Self-Selection Sampling Bias B

The reviews of restaurants, hotels, cafes, and so on that you read on

social media sites like Yelp are prone to bias because the people

submitting them are not randomly selected; rather, they themselves

have taken the initiative to write. This leads to self-selection bias—

the people motivated to write reviews may have had poor experiences,

may have an association with the establishment, or may simply

be a different type of person from those who do not write reviews.

Note that while self-selection samples can be unreliable indicators

of the true state of affairs, they may be more reliable in simply comparing

one establishment to a similar one; the same self-selection

bias might apply to each.

O

Regression to the mean, meaning to “go back,” is distinct from the

statistical modeling method of linear regression, in which a linear

relationship is estimated between predictor variables and an outcome

variable.

It is important to distinguish between the distribution of the individual

data points, known as *the data distribution*, and the distribution

of a sample statistic, known as the *sampling distribution*.

Standard Deviation Versus Standard Error O

Do not confuse standard deviation (which measures the variability

of individual data points) with standard error (which measures the

variability of a sample metric).

The bootstrap does not compensate for a small sample size; it does

not create new data, nor does it fill in holes in an existing data set.

It merely informs us about how lots of additional samples would

behave when drawn from a population like our original sample.

B

Of course, what we are really interested in when we have a sample

result is, “What is the probability that the true value lies within a

certain interval?” This is not really the question that a confidence

interval answers, but it ends up being how most people interpret

the answer.

The probability question associated with a confidence interval

starts out with the phrase “Given a sampling procedure and a population,

what is the probability that…” To go in the opposite direction,

“Given a sample result, what is the probability that

(something is true about the population)?” involves more complex

calculations and deeper imponderables.

For a data scientist, a confidence interval is a tool that can be used

to get an idea of how variable a sample result might be. Data scientists

would use this information not to publish a scholarly paper or

submit a result to a regulatory agency (as a researcher might) but

most likely to communicate the potential error in an estimate, and

perhaps to learn whether a larger sample is needed.

It is a common misconception that the normal distribution is

called that because most data follows a normal distribution—that

is, it is the normal thing. Most of the variables used in a typical data

science project—in fact, most raw data as a whole—are *not* normally

distributed: see “Long-Tailed Distributions” on page 73. The

utility of the normal distribution derives from the fact that many

statistics *are* normally distributed in their sampling distribution.

Even so, assumptions of normality are generally a last resort, used

when empirical probability distributions, or bootstrap distributions,

are not available. O

Converting data to *z*-scores (i.e., standardizing or normalizing the

data) does *not* make the data normally distributed. It just puts the

data on the same scale as the standard normal distribution, often

for comparison purposes. O

There is much statistical literature about the task of fitting statistical

distributions to observed data. Beware an excessively datacentric

approach to this job, which is as much art as science. Data is

variable, and often consistent, on its face, with more than one

shape and type of distribution. It is typically the case that domain

and statistical knowledge must be brought to bear to determine

what type of distribution is appropriate to model a given situation.

For example, we might have data on the level of internet traffic on a

server over many consecutive five-second periods. It is useful to

know that the best distribution to model “events per time period” is

the Poisson (see “Poisson Distributions” on page 83). B

What do data scientists need to know about the t-distribution and

the central limit theorem? Not a whole lot. The t-distribution is

used in classical statistical inference but is not as central to the purposes

of data science. Understanding and quantifying uncertainty

and variation are important to data scientists, but empirical bootstrap

sampling can answer most questions about sampling error.

However, data scientists will routinely encounter t-statistics in output

from statistical software and statistical procedures in *R*—for

example, in A/B tests and regressions—so familiarity with its purpose

is helpful. B