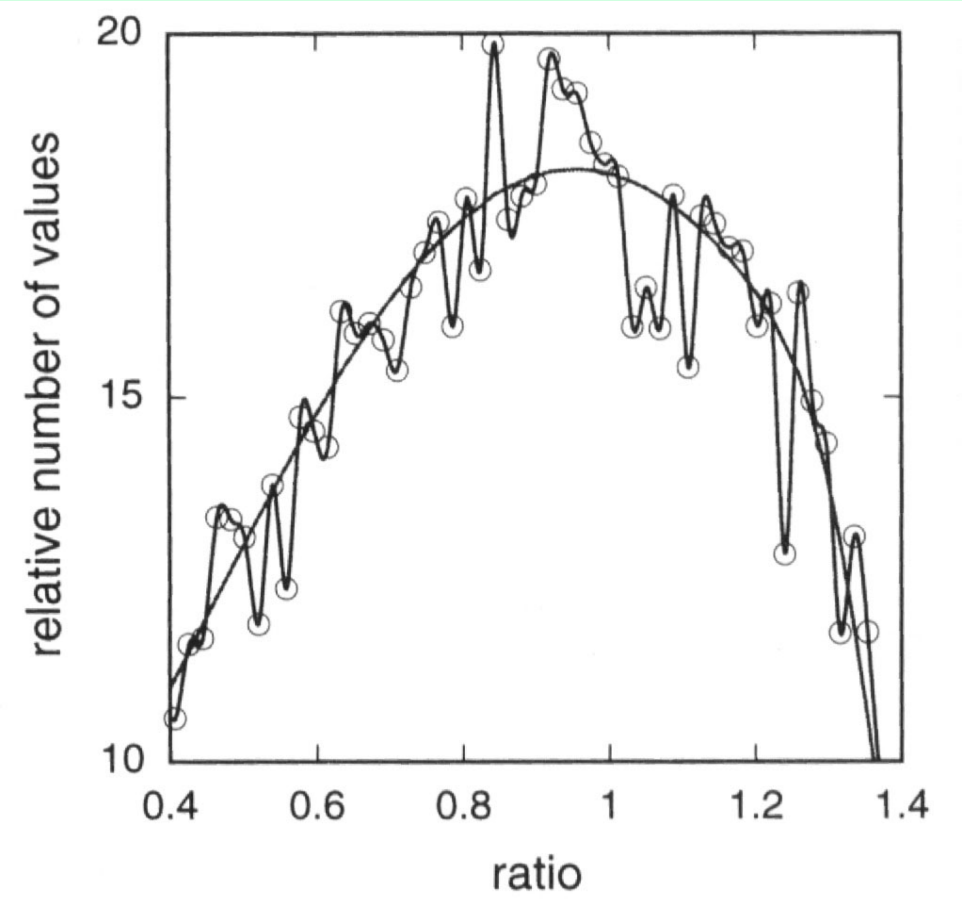
But is it just a matter of
oscillating high and low values
in adjacent bins ?

Runs are Possible!

A: integers, 0-73

B: integers, 0-108

R=ratio=A/B: 0-5
in 264 bins



Getting Fooled By the Statistical Artifact

But is the statistical artifact
in the ratio REALLY a problem?

Getting Fooled By the Statistical Artifact (con't)

Yes!

We're aware of 7 examples

at Los Alamos National Laboratory

of the artifact fooling scientists,

engineers, or technicians.

Getting Fooled By the Statistical Artifact -- example 1

<u>Application</u>

data acquisition software

<u>Artifact Misinterpreted As</u>

software bug

Getting Fooled By the Statistical Artifact -- example 2

<u>Application</u>

analog-to-digital converter electronics

<u>Artifact Misinterpreted As</u>

electronic noise

Getting Fooled By the Statistical Artifact -- example 3

<u>Application</u>

image processing (ratio of one image to another)

<u>Artifact Misinterpreted As</u>

video noise

Getting Fooled By the Statistical Artifact -- example 4

Application

computer modeling

Artifact Misinterpreted As

numeric non-convergence

Getting Fooled By the Statistical Artifact -- example 5

Application

light scattering (normalizing to laser intensity)

Artifact Misinterpreted As

instrument problems

Getting Fooled By the Statistical Artifact -- example 6

<u>Application</u>

fluorescence from biological cells during flow cytometry

<u>Artifact Misinterpreted As</u>

a new subset population of cells

Getting Fooled By the Statistical Artifact -- example 7

<u>Application</u>

finding data outliers

<u>Artifact Misinterpreted As</u>

excessive number of outliers

Recommendations for Not Getting Fooled by the Artifact

- Use the highest practical resolution (lots of bits) for the numerator & denominator but the lowest practical resolution for the ratio.

- Add a small amount of real random noise to the numerator and/or denominator.

Recommendations for Not Getting Fooled by the Artifact (con't)

- Smooth the ratio histogram
- Use analog electronics to measure the analog ratio of the numerator & denominator before digitizing.
- Model the artifact

Recommendations for Not Getting Fooled by the Artifact

(con't)

- If nothing else, at least be aware of the artifact so as not to get fooled!

Lexicon Problems

If you believe the Dictionary (usually a bad idea), then "ratio" is only a noun. Thus, these statements are not allowed:

"We are going to ratio 2 numbers." (verb)

"The artifact shows up during ratioing." (gerund)

"I promise to never get fooled again by the ratioing (or ratio) process." (adjective)

Lexicon Problems (con’t)

But the only important test of the appropriateness of a given (non-obscene) word or phrase in English is:

(1) is it unambiguous?

and

(2) is it concise?

Thus, we should surely allow “ratio” to be used as a verb, gerund, and adjective (not just as a noun) as is the case with many words in English and most technical words!

References

- Roger G. Johnston, Shayla D. Schroder, and A. Rajika Mallawaaratchy, “Statistical Artifacts in the Ratio of Discrete Quantities”, American Statistician 49, 285-291 (1995).

- Comments by Cornel G. Ormsby and Reply by Roger G. Johnston, American Statistician 50, 281 (1996).

- Argonne National Laboratory Vulnerability Assessment Team Home Page: http://www.ne.anl.gov/capabilities/vat/ (since October 2007)