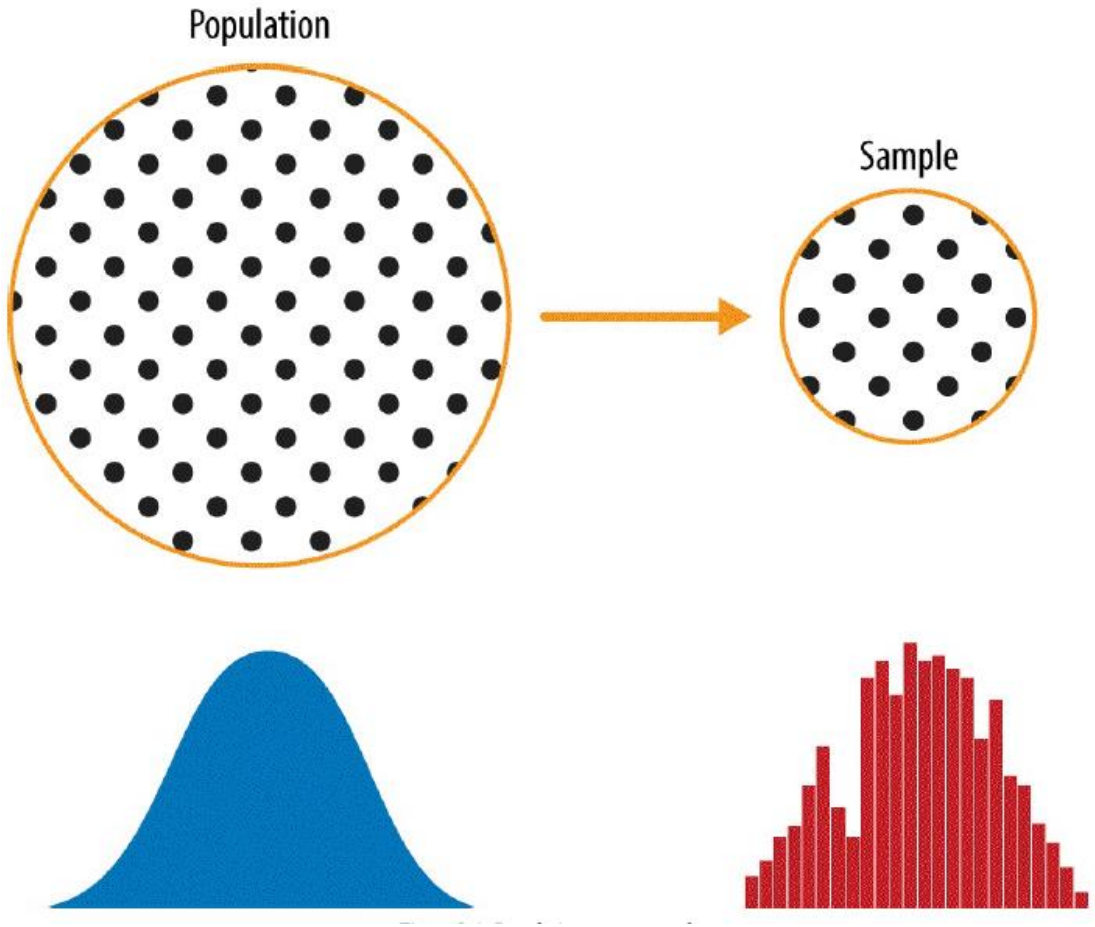
Data and Sampling Distributions



*This presentation is based on the book ‘Practical Statistics for Data Scientist’ by Peter Bruce & Andrew Bruce

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.

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.

Random Sampling

- *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*.
- 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

- The classic example is the *Literary Digest* poll of 1936 that predicted a victory of Al Landon against Franklin Roosevelt. The *Literary Digest*, a leading periodical of the day, polled its entire subscriber base, plus additional lists of individuals, a total of over 10 million, and predicted a landslide victory for Landon. George Gallup, founder of the Gallup Poll, conducted biweekly polls of just 2,000, and accurately predicted a Roosevelt victory. The difference lay in the selection of those polled.
- The Literary Digest opted for quantity, paying little attention to the method of selection. They ended up polling those with relatively high socioeconomic status (their own subscribers, plus those who, by virtue of owning luxuries like telephones and automobiles, appeared in marketers' lists).

- The result was ***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.

  - 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 be those who 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.

Bias

- Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.
  - 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
  - The results shown in secodn figure 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.
- When a result does suggest bias, it is often an indicator that a statistical or machine learning model has been misspecified, or an important variable left out.

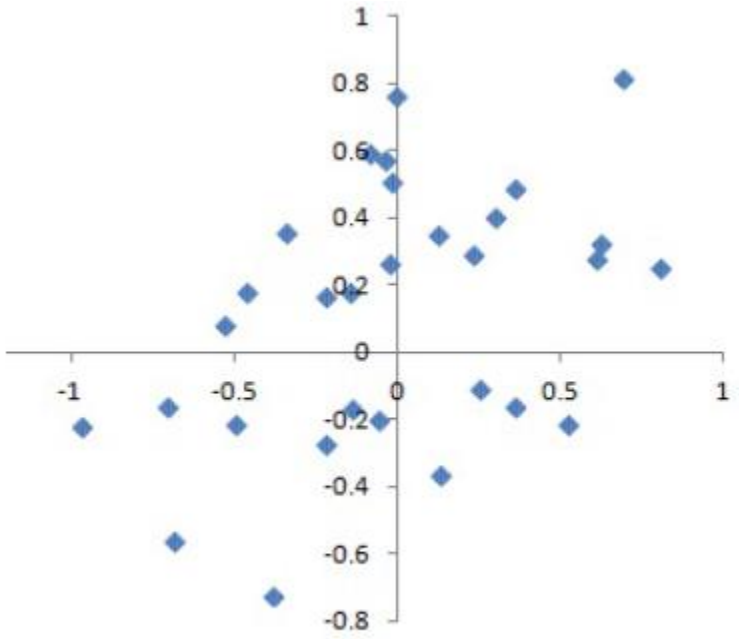


Figure 2-2. Scatterplot of shots from a gun with true aim

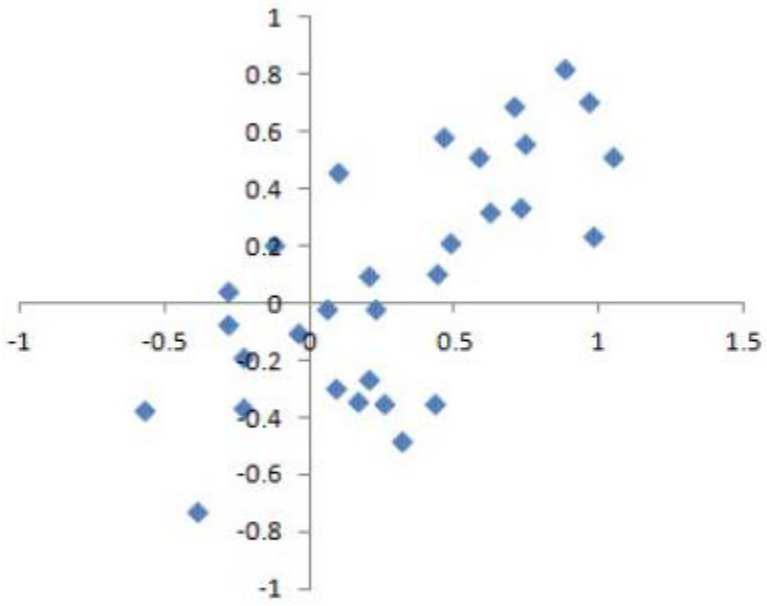


Figure 2-3. Scatterplot of shots from a gun with biased aim

- 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.

- 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

- To paraphrase Yogi Berra, “If you don’t know what you’re looking for, look hard enough and you’ll find it.”
- **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.

- If you specify a hypothesis and conduct a well-designed experiment to test it, you can have high confidence in the conclusion. Such is often not the case, however. Often, one looks at available data and tries to discern patterns. But is the pattern for real, or just the product of **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.”

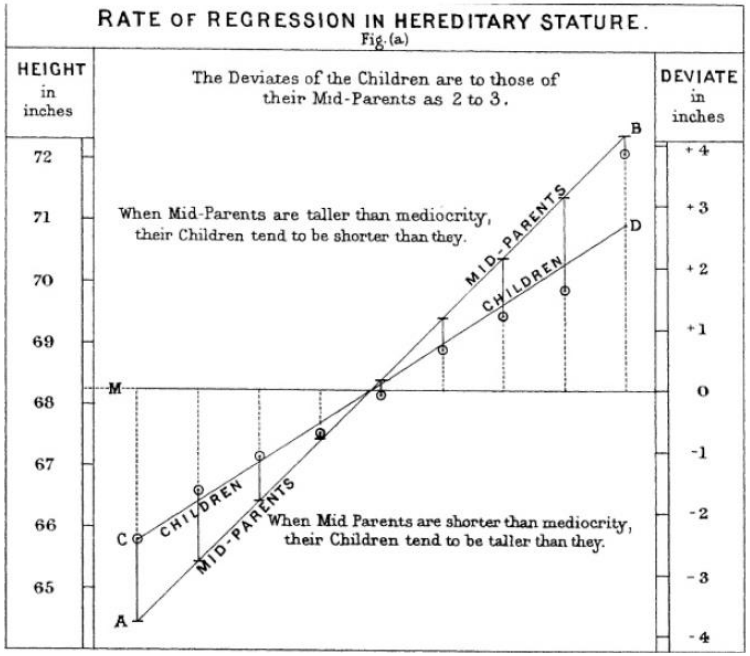
- Since repeated review of large data sets is a key value proposition in data science, selection bias is something to worry about. A form of selection bias of particular concern to data scientists is what John Elder calls the ***vast search effect***.
  - If you repeatedly run different models and ask 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?
  - We can guard against this by using a holdout set, and sometimes more than one holdout set, against which to validate performance.
  - 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 particular statistical effect
  - stopping an experiment when the results look “interesting.”

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.
- 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.

The children of extremely tall men tend not to be as tall as their father (Galton-1886)



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.

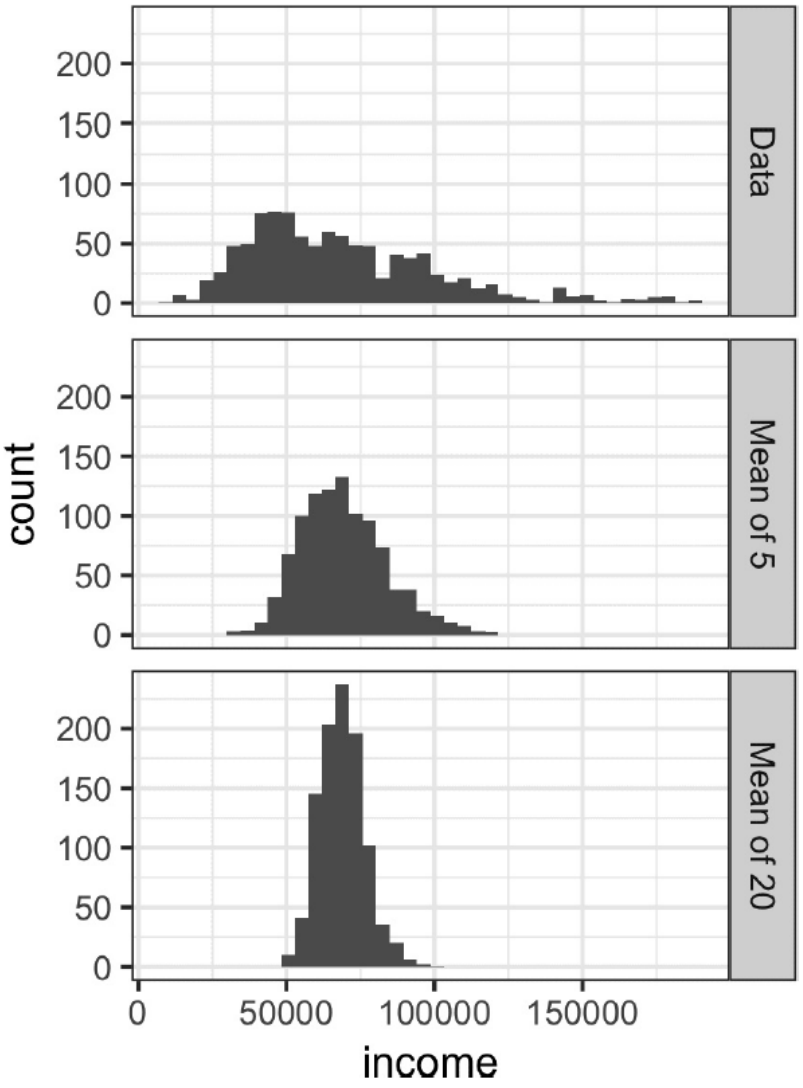
Standard error

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*).

- 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***.
- 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***.

- This is illustrated in an example using annual income for loan applicants to Lending Club
- 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

- This phenomenon is termed the **central limit theorem**.



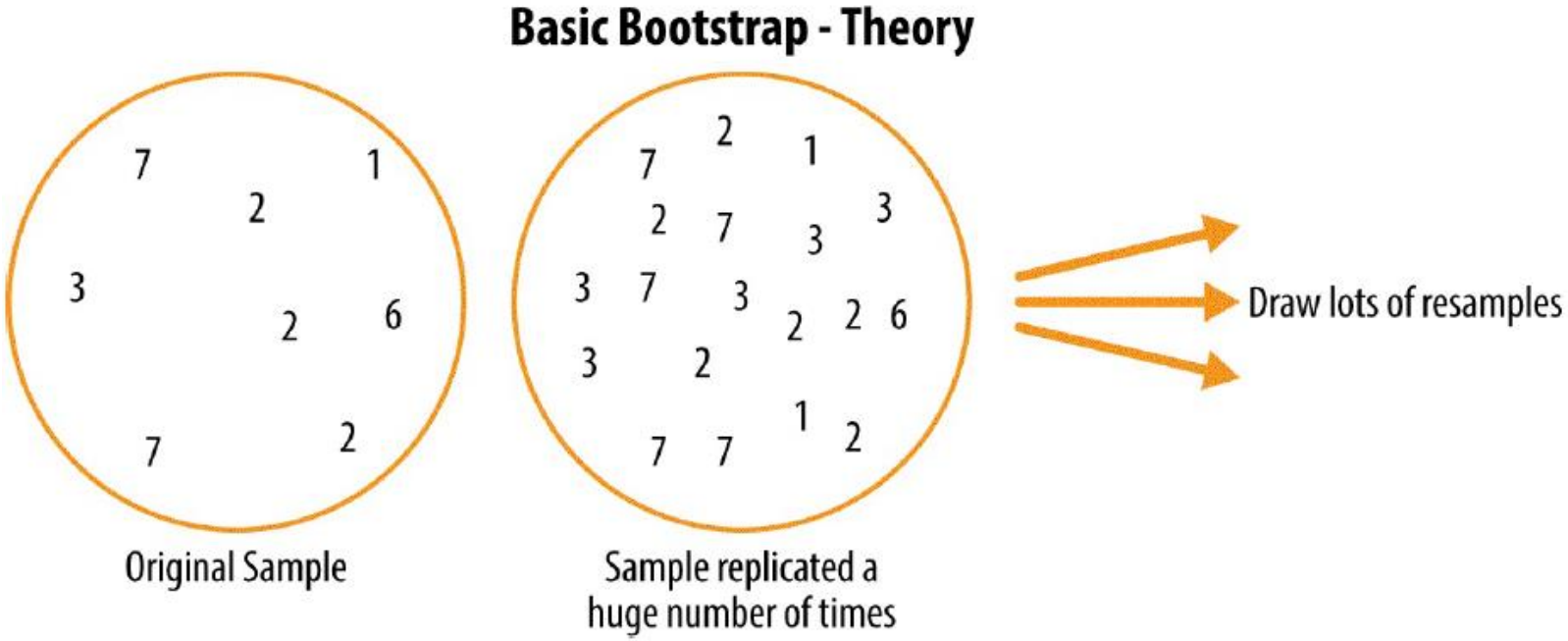
Central Limit Theorem

- It says that the means drawn from multiple samples will resemble the familiar bell-shaped normal curve (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 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.

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.
- 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).



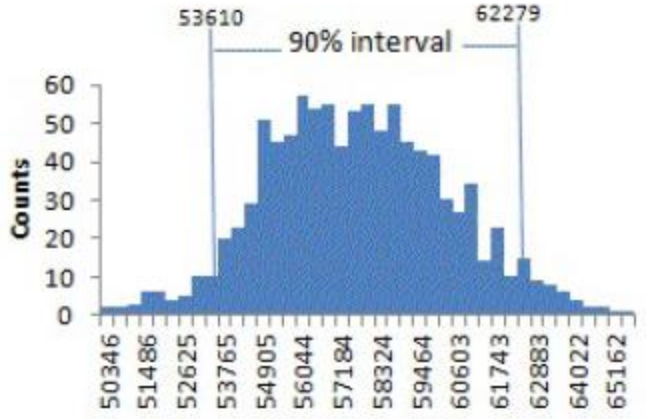
- The algorithm for a bootstrap resampling of the mean is as follows, for a sample of size n:
  1. Draw a sample value, record, replace it.
  2. Repeat n times.
  3. Record the mean of the n resampled values.
  4. Repeat steps 1–3 R times.
  5. Use the R results to:
  - Calculate their standard deviation (this estimates sample mean standard error).
  - Produce a histogram or boxplot.
  - Find a confidence interval.

- 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.

Confidence Intervals

- 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.
- 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.

90% confidence interval for the mean annual income of loan applicants, based on a sample of 20



- 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.

- For a data scientist, a confidence interval is a tool 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, learn whether a larger sample is needed.

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.

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 standardizing 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.

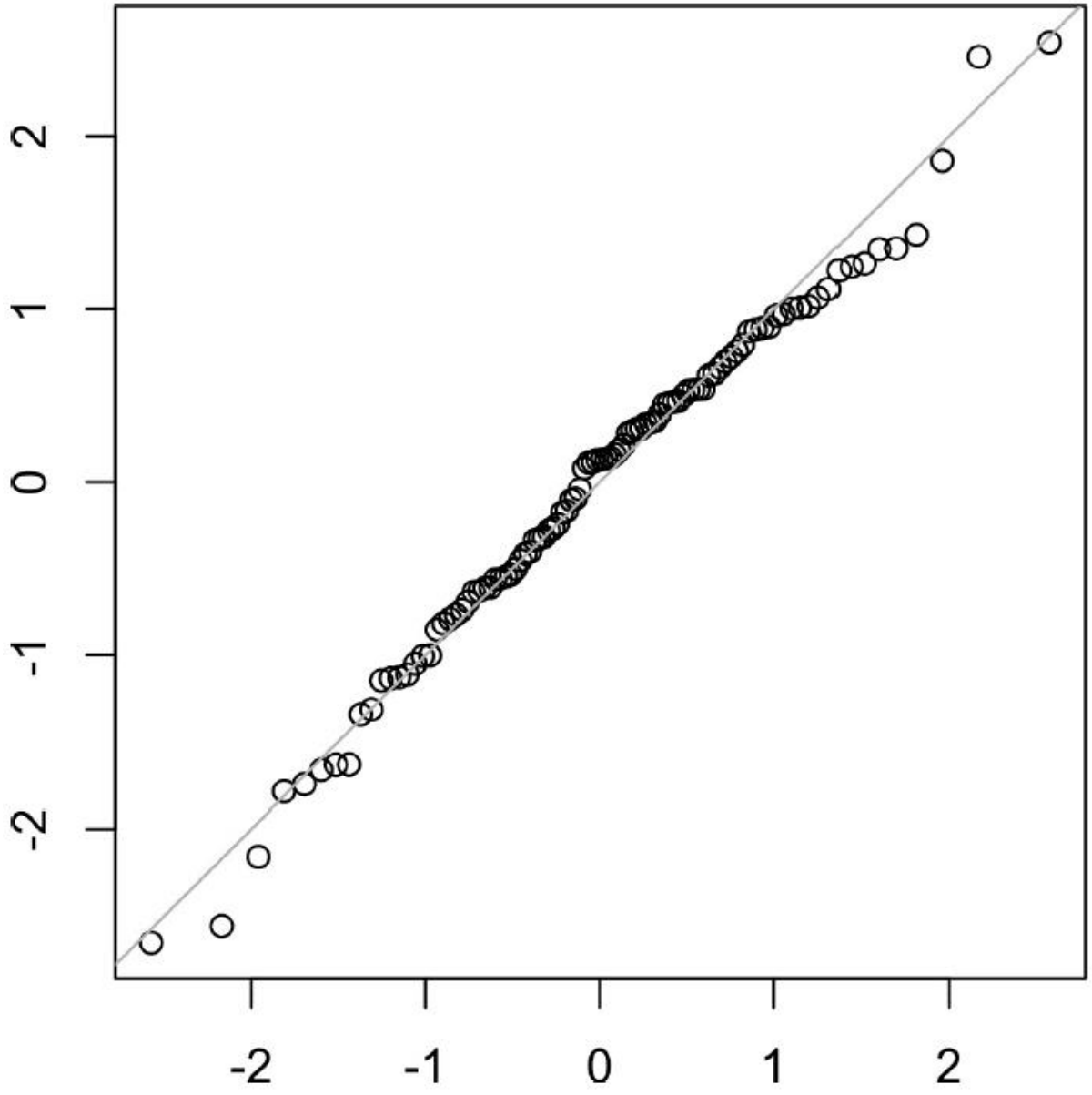
- It is a common misconception that the normal distribution is called that because most data follows a normal distribution — that is, it is the normal thing.
- Most of the variables used in a typical data science project — in fact most raw data as a whole — are not normally distributed: see “Long-Tailed Distributions”.

- The utility of the normal distribution derives from the fact that *many statistics are normally distributed in their sampling distribution*. (Central Limit Theorem)
- Even so, assumptions of normality are generally a last resort, used when empirical probability distributions, or bootstrap distributions, are not available.

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.**
- 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.
- If the points roughly fall on the diagonal line, then the **sample distribution** can be considered close to normal.

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



Long-Tailed Distributions

- 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***.
- Both symmetric and asymmetric distributions may have *long tails*.
- The tails of a distribution correspond to the extreme values (small and large).
- 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.

- In contrast to Figure above, 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.

- Figure on right 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

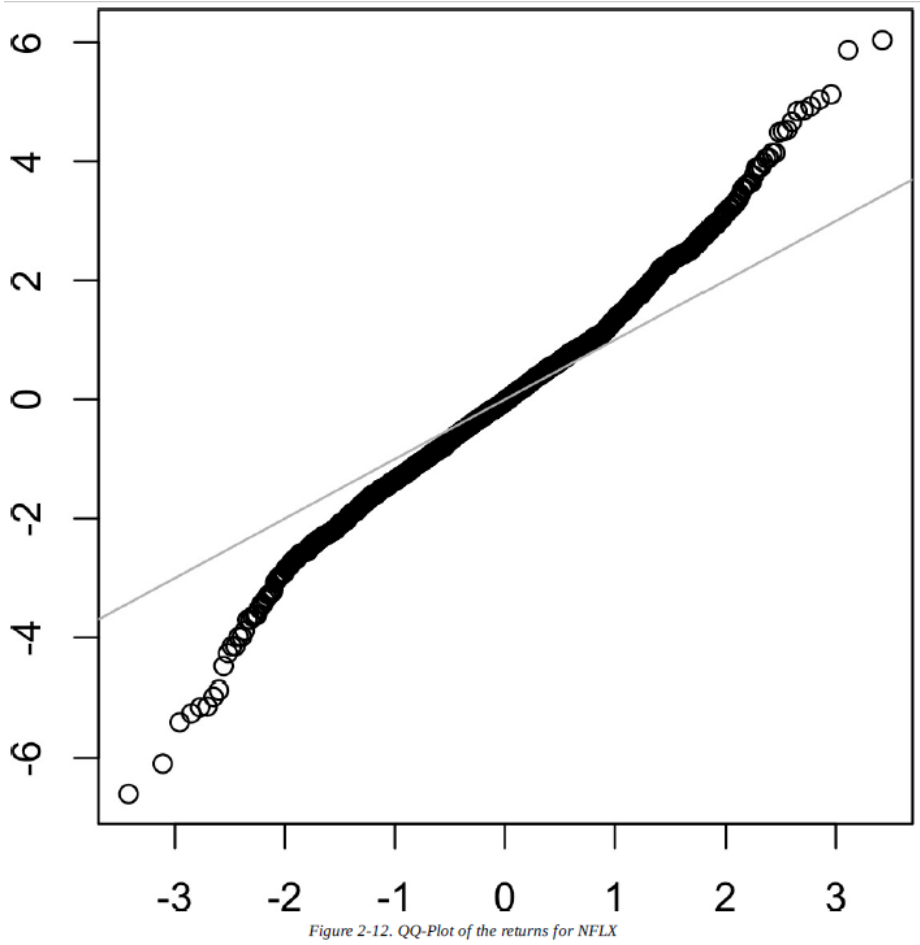


Figure 2-12. QQ-Plot of the returns for NFLX

Student's 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. (Central Limit Theorem)

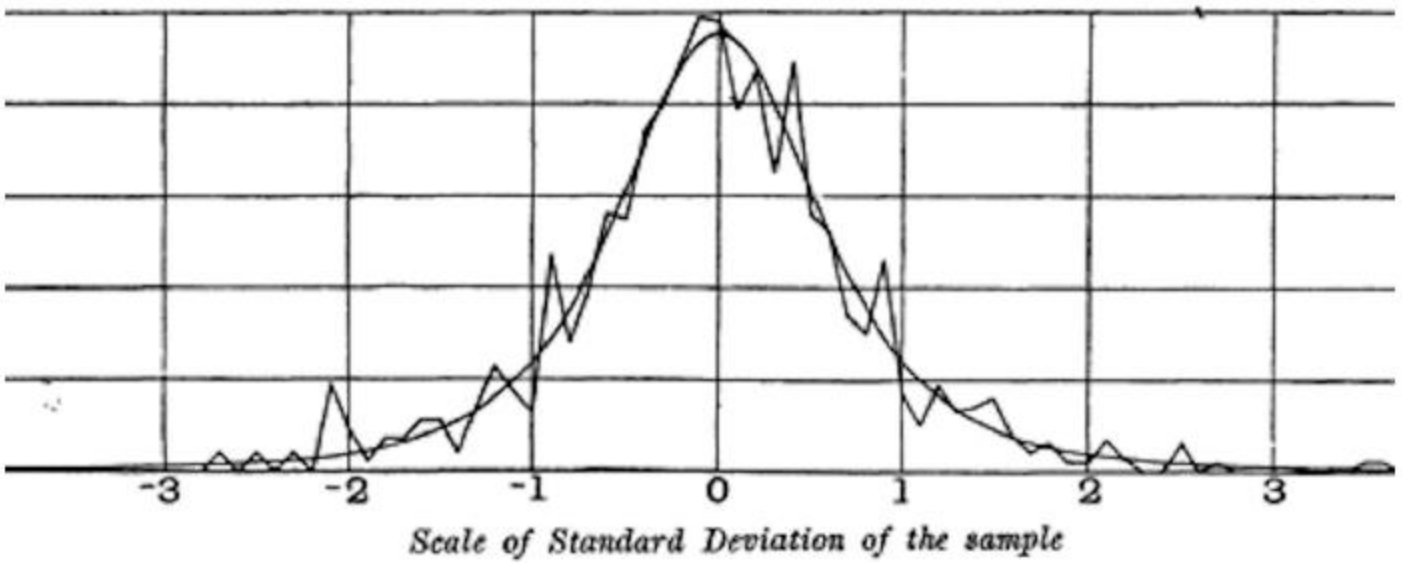


Figure 2-13. Gossett's resampling experiment results and fitted t-curve (from his 1908 Biometrika paper)

****

- Had computing power been widely available in 1908, statistics would no doubt have relied much more heavily on computationally intensive resampling methods from the start. Lacking computers, statisticians turned to mathematics and functions such as the t-distribution to approximate sampling distributions.
- 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.

- What do data scientists need to know about the t-distribution and the central limit theorem? Not a whole lot.
- These distributions are used in classical statistical inference, but are 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.

Binomial Distribution

- Central to understanding the binomial distribution is the idea of a set of trials, each trial having two possible outcomes with definite probabilities.

- 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.

- The binomial distribution would answer a question like:
  - If the probability of a click converting to a sale is 0.02, what is the probability of observing 0 sales in 200 clicks?

- 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

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

Exponential Distribution

- Using the same parameter that we used in the Poisson distribution, we can also model the distribution of the time between events
  - time between visits to a website
  - time between cars arriving at a toll plaza

- A key assumption in any simulation study for either the Poisson or exponential distribution is that the rate, , 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.

Weibull Distribution

- In many cases, the event rate does not remain constant over time.
- 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*, ß .
- If ß > 1, the probability of an event increases over time, if ß < 1, it decreases.

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***.