Unit #4: The Normal Distribution and the Central Limit Theorem

4.3, 6.4, 6.5

At the end of this unit, students should be able to:

1. Define the normal distribution, the standard normal distribution, and state the parameters that characterize the these distributions.

2. Identify how the normal distribution changes as a function of the parameters.

3. Describe situations where the normal distribution would be a good model.

4. Calculate and simulate probabilities involving the normal distribution using a table and/or R.

5. Define the critical value for a normal distribution

6. Convert a non-standard normal distribution to a standard normal distribution.

At the end of this unit, students should be able to:

1. Describe and provide examples of statistical inference.
2. Define independent and identically distributed and a simple random sample.
3. Define and provide examples of an *estimator.*
4. Describe why estimators (e.g., the sample mean) has a probability distribution (that is, describe why estimators are random variables).
5. Define the sampling distribution of an estimator.
6. Describe the three features upon which the sampling distribution of an estimator depends.
7. Define the standard error of an estimator.
8. Write R code that illustrates the fact that an estimator (e.g., the mean) has a sampling distribution.
9. Describe the mean and variance of the sample mean of a sample.
10. State the Central Limit Theorem (CLT), which characterizes the distribution of the sample mean.
11. Apply the central limit theorem to answer questions about the mean of a simple random sample.

Photo by Toby Elliott on Unsplash

The normal distribution (sometimes called the Gaussian distribution) is probably the most important distribution in all of probability and statistics.

Many populations have distributions that can be fit very closely by an appropriate normal curve.

Examples: height, weight, and other physical characteristics, scores on various tests, etc.

$Y_1 \sim Bin(n, p)$

$Y_2 \sim Exp(\lambda)$

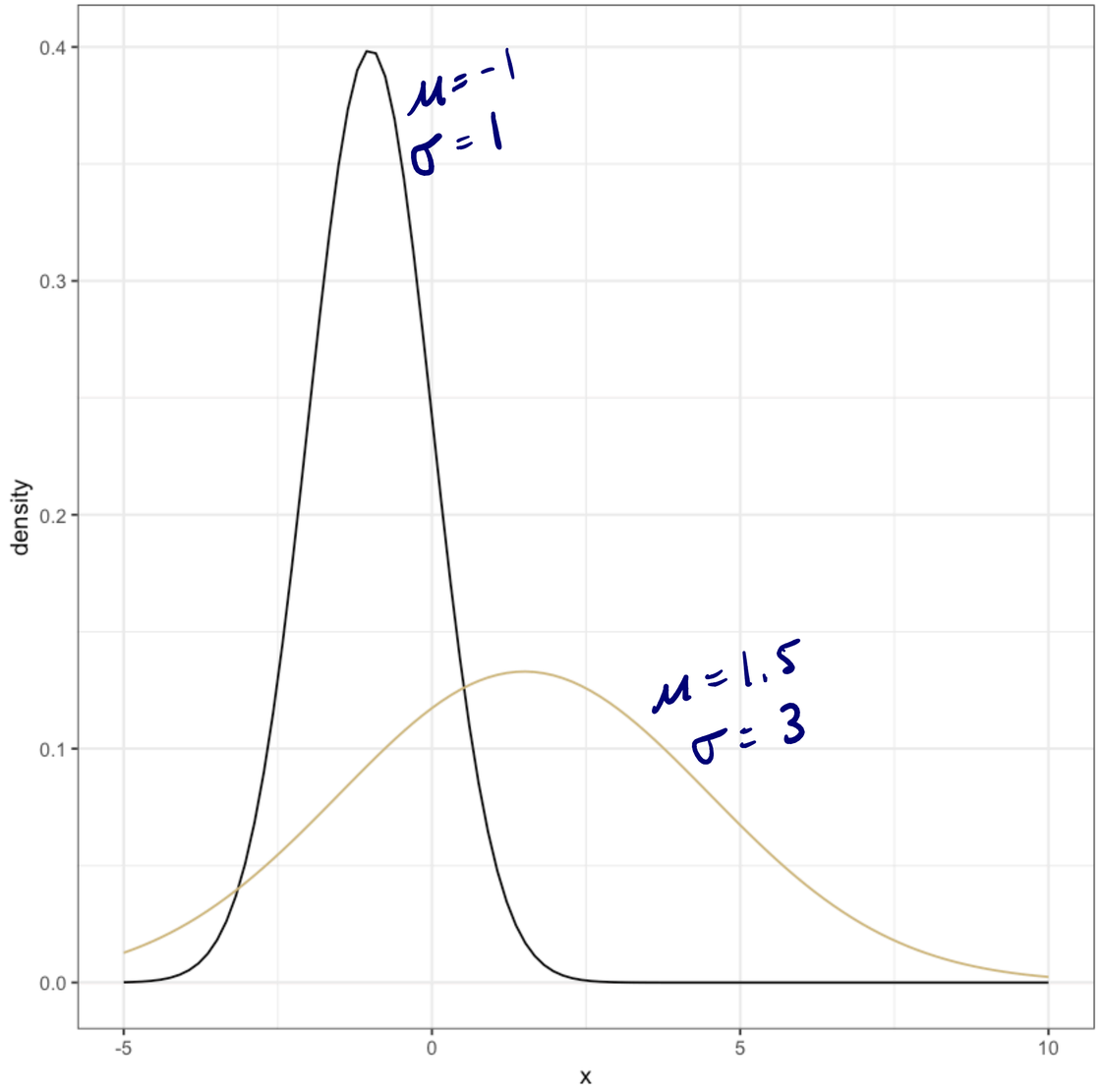
A continuous random variable X is said to have a **normal distribution** with parameters $E(X) = \mu$ and $Var(X) = \sigma^2$, if the pdf of X is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Notation:

$$X \sim N(\mu, \sigma^2)$$

The figure below presents graphs of $f(x)$ for different parameter pairs:



The normal distribution with parameter values $\mu = 0$ and $\sigma^2 = 1$ is called the **standard normal distribution**.

$$Z \sim N(0, 1)$$

A random variable with this distribution is called a standard normal random variable and is denoted by Z. Its pdf is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Note:

1. The standard normal distribution rarely occurs naturally.

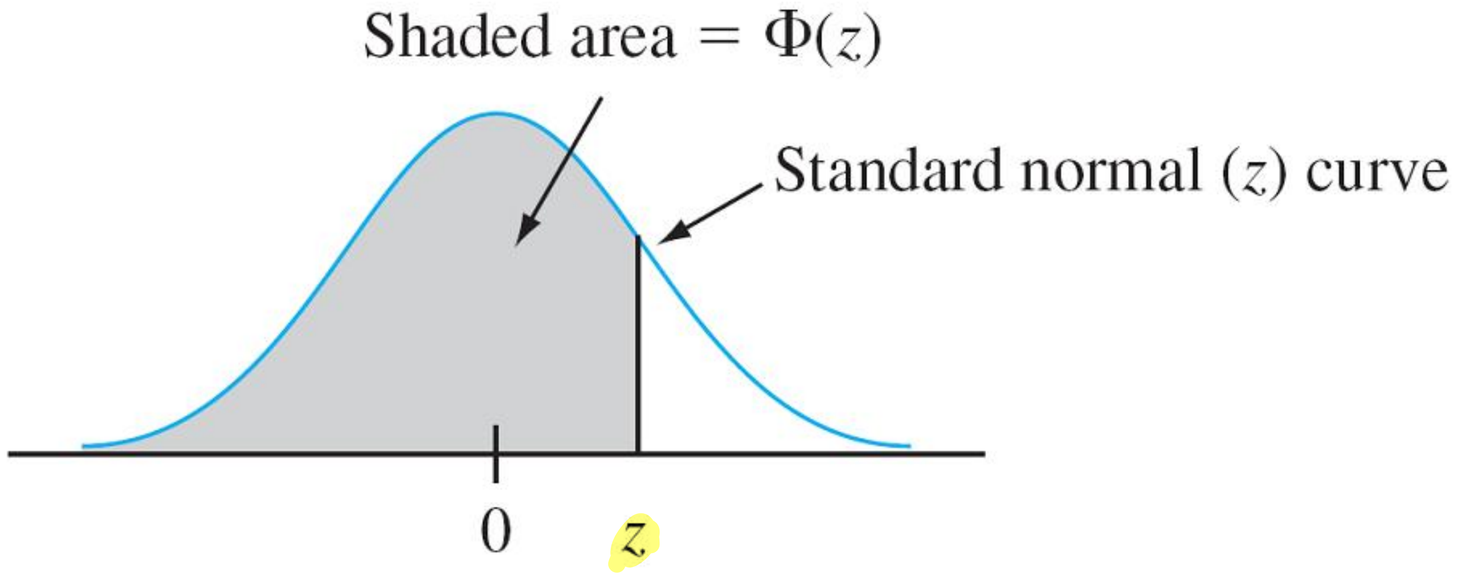
2. Instead, it is a *reference distribution* from which information about other normal distributions can be obtained via a simple formula.

3. These probabilities can then be found “normal tables”.

4. This can also be computed with a single command in R.

$$\text{pnorm}(x, \mu, \sigma) = P(X \leq x)$$

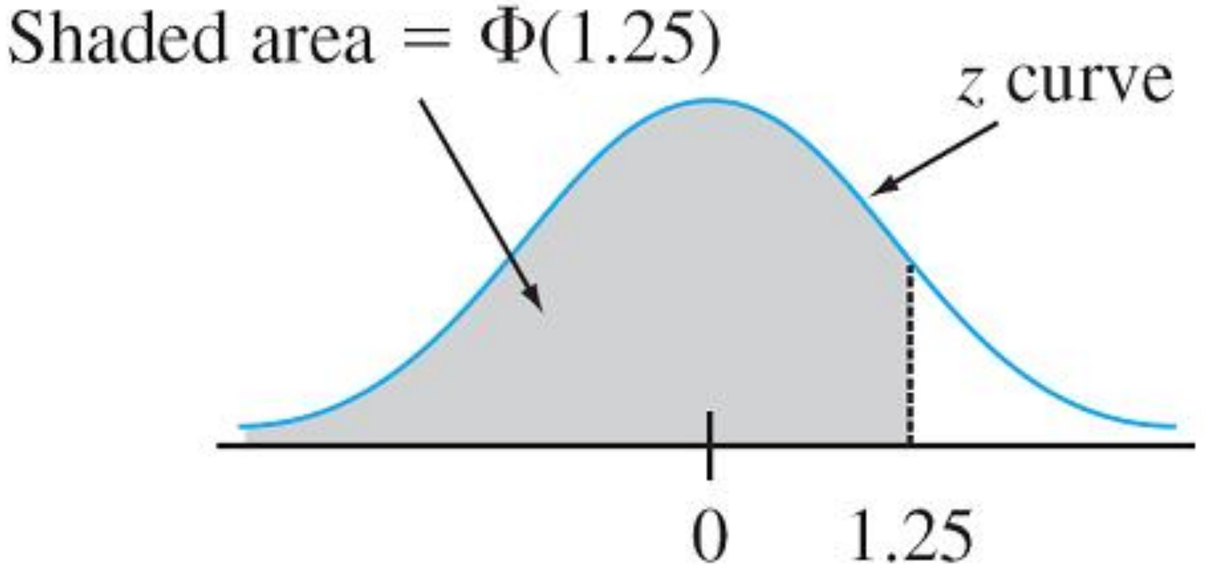
The figure below illustrates the probabilities found in a normal table (this can easily be found online):



$Z \sim N(0,1)$

$P(Z \leq z) = \Phi(z)$

Photo by Toby Elliott on Unsplash



$$Z \sim N(0, 1)$$

μ σ^2

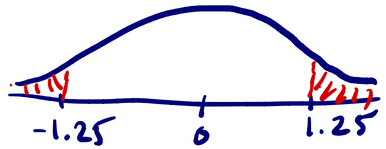
Examples:

pnorm(1.25, 0, 1)

1. $P(Z \leq 1.25) \approx 0.89$

2. Why does $P(Z \leq -1.25) = P(Z \geq 1.25)$?
   What is $\Phi(-1.25)$?

0.11



3. How do we calculate $P(-.38 \leq Z \leq 1.25)$?

$= P(Z \leq 1.25) - P(Z \leq -0.38) \approx 0.542$

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

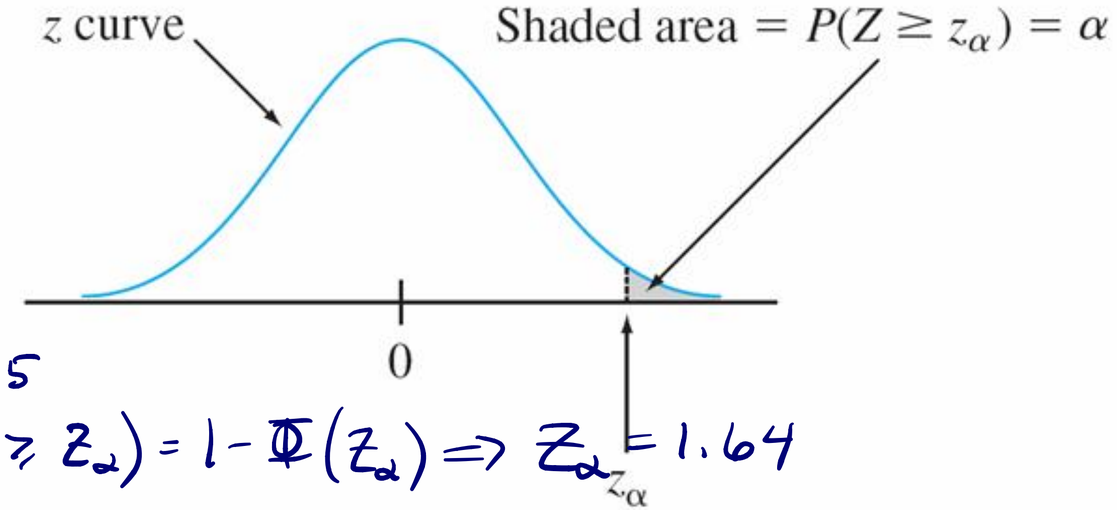
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997


Photo by Toby Elliott on Unsplash

In statistical inference, we often need the z values that give certain tail areas under the standard normal curve.

That is, we want to find (functions of) $F^{-1}(z) = \Phi^{-1}(z)$.

The **critical value** z_α will denote the z value for which α area under the standard normal lies to the right of z_α.



Let $\alpha = 0.05$

$0.05 = P(Z \geq z_\alpha) = 1 - \Phi(z_\alpha) \Rightarrow z_\alpha = 1.64$

Photo by Toby Elliott on Unsplash

When normal probabilities involving X are computed by "standardizing." The **standardized variable** is:

$X \sim N(\mu, \sigma^2)$

Proposition: If X has a normal distribution with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

is distributed standard normal.

The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.

$\text{pnorm}(1.75, 1.25, 0.46) - \text{pnorm}(1.00, 1.25, 0.46)$

Research suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec. $X \sim N(1.25, 0.46^2)$

What is the probability that reaction time is between 1.00 sec and 1.75 sec?

$$P(1 \le X \le 1.75) = P\left(\frac{1-1.25}{0.46} \le Z \le \frac{1.75-1.25}{0.46}\right)$$

$$= P(Z \le 1.08) - P(Z \le -0.543) \approx 0.568$$

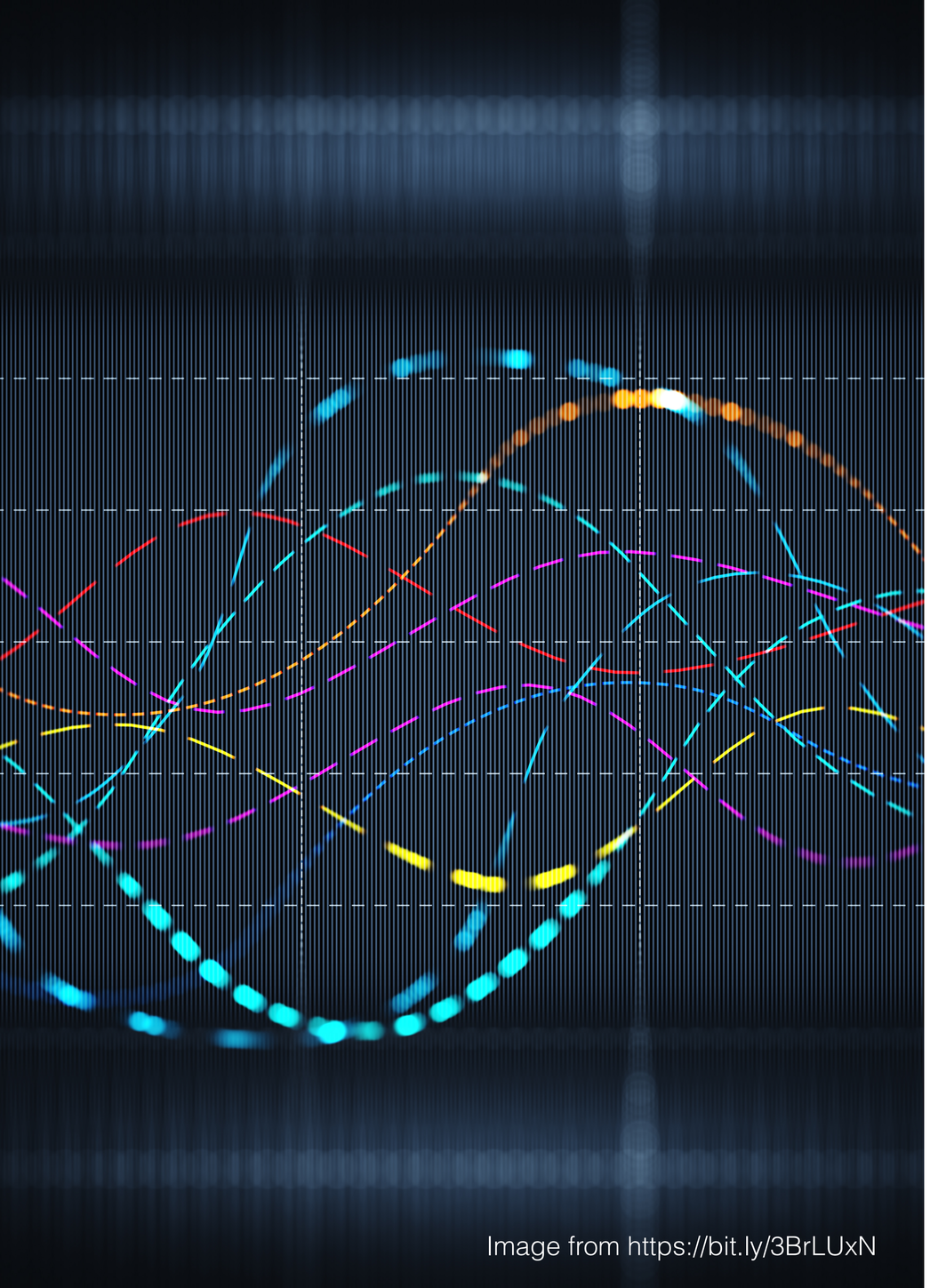


Image from https://bit.ly/3BrLUxN

Soon, we will be focusing on making “statistical inference” about the true mean of a population by using sample datasets. The normal distribution is widely used as a **statistical model** (and is often the distribution of a **statistic**—to be defined later).

$$M_\theta(\underline{x}) = \left(\underline{X} \;,\; f(x_i;\theta) \right)$$

$X_1, \ldots, X_n$

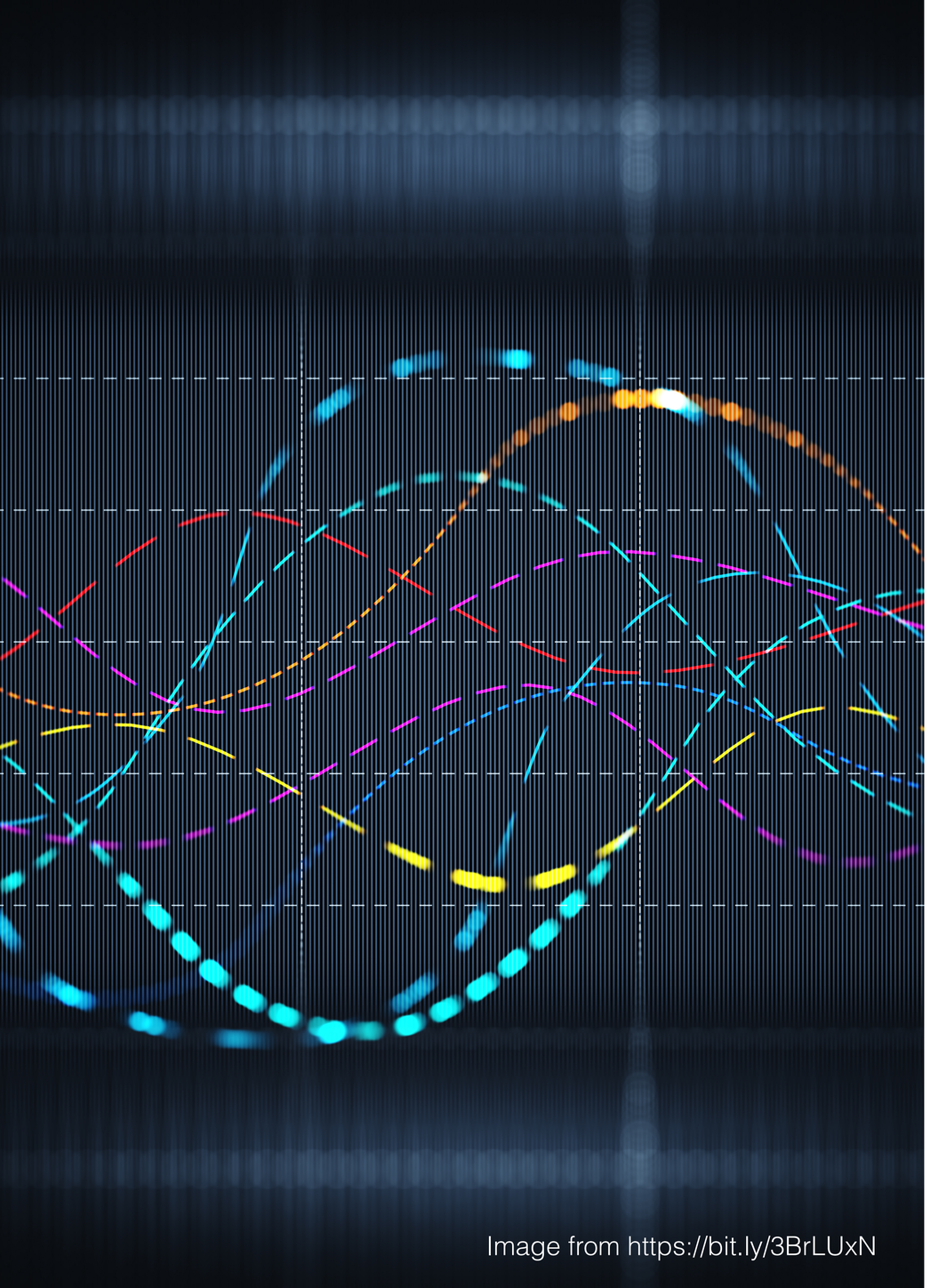


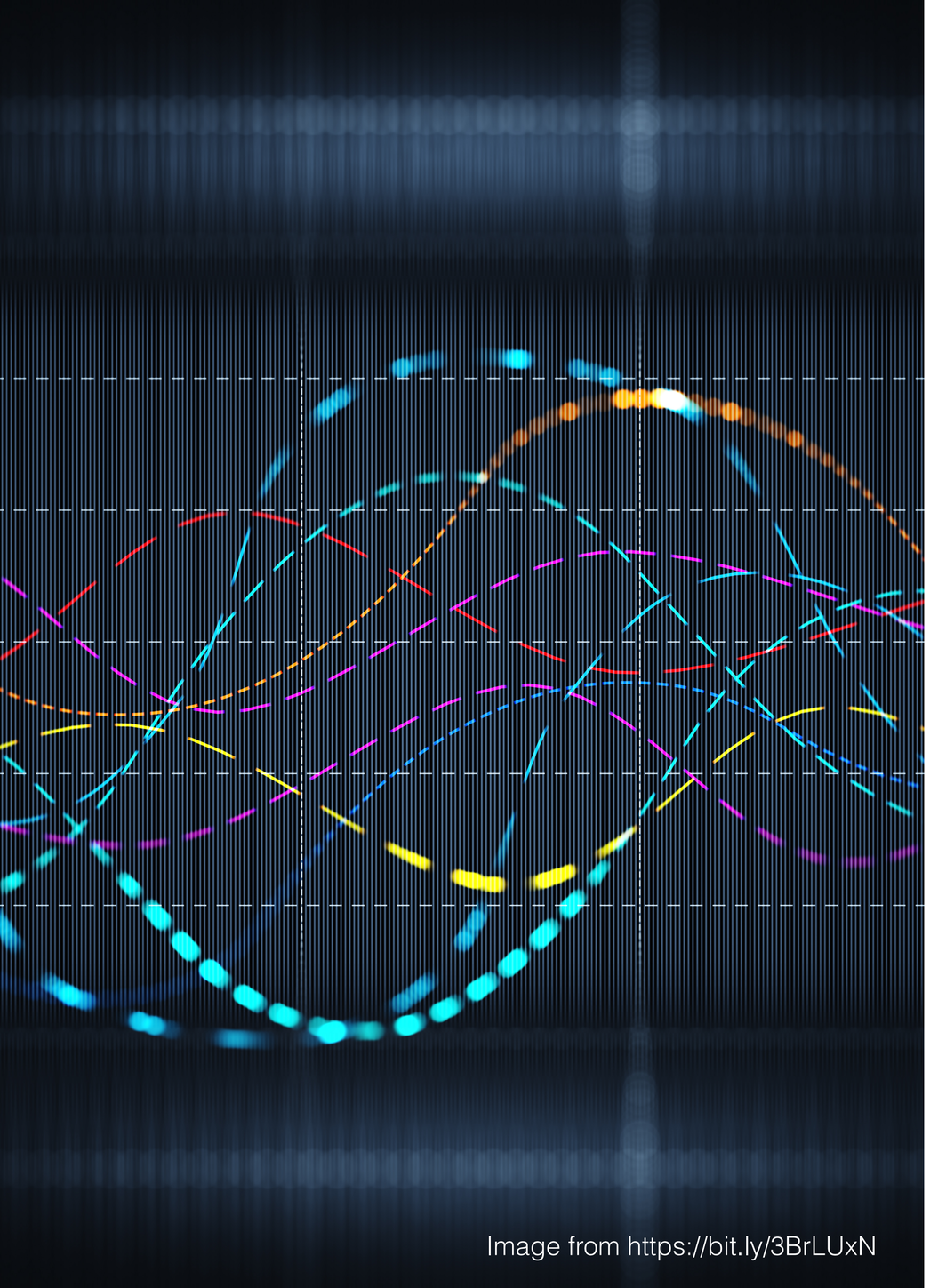
Image from https://bit.ly/3BrLUxN

The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) **random sample** of size n if:

1. X_i is independent from X_j , $i \neq j$

2. $X_1, \ldots, X_n$ are identically dist.

We say that $X_1, X_2, \ldots, X_n$ are **independent and identically distributed (iid).**



Data, $\mathbf{x}$, are assumed to be realizations of the random variables, $\mathbf{X}$, defined by a statistical model.

A **statistic** is a function of random variables ($\mathbf{X}$) or data ($\mathbf{x}$).

An **estimator** is a statistic put to the purpose of pinpointing or guessing a population parameter.

Examples: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of μ.
- $s(\underline{X}) = \frac{X_1 + X_2}{2}$ "

ID	x	y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n

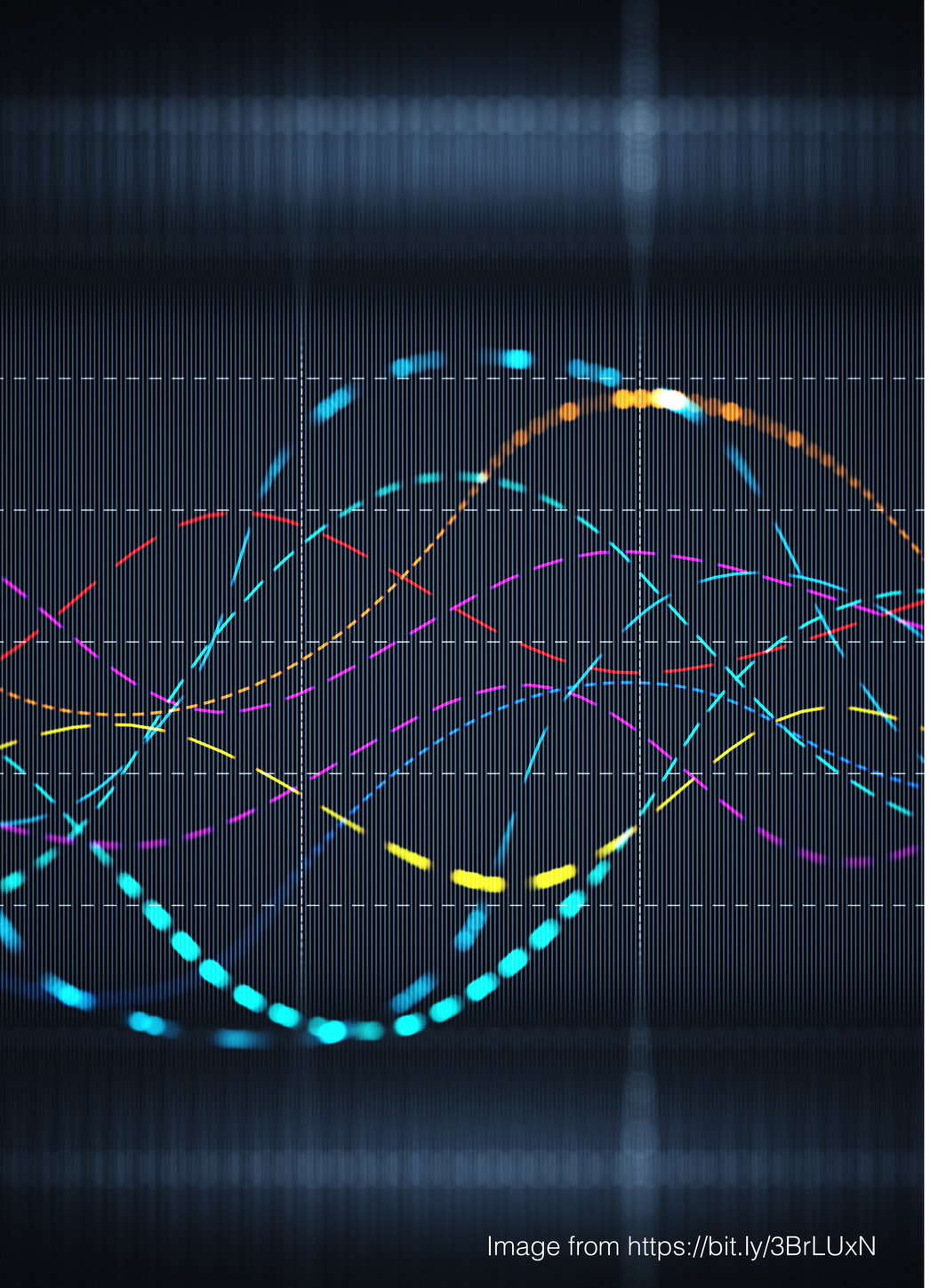


Image from https://bit.ly/3BrLUxN

We use estimators to summarize our iid sample. Any estimator, including the sample mean $\bar{X}$ is a random variable (since it is based on a random sample).

This means that $\bar{X}$ has a distribution of its own, which is referred to as **sampling distribution of the sample mean.** This sampling distribution depends on:

1.) The sample size, n

2.) The pop. dist. (dist. X_i, $i = 1, \ldots, n$)

3.) Method of sampling

The standard deviation of this distribution is called **the standard error of the estimator.**

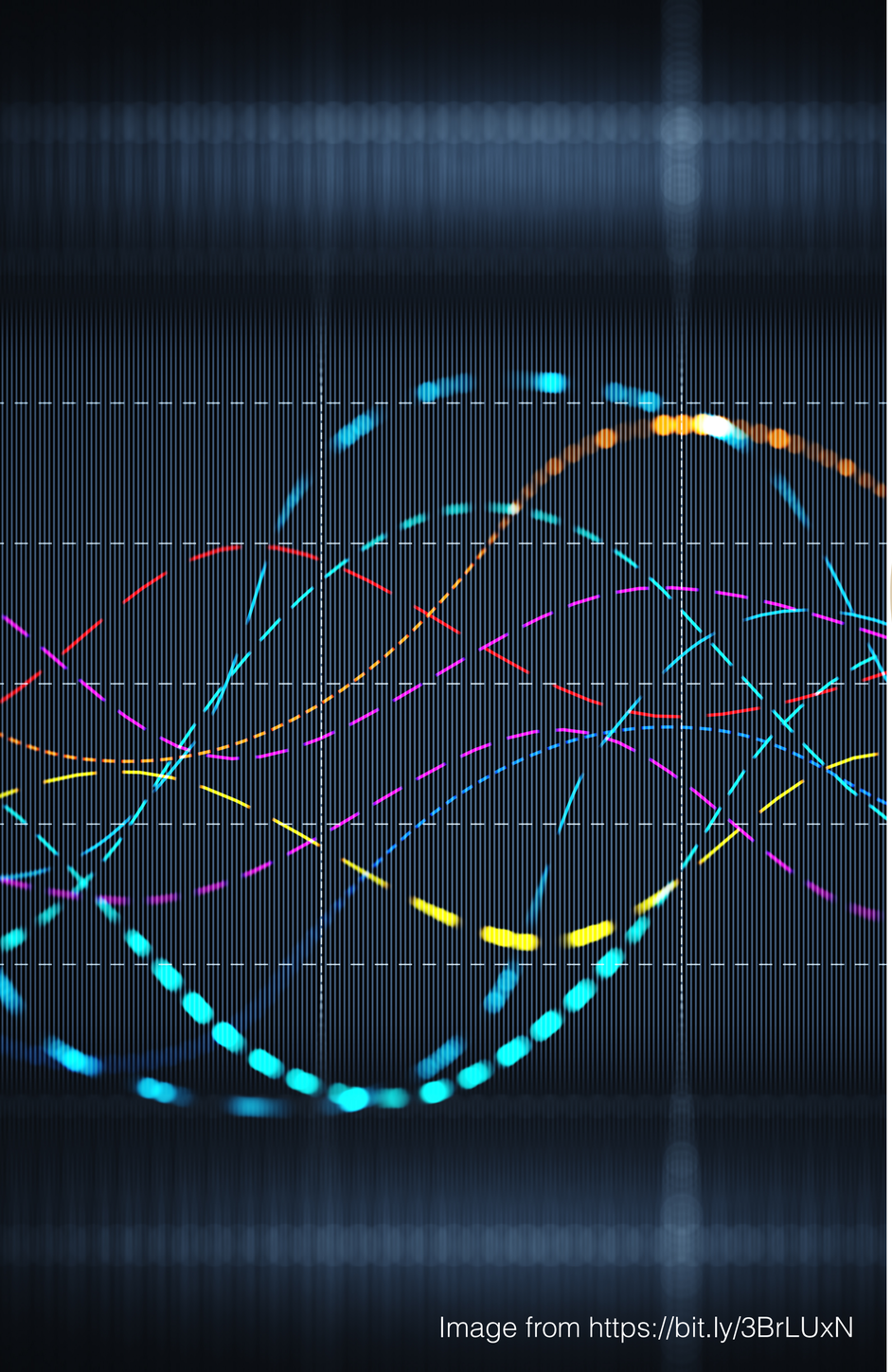


Image from https://bit.ly/3BrLUxN

We usually only have one sample

Population attribute modeled by normal random variables mean μ.

iid

sample #1 of size n

sample #2 of size n

sample #3 of size n

sample #m of size n

hypothetical

$\bar{X}_1$

$\bar{X}_2$

$\bar{X}_3$

$\bar{X}_n$

Sampling distribution of $\bar{X}$

(The standard deviation of this distribution is the "standard error of $\bar{X}$.")

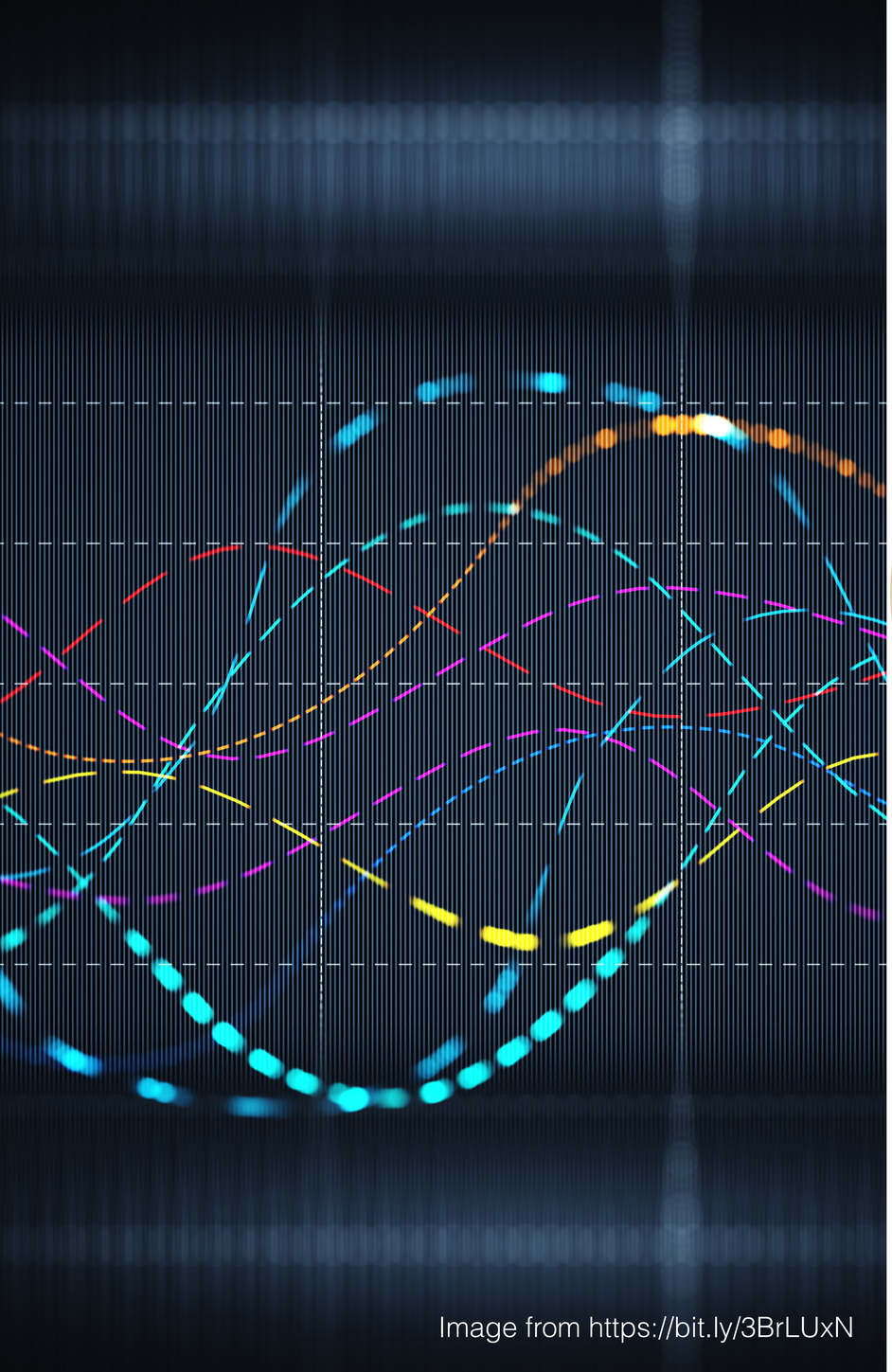


Image from https://bit.ly/3BrLUxN

Caffeine content (mg) for a 12 oz cup of coffee made by *Hello Coffee*. Modeled by normal distribution, with parameter μ

sample #1 of size n → $\bar{X}_1$

sample #2 of size n → $\bar{X}_2$

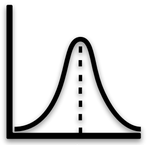
sample #3 of size n → $\bar{X}_3$

⋮ ⋮

sample #m of size n → $\bar{X}_n$

⋮ ⋮

Sampling distribution of $\bar{X}$



(The standard deviation of this distribution is the "standard error of $\bar{X}$.)

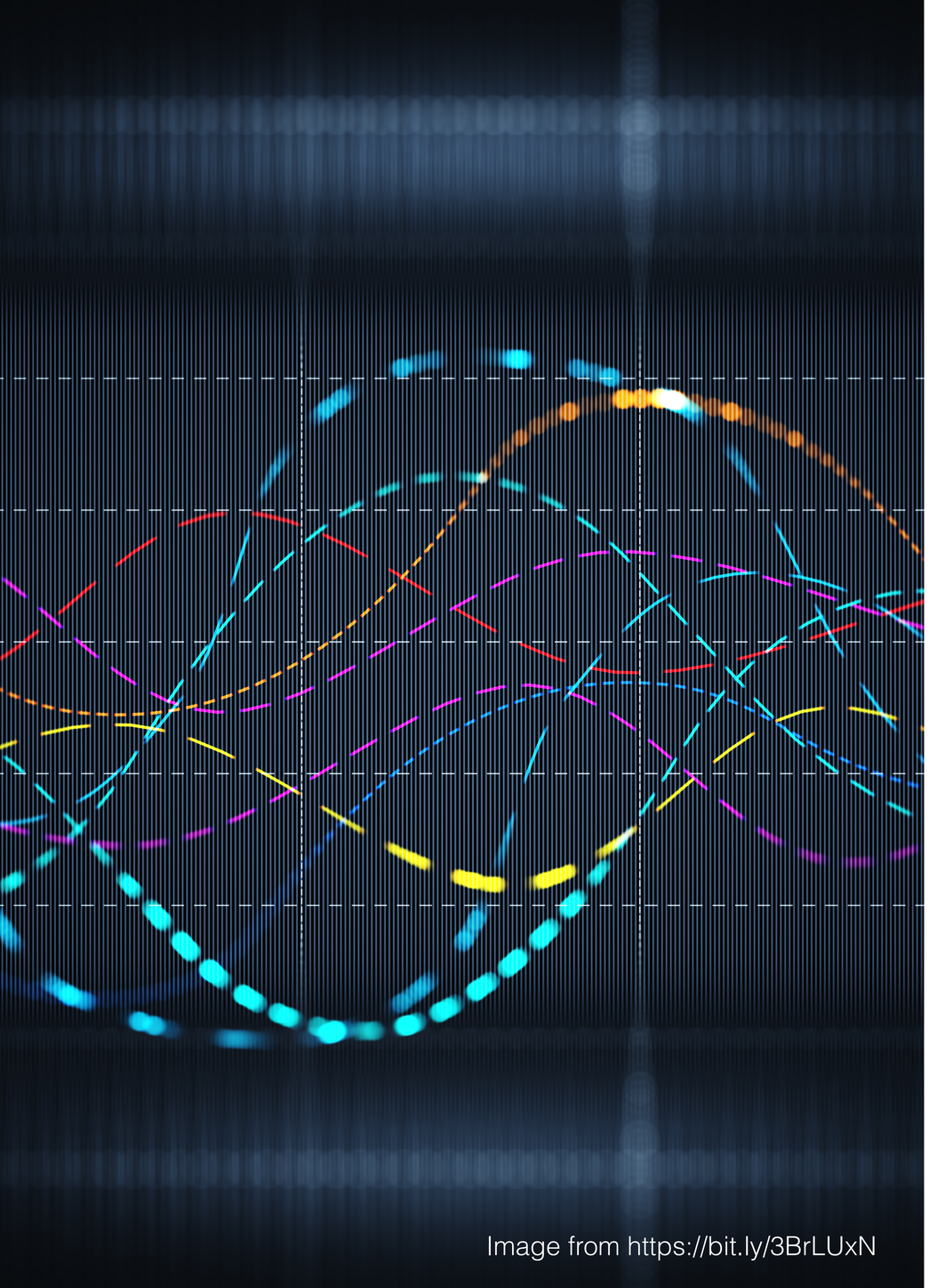


Image from https://bit.ly/3BrLUxN

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all $i = 1,...,n$. Then:

iid

- $E(\bar{X}) = \mu$
- $Var(\bar{X}) = \frac{\sigma^2}{n}$

The standard deviation of the sample mean is:

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

This is also called the **standard error** of the mean.

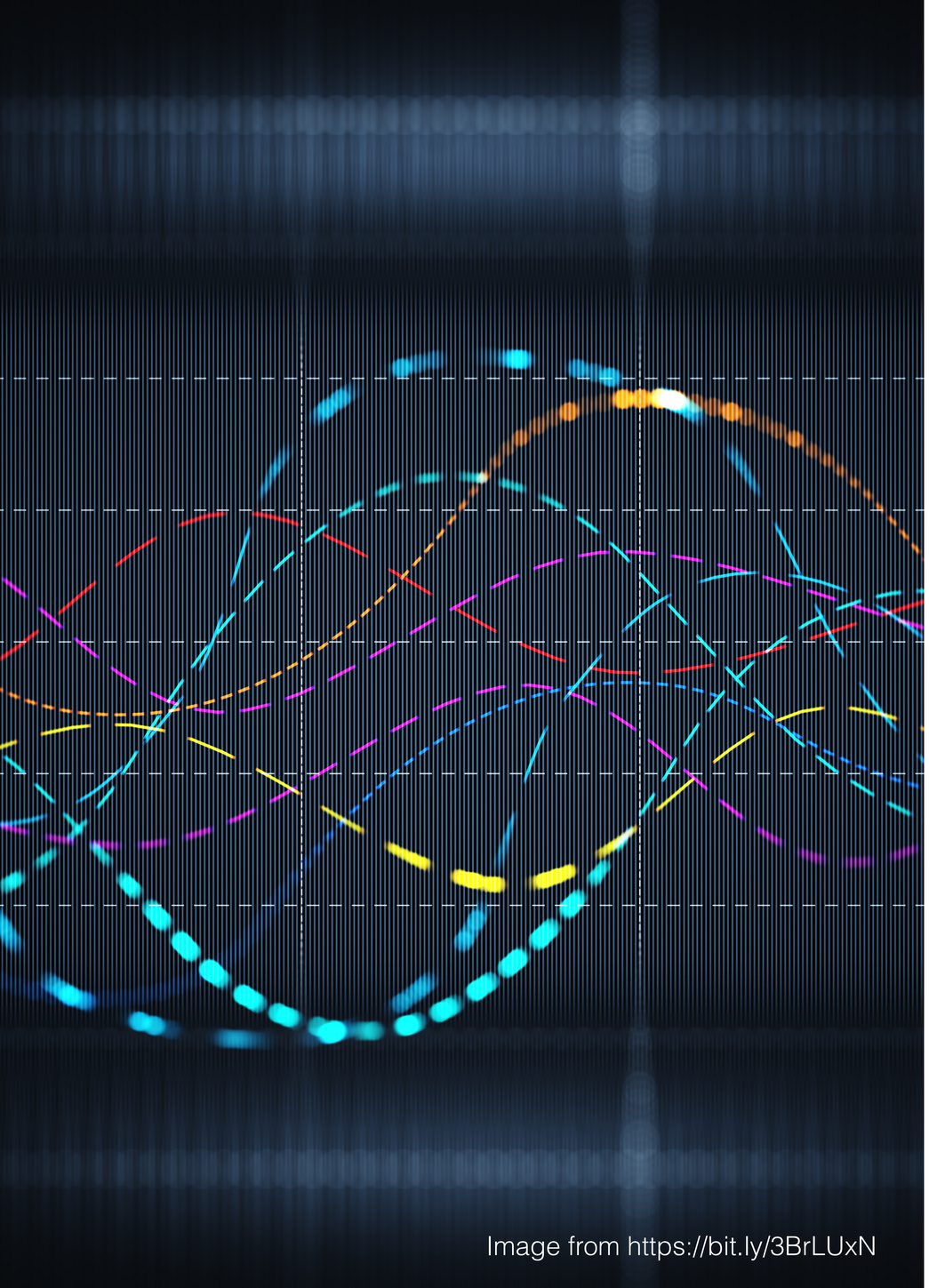


Image from https://bit.ly/3BrLUxN

Great, but what is the *distribution* of the sample mean?

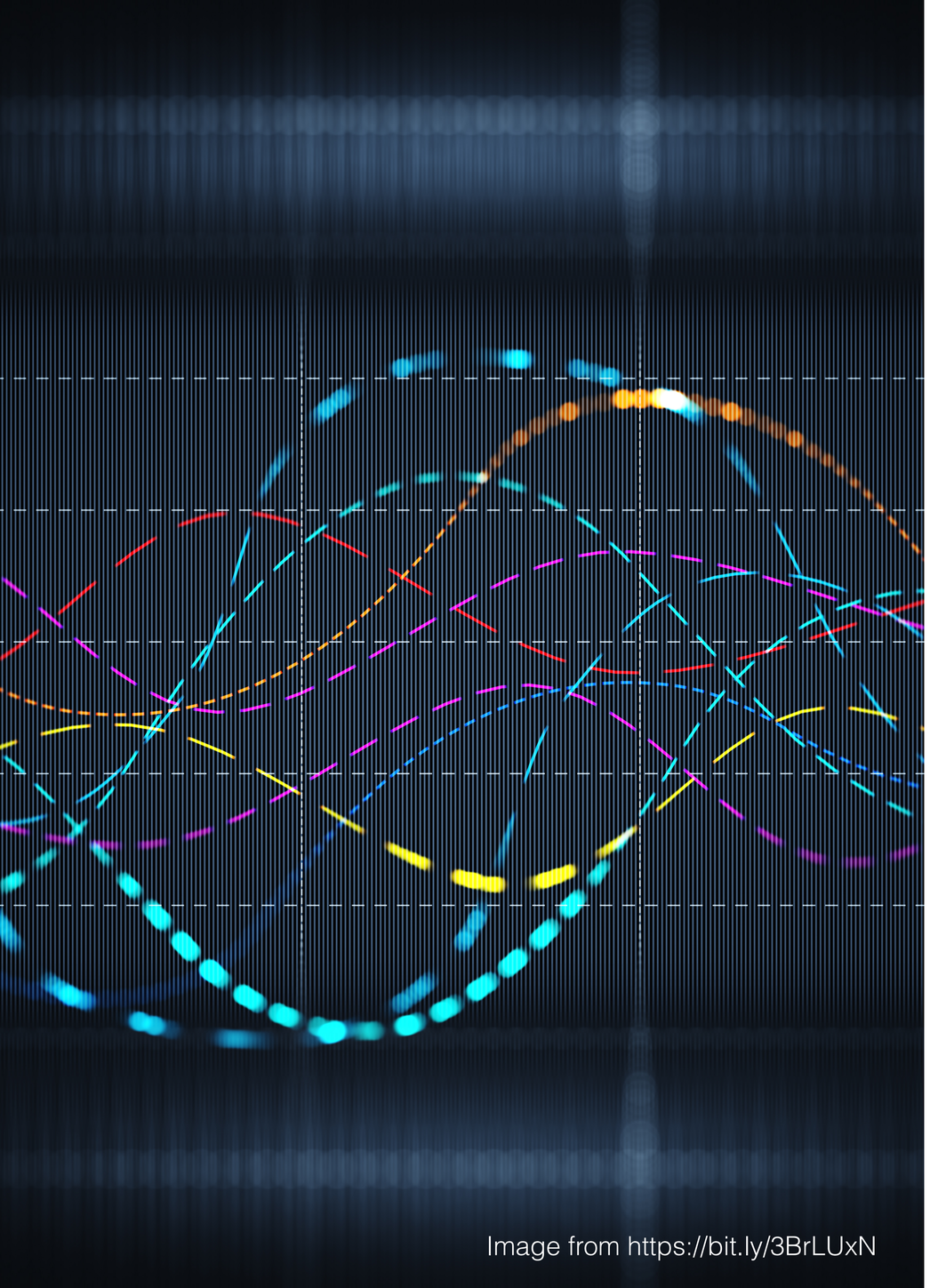


Image from https://bit.ly/3BrLUxN

If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

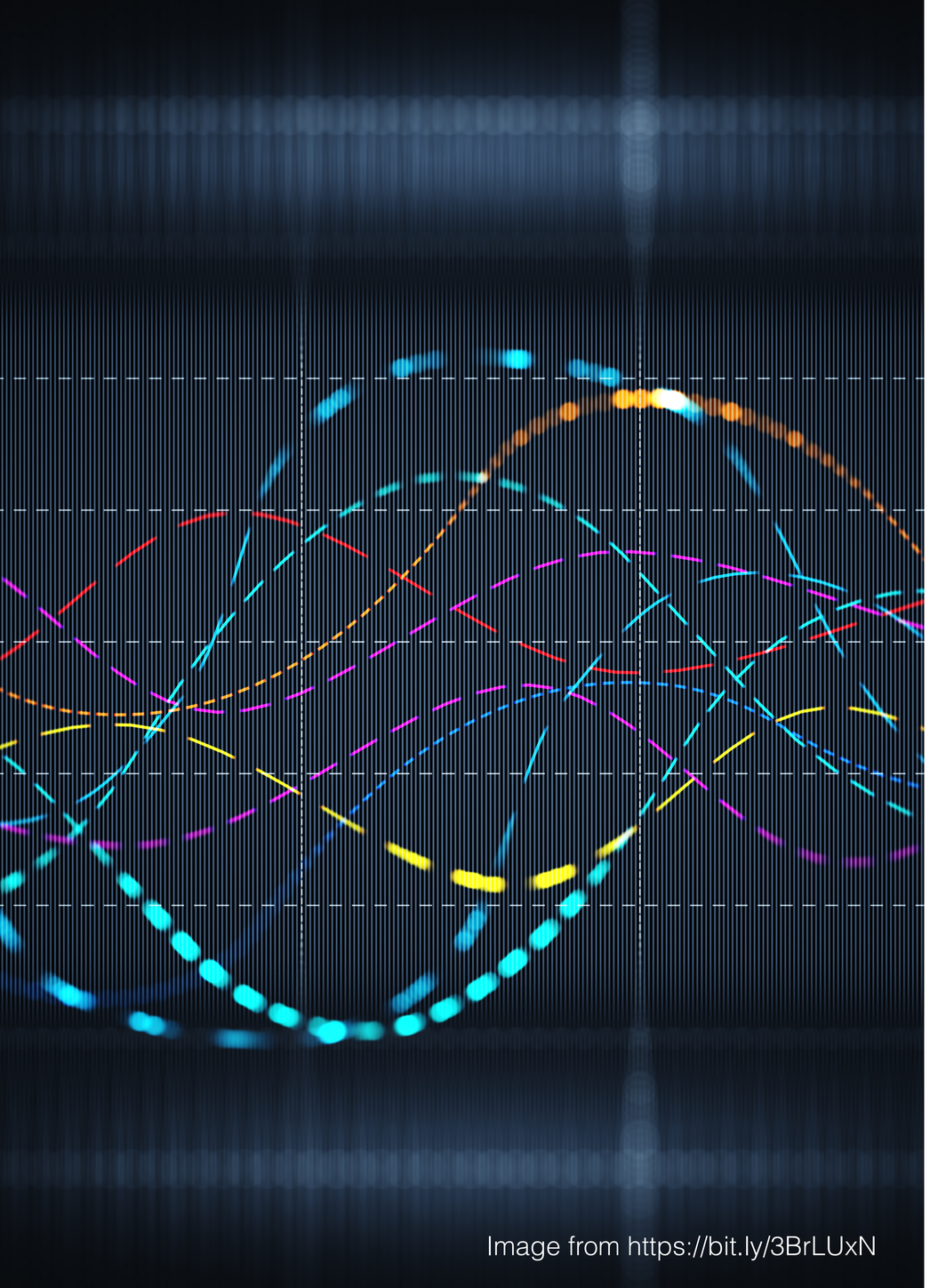
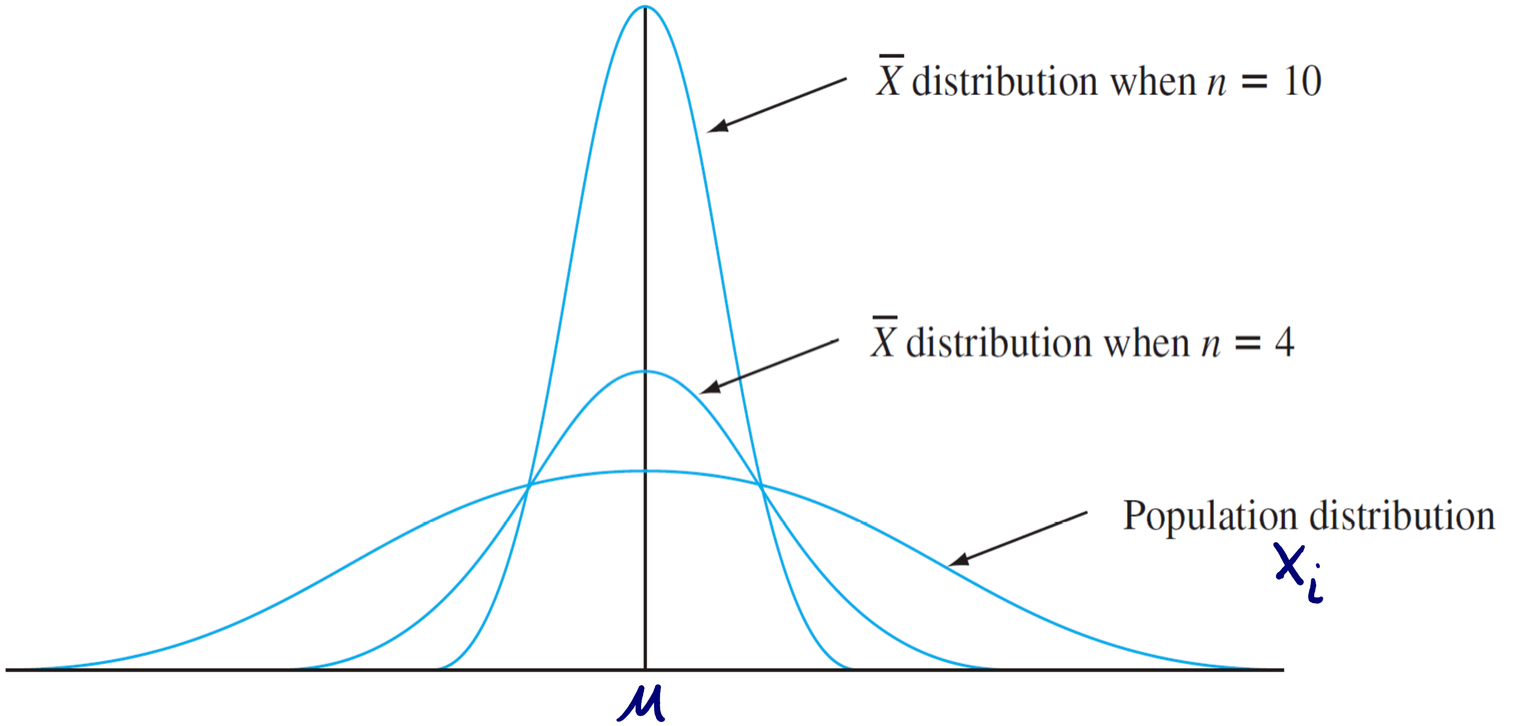


Image from https://bit.ly/3BrLUxN



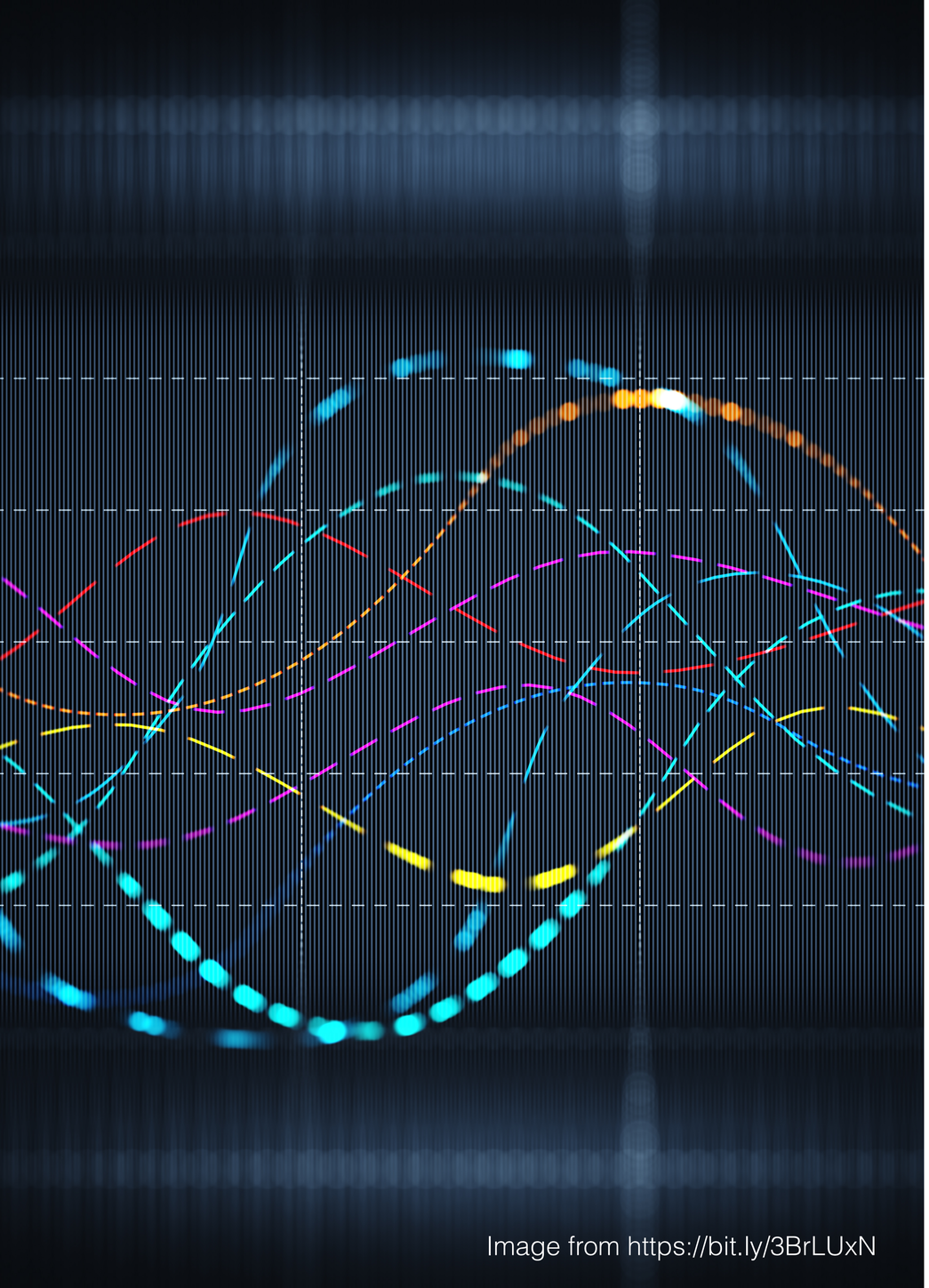


Image from https://bit.ly/3BrLUxN

But what if the underlying distribution of $\bar{\mathbf{X}}$ is not normal?

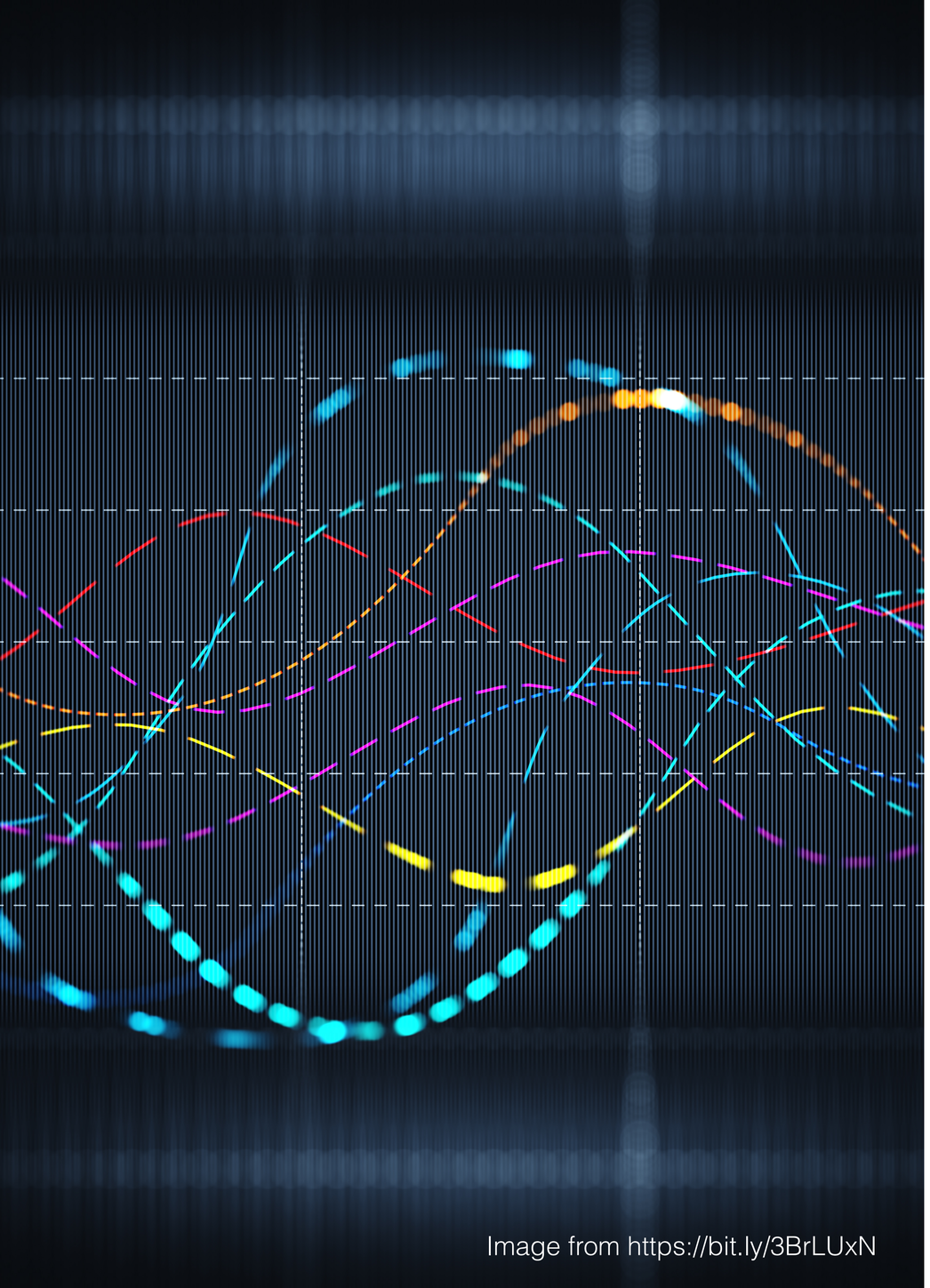


Image from https://bit.ly/3BrLUxN

Important: When the population distribution is non-normal, averaging produces a distribution more bell-shaped than the one being sampled.

A reasonable conjecture is that if n is large, a suitable normal curve will approximate the actual distribution of the sample mean.

The formal statement of this result is one of the most important theorems in probability and statistics: **Central Limit Theorem!**

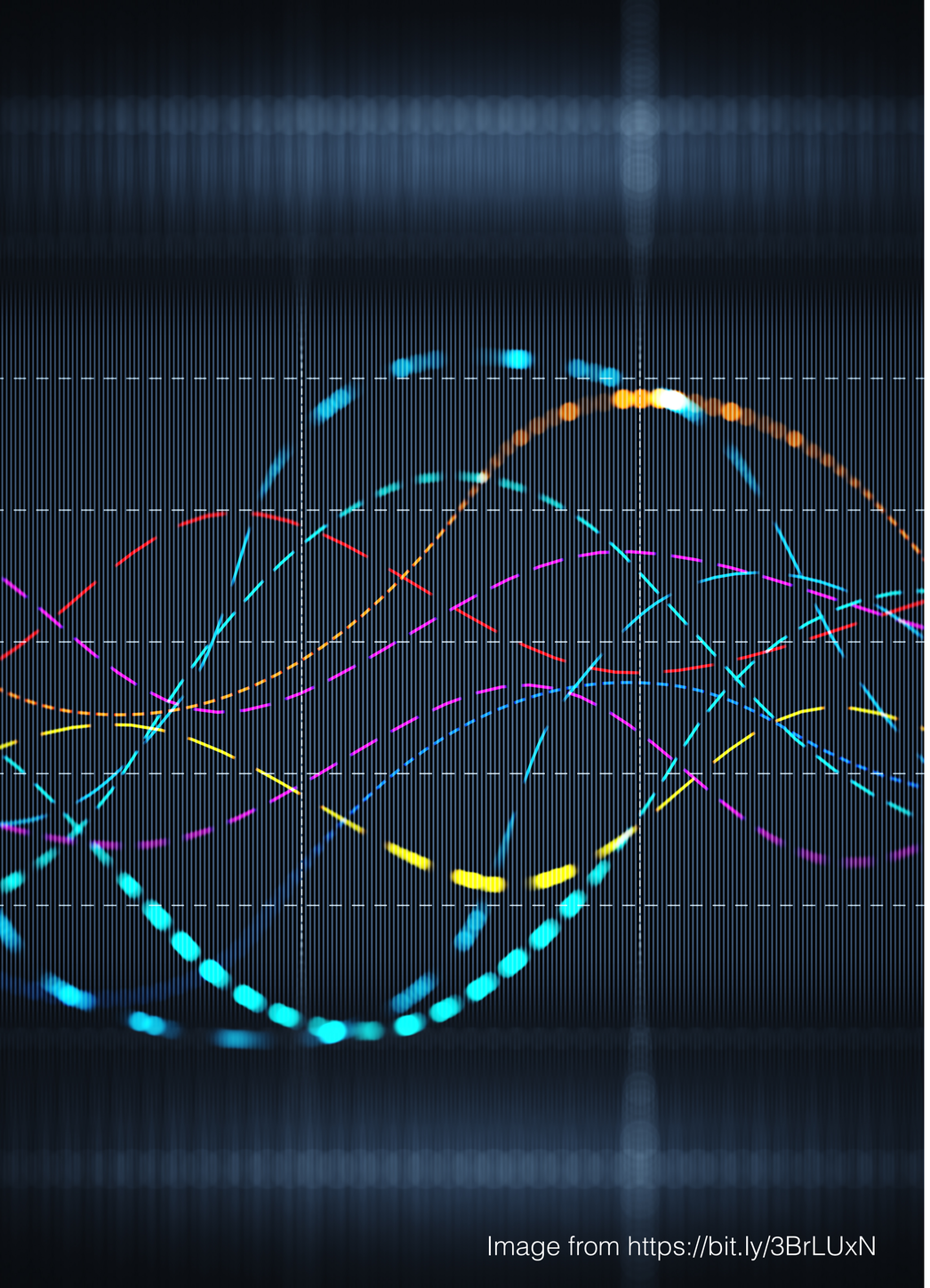


Image from https://bit.ly/3BrLUxN

The Central Limit Theorem (CLT): Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a set of iid random variables with $E(X_i) = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$ for all $i = 1,...,n$. Then:

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\left(\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)\right)$$

The larger the value of n, the better the approximation! Typical rough rule: $n \geq 30$.

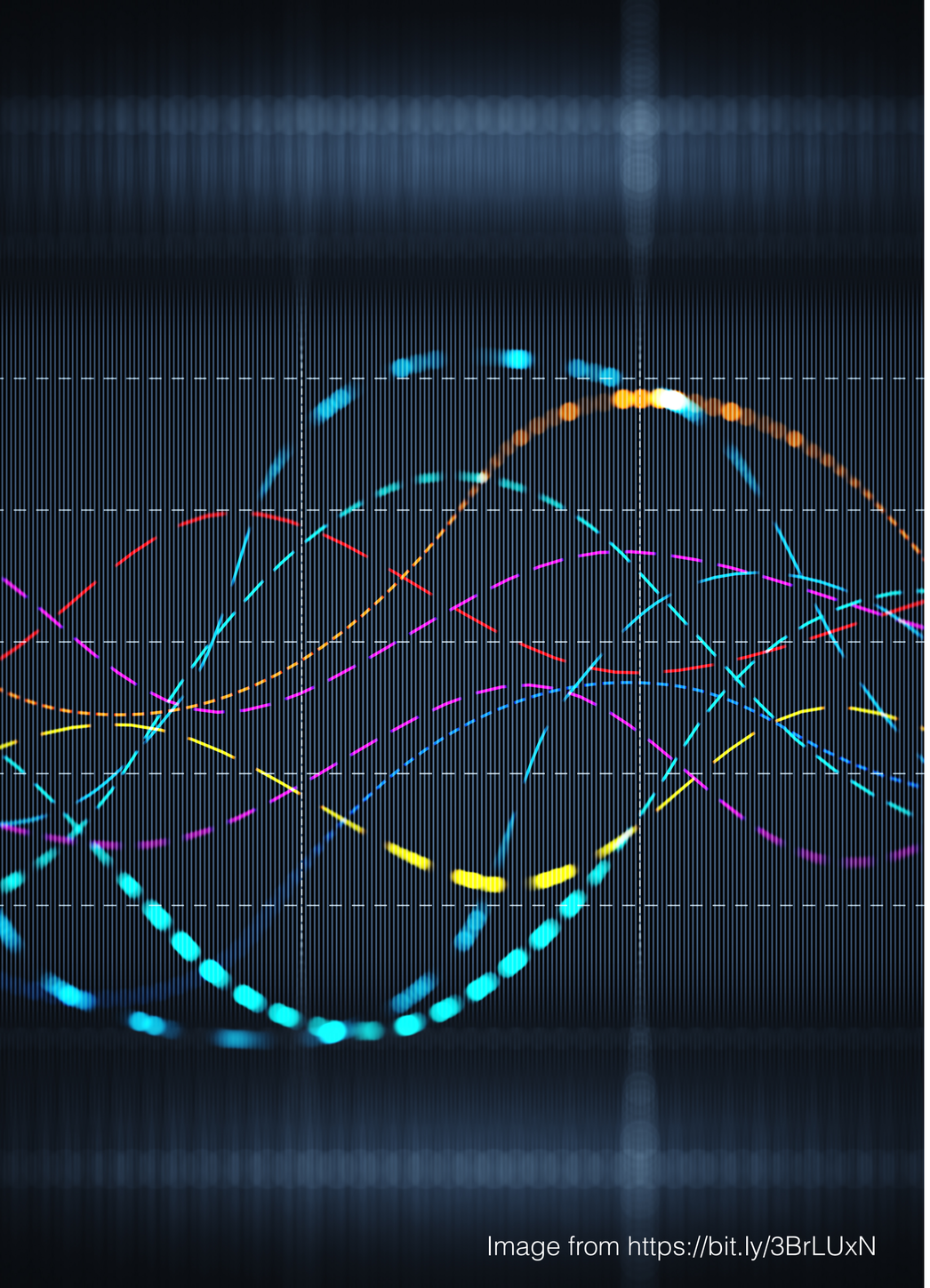
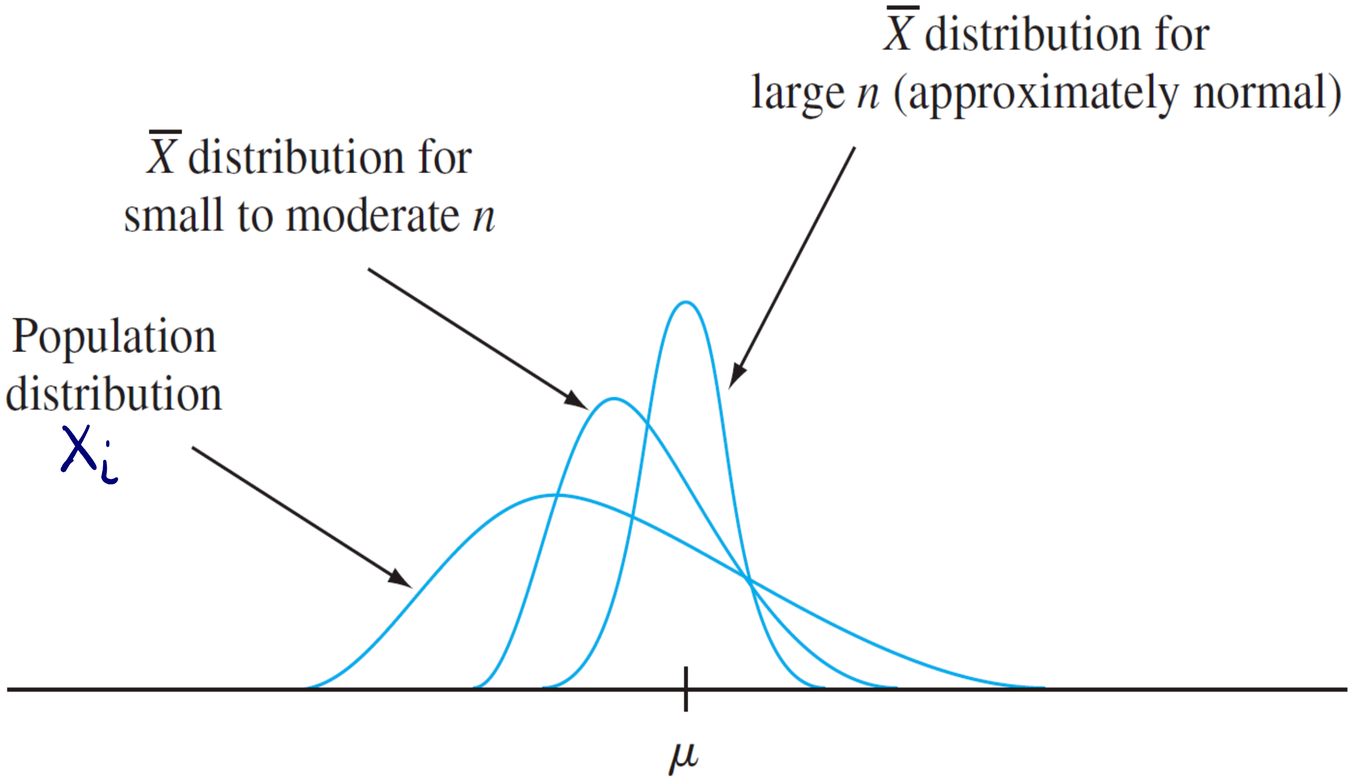


Image from https://bit.ly/3BrLUxN



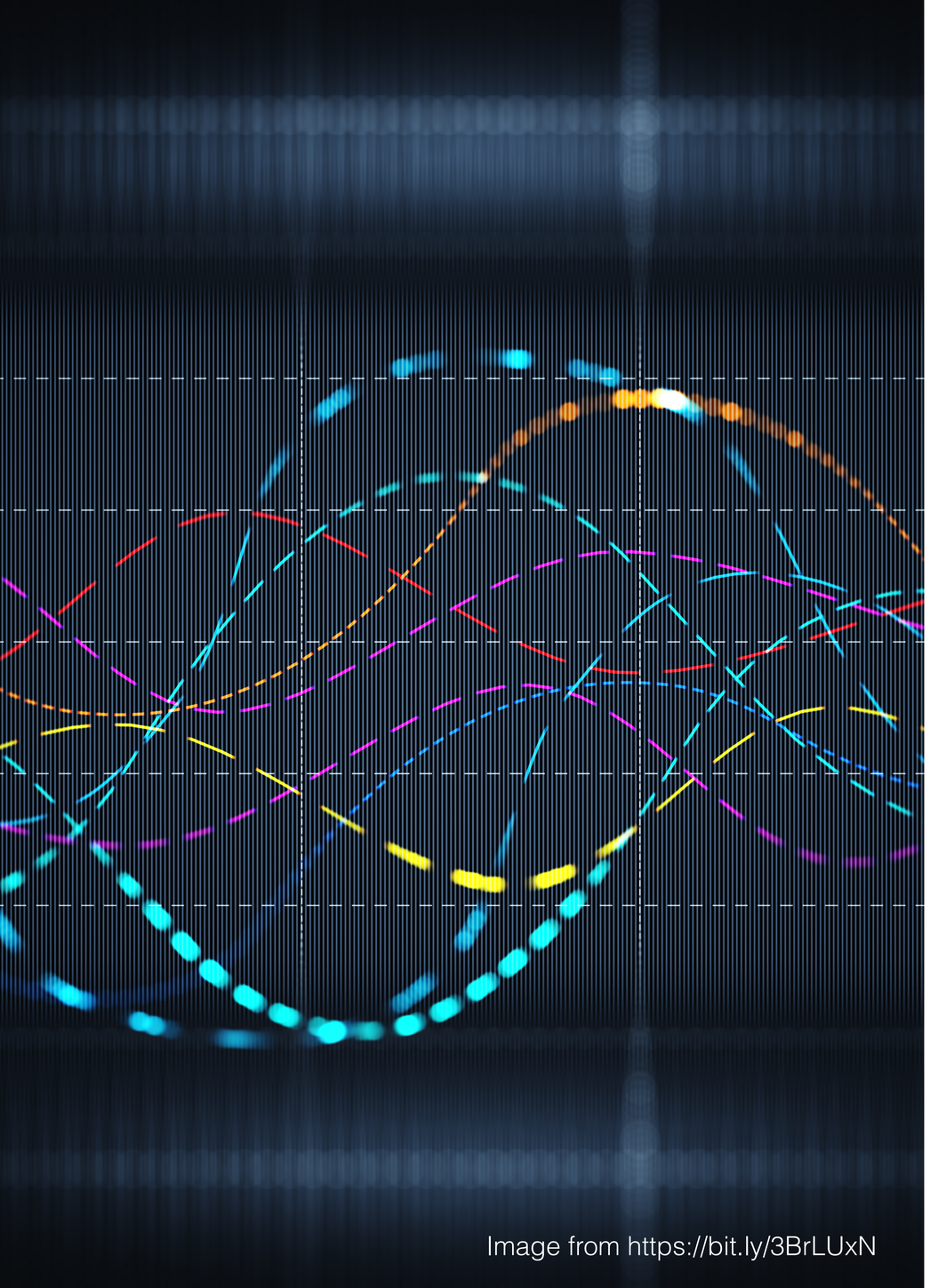


Image from https://bit.ly/3BrLUxN

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when n is sufficiently large. The problem is that the accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled.