Sample range as a substitute for standard deviation

Steve Paik
Santa Monica College

Motivation:

In a non-majors class, teachers often tell students to use (half) the sample range as an estimate of the width of a distribution. This idea is intuitive, much easier to explain than the notion of standard deviation, and trivial to compute. However, if one wishes to use the sample range in lieu of standard deviation as a measure of distribution width, then one cannot make the sample size, n, too large or risk sampling extreme values.

Question:

Let $X_1, \ldots, X_n$ be iid variables. Is there an optimal n such that (half) the sample range, R, is as close as possible to the standard deviation?

Caveats:

The answer depends on what one means by ‘optimal.’ But a simple criterion is that the most probable value for R/2 is as close as possible to sd(X).

The answer is also sensitive to the distribution for X. So one must already have some suspicion about its nature.

One must study the probability density for the sample range $R = \max\{X_i\} - \min\{X_i\}$.

General fact:

The same value of n is optimal for the entire location-scale family associated with X.

Proof: One must minimize $|R_{mp}/2 - sd(X)|$ over all n. But this condition transforms homogeneously under X -> a+bX.

X is normally distributed:

n = 5 is optimal. (4 or 6 aren't bad, either)

There is an 18% chance that R/2 is within 10% of sd(X); or a 51% chance that R/2 is within 30% of sd(X).

X is exponentially distributed:

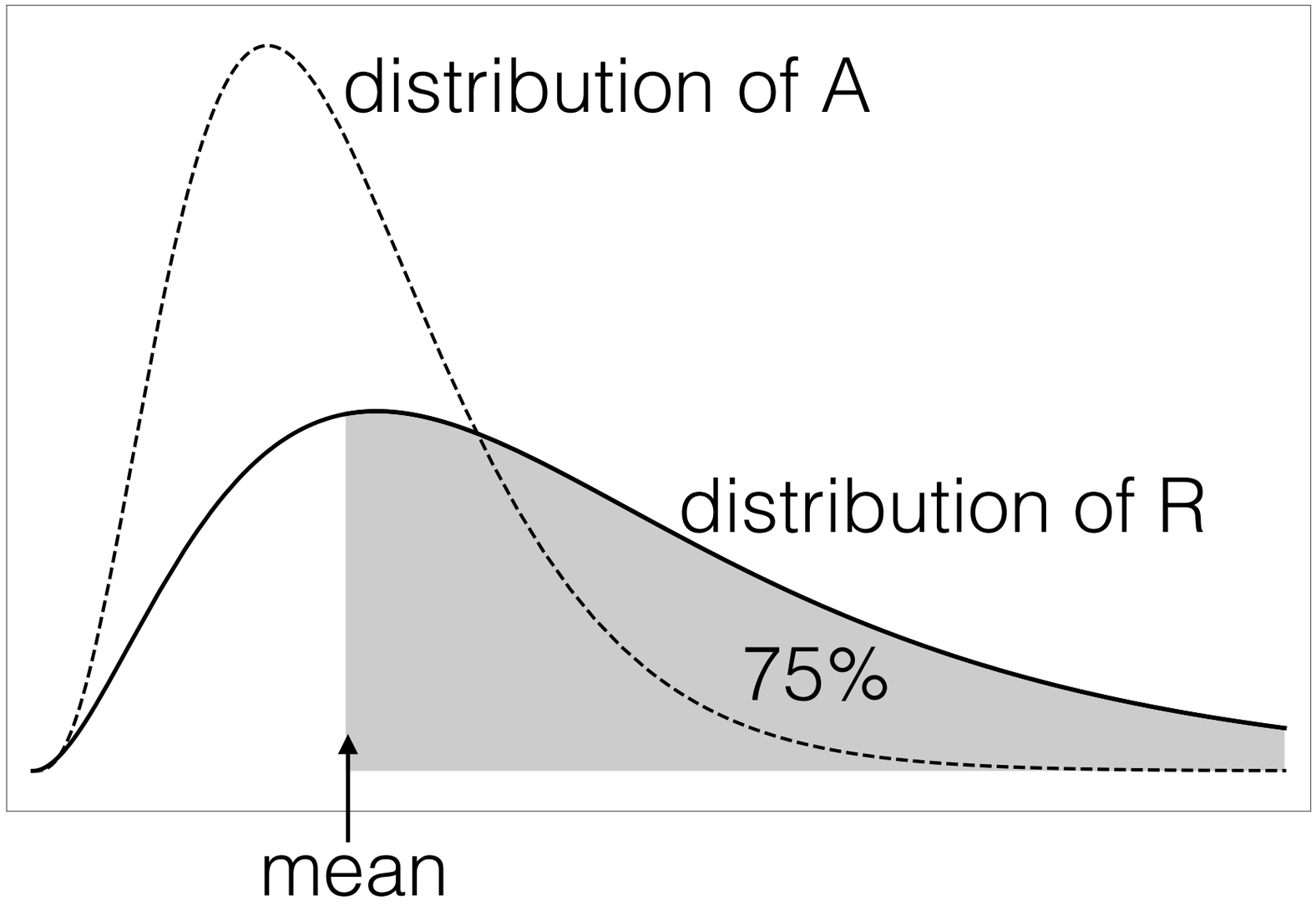
n = 4 is optimal (N.B. Here, I minimized $|R_{mp} - sd(X)|$).

There is only a single parameter — the rate r — that needs to be estimated. Since E(X) = sd(X) = 1/r one can take n large and use the sample average A. This is better than using the range (Law of Large Nos.).

Still, there is a nice economy to n = 4. And student bias often makes it difficult to take more measurements than the number of students in a group.

So in keeping with the theme of this talk, can we use the sample range to say anything useful, and with a reasonably high degree of confidence, about the rate?

Yes. For n = 4, Pr(R > 1/r) = 75%. So R serves as a *probable upper bound* on 1/r.



R and A are positively correlated. This means a comparison of their relative sizes can be used to enhance or degrade the bound.

Pr(R > 1/r | R > A) = 80%: enhancement; happens nine times out of ten.

Pr(R > 1/r | R < A) = 24%: reversal! R serves as a probable lower bound; happens once every ten times.