Introduction

Misconception: when the era of **big data** had come, the era of **sampling** has **ended**.

A lot of **data** has different **quality levels** and **relevance**, reinforcing the need for **sampling** as a **tool** to work **efficiently** with a **variety of data** and to **minimize bias**. Even in **big data projects**, **predictive models** are **developed** and **piloted** with **samples**. Also, **sampling** is used in **tests** for **various sorts**: comparing the **effect** of **web page designs** on **clicks.**

Figure 2-1: **left side** is **population**, **right side** is **sample**.

Population in **statistics** is assumed to follow an **underlying but unknown distribution**. All what we have in hand is the **sample data** and its **empirical distribution** on the **right side**.

Sampling is referred to by an **arrow.**

Traditional statistics focuses on **population** with **strong assumptions** about population. **Modern statistics** has moved to the **right side** where such **strong assumptions** are **not needed**.

Data scientists need not worry about the **theoretical nature** of the **left-hand side** and instead should focus on the **sampling procedures** and the **data at hand**.

Sometimes **data** is generated from a **physical process** that can be **modelled** (simple case, **flipping a coin**, **binomial distribution**).

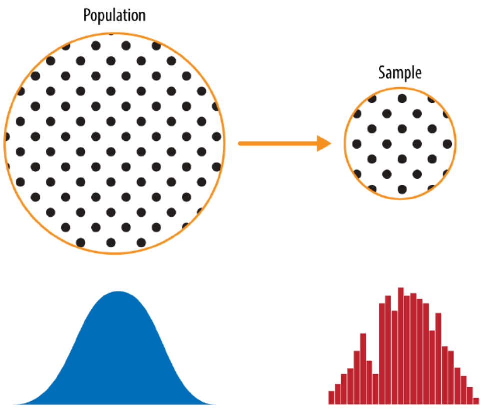


Figure 2-1. Population versus sample

I. Random Sampling and sample Bias

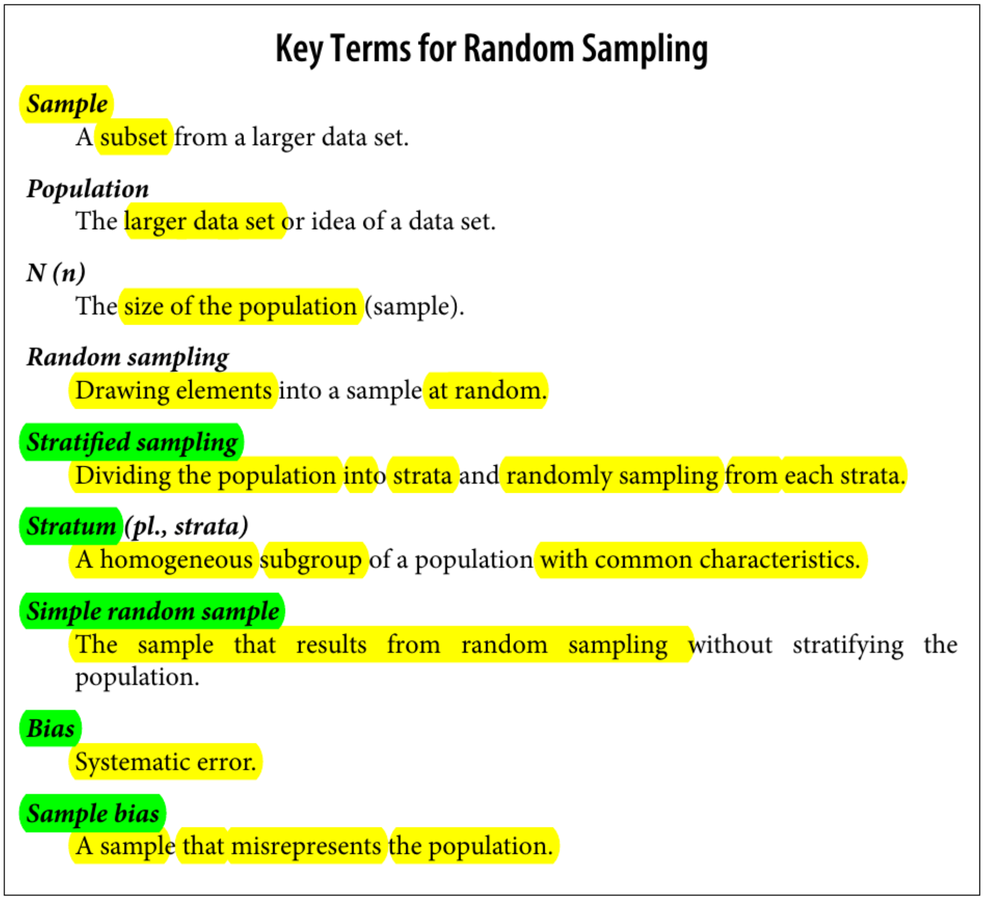
Population in **statistics** is a **large and defined** (sometimes **imaginary**) **set of data**.

Random sampling is a **process** in which each **available element** of the **population** being sampled has an **equal chance** of **being chosen** at each **draw**.
Result: **“simple random sample”.**

Sampling can be **with or without replacement** (مع ارجاع / بدون ارجاع).

Data quality is more important than **data quantity**; in **statistics**, that is called **representativeness**.

Data quality in **data science** involves: **completeness**, **consistency of format**, **cleanliness**, and **accuracy of individual 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.

***Stratum** (pl., strata)*
A homogeneous subgroup of a population with common characteristics.

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

Bias
Systematic error.

Sample bias
A sample that misrepresents the population.

A **classic example** is the **Literary Digest** magazine that **predicted** a victory of **Alf Landon** over **Franklin Roosevelt**. Out of a total of **over 10 million people**, it predicted a **landslide victory** for **Landon**.
On the **other hand**, **George Gallup** magazine conducted **biweekly polls** of just **2,000 people**, 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**. The **result** of this poll was **affected by sample bias**, which means the **sample** was **different** in some **meaningful and nonrandom way** from the **population**.
The term **nonrandom** is important: **hardly (rarely)** will any **sample**, including **random samples**, be **exactly representative** of the **population**.

Sample bias occurs when the **difference** is **meaningful**, and it can be expected to **continue** for other **samples** drawn in the **same way** as the first.

Self-selection sampling bias

Reviews of **restaurants**, **cafés**, and **hotels**, etc., on **social media** are **prone to bias**, because people **submitting** them are **not randomly selected**; rather, they have **taken the initiative** to write. This leads to **self-selection bias**. Because people who **write those comments** are **motivated** to write **reviews**, they may have a **poor experience** or may have an **association** with the **establishment**, etc.

While **self-selection samples** can be **unreliable indicators** of the **true state of affairs**, they can be **reliable** for **comparing** one **establishment** to a **similar one** (the same **self-selection bias** might apply to each).

I.1 Bias

Statistical bias refers to **sampling errors** that are **systematic** and produced by a **measurement** or **sampling process**.

An important **distinction** should be made between **errors due to randomness** and **errors due to bias**.

Example: **shooting a target with a gun** — bullets will not hit the **absolute center** every time, or even much at all.
An **unbiased process** will produce **errors** (**random errors**) and **does not tend strongly in any direction**.

Figure 2-2

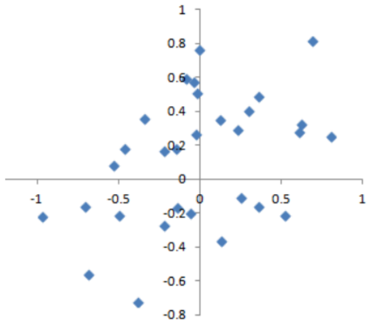


Figure 2-2. Scatterplot of shots from a gun with true aim

Instead, when we look at **Figure 2-3**, the figure shows **random error** in both **x** and **y directions**, but there is also a **bias**. **Shots tend to fall** in the **upper right quadrant**.

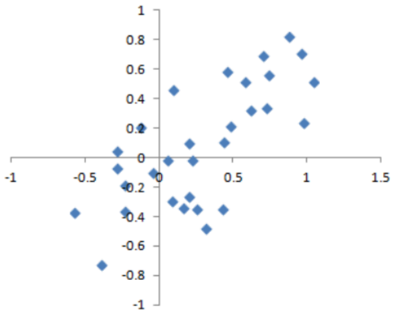


Figure 2-3. Scatterplot of shots from a gun with biased aim

Important note:
Bias comes in **different forms** and may be **observable** or **invisible**.
When a **result** does suggest **bias** (e.g., 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** has been **left out**.

I.2 Random Sampling

To avoid the **problem of sample bias** that led the **Literary Digest** to predict **Landon** over **Roosevelt**, **George Gallup** opted for more **scientifically chosen methods** to achieve a **sample** that was **representative** of the **voting electorate**.

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

To conduct a **representative customer survey**, it is first **essential** to clearly **define** who qualifies as a **customer** (e.g., **purchases greater than zero**, **inclusion or exclusion of past customers**, **refunds**, **test purchases**, **resellers**, or **billing agents**).
Next, a clear **sampling procedure** must be **established**, such as **selecting customers at random**, while **accounting for timing effects** when sampling from **continuous flows** like **online visitors** or **real-time transactions**.

Stratified sampling can improve **representativeness** by dividing the **population** into **meaningful subgroups (strata)** and drawing **random samples** from each.
For example, in **political polling**, voters may be divided into **whites**, **blacks**, and **Hispanics**; because a **simple random sample** would include **too few blacks and Hispanics**, these groups are intentionally **overweighted** in **stratified sampling** to obtain **balanced** and **comparable sample sizes** across all groups.

I.3 Size vs Quality: when does size matter?

In the **era of big data**, it is sometimes **shocking** that **smaller is better**, and spending more **time** and **effort** on **random samples** not only **reduces bias** but also allows greater **attention** to **data exploration** and **data quality**.

For example, **missing data** and **outliers** may contain **useful information.** When we have **millions of data points**, it becomes **prohibitively expensive** to track down **missing values** or evaluate **outliers**. Instead, with a **sample size of 10,000**, for example, this may be **feasible**.
Also, **data plotting** and **manual inspection** bog down if there is **too much data**.

I.3.1 So when are massive amounts of data needed?

A **classic scenario** for the value of **big data** is when the data is not only **big** but also **sparse**.

For example, in **Google search**, when a user submits a **query**, Google receives that query and represents it in a **matrix** where **columns** are **terms** and **rows** are the **search queries**. **Cell values** are **0**

or 1, depending on whether the query has a **specific term** or not.
Knowing that **English** has more than **150,000 words**, Google processes over **one trillion queries per year**. This yields a **huge matrix**, the **vast majority** of whose entries are **“0.”**

The **larger the data**, the **easier** it is to search for **anything**—even **very rare items**—and get **results**.

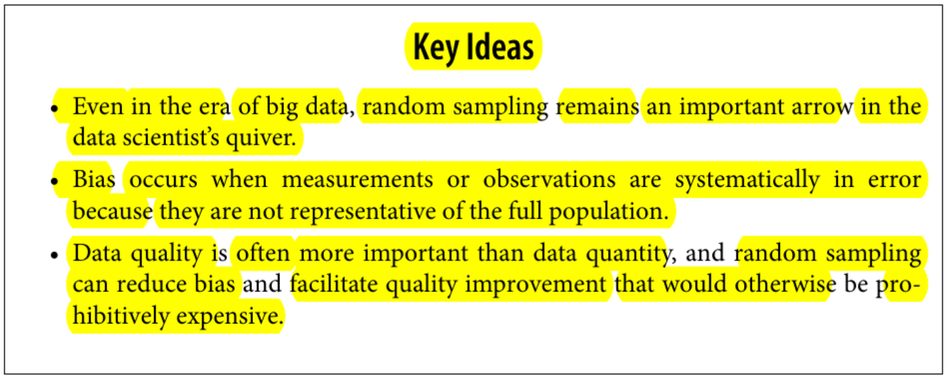
وذات الرداء الأحمر". في الأيام الأولى للإنترنت، من المحتمل أن تعطي هذه الاستعلامات **"Ricky Ricardo** ضع في اعتبارك عبارة البحث الذي ظهر فيه هذا الشخصية، و**قصة الأطفال ذات "I Love Lucy"** و**برنامج التلفزيون**، **Ricky Ricardo** نتائج عن **قائد الفرقة الموسيقية** **الرداء الأحمر**. كلا العنصرين الفرديين كان لهما العديد من عمليات البحث التي يمكن الرجوع إليها، لكن **الجمع بينهما كان نادرًا جدًا**

التي يقوم فيها **"I Love Lucy"** في وقت لاحق، بعد تراكم **تريليونات استعلامات البحث**، أصبح هذا الاستعلام يعطي **الحلقة الدقيقة من برنامج** بسرد قصة ذات الرداء الأحمر لابنه الرضيع بطريقة كوميدية تمزج بين الإنجليزية والإسبانية Ricky.

Keep in mind the **number of pertinent records** (ones in which the **exact search query** or something **similar** appears), plus a **link** that **most people** have **clicked on** afterward, might need only **thousands of clicks** to be **effective**. However, **trillions of data points** are needed to obtain these **pertinent records**, and **random sampling** will **not help**.

I.4 Sample mean vs Population mean

Symbol x̄ = **mean of sample** (**observed mean**)
Symbol μ = **mean of population** (**inferred mean**)

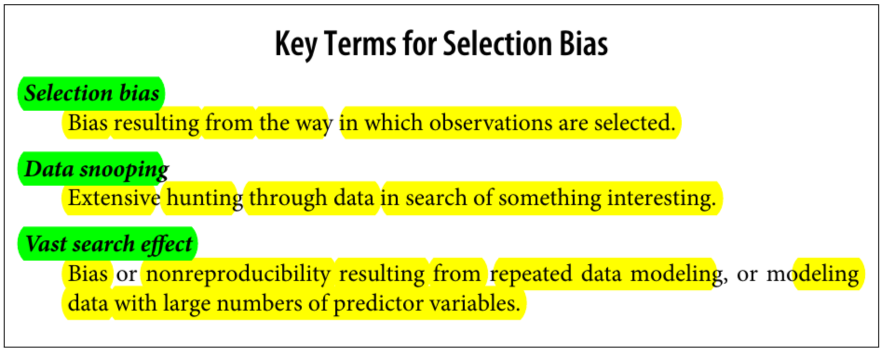


Key Ideas

- 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 otherwise be prohibitively expensive.

II. Selection Bias

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



Key Terms for Selection Bias

Selection bias
Bias resulting from the way in which observations are selected.

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.

If you **specify a hypothesis** and conduct a **well-designed experiment**, you can have **high confidence** in the **conclusion** (this **frequently does not occur**).

Often, one looks at **available data** and tries to **discern patterns**, but in this case, is it a **real pattern**, or just a **product of data snooping**? There is a saying among **statisticians**:
"If you torture the data long enough, sooner or later it will confess."

The **difference** between a **phenomenon** that you **verify** when **testing a hypothesis** using an **experiment** and a **phenomenon** that you **discover** by examining **available data** can be clarified with the following **thought experiment**:

Imagine you **challenge someone** to **flip a coin 10 times** and **every time get heads**, and he **does it**. In this case, you **clearly ascribe some special talent** to this person.
Now imagine **20,000 people** in a **stadium**, and an **usher** demands that they **flip a coin 10 times** and **report** if they get **heads 10 times in a row**. Here, the **probability** is **99%** that **someone** out of **20,000** gets **10 in a row**. In this case, it **does not indicate** that those persons who had **10 heads in a row** have **special talents**.

In **data science**, we often **analyze** the same **large dataset** many times, trying **different models** and asking **many different questions**. This is **powerful**—but also **dangerous**.
Because the **dataset** is so **large**, you are almost **guaranteed** to find something that **looks interesting**, even if there is **no real underlying pattern**. This problem is called the **"vast search effect"**, by **John Elder**.

The **key warning** is:
The **result** you discover may **not** be a **real, meaningful phenomenon**.
It could simply be a **chance outlier** that appears only because you **searched so much**.

There are **guards against this**, such as using a **holdout set**, and sometimes **many holdout sets**, for the **purpose** of **validating performance**.
Elder also advocates using what he calls **target shuffling** (it is a **permutation test** in its **essence**) to **test the validity** of **predictive associations** that a **data mining model** suggests.

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

II.1 Regression to the mean (الرجوع للاصل فضيلة)

Regression to the mean is a **phenomenon** where **extreme measurements** tend to be followed by more **average ones**. In other words, if someone or something performs **exceptionally high or low** on one **measurement**, the **next measurement** is likely to be **closer to the average**.

Example:
Imagine a **student** gets **100/100** on a test (**extremely high**). On the **next test**, they might score **90/100** (**closer to the average**).

Or, if someone **runs a race really slowly** one day, the **next time** they might run **faster**, closer to their **usual average speed**.

It's **not** because they did **better or worse** necessarily—it's just a **statistical tendency**: **extreme results** are often followed by more **normal results**.

Connection to Selection Bias

- When we **focus on extreme values**, like the **best rookie of the year**, we may **misinterpret** the **extreme performance** as being solely due to **skill**, ignoring the role of **luck** or **chance**. This is a type of **selection bias**: we **picked** someone because they were **extreme** in one **measurement**.
- The **"rookie of the year, sophomore slump"** is a **classic illustration**:
  A **rookie** has an **outstanding first season** → **skill + good luck**.
  In the **second season** → **skill remains**, but **luck might not** → **performance drops** → **regresses toward the average**.
- **Regression to the mean** was first **identified** by **Francis Galton** in **1886** in the context of **genetics**.
- **Example: extremely tall fathers** tend to have **children** who are **less extremely tall**, closer to the **population average**.

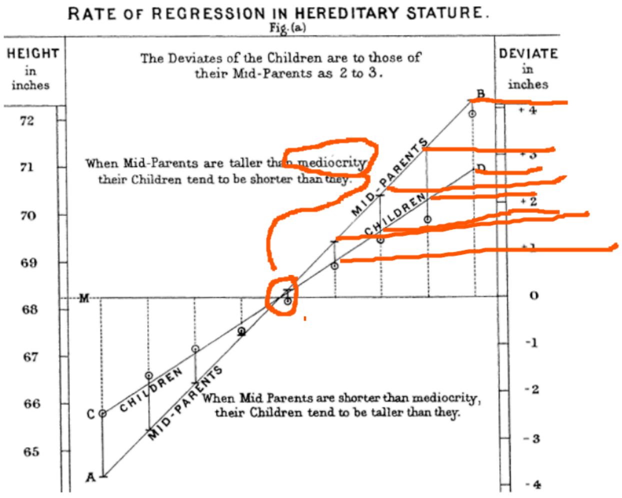
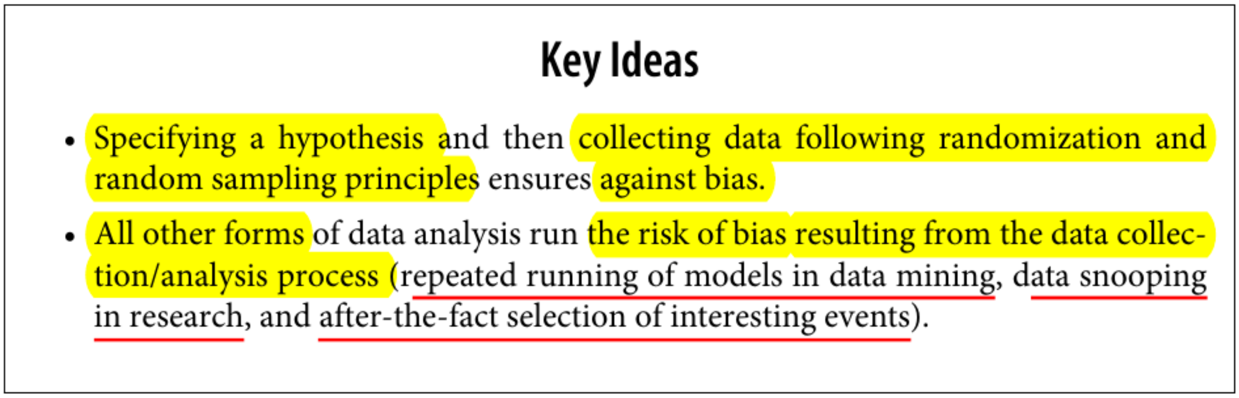


Figure 2-5. Galton's study that identified the phenomenon of regression to the mean

Important Clarification

Regression to the mean is **not** the same as **linear regression** in **statistics**.

- **Regression to the mean** is a **natural tendency** of **extreme values** to **move toward the average** in **successive measurements**.
- **Linear regression** is a **modeling technique** to **estimate relationships** between **predictor variables** and an **outcome variable**.



Key Ideas

- Specifying a hypothesis and 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).

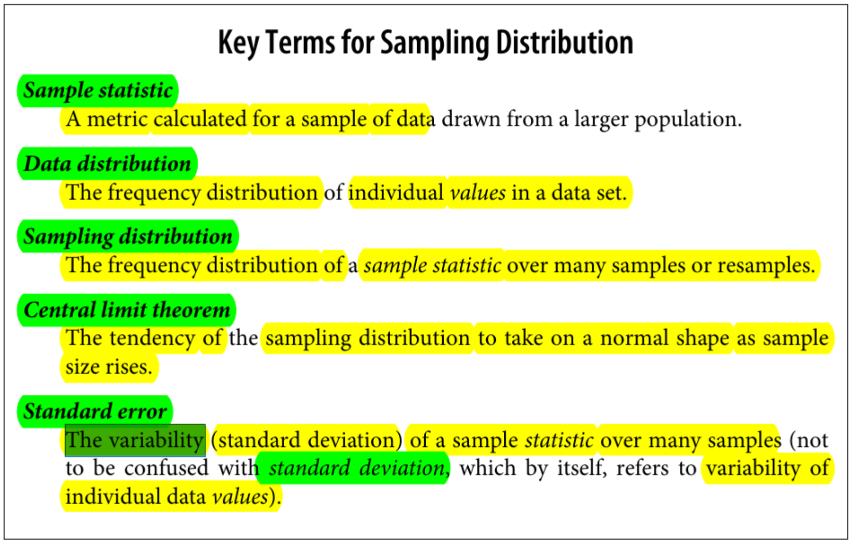
II.2 Sampling Distribution of a Statistics:

Sample distribution refers to the **study of the distribution** of some **sample statistic** over **many samples** drawn from the **same population**.

Example:

- **Population:** heights of **all students** in a school
- **Statistic: mean height**

Take **many samples** of **30 students each** → compute the **mean** for each sample → plot **all the means** → this **plot** is the **sampling distribution of the mean**.



Key Terms for Sampling Distribution

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.

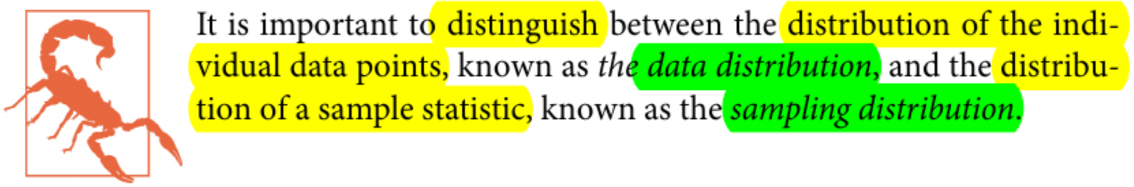
Central limit theorem
The tendency of the sampling distribution to take on a normal shape as sample size rises.

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*).

When we take a **sample**, we usually want to **measure something** (like the **mean**) or **build a model** (**statistical** or **machine learning**). However, our **estimate** or **model** is based on that **one sample**, so it might **not be exactly correct**. If we took a **different sample**, the **result** could be **different**. This **difference** is called **sampling variability** — it's a **key concern** in **statistics**.

If we had **lots of data**, we could take **many samples** and calculate the **statistic** for each. Then we could directly see the **distribution** of that **statistic** — this is the **sampling distribution**.

But there is a **practical limitation**:
In reality, we usually only have **one sample**, so we **can't easily draw additional samples** from the **population**. This is why **statisticians** use **theory** and **formulas** to **estimate the sampling distribution** instead of literally taking **all possible samples**.



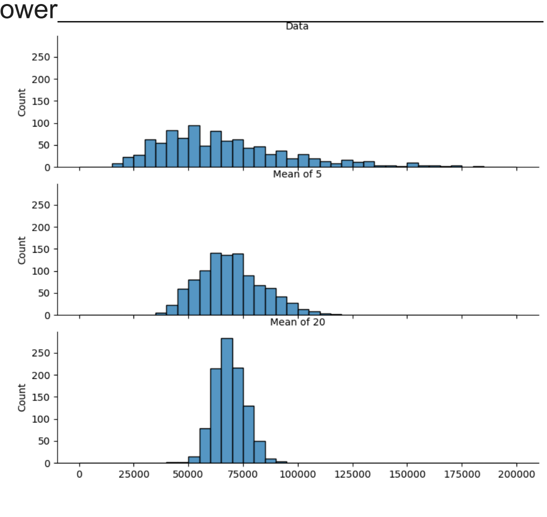
It is important to distinguish between the distribution of the individual data points, known as *the data distribution*, and the distribution of a sample statistic, known as the *sampling distribution*.

The **distribution of a sample** (like the **mean**) is usually **smoother and more bell-shaped** than the distribution of the individual data points. The **larger the sample** used to compute the statistic, the more it **approaches a normal (bell-shaped) distribution**. Also, **larger samples produce a narrower distribution** meaning the statistic varies less from sample to sample.

Example with LendingClub data

1. Suppose you have **annual incomes of loan applicants**.
2. The author suggests taking **three different types of samples**:
    1. **1,000 individual income values** → this shows the original data distribution
    2. **1,000 sample means of 5 incomes each** → this shows the distribution of the mean for small samples
    3. **1,000 sample means of 20 incomes each** → this shows the distribution of the mean for larger samples

3. Plotting histograms of these three samples shows:
    a. The **original data** might be irregular or skewed
    b. **Means of 5 values** are more bell-shaped
    c. **Means of 20 values** are even more bell-shaped and narrower



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**.

Explanation of Python code: `loans_income.sample(1000, random_state=42)`

`.sample(1000)` randomly selects **1000 rows** from the `loans_income` Series.By default, it **doesn't replace.**
`random_state` sets the seed for the random number generator.

Random number generators in Python are pseudo-random: they use a starting number (the seed) to produce a sequence of "random" numbers. fixing the same seed origin ensures getting the same random numbers every time.

Without `random_state`, every time you run `loans_income.sample(1000)`, you'd get a **different set of 1000 values**.

With `random_state=42`, **you always get the exact same 1000 rows**.

This is important for:

- **Reproducibility**: others can run your code and get the same results
- **Debugging**: you can be sure that any results you see are consistent

✅ **Example**

Suppose `loans_income = [50, 60, 70, 80, 90]`
`loans_income.sample(3, random_state=42)`
Might always return `[70, 50, 90]`.
If you ran it **without `random_state`**, it might return `[50, 80, 70]` next time — completely different.

how `random_state=42` actually produces the same values every time under the hood ?

Computers **cannot generate truly random numbers**; they use **algorithms** that produce sequences of numbers that **look random**. These sequences are called **pseudo-random numbers**. A **seed** is the starting point for that sequence.

Story of it : When you write `random_state=42`, Pandas passes **42 as the seed** to the random number generator. The generator uses a **deterministic algorithm**: given the seed 42, it produces the **same sequence of "random" numbers every time**.
These numbers are then used to **pick which rows to sample** from your dataset.

Same seed → same sequence → same sample

Different seed → different sequence → different sample

How a pseudo-random number generator (PRNG) works

- When you set a **seed** (e.g., `42`), the generator starts from an **initial state**.
- The PRNG uses a **deterministic algorithm** to produce a **sequence of numbers**, one after another.
- It does **not generate all numbers at once**; it produces each number **step by step**, based on the previous state.

□ **How it works in Pandas `.sample()`**

- You call `loans_income.sample(1000, random_state=42)`.
- Pandas asks the PRNG to **generate 1000 pseudo-random numbers** (or indices) **step by step**.
- These numbers are then used to **select rows from the dataset**.
- Because the seed is fixed, the sequence of indices is **always the same**, giving the same sample.

Like a chaine : seed → number1 → number2 → number3 → ... → number1000
Each number depends **mathematically** on the previous one(s) in the sequence.

```
sample_mean_05 = pd.DataFrame({ 'income': [loans_income.sample(5, random_state=i).mean() for i in range(1000)], 'type': 'Mean of 5',})
```

The goal is to build the sampling distribution of the mean for samples of size 5.

In statistics terms:

This line **simulates a sampling distribution of the mean**:

- You repeatedly take **samples of size 5**
- You compute the **mean** of each sample
- You do this **1000 times**
- You store the **1000 means** in a DataFrame

◆ **`loans_income.sample(5, random_state=i)`**

- Draws **5 incomes** from `loans_income`
- Uses `i` as the **seed**

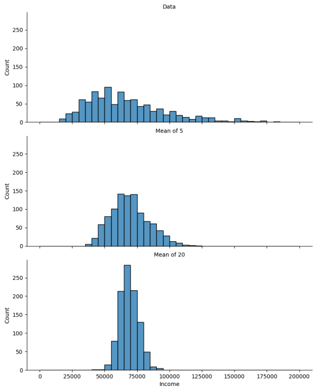
Important:

- When `i` changes → the sample changes
- When the code is rerun → the same `i` gives the same sample again

So:

- Iteration 0 → always the same 5 values
- Iteration 1 → always the same (but different) 5 values
- …
- Iteration 999 → always the same 5 values

II.3 Central Limit Theorem



The **phenomenon** described above is called the **Central Limit Theorem**.
 The **means drawn** from **multiple samples** will resemble the **bell-shaped normal curve**, even if the **population distribution** is **not normally distributed**, under the **conditions** that the **population** 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**); these **formulas** are used in **calculating sampling distributions** for **inference**.

What are "normal-approximation formulas"?

Many statistical tools are based on the **normal distribution**, such as:

- the **t-distribution**
- z-scores
- confidence intervals
- hypothesis tests

These tools **assume normality**. Without the CLT, we could **not** safely use them when the population distribution is unknown or non-normal.

The central limit theorem receives a lot of attention in traditional statistics texts because it underlies the machinery of hypothesis tests and confidence intervals, however, 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.

II.4 Standard Error

Standard error is a single metric that sums up the variability in samling distribution, measures **how much a sample statistic (usually the mean) varies from sample to sample**.

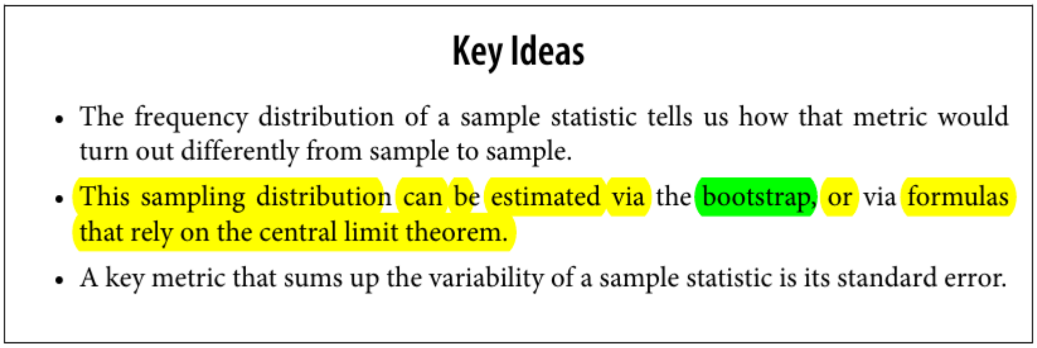
$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

What each symbol means:

- **SE** → Standard Error of the mean
- **s** → Sample standard deviation (variability inside one sample)
- **n** → Sample size
- √**n** → Square root of the sample size

Large s → data are very spread out → **SE is larger**
Large n → more data → **SE is smaller**

In practice, this approach of collecting new samples to estimate the standard error is typically not feasible (and statistically very wasteful). instead, you can use bootstrap resamples. 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.



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.

III.Bootstrap

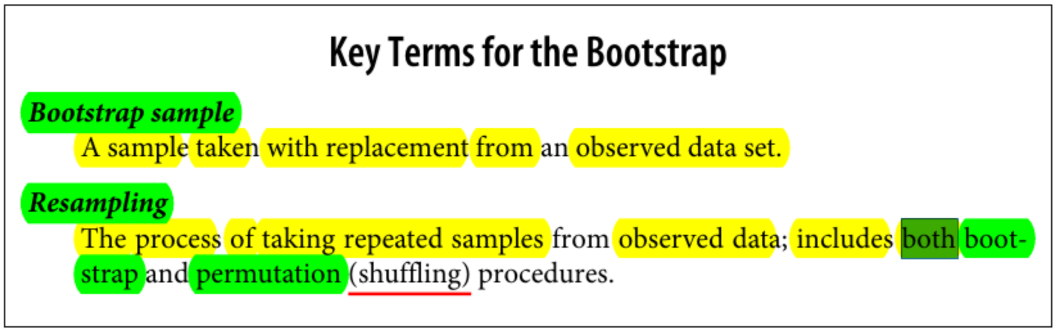
What problem are we trying to solve?
You have **one sample** from a population.
You compute a statistic from it (for example: **mean income**, **median**, **model coefficient**, etc.).
👉 But you know this statistic is **uncertain**, because:

- If you had drawn a **different sample**, the result would be slightly different.

So the real question is:

How variable is my statistic?
How much can it change if the sample were different?

The difficulty
In theory: To know the sampling distribution, you would need **many samples from the population**.
In reality: You usually have **only one sample**. You **cannot go back** and collect new data from the population.
The solution: Bootstrap Instead of resampling from the **population**, you resample from **your existing sample**.



Key Terms for the Bootstrap

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.

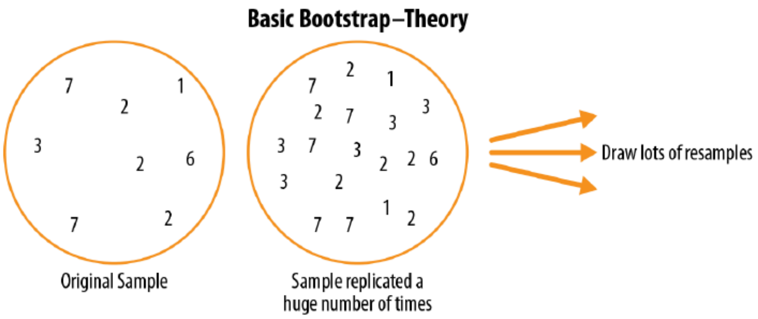
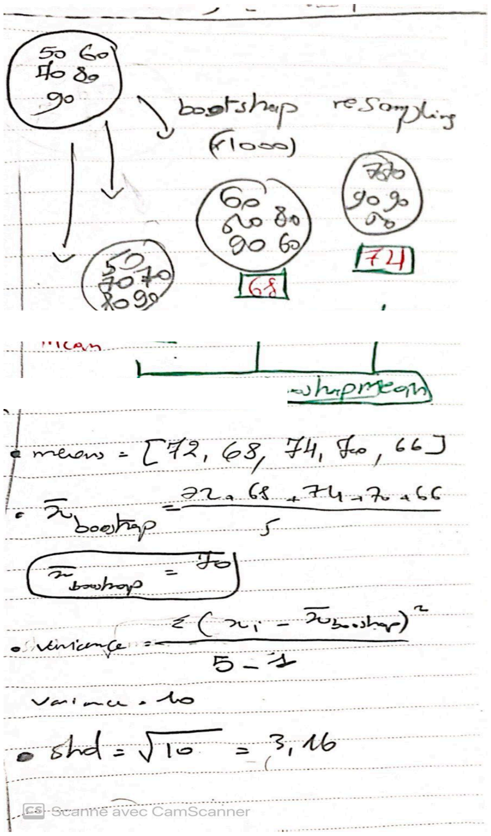


Figure 2-7. The idea of the bootstrap

You can then draw samples from this hypothetical population for the purpose of estimating a sampling distribution; see Figure 2-7.

- **Draw a sample with replacement** from your observed data (size = n).
- “With replacement” means after picking a value, you put it back so it **can appear again**.
- **Repeat n times** to create **one bootstrap sample** of size n.
- **Compute the mean** (or any statistic) of this bootstrap sample.
- **Repeat steps 1–3 R times** (e.g., 1000 or 10,000).
- You now have **R bootstrap estimates of the mean**.
- **Analyze the R bootstrap means**:
  a. Compute their **standard deviation** → estimates the **standard error** of the mean.
  b. Plot a **histogram or boxplot** → visualize the sampling distribution.
  c. Calculate **confidence intervals** (e.g., 2.5th to 97.5th percentile for a 95% CI).



- means = [72, 68, 74, 70, 66]
- $\bar{x}_{boostrap} = \frac{72 + 68 + 74 + 70 + 66}{5}$

$\boxed{\bar{x}_{bootstrap} = 70}$

- variance $= \frac{\Sigma(x_i - \bar{x}_{bootstrap})^2}{5-1}$

variance = 10

- std $= \sqrt{10} = 3,16$

The **standard deviation of the bootstrap means** (≈ 3.16) is an estimate of **how much the sample mean would vary if we repeatedly sampled from the population**.

SD = spread of **individual data points**
SE = spread of **sample means** (statistic)
Bootstrap SD of means ≈ SE of the mean

1. The Data

- **Bootstrap Sample Means ($n=5$):** [72, 68, 74, 70, 66]
- **Calculated Standard Error (SE):** 3.16
- **Average of Bootstrap Means:** 70

2. Approximate 95% Confidence Interval (CI)

To find the interval, we use the **Normal Approximation** formula. This is valid because the Central Limit Theorem tells us that our distribution of means will be approximately normal.

The Formula: Confidence Interval = Mean ± (1.96 × SE)

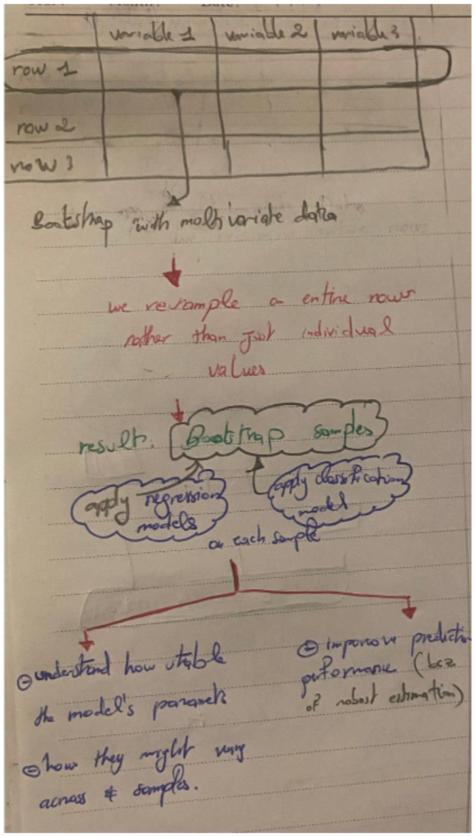
1. **Plug in the numbers:** 70 ± (1.96 × 3.16)
2. **Calculate the Margin of Error:** 1.96 × 3.16 = 6.19
3. **Find the Range:** 70 - 6.19 to 70 + 6.19

The Final Interval: CI = [63.81, 76.19]

3. Final Interpretation

We are **95% confident** that the true population mean (the actual average income of all applicants) lies between approximately **63.8** and **76.2**.

We are 95% confident that the true population mean lies between ≈ 63.8 and 76.2. This interval comes directly from the variability of the bootstrap sample means.

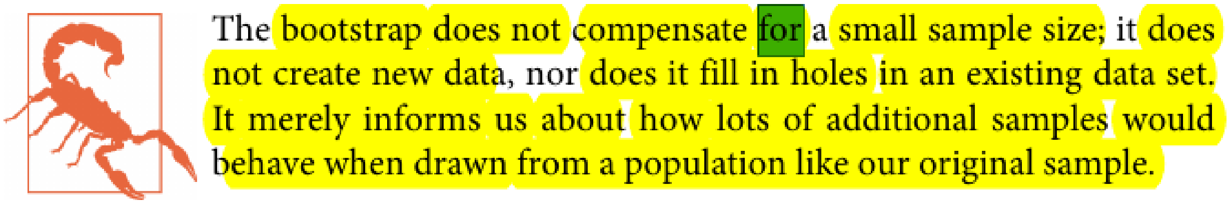

variable 1
variable 2
variable 3
row 1
row 2
row 3
Bootstrap with multivariate data
we resample a entire rows rather than just individual values
result: Bootstrap samples
apply regression models
apply classification model
on each sample
understand how stable the model's parameters
how they might vary across ≠ samples.
improve prediction performance (bcz of robust estimation)

Decision Trees and Bagging

- **Decision trees** are models that split the data into branches to make predictions.
- Instead of relying on just one decision tree, we can build **multiple trees** on different bootstrap samples.
- After building these multiple trees, we combine their predictions:
    - For **regression**, we take the **average** of all tree predictions.
    - For **classification**, we take a **majority vote** to decide the final class.

Bagging (Bootstrap Aggregating)

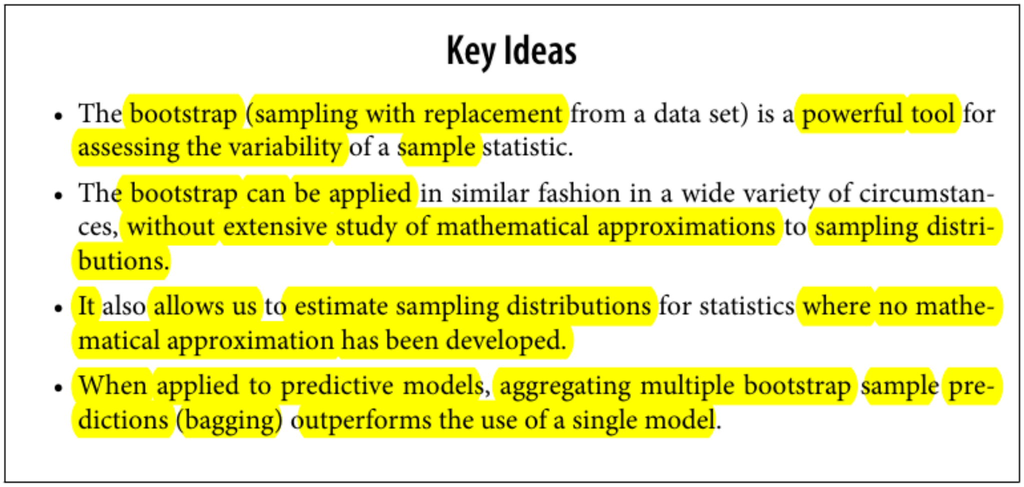
- This combined approach is known as **bagging**, which stands for **bootstrap aggregating**.
- The main idea is that by averaging multiple models (or taking majority votes), we **reduce variance** and **improve the overall accuracy** of the model.
- Bagging generally **outperforms a single decision tree** because it helps to **prevent overfitting** and makes the model more stable and reliable.



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.

III.1 Resampling Versus Bootstrapping

- **Resampling** = general family of methods
- **Bootstrap** = specific method → **always with replacement *(bootstrapping)***
- **Permutation** = different method → often **without replacement**



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.

IV. 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.

How to calculate CI:
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 [(100-x) / 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.

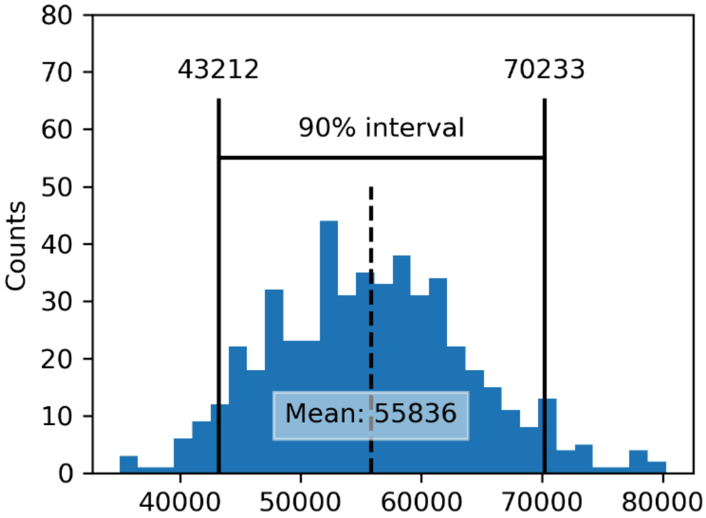


Figure 2-9. Bootstrap confidence interval for the annual income of loan applicants, based on a sample of 20

The bootstrap is a general tool that can be used to generate confidence intervals for most statistics, or model parameters.

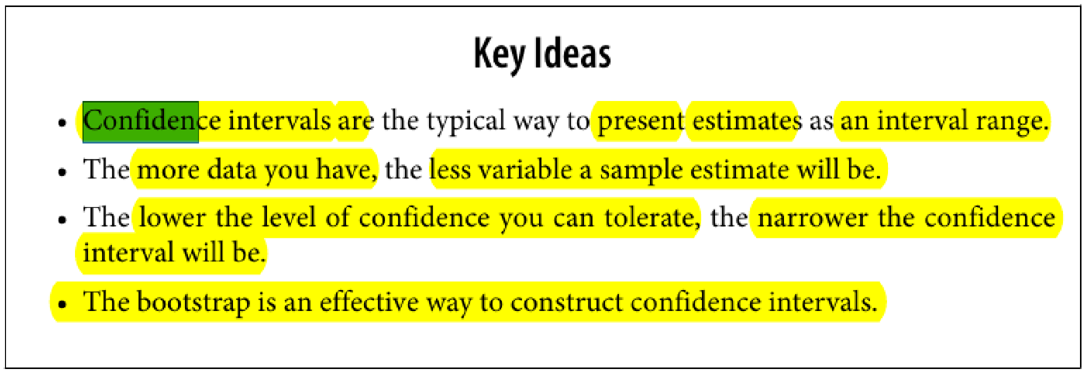
The probability question associated with a confidence interval starts out with the phrase “Given a sampling procedure and a population, what is the probability that...”

The percentage linked to a confidence interval is called the **confidence level**. A higher confidence level the **wider interval is**. Likewise, **smaller sample sizes lead to wider confidence intervals**, reflecting greater uncertainty.

In short, when you want more confidence or have less data, you must accept a wider interval to be reasonably sure it contains the true value.

Very important :

> For a data scientist, a confidence interval is mainly a way to understand how uncertain a sample result is. Rather than using it for formal reporting or regulation, it is used to communicate the possible error in an estimate and to **decide whether more data may be needed to improve reliability.**

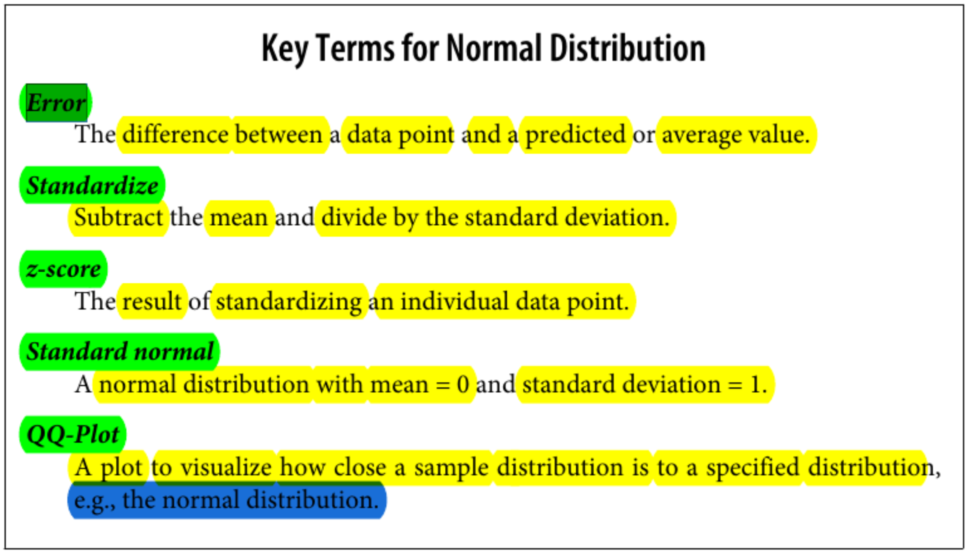


Key Ideas

- 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.

V. Normal Distribution

The bell-shaped normal distribution is a powerful tool in the development of mathematical formulas that approximate those distributions.



Key Terms for Normal Distribution

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 specified distribution, e.g., the normal distribution.

- About 68% of the data points are within ±1 standard deviation from the mean.
- About 95% of the data points are within ±2 standard deviations from the mean.

For example, if:

- Mean = 50
- SD = 5

Then ±1 SD from the mean = 50 − 5 to 50 + 5 = **45 to 55**.

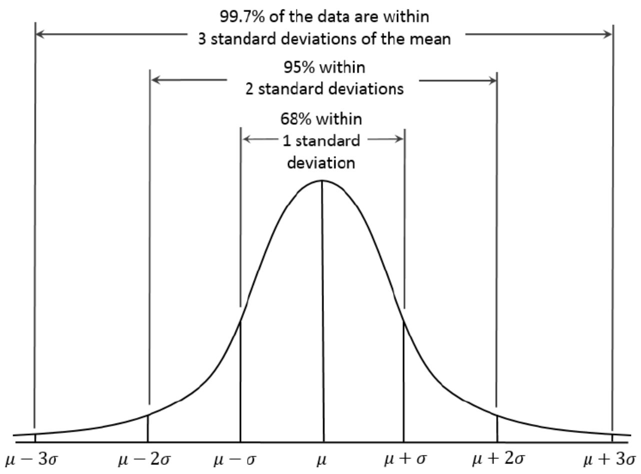


Figure 2-10. Normal curve

V.1. Standard Normal and QQ-Plots

A **standard normal distribution** expresses values in **units of standard deviations from the mean.** To convert a value to this scale, you **subtract the mean** and **divide by the standard deviation**; this process is called **standardization** or **normalization.**

- The resulting value is called a **z-score.**
- This allows comparison of data on the same scale, regardless of original units.
- The standard normal distribution is also called the **z-distribution.**

✅ Simple version: **z-score = how many standard deviations a value is from the mean.**

QQ-Plot is used to **visually determine** how **close a sample** is to a **specified distribution**—in this case, 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 (الدوائر)** correspond to the **number of standard deviations away from the mean**.

If the **points roughly fall** on the **diagonal line**, then the **sample distribution** can be considered **close to normal**.
Figure 2-11 shows a **QQ-Plot** for a **sample of 100 values** randomly **generated from a normal distribution**; as expected, the **points closely follow the line**.

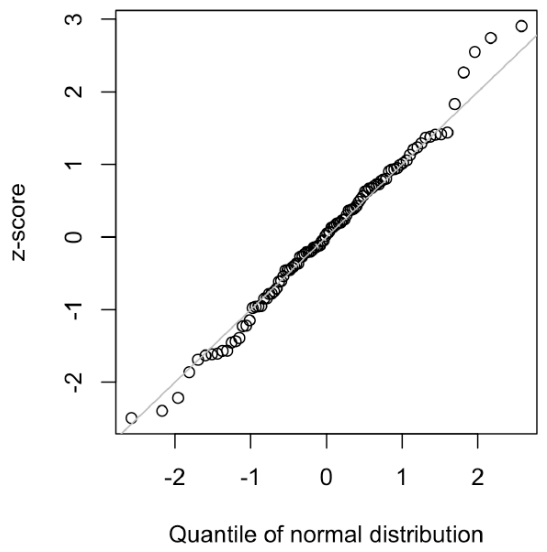


Figure 2-11. QQ-Plot of a sample of 100 values drawn from a standard normal distribution

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.

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.
- 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.

VI. Long-Tailed Distributions

Key Terms for Long-Tailed Distributions

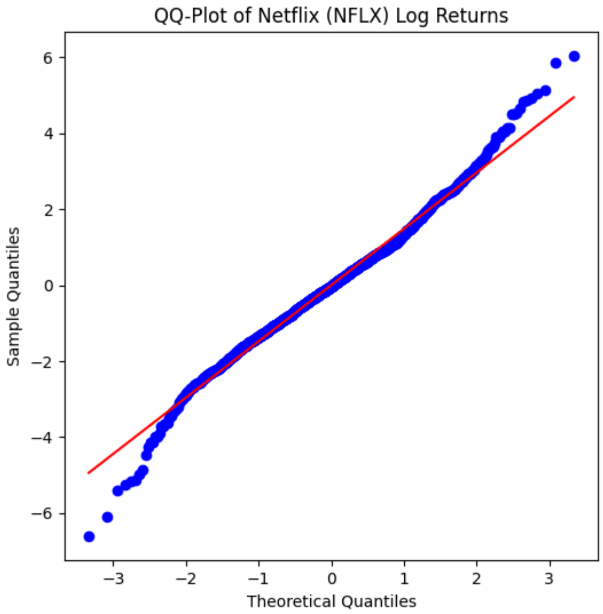
Tail

The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.

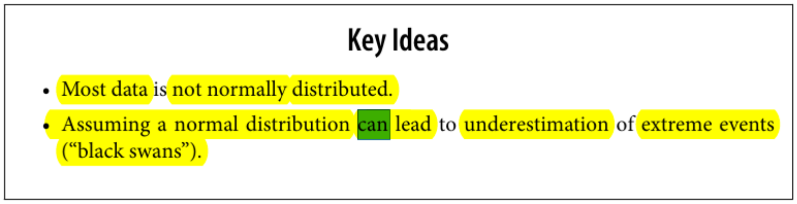
Skew

Where one tail of a distribution is longer than the other.

- **Normal distribution** is often useful for **errors and sample statistics**, but **raw data** often do not follow it.
- Data can be **skewed** (e.g., income) or **discrete** (e.g., binomial outcomes).
- **Tails** of a distribution represent **extreme values (large or small ones),** long tails are important to consider in practice.
- **Black Swan theory (Taleb)**: extreme, rare events (like stock market crashes) are **more likely than predicted by a normal distribution**.



- When points on a Q-Q plot **deviate below the line for low values and above for high values**, the data **have heavy tails** and are **not normally distributed**.
- This means **extreme values occur more often than a normal distribution predicts**.
- Sometimes, the data **look normal near the mean** but have **long tails**, a phenomenon Tukey calls **"normal in the middle"**.

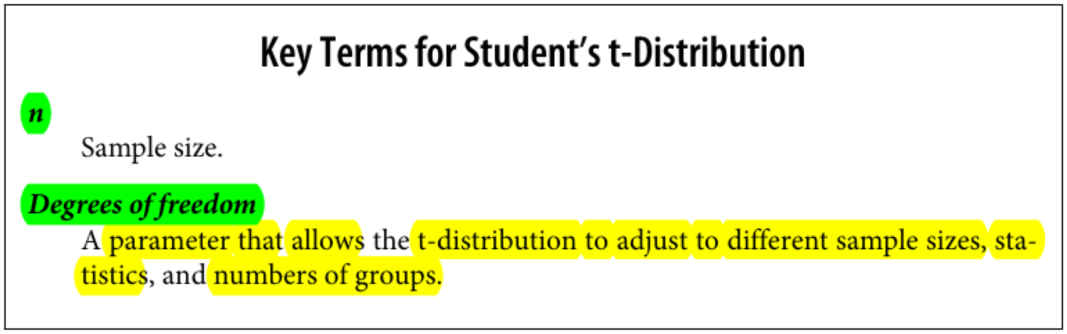


Key Ideas

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

VII. Student's t-Distribution

The **t-distribution** looks similar to the normal (bell-shaped) distribution but has **heavier and longer tails**, meaning it allows for more extreme values. It is mainly used to describe the distribution of **sample statistics**, especially **sample means**, when the sample size is small. There are different t-distributions depending on the **sample size**: with **small samples**, the tails are thicker, and as the **sample size increases**, the t-distribution becomes closer to the normal distribution.



Key Terms for Student's t-Distribution

n

Sample size.

Degrees of freedom

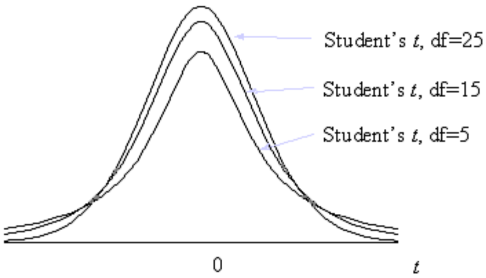
A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and numbers of groups.

This section explains **why and how the t-distribution is used to build confidence intervals** when working with sample statistics, especially the **sample mean**.
When we take a sample of size **n**, we compute:

- the sample mean **x̄**
- the sample standard deviation **s**

Because the sample mean varies from one sample to another, we need a way to account for this **sampling variability**. The **t-distribution** is used for this purpose.
A **90% confidence interval** for the sample mean is given by:

$$\bar{x} \pm t_{n-1}(0.05) \cdot \frac{s}{\sqrt{n}}$$



- tn−1, 0.05t_{n-1,\,0.05}tn−1,0.05 is a value from the t-distribution
- n−1n - 1n−1 is the **degrees of freedom**
- The value cuts off 5% of the distribution in each tail (total 10%)

The t-distribution is used not only for sample means, but also for:

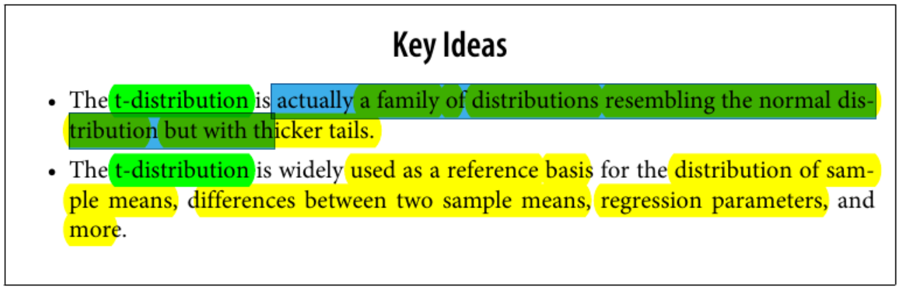
- differences between means
- regression coefficients
- other standardized statistics

Historically, statisticians relied on the t-distribution because **computers were not available**, so mathematical formulas were needed to approximate sampling distributions. When computers became widespread, **resampling methods** (like the bootstrap) became practical, but the t-distribution remained deeply embedded in statistics.

The t-distribution works well **when the sampling distribution of the statistic is approximately normal**. Thanks to the **central limit theorem**, many sample statistics (especially means) are close to normally distributed **even if the original data is not**. This is why the t-distribution has been so widely and successfully used.

Important Note: Data scientists do **not need deep theoretical knowledge** of the *t-distribution* or the *central limit theorem*. These concepts are central to **classical statistics**, but in practical data science, uncertainty and variability are often better handled using **empirical methods like the bootstrap**.

That said, data scientists **will frequently encounter t-statistics** in the output of statistical software (especially in **A/B testing, regressions, and hypothesis testing**). For this reason, it is useful to understand **what the t-distribution represents** and **why it appears**, even if it is not heavily used for decision-making.



Key Ideas

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

VIII. Binomial Distribution

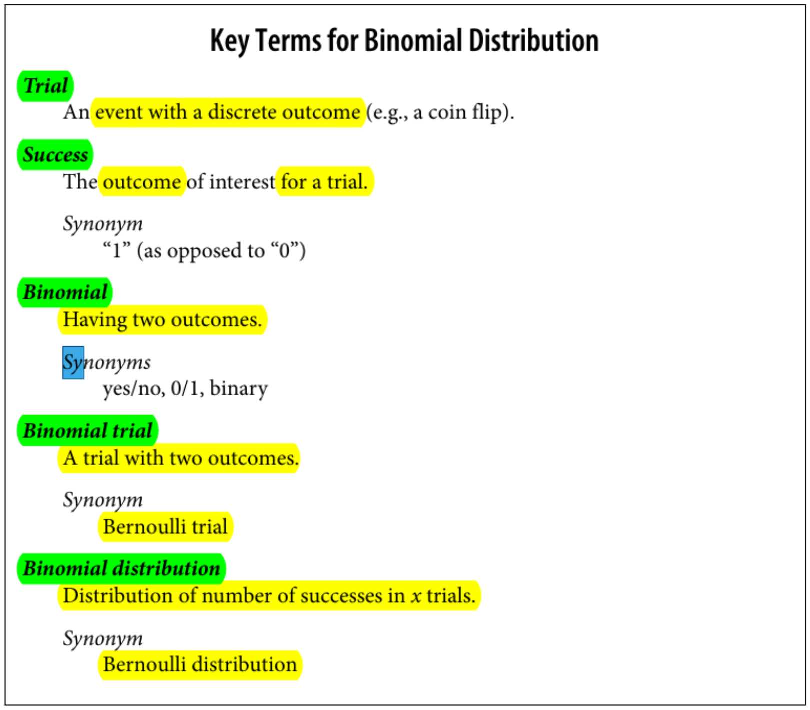
Many real-world data science problems end in a **yes/no decision**, such as *buy or not buy*, *click or not click*, or *default or not default*. These outcomes are called **binary outcomes** and are modeled using the **binomial distribution**.

A binomial situation involves:

- A fixed number of **trials**
- Each trial has **two possible outcomes** (0/1, yes/no)
- Each outcome has a **fixed probability**

For example, flipping a coin 10 times is a binomial experiment with 10 trials and two outcomes (heads or tails). The probabilities do not have to be 50/50.

In statistics, the outcome labeled **“1”** is called a **success**, but this does **not** mean something good. It simply refers to the outcome of interest, often the **rarer event**, such as loan default or fraud, which data scientists want to predict.



Key Terms for Binomial Distribution

Trial

An event with a discrete outcome (e.g., a coin flip).

Success

The outcome of interest for a trial.

Synonym

"1" (as opposed to "0")

Binomial

Having two outcomes.

Synonyms

yes/no, 0/1, binary

Binomial trial

A trial with two outcomes.

Synonym

Bernoulli trial

Binomial distribution

Distribution of number of successes in *x* trials.

Synonym

Bernoulli distribution

The **binomial distribution** describes how many times a specific outcome (called a **success**) occurs in a fixed number of independent trials, when each trial has the same probability of success.

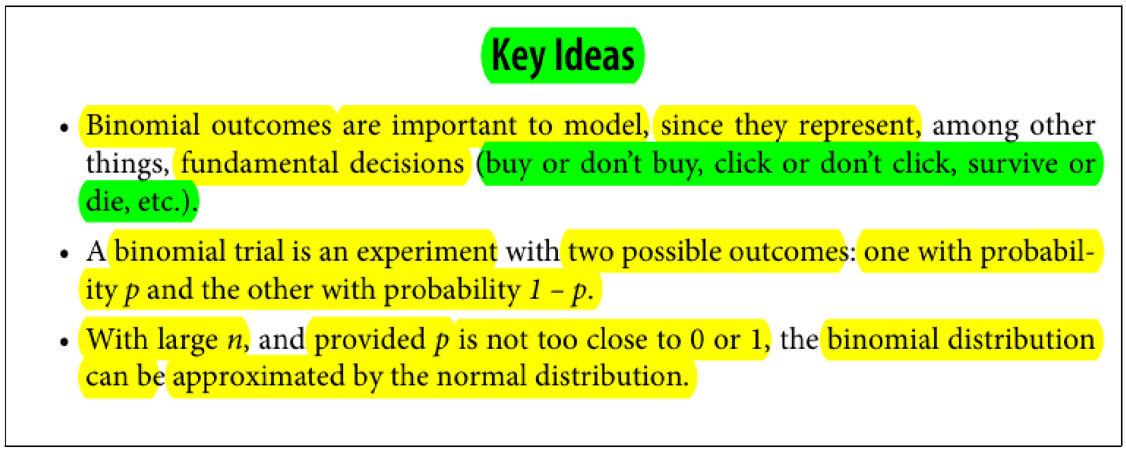
It is defined by three parameters:

- **n**: number of trials
- **p**: probability of success in each trial
- **x**: number of successes observed

For example, if the probability that a click leads to a sale is **p = 0.02**, and you observe **200 clicks**, the binomial distribution can tell you the probability of getting **0 sales**, **1 sale**, **2 sales**, etc.

- **Mean (expected value)** of a binomial distribution is
  n × p
  → It represents the **average number of successes** you expect in *n* trials when each trial has probability *p* of success.
- **Variance** is
  n × p × (1 − p)
  → It measures how much the number of successes varies around the mean.
- When the **number of trials n is large** (especially if *p* is near 0.5), the **binomial distribution looks very similar to a normal distribution**.
- Because calculating exact binomial probabilities for large *n* is **computationally expensive**, statisticians often **approximate the binomial distribution using a normal distribution** with:
  - Mean = n × p
  - Variance = n × p × (1 − p)

👉 This approximation makes calculations faster and easier, and it is widely used in practice.



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.

IX Chi-Square Distribution

In statistics, we often want to know whether the data behaves **as expected** or shows something unusual.
This idea of "expectation" means **what we would see if nothing special were happening**—for example, if two variables were completely independent. This assumption is called the **null hypothesis** (or null model).
To test departures from this expectation—especially with **category counts**—we use the **chi-square (χ^2) statistic**.
The chi-square statistic compares:

- **Observed counts** (what we actually see in the data)
- **Expected counts** (what we would expect if the null hypothesis were true)

It measures how far the observed values deviate from the expected ones, in a standardized way, and then sums these deviations across all categories.
In simple terms:

- If observed ≈ expected → data fits the expectation
- If observed ≠ expected → data departs from expectation

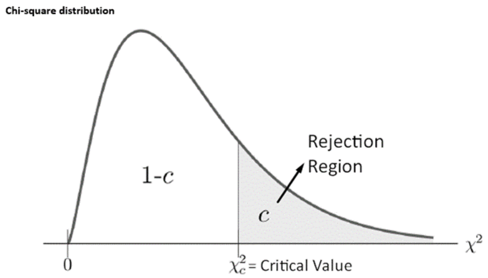
The chi-square test is commonly used to:

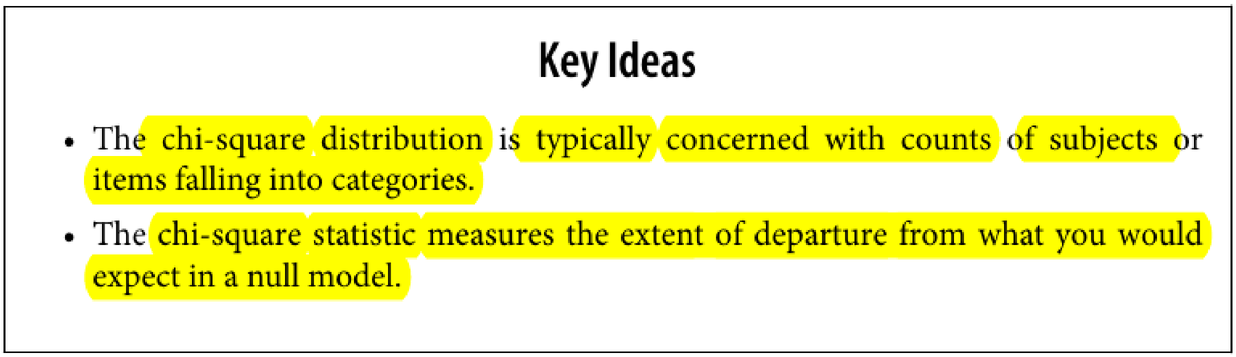
- Test **independence** between variables (e.g., gender vs promotion)
- Perform **goodness-of-fit tests**
- Compare multiple groups (A/B/C tests)

The **chi-square distribution** describes the behavior of the chi-square statistic under the **null hypothesis** (expected distribution).

Low chi-square values → observed counts are close to expected counts.
High chi-square values → observed counts deviate strongly from expectations.

Different **degrees of freedom** change the shape of the chi-square distribution.





Key Ideas

- The chi-square distribution is typically concerned with counts of subjects or items falling into categories.
- The chi-square statistic measures the extent of departure from what you would expect in a null model.

X. F-distribution

- **F-statistic** compares **variability between group means** to **variability within groups**.
- Used in **ANOVA** to test if group means differ more than expected by chance.
- The **F-distribution** shows the range of F values under the **null hypothesis** (all group means equal).
- **Different degrees of freedom** (number of groups, sample sizes) change the shape of the F-distribution.
- Also used in **linear regression** to assess how much variation the model explains.

✅ In short: F-statistic = signal (between-group variation) ÷ noise (within-group variation).

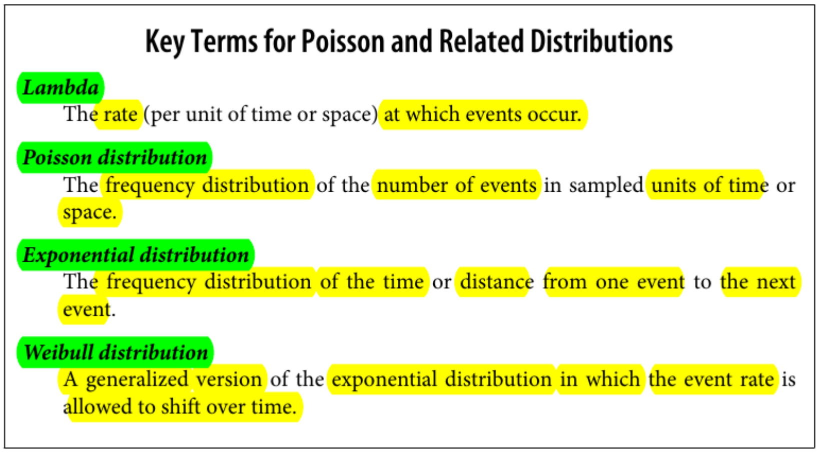
F-statistic measures **how much of the total variation in your data can be explained by the factor(s) you're testing** (e.g., different treatments or groups) compared to the variation that is just random or unexplained within the groups.

- **Numerator (variation due to factors)** → differences between group means.
- **Denominator (overall or residual variation)** → differences within each group.

✅ A **high F value** indicates that the factor explains a large part of the total variation, suggesting the group means are significantly different.
A **low F value** indicates most variation is just random, so the factor has little effect.

XI Poisson and Related Distributions

Many processes produce events randomly at a given overall rate—visitors arriving at a website, or 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).



Key Terms for Poisson and Related Distributions

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 distribution in which the event rate is allowed to shift over time.

XIII. Poisson Distributions

The Poisson distribution models the number of events occurring in a fixed unit of time or space. Its key parameter, **λ (lambda)**, represents both the **average number of events** and the **variance**. It's useful for estimating variability in counts, like daily flu cases or server requests, and for planning capacity to handle expected website traffic.

Python output:
```
[1 1 3 1 1 3 0 0 3 0 0 1 3 1 2 0 1 1 3 1 1 1 3 1 1 1 4 1 2 2 1 3 3 1 4 3 1
 1 1 2 2 1 1 2 1 2 1 6 2 5 4 3 3 2 2 1 1 5 2 1 2 2 2 2 0 4 4 3 3 3 2 2 2 3
 2 2 1 2 3 4 1 3 1 2 3 3 3 4 2 1 2 0 2 1 1 1 3 3 3 4]
```

XIV. Exponential Distribution

The exponential distribution models the **time between events** when events occur randomly at a constant average rate (λ).

- **Applications:** time between website visits, cars at a toll, time to equipment failure, or time per service call.
- **Parameters:**
  - `n`: number of random values to generate
  - `rate` (λ): average number of events per time period

It is essentially the **continuous counterpart of the Poisson distribution**: Poisson counts events per interval, exponentially measures the time between those events.

Python output:
```
[0.03492931 0.02087872 0.13192011 0.04412971 0.21908228 0.074703
 0.22074579 0.45897324 0.21534881 0.21756887 0.20268386 0.10679318
 1.07978488 1.62101198 0.00805058 0.07386227 0.51318852 0.77611487
 1.2306378  0.10324942 0.74853594 0.02961869 0.25864479 0.16524031
 0.28470914 0.2032431  0.99026981 0.33849935 0.74170067 0.0888013
 0.82564269 0.70583424 0.65391513 0.42302451 0.04448321 0.6230653
 0.06678464 0.12225972 0.48462161 0.47846282 1.06449161 0.50799646
 0.21066642 0.01770627 0.14694236 1.06999071 0.38838534 0.09376164
 0.23573631 2.02098016 1.20987838 0.49324833 0.02295803 0.61774713
 1.21935158 0.04994408 0.10451704 0.31311311 0.00955289 0.28564501
 2.33189753 0.91265304 0.60055928 1.21790812 0.40675911 0.82307133
 0.36675318 0.02811946 0.43429418 0.0808463  0.13567416 0.24260777
 0.39452232 0.67076033 1.92676298 2.00190653 0.82565473 0.72526401
 0.66023056 1.51933876 0.52601528 0.0984964  0.10953658 0.07590569
 1.61666299 0.07995985 0.09251578 0.0730027  0.40518364 0.39834693
 0.12531253 0.06754758 0.48453396 1.77083645 0.30062365 1.3533515
 0.89070924 0.18735084 0.48330509 0.85020263]
```

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

XV. Estimating the Failure Rate

Unknown event rate for rare events: Suppose you are monitoring **aircraft engine failures**, and a new engine model has flown **20 hours without failure**. You cannot assume the failure rate is 1 per hour—it's much lower. You can **simulate different rates** (e.g., 0.01 per hour, 0.05 per hour) to see which rates are plausible given no failures observed.

Limited data scenario: Imagine you have data on **server crashes**, but only 3 crashes in 1,000 hours. You can use a **chi-square goodness-of-fit test** to check which Poisson rates (e.g., λ = 0.002, λ = 0.005 per hour) are consistent with the observed crash counts.

Practical takeaway: For rare events, you **cannot rely solely on observed frequencies**. Simulation and statistical testing help **estimate and validate reasonable rates**.

XVI Weibull Distribution

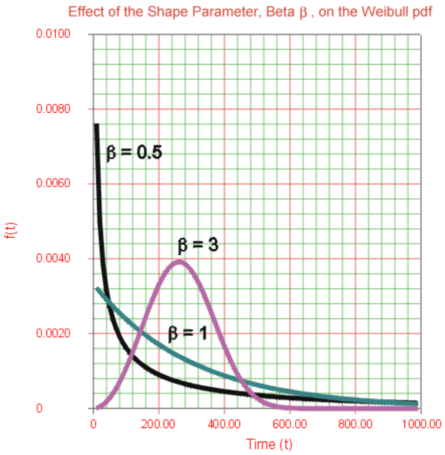
- **Problem:** Poisson and exponential distributions assume a **constant event rate (λ)**. But in reality, rates often change over time—for example, **mechanical parts wear out**, so failure becomes more likely as time passes.

- **Solution:** The **Weibull distribution** extends the exponential by allowing a changing event rate, controlled by a **shape parameter β**:

  - **β > 1:** event probability increases over time (e.g., engine parts wearing out).
  - **β < 1:** event probability decreases over time (e.g., early "infant mortality" failures in electronics).
- **Scale parameter η (eta):** Represents the **characteristic life**, or the time by which a certain proportion of items are expected to fail.

Example 1: A factory monitors **bearings**: early on, failures are rare, but as bearings age, failures increase → model with Weibull, β > 1, η = 5 years.
Example 2: Electronic devices often fail early due to defects, then stabilize → Weibull with β < 1.

This makes Weibull suitable for **time-to-failure analysis** where the event rate is not constant.

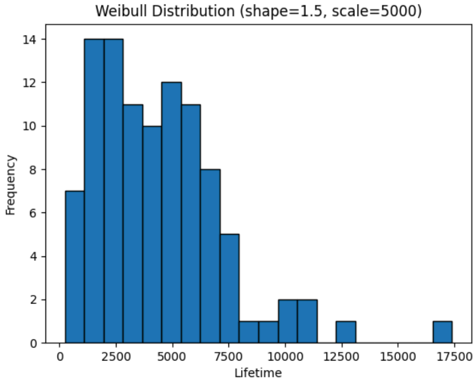


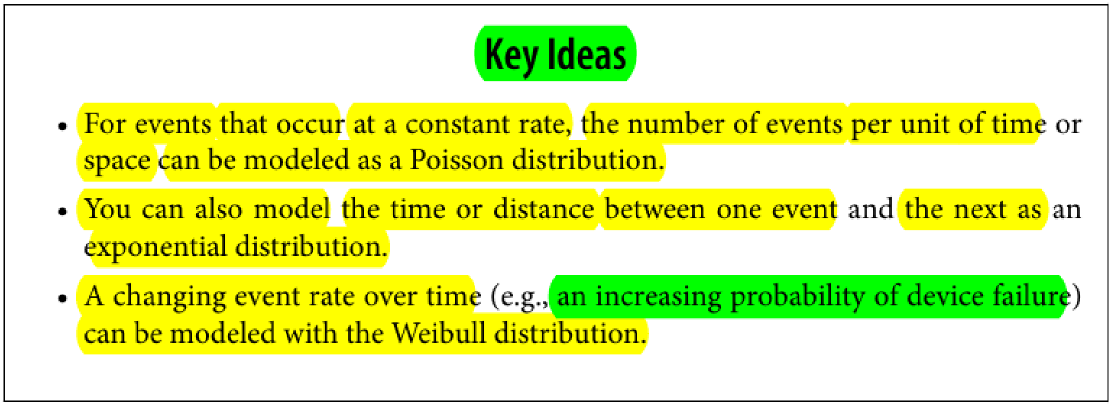
Output code python 10:

```
[6148.50 1009.55  4380.92 4217.80  4705.08
 2525.85 2788.01  921.78 813.86 7840.07]
```

- `1.5` → the **shape parameter β**, which controls how the failure rate changes over time (β > 1 → increasing failure probability).
- `scale=5000` → the **scale parameter η**, representing the characteristic life (typical lifetime of items).
- `size=100` → generate 100 random lifetimes from this Weibull distribution.

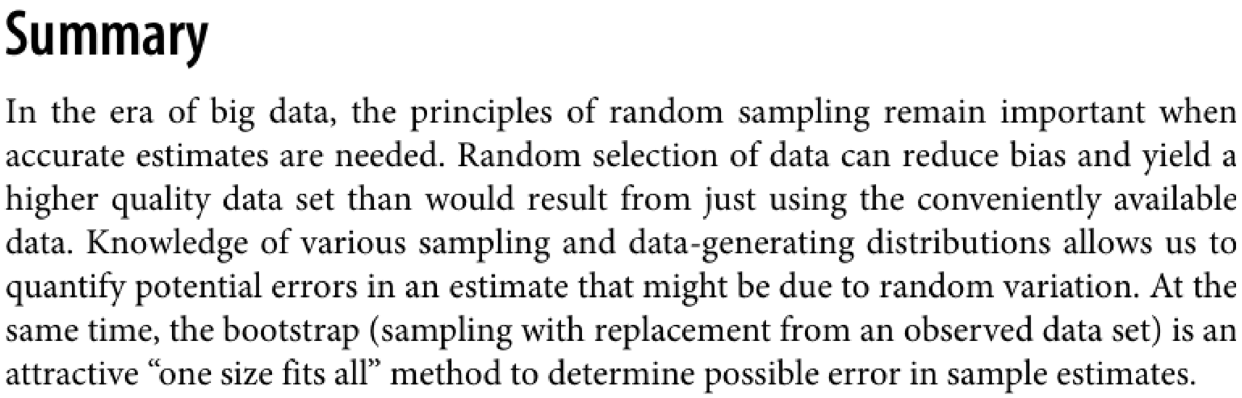
Meaning: This simulates 100 items' lifetimes assuming failures increase over time, with a typical life around 5,000 units (hours, cycles, etc.).





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.
- 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.



Summary

In the era of big data, the principles of random sampling remain important when accurate estimates are needed. Random selection of data can reduce bias and yield a higher quality data set than would result from just using the conveniently available data. Knowledge of various sampling and data-generating distributions allows us to quantify potential errors in an estimate that might be due to random variation. At the same time, the bootstrap (sampling with replacement from an observed data set) is an attractive "one size fits all" method to determine possible error in sample estimates.