Data And Sampling Distributions

Trevor Barnes

8/27/2020

A popular misconception holds that the era of big data means the end of a need for sampling. In fact, the proliferation of data of varying quality and relevance reinforces the need for sampling as a tool to work efficiently with a variety of data and to minimize bias. Even in a big data project, predictive models are typically developed and piloted with samples. Samples are also used in tests of various sorts (e.g., pricing, web treatments).

Sometimes data is generated from a physical process that can be modeled. The simplest example is flipping a coin: this follows a binomial distribution. Any real-life binomial situation (buy or don't buy, fraud or no fraud, click or don't click) can be modeled effectively by a coin (with modified probability of landing heads, of course). In these cases, we can gain additional insight by using our understanding of the population.

Random Sampling and Sample Bias

A ***sample*** is a subset of data from a larger data set; statisticians call this larger data set the ***population***. A population in statistics is not the same thing as in biology—it is a large, defined but sometimes theoretical or imaginary, set of data.

Table 1: KEY TERMS FOR RANDOM SAMPLING

Term	Definition
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 Sampling	The sample that results from random sampling without stratifying the population.
Sample Bias	A sample that misrepresents the population.

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***.

Sample bias; that is, the sample was different in some meaningful nonrandom way from the larger population it was meant to represent. The term ***nonrandom*** is important—hardly any sample, including random samples, will be exactly representative of the population. Sample bias occurs when the difference is meaningful, and can be expected to continue for other samples drawn in the same way as the first.

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. An unbiased process will produce error, but it is random and does not tend strongly in any direction. 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

There are now a variety of methods to achieve representativeness, but at the heart of all of them lies random sampling.

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 overweighted in stratified sampling to yield equivalent sample sizes.

Size versus Quality

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. The classic scenario for the value of big data is when the data is not only big, but sparse as well.

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 μ is used to represent the mean of a population.

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.

Selection Bias

Selection bias refers to the practice of selectively choosing data—consciously or unconsciously—in a way that that leads to a conclusion that is misleading or ephemeral.

Term	Definition
Bias	Systemic 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.

Data snooping—that is, extensive hunting through the data until something interesting emerges. There is a saying among statisticians: "If you torture the data long enough, sooner or later it will confess."

A form of selection bias of particular concern to data scientists is what John Elder (founder of Elder Research, a respected data mining consultancy) calls the vast search effect. If you repeatedly run different models and ask different questions with a large data set, you are bound to find something interesting. Is the result you found truly something interesting, or is it the chance outlier?

Elder also advocates the use of what he calls target shuffling (a permutation test, in essence) to test the validity of predictive associations that a data mining model suggests.

Typical forms of selection bias in statistics, in addition to the vast search effect, include nonrandom sampling (see sampling bias), cherry-picking data, selection of time intervals that accentuate a partiular statistical effect, and stopping an experiment when the results look "interesting."

Regression to 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.

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).

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.

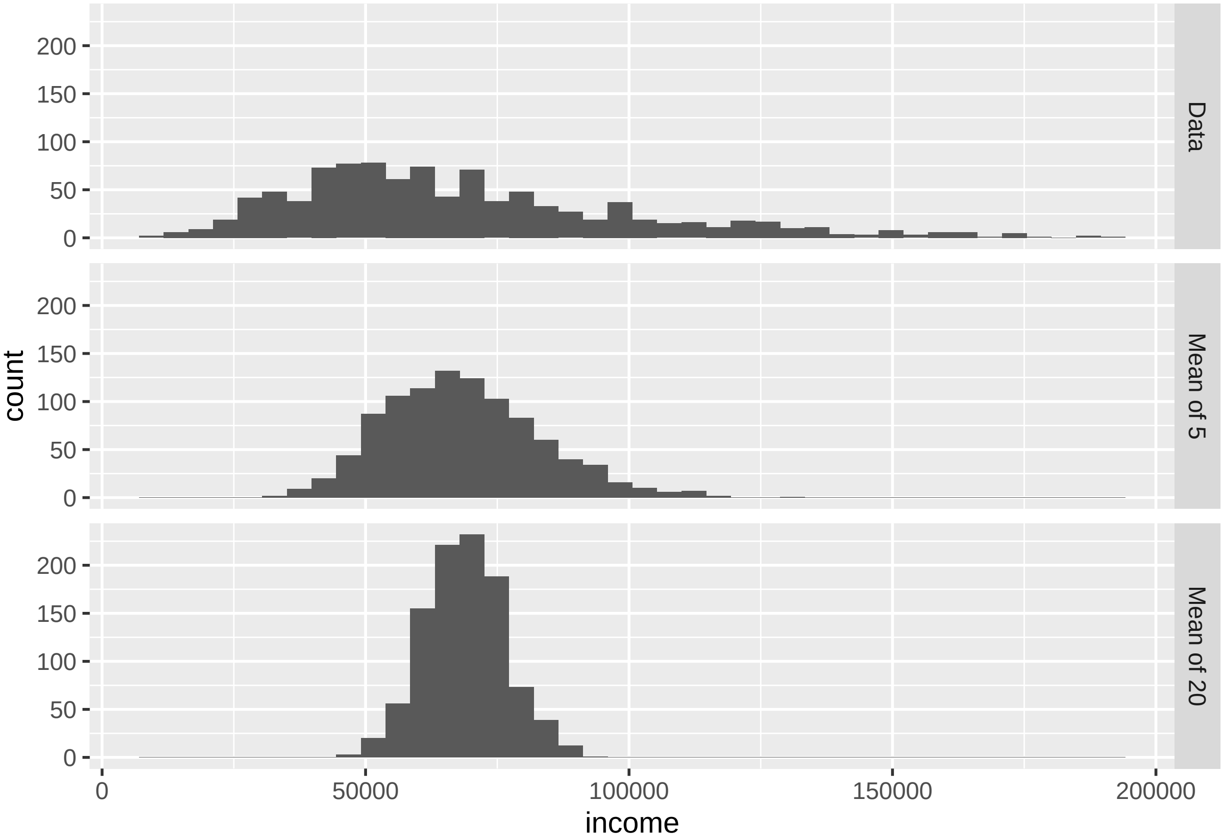
Term	Definition
Sample Statistic	A metric calculated for a sample of data drawn from a larger population.
Data Distribution	The frequency distribution of individual values in a data set.
Sampling Distribution	The frequency distribution of a *sample statistic* over many samples or resamples.
Central Limit 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*).

Typically, a sample is drawn with the goal of measuring something (with a ***sample statistic***) or modeling something (with a statistical or machine learning model). Since our estimate or model is based on a sample, it might be in error; it might be different if we were to draw a different sample. We are therefore interested in how different it might be—a key concern is ***sampling variability***. If we had lots of data, we could draw additional samples and observe the distribution of a sample statistic directly. Typically, we will calculate our estimate or model using as much data as is easily available, so the option of drawing additional samples from the population is not readily available.

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 that the statistic is based on, the more this is true. Also, the larger the sample, the narrower the distribution of the sample statistic.

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.

```r
#take a simple random sample
samp_data <- data.frame(income=sample(loans_income[[1]], 1000),
                        type="data_dist")
# take a sample of means of 5 values
samp_mean_05 <- data.frame(income=tapply(sample(loans_income[[1]], 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[[1]], 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 , 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.

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:

$StandardError = SE = \frac{s}{\sqrt{n}}$

As the sample size increases, the standard error decreases. 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.

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 modern statistics, the bootstrap has become the standard way to 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.

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.

Table 7: **Key Terms**

Term	Definition
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.

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.

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.

    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 result from this procedure is a bootstrap set of sample statistics or estimated model parameters, which you can then examine to see how variable they are.

The R package boot combines these steps in one function. For example, the following applies the bootstrap to the incomes of people taking out loans:

```
stat_fun <- function(x, idx) median(x[idx])
boot_obj <- boot(loans_income$x, R=1000, statistic = stat_fun)
boot_obj
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = loans_income$x, statistic = stat_fun, R = 1000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*    62000 -77.309    223.7771
```

The original estimate of the median is $62,000. The bootstrap distribution indicates that the estimate has a bias of about –$70 and a standard error of $209.

The bootstrap can be used with multivariate data, where the rows are sampled as units. 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*** (short for "bootstrap aggregating".

Resampling versus Bootstrapping

Sometimes the term ***resampling*** is used synonymously with the term ***bootstrapping***, as just outlined. More often, the term resampling also includes permutation procedures, where multiple samples are combined and the sampling may be done without replacement. In any case, the term ***bootstrap*** always implies sampling with replacement from an observed data set.

Key Ideas
The bootstrap (sampling with replacement from a data set) is a powerful tool for assessing the variability of a sample statistic. The bootstrap can be applied in similar fashion in a wide variety of circumstances, without extensive study of mathematical approximations to sampling distributions. It also allows us to estimate sampling distributions for statistics where no mathematical approximation has been developed. When applied to predictive models, aggregating multiple bootstrap sample predictions (bagging) outperforms the use of a single model.

Confidence Intervals

Frequency tables, histograms, boxplots, and standard errors are all ways to understand the potential error in a sample estimate. Confidence intervals are another.

|Term|Definition| |**Confidence Interval**|The percentage of confidence intervals, constructed in the same way from the same population, expected to contain the statistic of interest.| |**Interval Endpoints**|The top and bottom of the confidence interval.|

Analysts and managers, while acknowledging uncertainty, nonetheless place undue faith in an estimate when it is presented as a single number (a ***point estimate***). Presenting an estimate not as a single number but as a range is one way to counteract this tendency. Confidence intervals do this in a manner grounded in statistical sampling principles.

Confidence intervals always come with a coverage level, expressed as a (high) percentage, say 90% or 95%. One way to think of a 90% confidence interval is as follows: it is the interval that encloses the central 90% of the bootstrap sampling distribution of a sample statistic. More generally, an $x\%$ confidence interval around a sample estimate should, on average, contain similar sample estimates $x\%$ of the time (when a similar sampling procedure is followed).

Given a sample of size n, and a sample statistic of interest, the algorithm for a bootstrap confidence interval is as follows:

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.

The bootstrap is a general tool that can be used to generate confidence intervals for most statistics, or model parameters. Statistical textbooks and software, with roots in over a half-century of computerless statistical analysis, will also reference confidence intervals generated by formulas, especially the t-distribution.

The percentage associated with the confidence interval is termed the ***level of confidence***. The higher the level of confidence, the wider the interval. Also, the smaller the sample, the wider the interval (i.e., the more uncertainty). Both make sense: the more confident you want to be, and the less data you have, the wider you must make the confidence interval to be sufficiently assured of capturing the true value.

Key Ideas
Confidence intervals are the typical way to present estimates as an interval range.
The more data you have, the less variable a sample estimate will be.
The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
The bootstrap is an effective way to construct confidence intervals.

Normal Distribution

The bell-shaped normal distribution is iconic in traditional statistics. The fact that distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.

|Term|Definition| |**Error**|The difference between a data point and a predicted or average value.| |**Standardize**|Subtract the mean and divide by the standard deviation.| |**z-score**|The result of standard-

izing an individual data point.| |**Standard Normal**|A normal distribution with mean = 0 and standard deviation = 1.| |**QQ-Plot**|A plot to visualize how close a sample distribution is to a normal distribution|

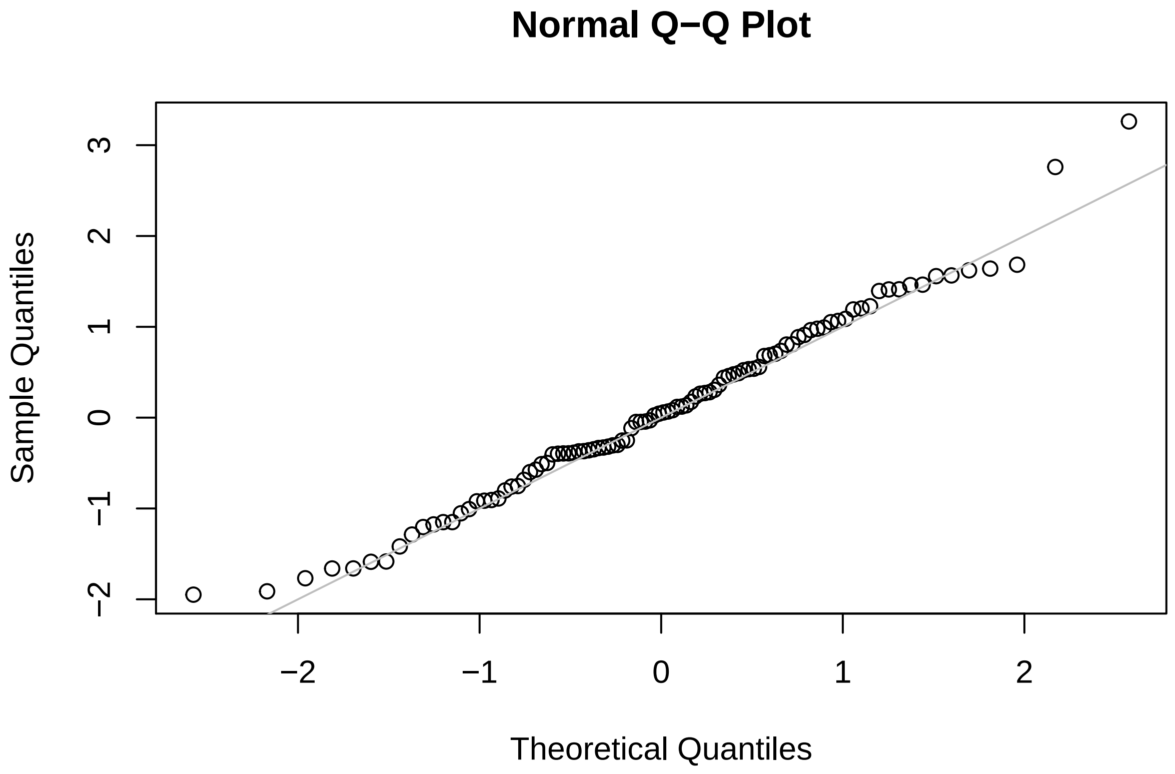
In a normal distribution, 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations.

Standard Normal and QQ-Plots

A ***standard normal*** distribution is one in which the units on the x-axis are expressed in terms of standard deviations away from the mean. To compare data to a standard normal distribution, you subtract the mean then divide by the standard deviation; this is also called ***normalization*** or ***standardization***. The transformed value is termed a ***z-score***, and the normal distribution is sometimes called the ***z-distribution***.

A QQ-Plot is used to visually determine how close a sample is to the normal distribution. The QQ-Plot orders the z-scores from low to high, and plots each value's z-score on the y-axis; the x-axis is the corresponding quantile of a normal distribution for that value's rank. Since the data is normalized, the units correspond to the number of standard deviations away of the data from the mean. If the points roughly fall on the diagonal line, then the sample distribution can be considered close to normal.

```r
norm_samp <- rnorm(100)
qqnorm(norm_samp)
abline(a=0, b=1, col='grey')
```



Key Ideas
The normal distribution was essential to the historical development of statistics, as it permitted mathematical approximation of uncertainty and variability.
While raw data is typically not normally distributed, errors often are, as are averages and totals in large samples.

Key Ideas
To convert data to z-scores, you subtract the mean of the data and divide by the standard deviation; you can then compare the data to a normal distribution.

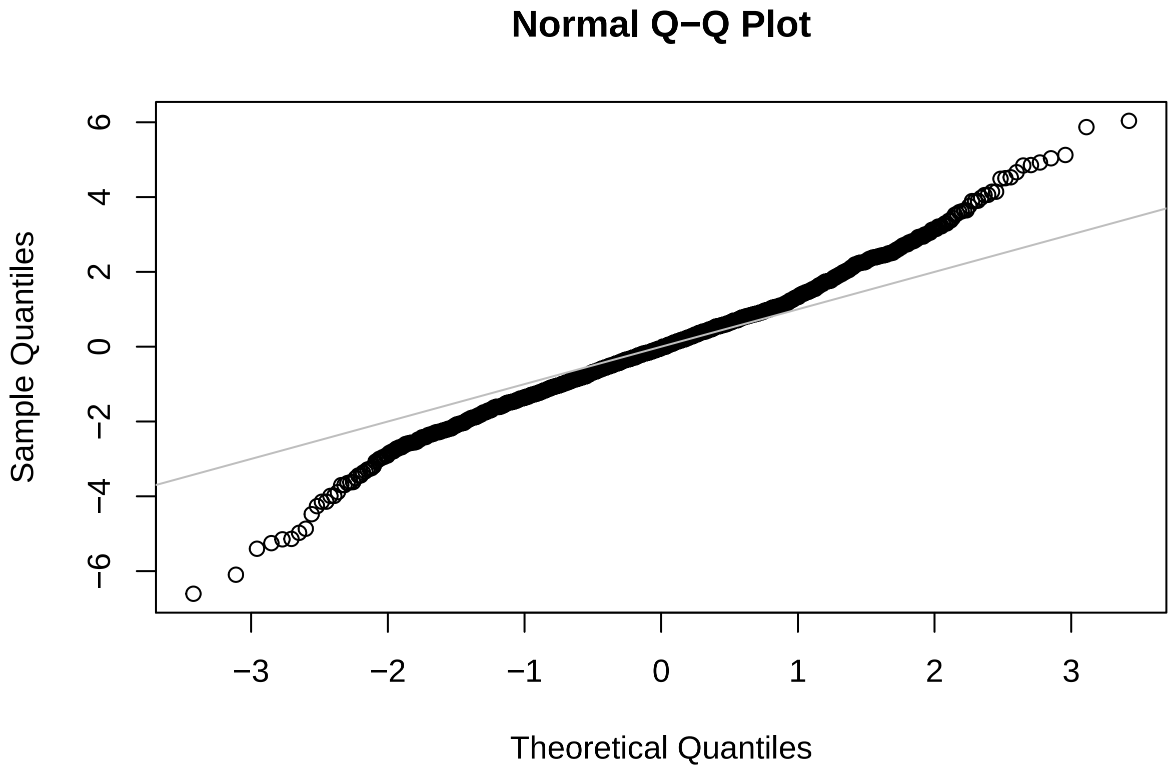
Long-Tailed Distributions

Despite the importance of the normal distribution historically in statistics, and in contrast to what the name would suggest, data is generally not normally distributed.

Term	Definition
Tail	The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.
Skew	Where one tail of a distribution is longer than the other.

While the normal distribution is often appropriate and useful with respect to the distribution of errors and sample statistics, it typically does not characterize the distribution of raw data. Sometimes, the distribution is highly ***skewed*** (asymmetric), such as with income data, or the distribution can be discrete, as with binomial data. Both symmetric and asymmetric distributions may have ***long tails***. The tails of a distribution correspond to the extreme values (small and large). Long tails, and guarding against them, are widely recognized in practical work. Nassim Taleb has proposed the ***black swan theory***, which predicts that anamolous events, such as a stock market crash, are much more likely to occur than would be predicted by the normal distribution.

```
nflx <- sp500_px[,'NFLX']
nflx <- diff(log(nflx[nflx>0]))
qqnorm(nflx)
abline(a=0, b=1, col='grey')
```



In contrast, the points are far below the line for low values and far above the line for high values. This means that we are much more likely to observe extreme values than would be expected if the data had a normal distribution. NFLX shows another common phenomena: the points are close to the line for the data within one standard deviation of the mean. Tukey refers to this phenomenon as data being "normal in the middle," but having much longer tails.

Key Ideas
Most data is not normally distributed. Assuming a normal distribution can lead to underestimation of extreme events ("black swans").

Student's t-Distribution

The ***t-distribution*** is a normally shaped distribution, but a bit thicker and longer on the tails. It is used extensively in depicting distributions of sample statistics. 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.

Term	Definition
n	Sample size.
Degrees of Freedom	A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.

A number of different statistics can be compared, after standardization, to the t- distribution, to estimate confidence intervals in light of sampling variation. Consider a sample of size n for which the sample mean has been calculated. If s is the sample standard deviation, a 90% confidence interval around the sample mean is given by:

$\bar{x} \pm t_{n-1}(.05) * \frac{s}{n}$

where $t_{n-1}(.05)$is the value of the t-statistic, with (n˘1) degrees of freedom, that "chops off" 5% of the t-distribution at either end. The t-distribution has been used as a reference for the distribution of a sample mean, the difference between two sample means, regression parameters, and other statistics.

It turns out that sample statistics are often normally distributed, even when the underlying population data is not (a fact which led to widespread application of the t-distribution). This phenomenon is termed the ***central limit theorem***.

|Key Ideas| |The t-distribution is actually a family of distributions resembling the normal distribution, but with thicker tails.| |It is widely used as a reference basis for the distribution of sample means, differerences between two sample means, regression parameters, and more.|

Binomial Distribution

Table 14: **KEY TERMS FOR BINOMIAL DISTRIBUTION**

Term	Definition	Synonym
Trial	An event with a discrete outcome (e.g., a coin flip).	
Success	The outcome of interest for a trial.	"1" (as opposed to "0")
Binomial	Having two outcomes	yes/no, 0/1, binary
Binomial Trial	A trial with two outcomes.	Bernoulli trial
Binomial Distribution	Distribution of number of successes in x trials.	Bernoulli distribution

Yes/no (binomial) outcomes lie at the heart of analytics since they are often the culmination of a decision or other process; buy/don't buy, click/don't click, survive/die, and so on. Central to understanding the binomial distribution is the idea of a set of ***trials***, each trial having two possible outcomes with definite probabilities. Such yes/no or 0/1 outcomes are termed ***binary*** outcomes, and they need not have 50/50 probabilities. Any probabilities that sum to 1.0 are possible. It is conventional in statistics to term the "1" outcome the ***success*** outcome; it is also common practice to assign "1" to the more rare outcome. Use of the term ***success*** does not imply that the outcome is desirable or beneficial, but it does tend to indicate the outcome of interest. For example, loan defaults or fraudulent transactions are relatively uncommon events that we may be interested in predicting, so they are termed "1s" or "successes."

The binomial distribution is the frequency distribution of the number of successes (x) in a given number of trials (n) with specified probability (p) of success in each trial. There is a family of binomial distributions, depending on the values of x, n, and p.

```
dbinom(x=2, size = 5, p=0.1)
```

```
## [1] 0.0729
```

would return 0.0729, the probability of observing exactly x = 2 successes in n = 5 trials, where the probability of success for each trial is p = 0.1.

Often we are interested in determining the probability of x or fewer successes in n trials. In this case, we use the function `pbinom`:

```
pbinom(2, 5, 0.1)
```

```
## [1] 0.99144
```

This would return 0.9914, the probability of observing two or fewer successes in five trials, where the probability of success for each trial is 0.1.

The mean of a binomial distribution is $n * p$; you can also think of this as the expected number of successes in n trials, for success probability $= p$.

The variance is $n * p(1 - p)$. With a large enough number of trials (particularly when p is close to 0.50), the binomial distribution is virtually indistinguishable from the normal distribution. In fact, calculating binomial probabilities with large sample sizes is computationally demanding, and most statistical procedures use the normal distribution, with mean and variance, as an approximation.

Key Ideas
Binomial outcomes are important to model, since they represent, among other things, fundamental decisions (buy or don't buy, click or don't click, survive or die, etc.).
A binomial trial is an experiment with two possible outcomes: one with probability p and the other with probability 1˘p.
With large n, and provided p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

Poisson and Related Distributions

Many processes produce events randomly at a given overall rate—visitors arriving at a website, cars arriving at a toll plaza (events spread over time), imperfections in a square meter of fabric, or typos per 100 lines of code (events spread over space).

Table 16: **KEY TERMS FOR POISSON AND RELATED DISTRIBUTIONS** ## Poisson Distributions

Term	Definition
Lambda	The rate (per unit of time or space) at which events occur.
Poisson Distribution	The frequency distribution of the number of events in sampled units of time or space.
Exponential Distribution	The frequency distribution of the time or distance from one event to the next event.
Weibull Distribution	A generalized version of the exponential, in which the event rate is allowed to shift over time.

From prior data we can estimate the average number of events per unit of time or space, but we might also want to know how different this might be from one unit of time/space to another. The Poisson distribution tells us the distribution of events per unit of time or space when we sample many such units. It is useful when addressing queuing questions like "How much capacity do we need to be 95% sure of fully processing the internet traffic that arrives on a server in any 5- second period?"

The key parameter in a Poisson distribution is λ, or lambda. This is the mean number of events that occurs in a specified interval of time or space. The variance for a Poisson distribution is also λ.

A common technique is to generate random numbers from a Poisson distribution as part of a queuing simulation. The rpois function in R does this, taking only two arguments—the quantity of random numbers sought, and lambda:

```
rpois(100, lambda=2)
```

```
##   [1] 2 1 4 2 1 2 2 0 2 2 2 3 2 0 1 2 3 3 2 3 2 2 2 2 3 1 1 1 5 1 3 5 2 3 3 1 0
##  [38] 1 0 3 3 3 1 1 3 1 4 2 2 5 1 4 2 1 1 4 3 1 1 0 2 3 4 4 2 2 4 0 1 4 3 3 1 0
##  [75] 1 2 1 3 2 1 2 0 3 2 3 0 4 3 0 5 1 3 3 1 2 1 3 1 2 2
```

This code will generate 100 random numbers from a Poisson distribution with $\lambda = 2$. For example, if incoming customer service calls average 2 per minute, this code will simulate 100 minutes, returning the number of calls in each of those 100 minutes.

Exponential Distribution

Using the same parameter λ that we used in the Poisson distribution, we can also model the distribution of the time between events: time between visits to a website or between cars arriving at a toll plaza. It is also used in engineering to model time to failure, and in process management to model, for example, the time required per service call. The R code to generate random numbers from an exponential distribution takes two arguments, n (the quantity of numbers to be generated), and ***rate***, the number of events per time period. For example:

```
rexp(n = 100, rate = .2)
```

```
##   [1]  7.8249540  1.9606647  2.9729312  5.7722935  5.9474265  1.1852879
##   [7]  0.6226817  1.2733655  0.5055869  9.9987730  2.1758023  1.6526729
##  [13]  2.8116733 14.0858851  7.5713256  3.4625258 11.0727074  2.4193373
##  [19]  2.4747339  2.0840537  4.0941971  0.1766453  1.6970376 29.1654622
##  [25]  0.4906194  6.4535360  4.0061364  1.7783524 16.7644225  3.6998871
##  [31]  1.5902482  0.4126561  5.4049608  5.8617126  1.6656084  9.5908584
##  [37]  6.2894686  0.8462398 16.4125442  4.6087793  1.4526327  2.1806273
##  [43]  0.4779252  0.4176100  3.8195717  1.6655504  1.5567965  2.4224045
##  [49]  5.8340169  3.0545686 12.2665459  5.2909811 17.3341014  7.6824742
##  [55]  3.0912138  3.2811445  8.8855207  8.1731283  7.6382499  3.0072771
##  [61]  2.3335638  3.4323222  2.3730885  7.4768071 12.4339489  1.1910142
##  [67]  0.8022172 23.5795884  6.9373898  0.3465583  0.6027417  0.5644035
##  [73]  4.4492158  0.7210473  6.5904036  1.0310292  2.2597339  7.7552814
##  [79]  4.1363150  0.9682523 33.7632405  2.0012371  0.7473122  5.9566148
##  [85]  3.9239735  3.1324395  2.3131921  2.5957007  4.4242151  4.2194296
##  [91]  4.7922933  0.3458673  1.8903330 15.8476286  3.4424582 11.5366371
##  [97]  1.8905567  6.4895508  3.4619704 13.7480172
```

This code would generate 100 random numbers from an exponential distribution where the mean number of events per time period is 2. So you could use it to simulate 100 intervals, in minutes, between service calls, where the average rate of incoming calls is 0.2 per minute.

A key assumption in any simulation study for either the Poisson or exponential distribution is that the rate, λ, remains constant over the period being considered. This is rarely reasonable in a global sense; for example, traffic on roads or data networks varies by time of day and day of week. However, the time periods, or areas of space, can usually be divided into segments that are sufficiently homogeneous so that analysis or simulation within those periods is valid.

Estimating the Failure Rate

In many applications, the event rate, λ, is known or can be estimated from prior data. However, for rare events, this is not necessarily so. Aircraft engine failure, for example, is sufficiently rare (thankfully) that, for a given engine type, there may be little data on which to base an estimate of time between failures. With no data at all, there is little basis on which to estimate an event rate. However, you can make some guesses: if no events have been seen after 20 hours, you can be pretty sure that the rate is not 1 per hour. Via simulation, or direct calculation of probabilities, you can assess different hypothetical event rates and estimate threshold values below which the rate is very unlikely to fall. If there is some data but not enough to provide a precise, reliable estimate of the rate, a goodness-of-fit test (see "Chi-Square Test") can be applied to various rates to determine how well they fit the observed data.

Weibull Distribution

In many cases, the event rate does not remain constant over time. If the period over which it changes is much longer than the typical interval between events, there is no problem; you just subdivide the analysis into the segments where rates are relatively constant, as mentioned before. If, however, the event rate changes over the time of the interval, the exponential (or Poisson) distributions are no longer useful. This is likely to be the case in mechanical failure—the risk of failure increases as time goes by. The Weibull distribution is an extension of the exponential distribution, in which the event rate is allowed to change, as specified by a ***shape parameter***, β. If $\beta > 1$, the probability of an event increases over time, if $\beta < 1$, it decreases. Because the Weibull distribution is used with time-to-failure analysis instead of event rate, the second parameter is expressed in terms of characteristic life, rather than in terms of the rate of events per interval. The symbol used is η, the Greek letter eta. It is also called the scale parameter.

With the Weibull, the estimation task now includes estimation of both parameters, β and η. Software is used to model the data and yield an estimate of the best-fitting Weibull distribution.

The R code to generate random numbers from a Weibull distribution takes three arguments, n (the quantity of numbers to be generated), shape, and scale. For example, the following code would generate 100 random numbers (lifetimes) from a Weibull distribution with shape of 1.5 and characteristic life of 5,000:

```r
rweibull(100, 1.5, 5000)
```

```
##   [1]  4166.3654  1976.5156  4282.4492  5591.2640  3719.9237   948.5159
##   [7]  9885.6024  1190.1851  4639.3029  6040.2727  3858.8845  1474.0978
##  [13] 10217.5514  3550.2521  3608.8457  2424.7167  1178.2341  5538.7292
##  [19]   657.2012  4075.7282  3923.1830  3403.1246   781.3662  1582.1656
##  [25]  2991.3687  1347.5663  7817.4292 12244.6043 11031.0106  2567.4961
##  [31]  4825.9657 10506.1711  6545.4244  1258.0880  3251.4297  7640.5897
##  [37]  4008.4876  4621.6942  6257.8242  3184.9886  1370.8151  6435.2138
##  [43]  3134.0482  4941.3086   115.1612  9535.8172  2706.8378  7148.1117
##  [49]  1247.5519  1307.6816   600.6575  3875.6774  8029.3057  7517.6278
##  [55]  5222.5849  2084.4018 13686.3134  6379.7659   293.9250 11879.5908
##  [61] 10137.6601  5263.2019  1655.7477  6051.7632  3732.4527   972.6355
##  [67]  9799.2488  2622.4853  6001.0184  2741.5790  2460.3631  2979.7866
##  [73]  4992.6474  5761.1510  3867.0339 12108.9511   588.5548  4072.9523
##  [79]  6746.1294   942.3757  6261.0054  3262.7028  5721.2768   771.6690
##  [85]  8537.4791  5432.2781 14272.8435  8437.2651 11021.8829  4262.6282
##  [91]  8898.4076   706.1005  2034.4467  5407.7208  4359.5866  5105.1347
##  [97]  5521.9112  1521.8859  3391.1119   656.9593
```

Key Ideas
For events that occur at a constant rate, the number of events per unit of time or space can be modeled as a Poisson distribution.
In this scenario, you can also model the time or distance between one event and the next as an exponential distribution.
A changing event rate over time (e.g., an increasing probability of device failure) can be modeled with the Weibull distribution.