**Chapter 2. Data and Sampling Distributions**

**Bias**

Random sampling can reduce bias

**Regression to the Mean**

It refers to a phenomenon involving successive measurements on a given variable: extreme observations tend to be followed by more central ones

**Sampling Distribution of a Statistic**

Sample statistic: A metric calculated for a sample of data drawn from a larger population.

Sampling distribution: The frequency distribution of a sample statistic over many samples from a larger population

Standard error: The variability of a sample statistic over many samples.

Standard deviation: variability of individual data points.

**Bootstrap**

Bootstrap sample: a sample taken with replacement from an observed dataset.

Bootstrap is a powerful tool for assessing the variability of a sample statistic. Eg: bootstrap confidence interval:

1. Draw a random sample of size *n* with replacement from the data (a resample).
2. Record the statistic of interest for the resample.
3. Repeat steps 1–2 many (*R*) times.
4. For an *x*% confidence interval, trim [(1 – [*x*/100]) / 2]% of the *R* resample results from either end of the distribution.
5. The trim points are the endpoints of an *x*% bootstrap confidence interval.

**Normal Distribution**

In a normal distribution ([Figure 2-10](https://www.safaribooksonline.com/library/view/practical-statistics-for/9781491952955/ch02.html#normal-curve)), 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations.

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.

**Long-Tailed distributions**

Tail: where relatively extreme values occur at low frequency.

**T-distribution**

The t-distribution is a normally shaped distribution, but a bit thicker and longer on the tails. Distributions of sample means are typically shaped like a t-distribution, and there is a family of t-distributions that differ depending on how large the sample is.The larger the sample, the more normally shaped the t-distribution becomes.

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.

**Binominal Distribution**

**Poisson Distribution**