



# Statistical Machine Learning for Data Science- BAD702

**Prepared By,  
Dr. Anitha DB  
Associate Professor & Head  
Department of CSE-Data Science  
ATME College of Engineering, Mysuru**

[GitHub - gedeck/practical-statistics-for-data-scientists: Code repository for O'Reilly book](https://github.com/gedeck/practical-statistics-for-data-scientists)

## Module-1

**Exploratory Data Analysis:** Estimates of locations and variability, Exploring data distributions, Exploring binary and categorical data, Exploring two or more variables.

**Textbook: Chapter 1**

## Module-2

**Data and Sampling Distributions:** Random sampling and bias, selection bias, sampling distribution of statistic, bootstrap, confidence intervals, data distributions: normal, long tailed, student's-t, binomial, Chi-square, F distribution, Poisson and related distributions.

**Textbook: Chapter 2**

## Module-3

**Statistical Experiments and Significance Testing:** A/B testing, hypothesis testing, resampling, statistical significance & p-values, t-tests, multiple testing, degrees of freedom.

**Textbook: Chapter 3**

## Module-4

Multi-Arm Bandit algorithm, power and sample size, factor variables in regression, interpreting the regression equation, Regression diagnostics, Polynomial and Spline Regression.

**Textbook: Chapter 3 & 4**

## Module-5

**Discriminant Analysis:** Covariance Matrix, Fisher's Linear discriminant, Generalized Linear Models, Interpreting the coefficients and odd ratios, Strategies for Imbalanced Data.

**Textbook: Chapter 5**

**Topics**

1. Random sampling and Sample Bias: Bias, Random Selection, Size Versus Quality, Sample Mean Versus Population Mean
2. Selection bias: Regression to the Mean
3. Sampling distribution of statistic: Central Limit Theorem, Standard Error
4. Bootstrap: Resampling versus Bootstrapping
5. Confidence intervals
6. Data distributions:
  - Normal distributions: Standard Normal and QQ-Plots
  - Long tailed distributions,
  - Student's-t distributions,
  - Binomial distributions,
  - Chi-square distributions,
  - F distribution,
7. Poisson and Related distributions: Poisson distributions, Exponential Distribution, Estimating the Failure Rate, Weibull Distribution

**Textbook: Peter Bruce, Andrew Bruce and Peter Gadeck, “Practical Statistics for Data Scientists”, 2nd edition, O'Reilly Publications, 2020. Chapter 2**

[GitHub - gedeck/practical-statistics-for-data-scientists](https://github.com/gedeck/practical-statistics-for-data-scientists): Code repository for O'Reilly book

# Random Sampling and Sample Bias

## KEY TERMS FOR RANDOM SAMPLING

### *Sample*

A subset from a larger data set.

### *Population*

The larger data set or idea of a data set.

### *$N$ ( $n$ )*

The size of the population (sample).

### *Random sampling*

Drawing elements into a sample at random.

### *Stratified sampling*

Dividing the population into strata and randomly sampling from each strata.

### *Simple random sample*

The sample that results from random sampling without stratifying the population.

### *Sample bias*

A sample that misrepresents the population.

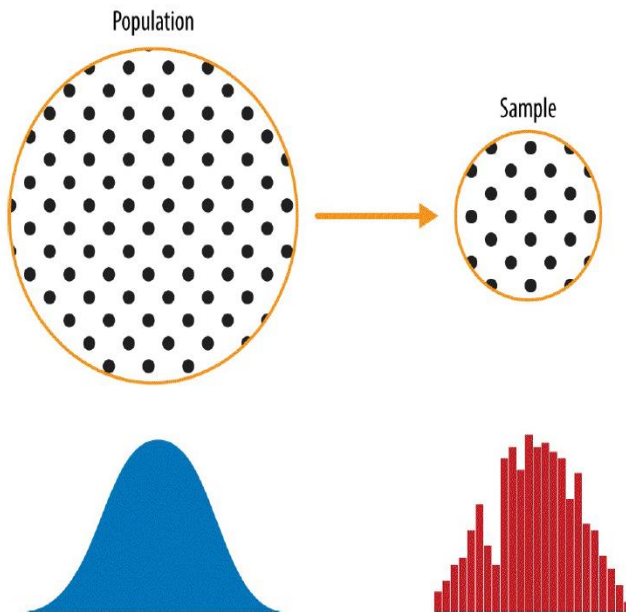


Figure 2-1. Population versus sample

*Random sampling* is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw. The sample that results is called a *simple random sample*.

Sampling can be done *with replacement*, in which observations are put back in the population after each draw for possible future reselection. Or it can be done *without replacement*, in which case observations, once selected, are unavailable for future draws.

Data quality often matters more than data quantity when making an estimate or a model based on a sample. Data quality in data science involves completeness, consistency of format, cleanliness, and accuracy of individual data points. Statistics adds the notion of *representativeness*.

### Bias

Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process. An important distinction should be made between errors due to random chance, and errors due to bias.

Consider the physical process of a gun shooting at a target. It will not hit the absolute center of the target every time, or even much at all. An unbiased process will produce error, but it is random and does not tend strongly in any direction(see Figure 2-2) . The results shown in Figure 2-3 show a biased process — there is still random error in both the x and y direction, but there is also a bias. Shots tend to fall in the upper-right quadrant.

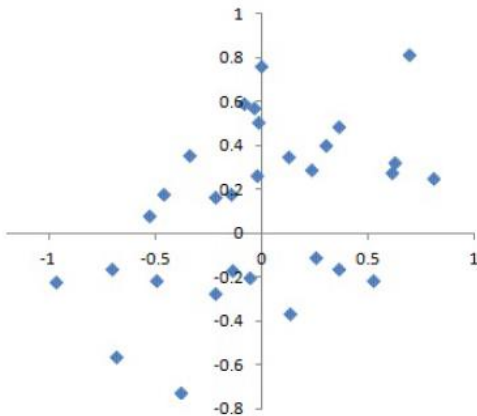


Figure 2-2. Scatterplot of shots from a gun with true aim

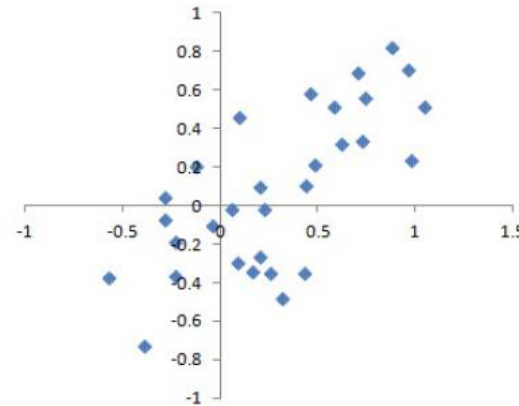


Figure 2-3. Scatterplot of shots from a gun with biased aim

Bias comes in different forms, and may be observable or invisible. When a result does suggest bias (e.g., by reference to a benchmark or actual values), it is often an indicator that a statistical or machine learning model has been misspecified, or an important variable left out.



### Random Selection

To avoid the problem of sample bias that led the *Literary Digest* to predict Landon over Roosevelt, George Gallup (shown in Figure 2-4) opted for more scientifically chosen methods to achieve a sample that was representative of the US voter. There are now a variety of methods to achieve representativeness, but at the heart of all of them lies *random sampling*.



Figure 2-4. George Gallup, catapulted to fame by the *Literary Digest*'s "big data" failure

Random sampling is not always easy. Proper definition of an accessible population is key. Suppose we want to generate a representative profile of customers and we need to conduct a pilot customer survey. The survey needs to be representative but is labor intensive.

First we need to define who a customer is. We might select all customer records where purchase amount  $> 0$ . Do we include all past customers? Do we include refunds? Internal test purchases? Resellers? Both billing agent and customer? Next we need to specify a sampling procedure. It might be "select 100 customers at random." Where a sampling from a flow is involved (e.g., real-time customer transactions or web visitors), timing considerations may be important (e.g., a web visitor at 10 a.m. on a weekday may be different from a web visitor at 10 p.m. on a weekend).

In *stratified sampling*, the population is divided up into *strata*, and random samples are taken from each stratum. Political pollsters might seek to learn the electoral preferences of whites, blacks, and Hispanics. A simple random sample taken from the population would yield too few blacks and Hispanics, so those strata could be over weighted in stratified sampling to yield equivalent sample sizes.

## Size versus Quality: When Does Size Matter?

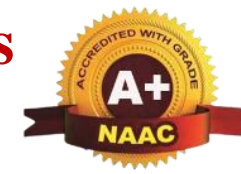
In the era of big data, it is sometimes surprising that smaller is better. Time and effort spent on random sampling not only reduce bias, but also allow greater attention to data exploration and data quality. For example, missing data and outliers may contain useful information. It might be prohibitively expensive to track down missing values or evaluate outliers in millions of records, but doing so in a sample of several thousand records may be feasible. Data plotting and manual inspection bog down if there is too much data.

So when *are* massive amounts of data needed?

The classic scenario for the value of big data is when the data is not only big, but sparse as well. Consider the search queries received by Google, where columns are terms, rows are individual search queries, and cell values are either 0 or 1, depending on whether a query contains a term. The goal is to determine the best predicted search destination for a given query. There are over 150,000 words in the English language, and Google processes over 1 trillion queries per year. This yields a huge matrix, the vast majority of whose entries are “0.”

This is a true big data problem — only when such enormous quantities of data are accumulated can effective search results be returned for most queries. And the more data accumulates, the better the results. For popular search terms this is not such a problem — effective data can be found fairly quickly for the handful of extremely popular topics trending at a particular time. The real value of modern search technology lies in the ability to return detailed and useful results for a huge variety of search queries, including those that occur only with a frequency, say, of one in a million.





Consider the search phrase “Ricky Ricardo and Little Red Riding Hood.” In the early days of the internet, this query would probably have returned results on Ricky Ricardo the band leader, the television show *I Love Lucy* in which he starred, and the children’s story *Little Red Riding Hood*. Later, now that trillions of search queries have been accumulated, this search query returns the exact *I*

*Love Lucy* episode in which Ricky narrates, in dramatic fashion, the Little Red Riding Hood story to his infant son in a comic mix of English and Spanish.

Keep in mind that the number of actual *pertinent* records — ones in which this exact search query, or something very similar, appears (together with information on what link people ultimately clicked on) — might need only be in the thousands to be effective. However, many trillions of data points are needed in order to obtain these pertinent records (and random sampling, of course, will not help).

## Sample Mean versus Population Mean

The symbol  $\bar{x}$  (pronounced x-bar) is used to represent the mean of a sample from a population, whereas  $\mu$  is used to represent the mean of a population. Why make the distinction? Information about samples is observed, and information about large populations is often inferred from smaller samples. Statisticians like to keep the two things separate in the symbology.

### KEY IDEAS

- Even in the era of big data, random sampling remains an important arrow in the data scientist's quiver.
- Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would be prohibitively expensive.

### KEY TERMS

***Bias***

Systematic error.

***Data snooping***

Extensive hunting through data in search of something interesting.

***Vast search effect***

Bias or nonreproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.

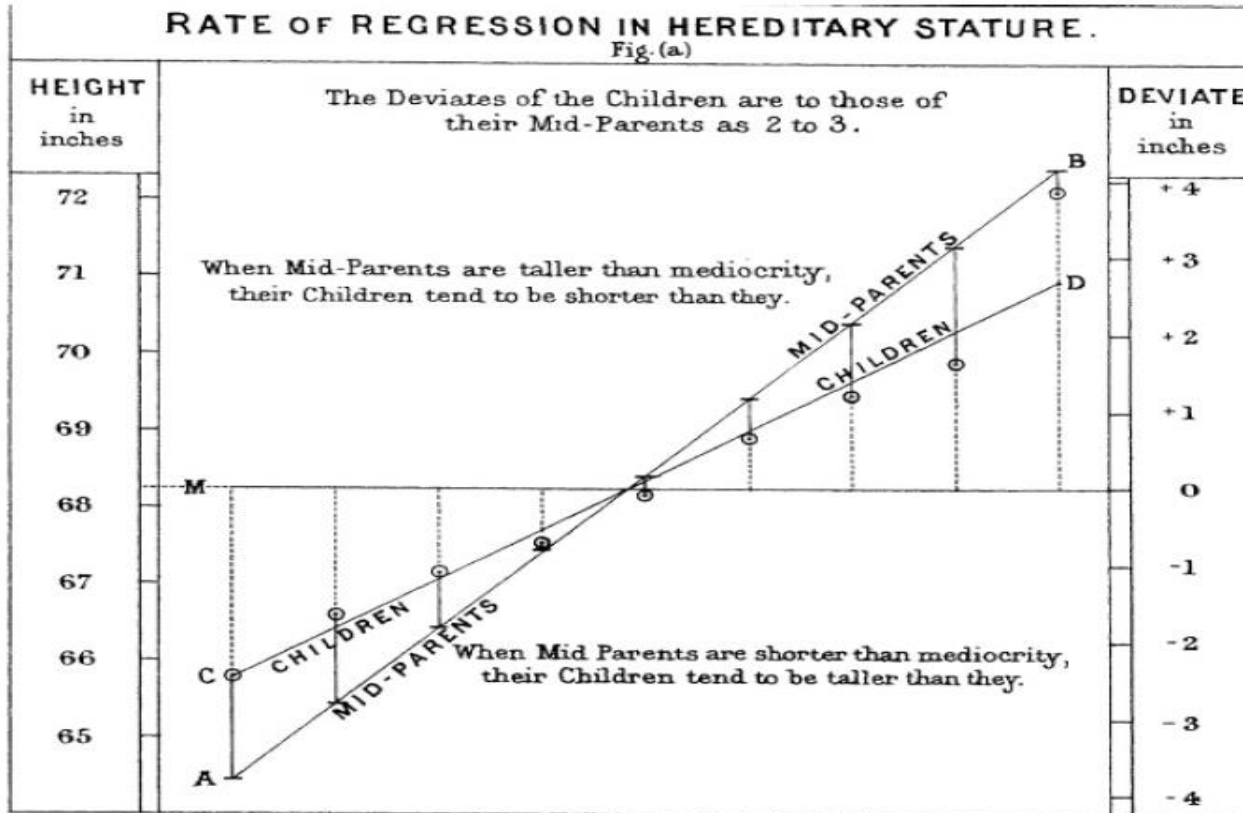
## Regression to the Mean

*Regression to the mean* refers to a phenomenon involving successive measurements on a given variable: extreme observations tend to be followed by more central ones. Attaching special focus and meaning to the extreme value can lead to a form of selection bias.

Sports fans are familiar with the “rookie of the year, sophomore slump” phenomenon. Among the athletes who begin their career in a given season (the rookie class), there is always one who performs better than all the rest. Generally, this “rookie of the year” does not do as well in his second year. Why not? In nearly all major sports, at least those played with a ball or puck, there are two elements that play a role in overall performance:

- Skill
- Luck

Regression to the mean is a consequence of a particular form of selection bias. When we select the rookie with the best performance, skill and good luck are probably contributing. In his next season, the skill will still be there but, in most cases, the luck will not, so his performance will decline — it will regress. The phenomenon was first identified by Francis Galton in 1886 [Galton-1886], who wrote of it in connection with genetic tendencies; for example, the children of extremely tall men tend not to be as tall as their father (see Figure 2-5).



### KEY IDEAS

- Specifying a hypothesis, then collecting data following randomization and random sampling principles, ensures against bias.
- All other forms of data analysis run the risk of bias resulting from the data collection/analysis process (repeated running of models in data mining, data snooping in research, and after-the-fact selection of interesting events).

Figure 2-5. Galton's study that identified the phenomenon of regression to the mean



**Topic 3: Sampling Distribution of a Statistic****KEY TERMS*****Sample statistic***

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

***Data distribution***

The frequency distribution of individual *values* in a data set.

***Sampling distribution***

The frequency distribution of a *sample statistic* over many samples or resamples.

***Central limit theorem***

The tendency of the sampling distribution to take on a normal shape as sample size rises.

***Standard error***

The variability (standard deviation) of a *sample statistic* over many samples (not to be confused with *standard deviation*, which, by itself, refers to variability of individual data *values*).

### Topic 3: Sampling Distribution of a Statistic

The term *sampling distribution* of a statistic refers to the distribution of some sample statistic, over many samples drawn from the same population. Much of classical statistics is concerned with making inferences from (small) samples to (very large) populations.

The distribution of a *sample statistic* such as the mean is likely to be more regular and bell-shaped than the distribution of the data itself. The larger the sample, the narrower the distribution of the sample statistic.

This is illustrated in an *example* using annual income for loan applicants to Lending Club (see “A Small Example: Predicting Loan Default” for a description of the data). Take three samples from this data: a sample of 1,000 values, a sample of 1,000 means of 5 values, and a sample of 1,000 means of 20 values.

Then plot a histogram of each sample to produce Figure 2-6.

The histogram of the individual data values is broadly spread out and skewed toward higher values as is to be expected with income data.

The histograms of the means of 5 and 20 are increasingly compact and more bell-shaped.

Here is the R code to generate these histograms, using the visualization package ggplot2.

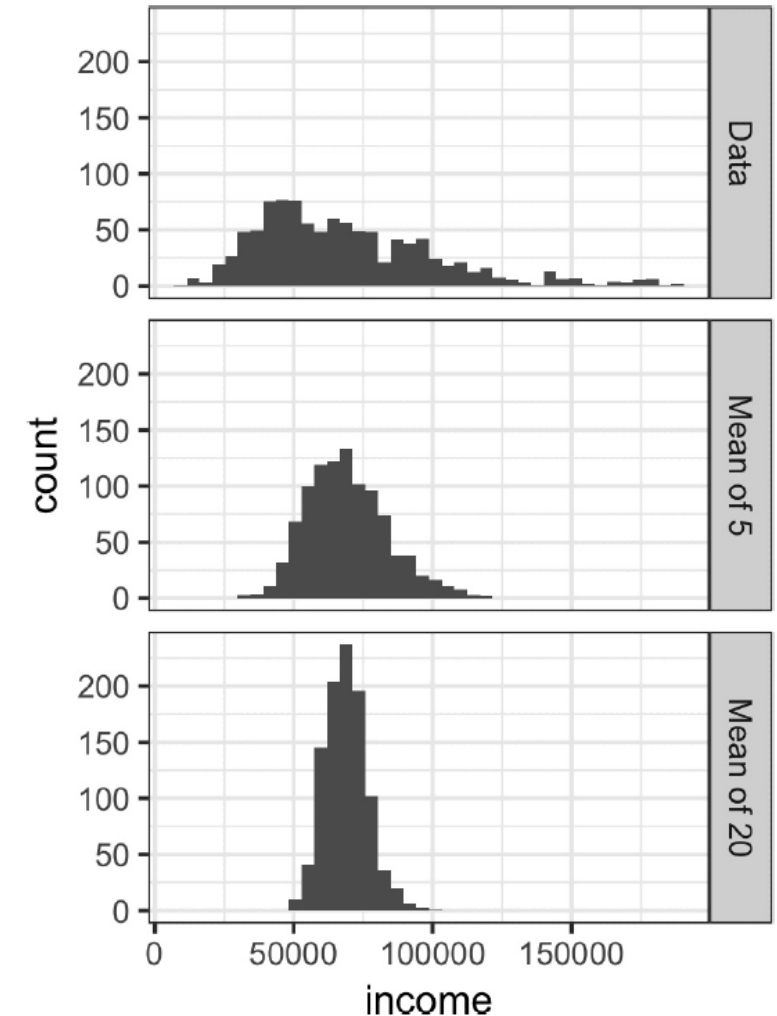
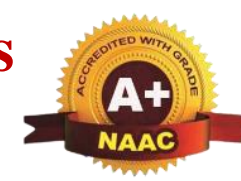


Figure 2-6. Histogram of annual incomes of 1,000 loan applicants (top), then 1000 means of  $n=5$  applicants (middle), and  $n=20$  (bottom)

## Topic 3: Sampling Distribution of a Statistic

```
library(ggplot2)
# take a simple random sample
samp_data <- data.frame(income=sample(loans_income, 1000),
                        type='data_dist')
# take a sample of means of 5 values
samp_mean_05 <- data.frame(
  income = tapply(sample(loans_income, 1000*5),
                  rep(1:1000, rep(5, 1000)), FUN=mean),
  type = 'mean_of_5')
# take a sample of means of 20 values
samp_mean_20 <- data.frame(
  income = tapply(sample(loans_income, 1000*20),
                  rep(1:1000, rep(20, 1000)), FUN=mean),
  type = 'mean_of_20')
# bind the data.frames and convert type to a factor
income <- rbind(samp_data, samp_mean_05, samp_mean_20)
income$type = factor(income$type,
                     levels=c('data_dist', 'mean_of_5', 'mean_of_20'),
                     labels=c('Data', 'Mean of 5', 'Mean of 20'))
# plot the histograms
ggplot(income, aes(x=income)) +
  geom_histogram(bins=40) +
  facet_grid(type ~ .)
```



## Central Limit Theorem

This phenomenon is termed the *central limit theorem*. It says that the means drawn from multiple samples will resemble the familiar bell-shaped normal curve (see “Normal Distribution”), even if the source population is not normally distributed, provided that the sample size is large enough and the departure of the data from normality is not too great. The central limit theorem allows normal-approximation formulas like the t-distribution to be used in calculating sampling distributions for inference — that is, confidence intervals and hypothesis tests.

The central limit theorem receives a lot of attention in traditional statistics texts because it underlies the machinery of hypothesis tests and confidence intervals, which themselves consume half the space in such texts. Data scientists should be aware of this role, but, since formal hypothesis tests and confidence intervals play a small role in data science, and the bootstrap is available in any case, the central limit theorem is not so central in the practice of data science.



## Topic 3: Sampling Distribution of a Statistic

### Standard Error

The *standard error* is a single metric that sums up the variability in the sampling distribution for a statistic. The standard error can be estimated using a statistic based on the standard deviation  $s$  of the sample values, and the sample size  $n$ : As the sample size increases, the standard error decreases, corresponding to what was observed in Figure 2-6. The relationship between standard error and sample size is sometimes referred to as the *square-root of  $n$*  rule: in order to reduce the standard error by a factor of 2, the sample size must be increased by a factor of 4.

The validity of the standard error formula arises from the central limit theorem (see “Central Limit Theorem”). In fact, you don’t need to rely on the central limit theorem to understand standard error. Consider the following approach to measure standard error:

1. Collect a number of brand new samples from the population.
2. For each new sample, calculate the statistic (e.g., mean).
3. Calculate the standard deviation of the statistics computed in step 2; use this as your estimate of standard error.

In practice, this approach of collecting new samples to estimate the standard error is typically not feasible (and statistically very wasteful). Fortunately, it turns out that it is not necessary to draw brand new samples; instead, you can use *bootstrap* resamples (see “The Bootstrap”). In modern statistics, the bootstrap has become the standard way to estimate standard error. It can be used for virtually any statistic and does not rely on the central limit theorem or other distributional assumptions.



## Topic 3: Sampling Distribution of a Statistic

### STANDARD DEVIATION VERSUS STANDARD ERROR

Do not confuse standard deviation (which measures the variability of individual data points) with standard error (which measures the variability of a sample metric).

### KEY IDEAS

- The frequency distribution of a sample statistic tells us how that metric would turn out differently from sample to sample.
- This sampling distribution can be estimated via the bootstrap, or via formulas that rely on the central limit theorem.
- A key metric that sums up the variability of a sample statistic is its standard error.

## The Bootstrap

One easy and effective way to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample. This procedure is called the *bootstrap*, and it does not necessarily involve any assumptions about the data or the sample statistic being normally distributed.

### KEY TERMS

***Bootstrap sample***

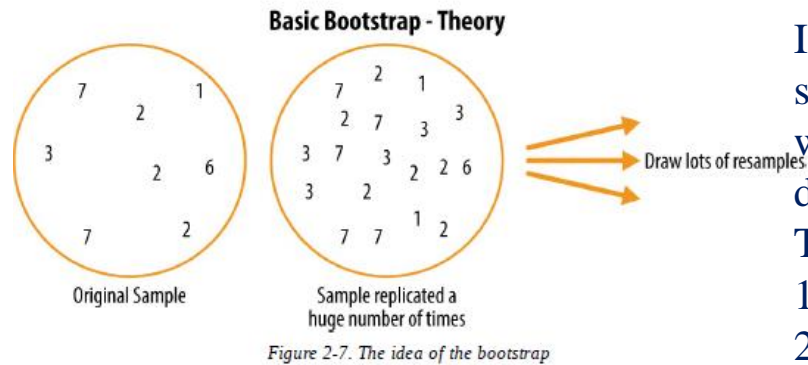
A sample taken with replacement from an observed data set.

***Resampling***

The process of taking repeated samples from observed data; includes both bootstrap and permutation (shuffling) procedures.

## The Bootstrap

Conceptually, you can imagine the bootstrap as replicating the original sample thousands or millions of times so that you have a hypothetical population that embodies all the knowledge from your original sample (it's just larger). You can then draw samples from this hypothetical population for the purpose of estimating a sampling distribution. See Figure 2-7.



In practice, it is not necessary to actually replicate the sample a huge number of times. We simply replace each observation after each draw; that is, we *sample with replacement*. In this way we effectively create an infinite population in which the probability of an element being drawn remains unchanged from draw to draw.

The algorithm for a bootstrap resampling of the mean is as follows, for a sample of size  $n$ :

1. Draw a sample value, record, replace it.
2. Repeat  $n$  times.
3. Record the mean of the  $n$  resampled values.
4. Repeat steps 1–3  $R$  times.
5. Use the  $R$  results to:
  - a. Calculate their standard deviation (this estimates sample mean standard error).
  - b. Produce a histogram or boxplot.
  - c. Find a confidence interval.

$R$ , the number of iterations of the bootstrap, is set somewhat arbitrarily. The more iterations you do, the more accurate the estimate of the standard error, or the confidence interval.

# The Bootstrap

The bootstrap can be used with multivariate data, where the rows are sampled as units (see Figure 2-8). A model might then be run on the bootstrapped data, for example, to estimate the stability (variability) of model parameters, or to improve predictive power.

With classification and regression trees (also called *decision trees*), running multiple trees on bootstrap samples and then averaging their predictions (or, with classification, taking a majority vote) generally performs better than using a single tree. This process is called *bagging*.

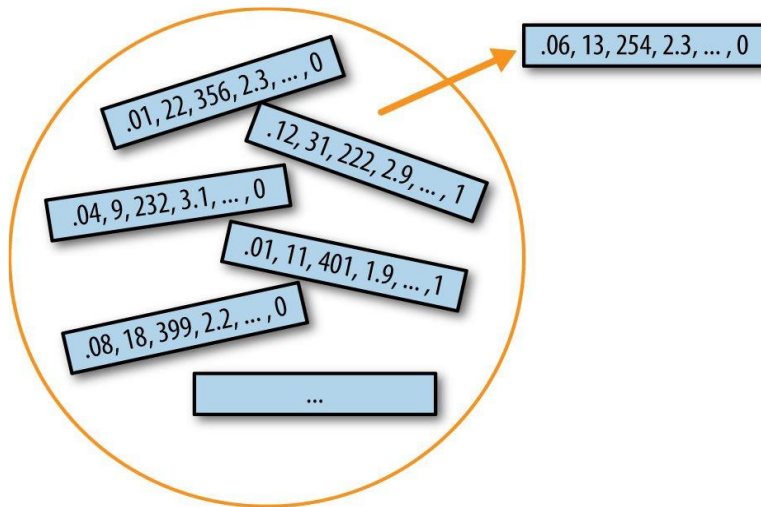


Figure 2-8. Multivariate bootstrap sampling