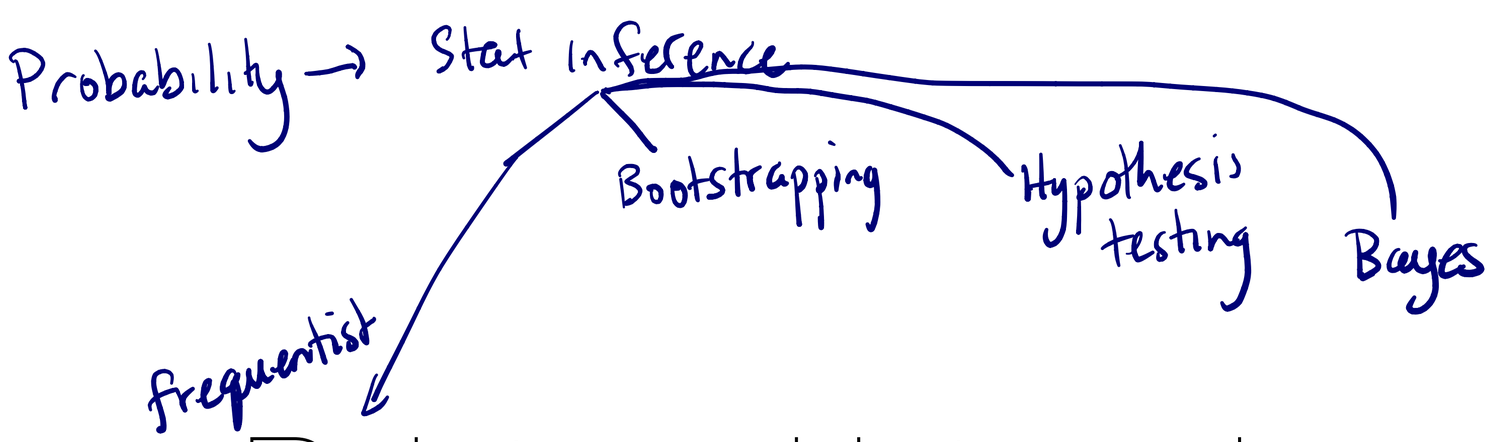
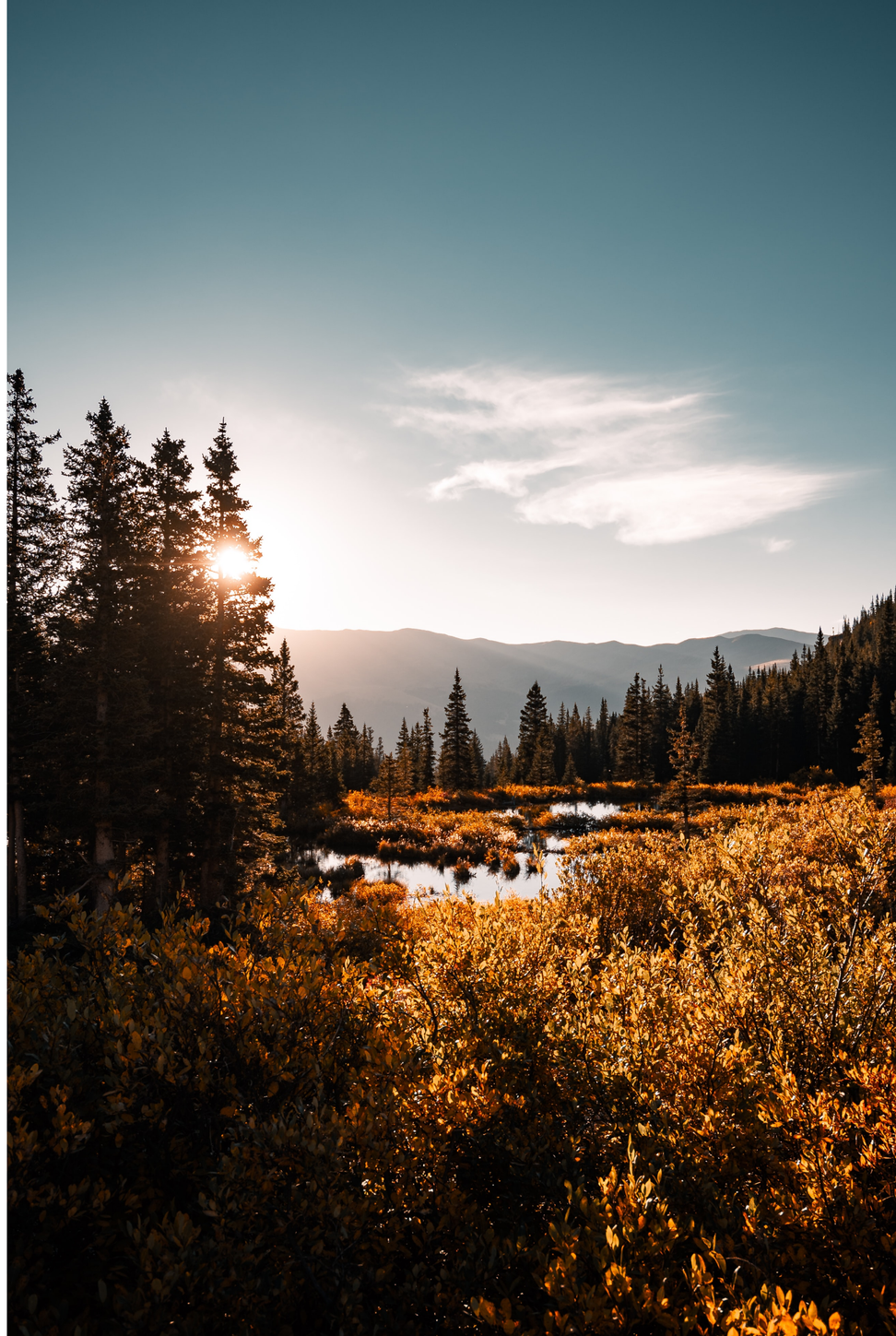


Point and Interval Estimation

Point estimation: Chapter 7 (7.2.2, 7.2.3, 7.3.2)
Confidence intervals: Chapter 8 (8.1, 8.2.1, 8.2.2, 8.4.1)



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

1. Define the terms **point estimator** and **interval estimator**, and differentiate between an *estimator* and an *estimate.*
2. Differentiate between a *frequentist* estimator and a *Bayesian* estimator.
3. Identify properties that make a frequentist point estimator *good*.
4. Define, derive, calculate, and interpret maximum likelihood estimators.
5. Derive the $X\%$ confidence interval for the mean of a population when the sample standard deviation is **known** (z-confidence interval).
6. Identify the center and the length of a confidence interval for the mean.
7. Properly interpret confidence intervals and identify common misinterpretations.
8. Find the sample size necessary to ensure that a resulting confidence interval has width of at most *w*.
9. Describe why the (pre-calculated) confidence interval endpoints are random (for all CIs).
10. Derive the $X\%$ confidence interval for the mean of a population when the **sample is large** and sample standard deviation is **unknown**. (z- confidence interval)
11. Describe the t-distribution, its properties, its parameter, and its use in confidence interval calculations. Use R to calculate critical values/percentiles of a t-distribution.
12. State the assumptions of the t-confidence interval for the mean.
13. Calculate the $X\%$ t-confidence interval for the mean of a normal population.
14. State the normal approximation of a binomial distribution.
15. Derive, calculate, and interpret the $X\%$ confidence interval for a population proportion.
16. Derive, calculate, and interpret the $X\%$ confidence interval for a population variance.

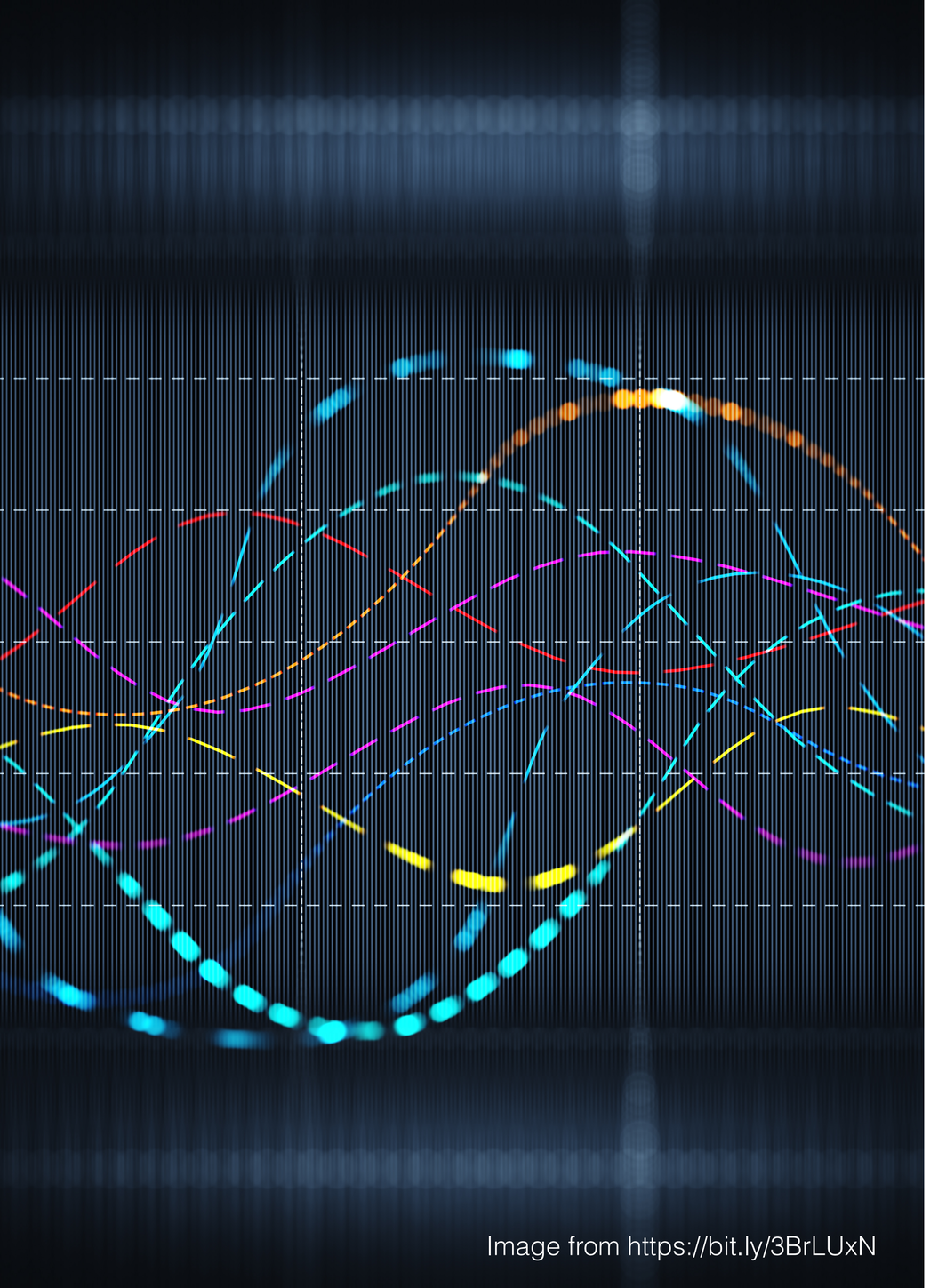


Image from https://bit.ly/3BrLUxN

Data, $\mathbf{x}$, are assumed to be realizations of the random variables, $\mathbf{X} = (X_1, \ldots, X_n)$, defined by a statistical model.

A statistical model will have a set of unknown **parameters**, or constants that define the probability distribution that generates the data.

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.

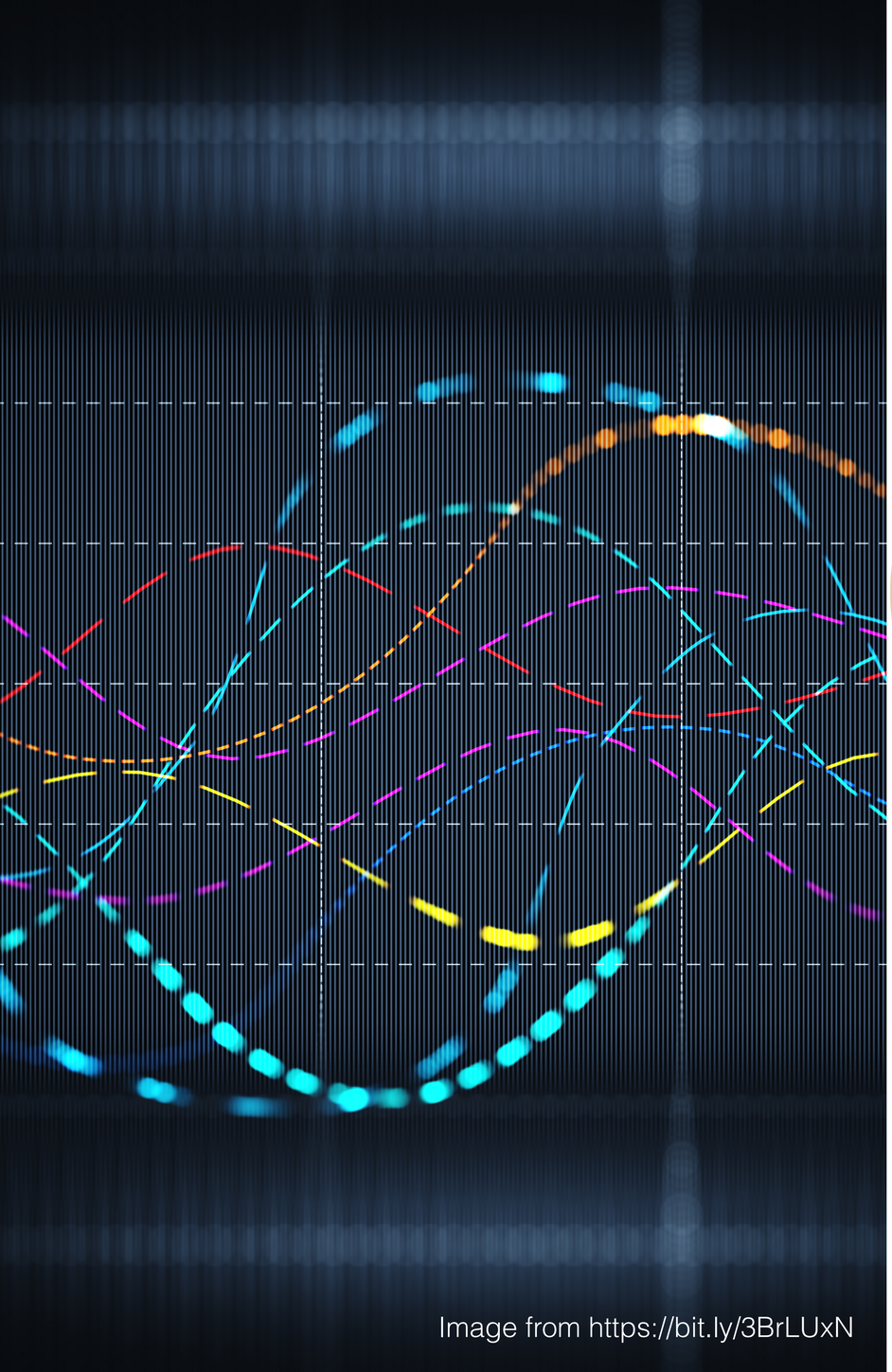


Image from https://bit.ly/3BrLUxN

$\hat{\theta}$ = estimator of θ

Population attribute modeled by random variables with probability distribution parameterized by θ.

sample #1 of size n → $\hat{\theta}_1$

sample #2 of size n → $\hat{\theta}_2$

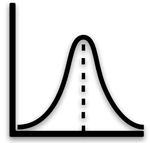
sample #3 of size n → $\hat{\theta}_3$

⋮ ⋮

sample #m of size n → $\hat{\theta}_m$

⋮ ⋮

Sampling distribution of $\hat{\theta}$



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

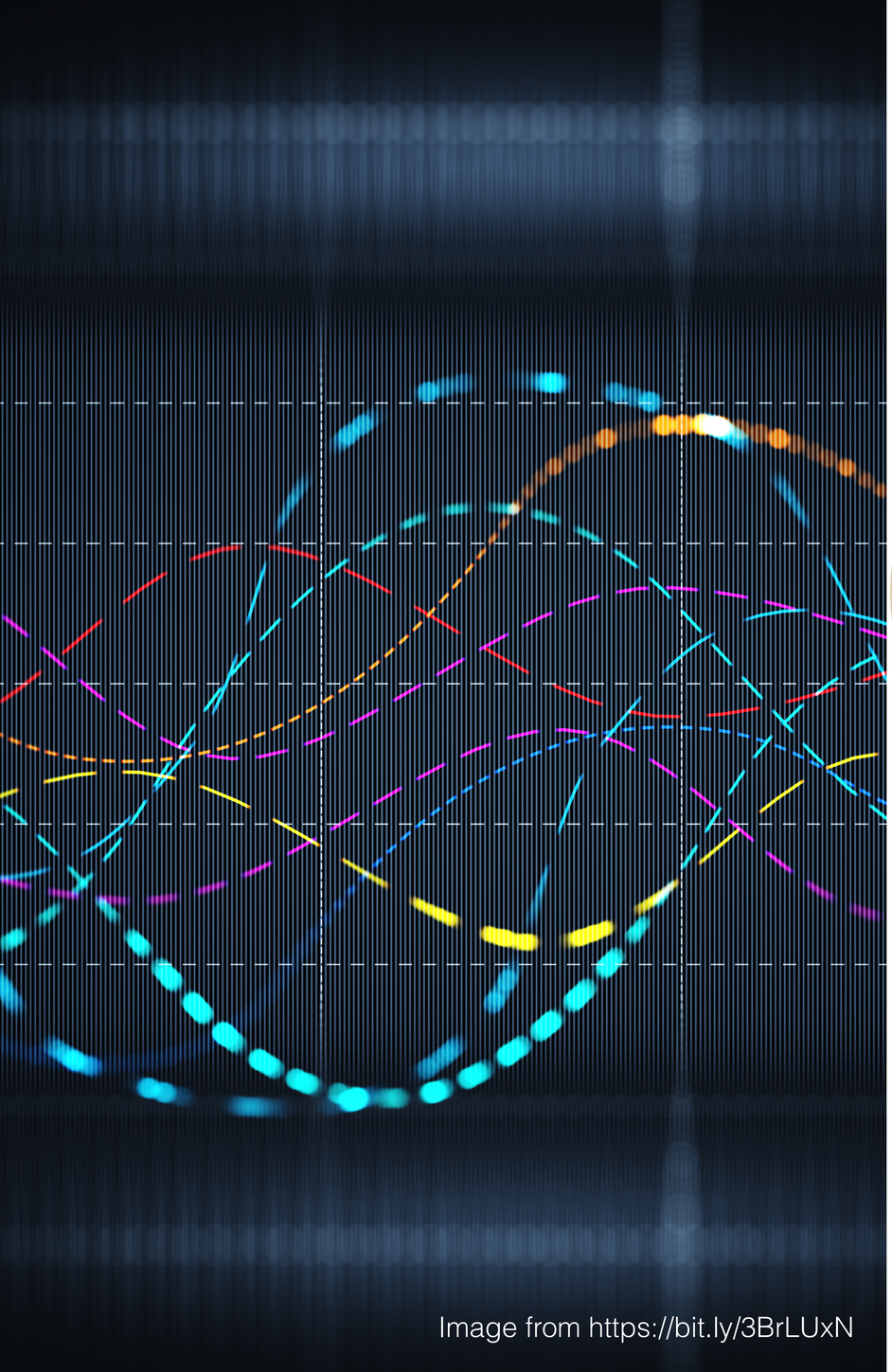


Image from https://bit.ly/3BrLUxN

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

$\mu = E(X)$

sample #1 of size n → $\hat{\mu}_1$

sample #2 of size n → $\hat{\mu}_2$

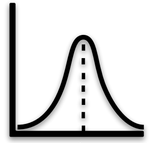
sample #3 of size n → $\hat{\mu}_3$

⋮ ⋮

sample #m of size n → $\hat{\mu}_m$

⋮ ⋮

Sampling distribution of $\hat{\mu}$



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

· $Cov(aX_1 + bX_2)$

A.4 (1) · $Var(Z^2) = E(Z^4) - [E(Z^2)]^2 = 3 - 1 = 2$

· B.1(b) not u

· Datasets

Frequentist vs Bayesian Inference

Broadly speaking, there are two different paradigms for finding estimators for parameters. The different paradigms result from different interpretations of probability.

Frequentist: (Popper)

- Parameters are unknown and fixed.
- Estimators for parameters are derived from sample data.
- We assess the *frequentist properties* of the estimator from the sampling distribution: that is, is the estimator unbiased? Consistent? Efficient?
- These properties have to do w/*frequency,* i.e., repeated sampling.
  - E.g., if I chose a sample of size n over and over from this population, would my estimator be correct, on average?

Bayesian:

- Parameters are unknown and fixed, but we *model* them as random variables.
  - This model is usually thought of as representing our *degree of belief* in the value of the parameter.
  - We might think (before sampling): I don't know the (fixed) mean height of all CU students, but I believe that it's definitely between 5 and 6 ft. (Prior Belief).
- We use Bayes' theorem to combine our prior belief about a parameter with evidence from a sample.
- The result is a new probability distribution over the parameter. It represents our belief in the value of the parameter after accounting for evidence (the sample). The mean or mode might serve as an est.

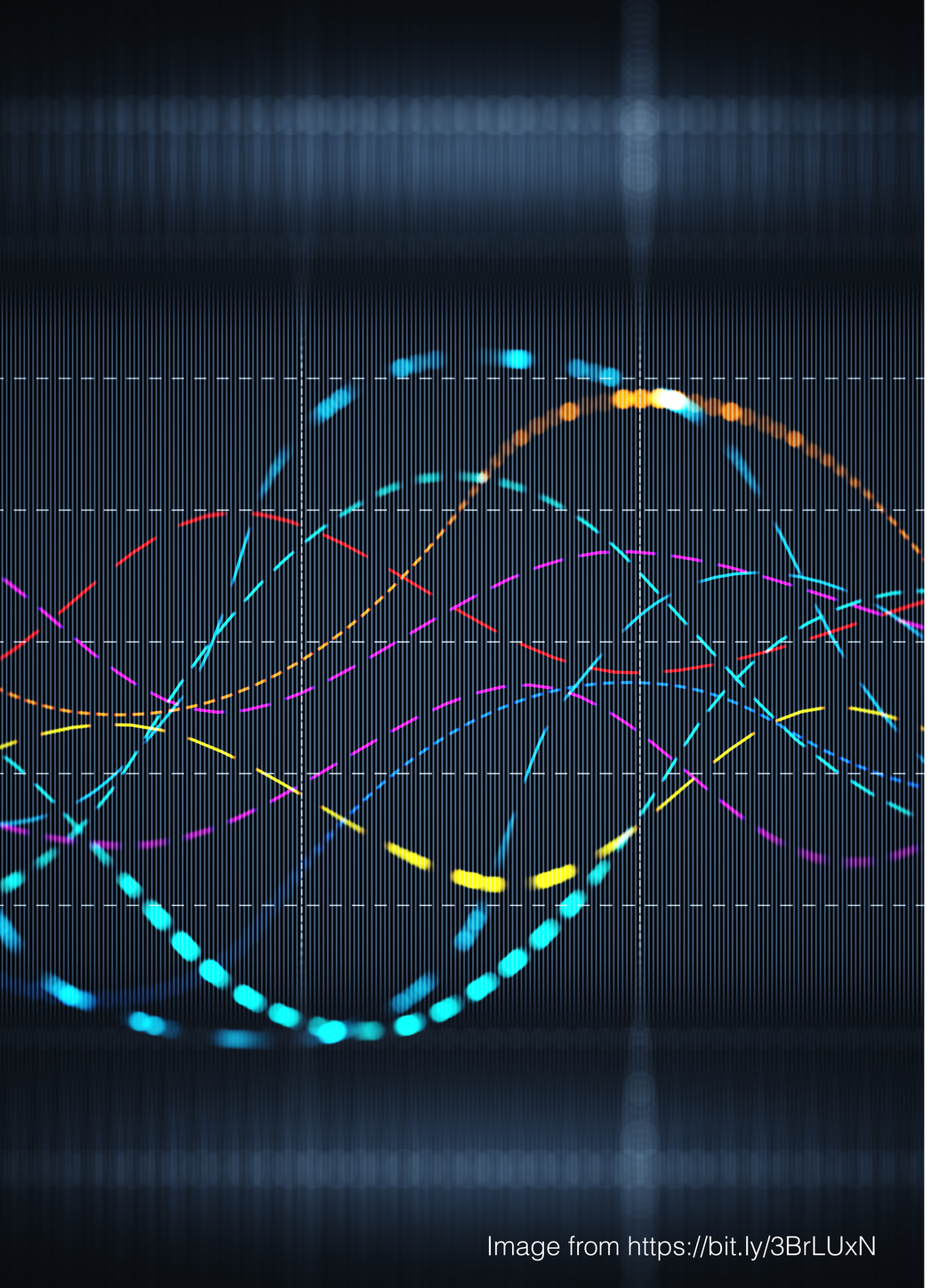


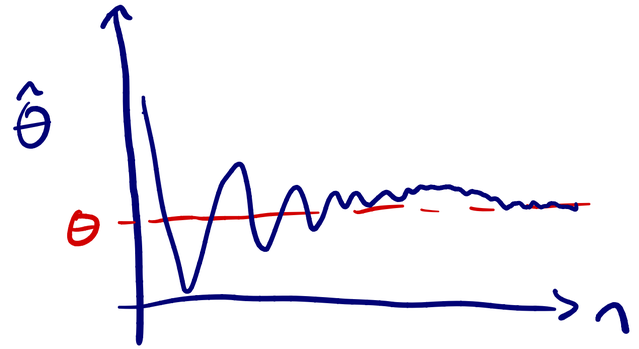
Image from https://bit.ly/3BrLUxN

Here are some important (frequentist) properties of estimators:

An estimator is **consistent** if, as $n \to \infty$,

$$P\left(|\hat{\theta}_n - \theta| \leq \epsilon \right) = 1.$$

"As the sample size gets bigger, the estimator converges to the true value."



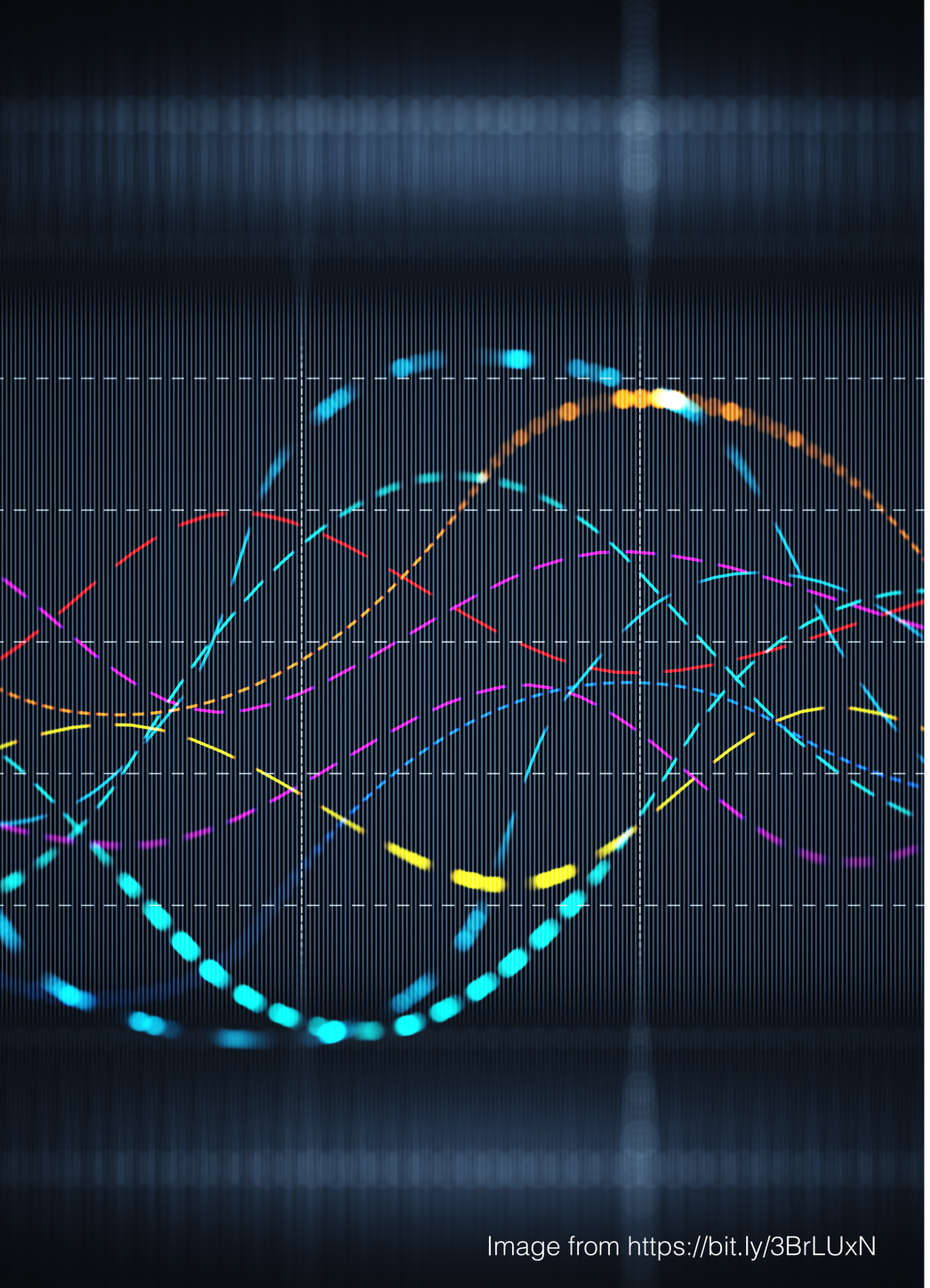


Image from https://bit.ly/3BrLUxN

Here are some important (frequentist) properties of estimators:

The **bias** of an estimator, $\hat{\theta}$, of parameter θ, is defined as $B\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$.

An estimator is **unbiased** if $E\left(\hat{\theta}\right) = \theta$. Unbiased or low-bias estimators are desirable.

An estimator $\hat{\theta}$, of θ, should have a small **variance**, $Var\left(\hat{\theta}\right)$.

The *bias-variance tradeoff* states that we can reduce the variance of an estimator by increasing its bias (or vice-versa).

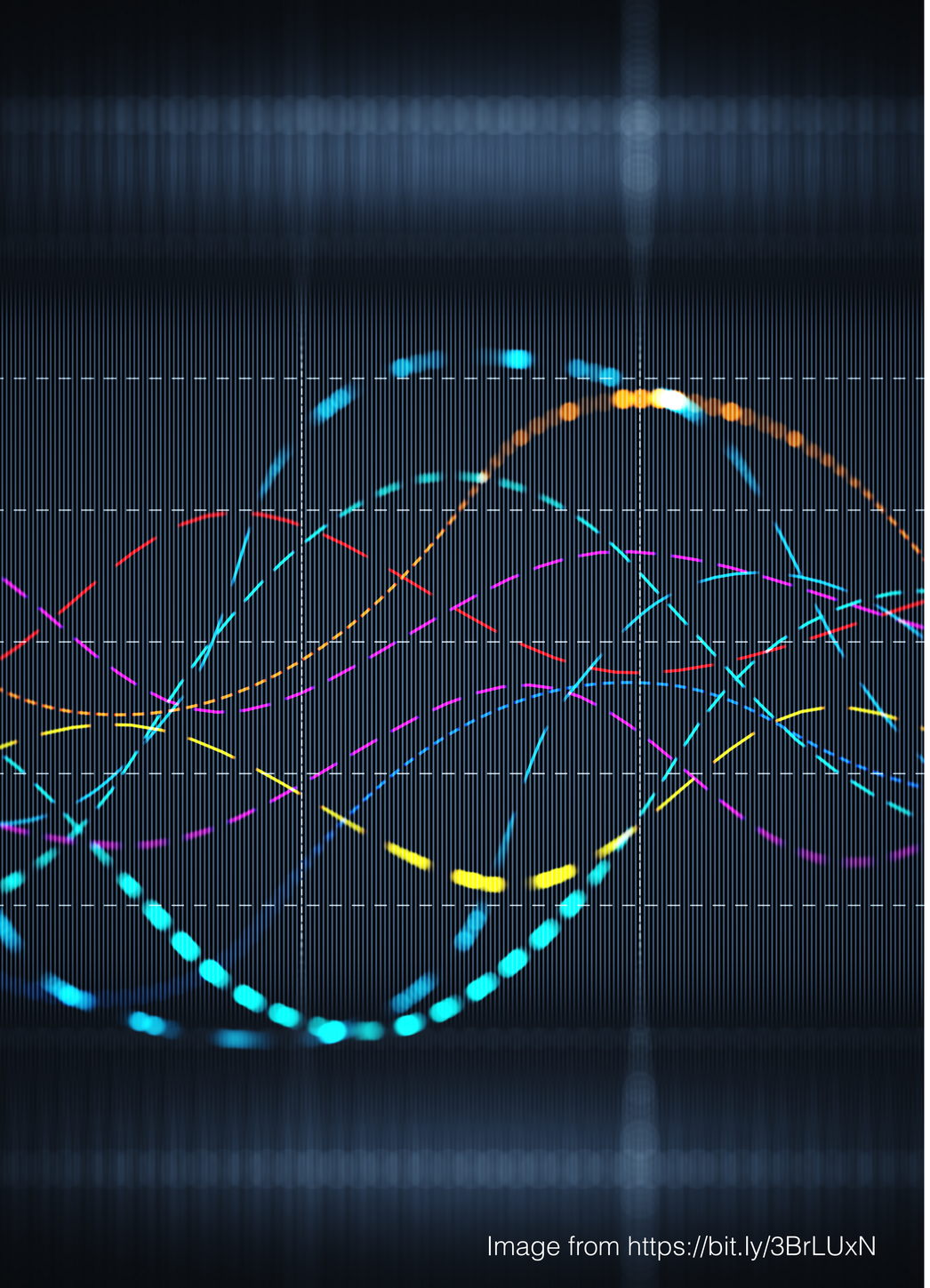


Image from https://bit.ly/3BrLUxN

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a "good" estimator of μ.

    1. $\bar{X}$ is consistent (weak law of large numbers).

    2. $\bar{X}$ is unbiased (prove it!)

    3. $\bar{X}$ has a variance that gets smaller as the sample size gets larger (in fact, $\bar{X}$ has the lowest variance among all unbiased estimators of μ…but we won't prove this here!)

$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

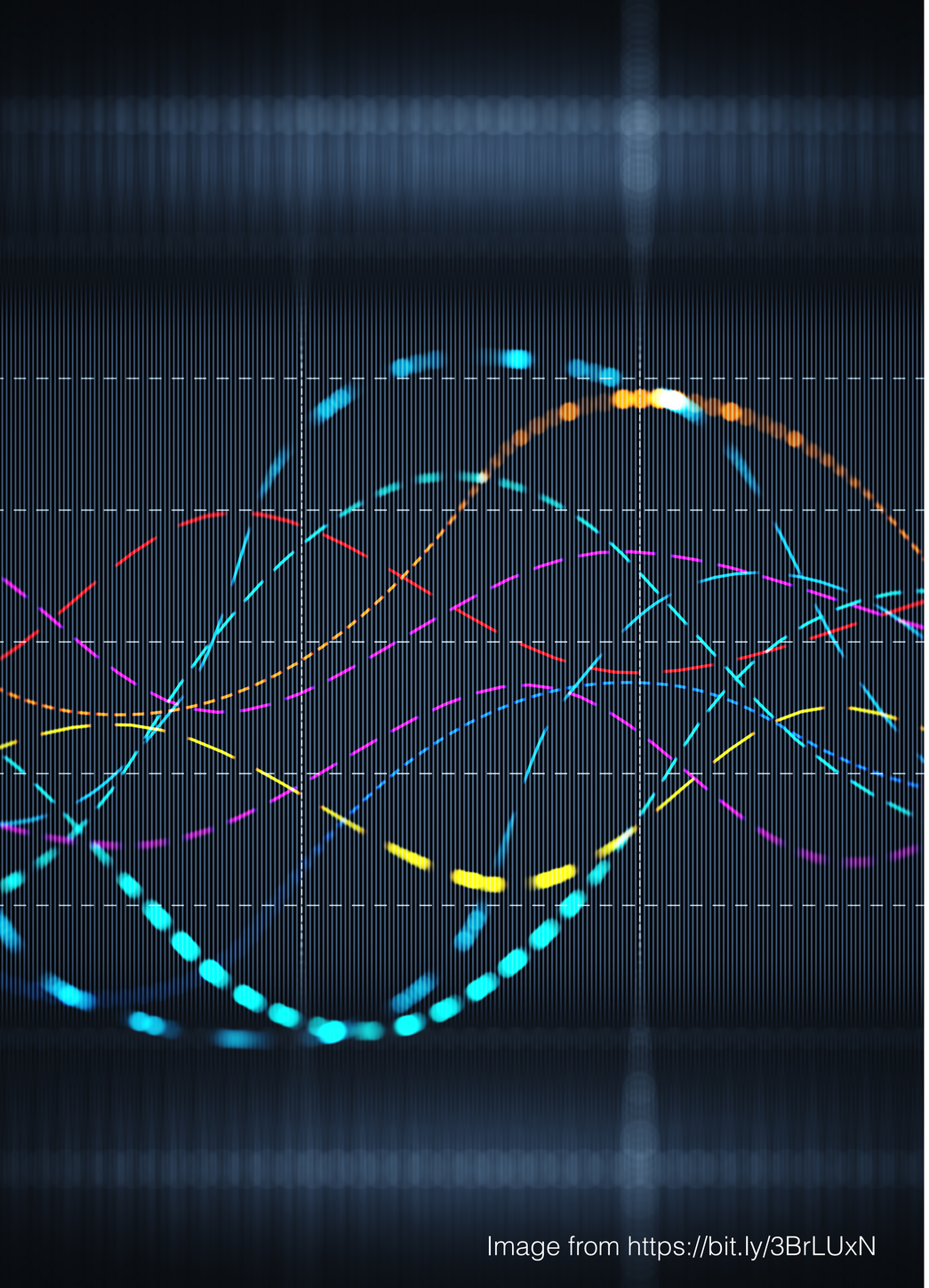


Image from https://bit.ly/3BrLUxN

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

1. $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ is a "good" estimator of σ^2.

    1. S^2 is consistent.

    2. S^2 is unbiased (prove it!)

    3. S^2 has a variance that gets smaller as the sample size gets larger (in fact, S^2 has the lowest variance among all unbiased estimators of σ^2…but we won't prove this here!)

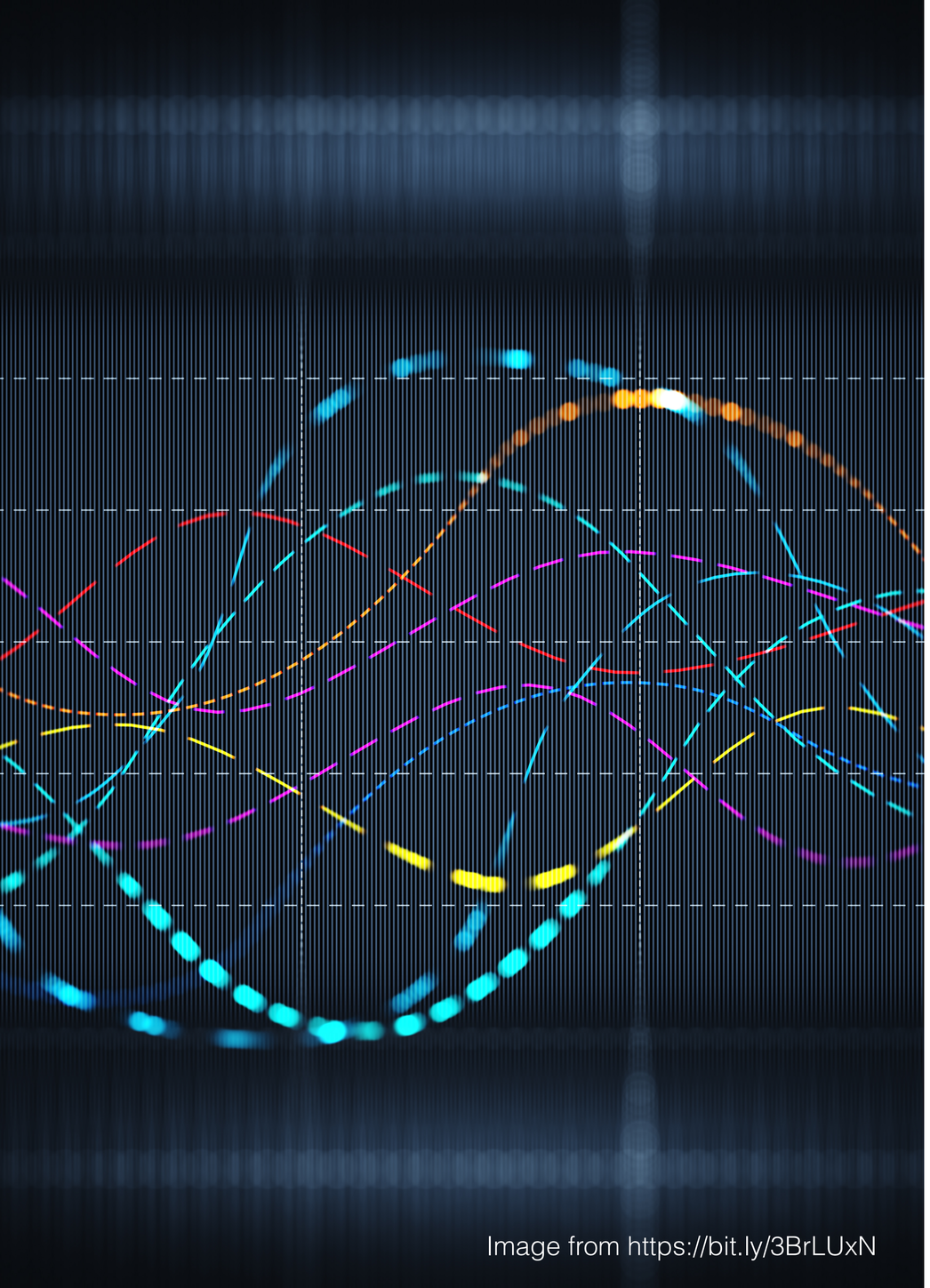


Image from https://bit.ly/3BrLUxN

Maximum likelihood estimation (MLE) is a procedure for finding an estimator of a parameter.

The **maximum likelihood estimator (also, MLE)** of θ is an estimator $\hat{\theta}$ that renders the data most likely.

In order to find the MLE, we'll need to define the likelihood function.

Photo by Eduardo Soares on Unsplash

Motivating Example: Suppose you have an unfair coin where the parameter of interest, $p = P(\text{"Heads"})$, is known to be one of $0.2, 0.3,$ or 0.8.

Let's suppose that we toss the coin twice and use the results to try to estimate p.

What is the distribution modeling the data?

$X_1, X_2 \overset{iid}{\sim} Bin(1, p)$

Photo by Eduardo Soares on Unsplash

Motivating Example: Suppose you have an unfair coin where the parameter of interest, $p = P(\text{"Heads"})$, is known to be one of $0.2, 0.3$, or 0.8. Let's suppose that we toss the coin twice and use the results to try to estimate p.

		(x_1, x_2)			
		(0,0)	(0,1)	(1,0)	(1,1)
p	0.2	0.64	0.16	0.16	0.04
	0.3	0.49	0.21	0.21	0.09
	0.8	0.04	0.16	0.16	0.64

$\hat{p}_{ML} = 0.3$

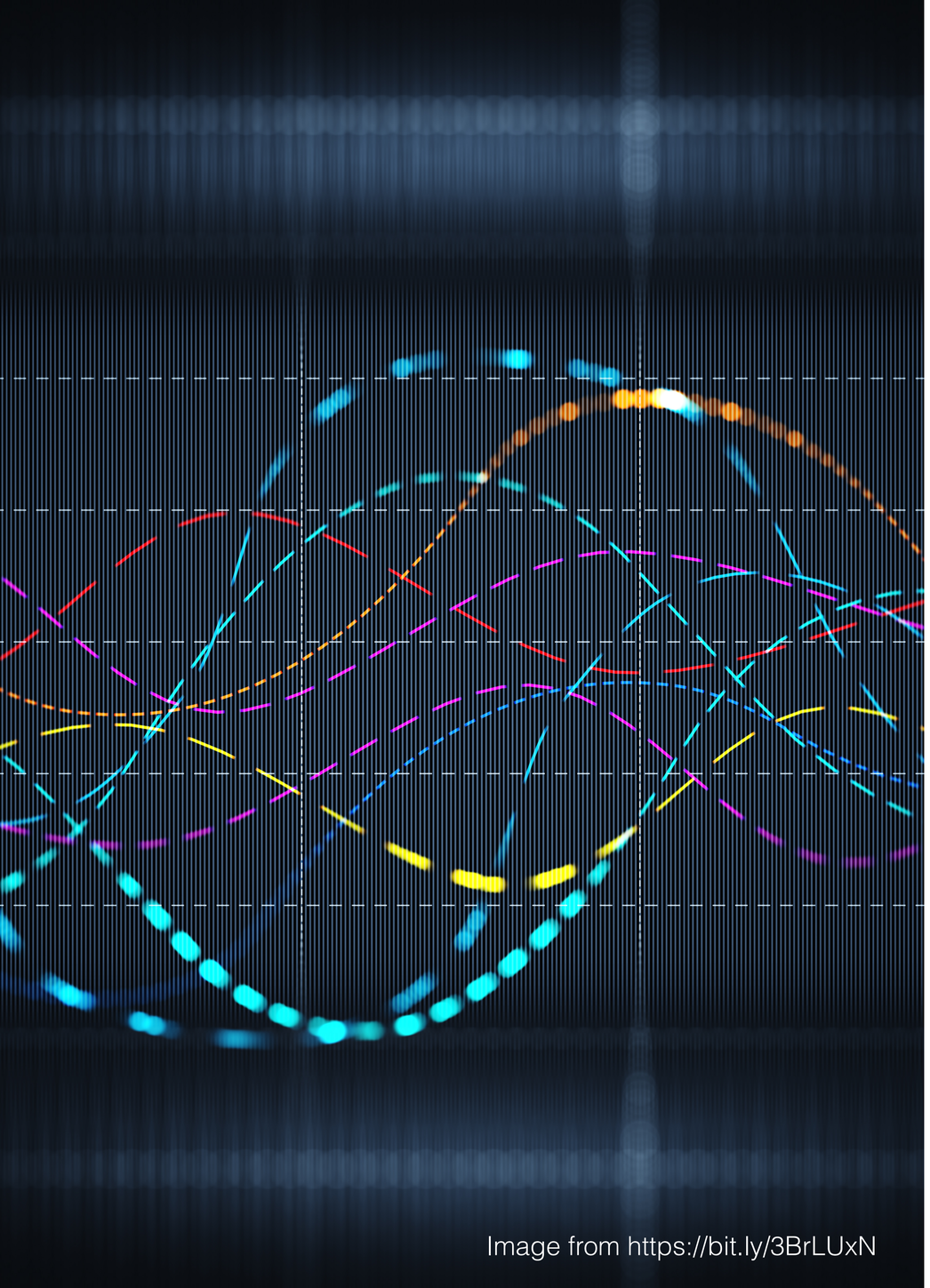


Image from https://bit.ly/3BrLUxN

More generally, the maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function.

A **likelihood function**, denoted by $L(\theta)$, is defined as any function proportional to the joint pdf $f(\mathbf{x}\,;\,\theta)$, but thought of as a function of θ.

Note: for our purposes (when the data are independent), $f(\mathbf{x}\,;\,\theta) = \prod_{i=1}^{n} f(x_i;\theta)$.

A **log likelihood function**, denoted by $\ell(\theta)$, is the log of a likelihood function. $\ell(\theta) = \log(L(\theta))$

Often, it is more convenient to maximize the log likelihood (and the maximizer will always be the same!).

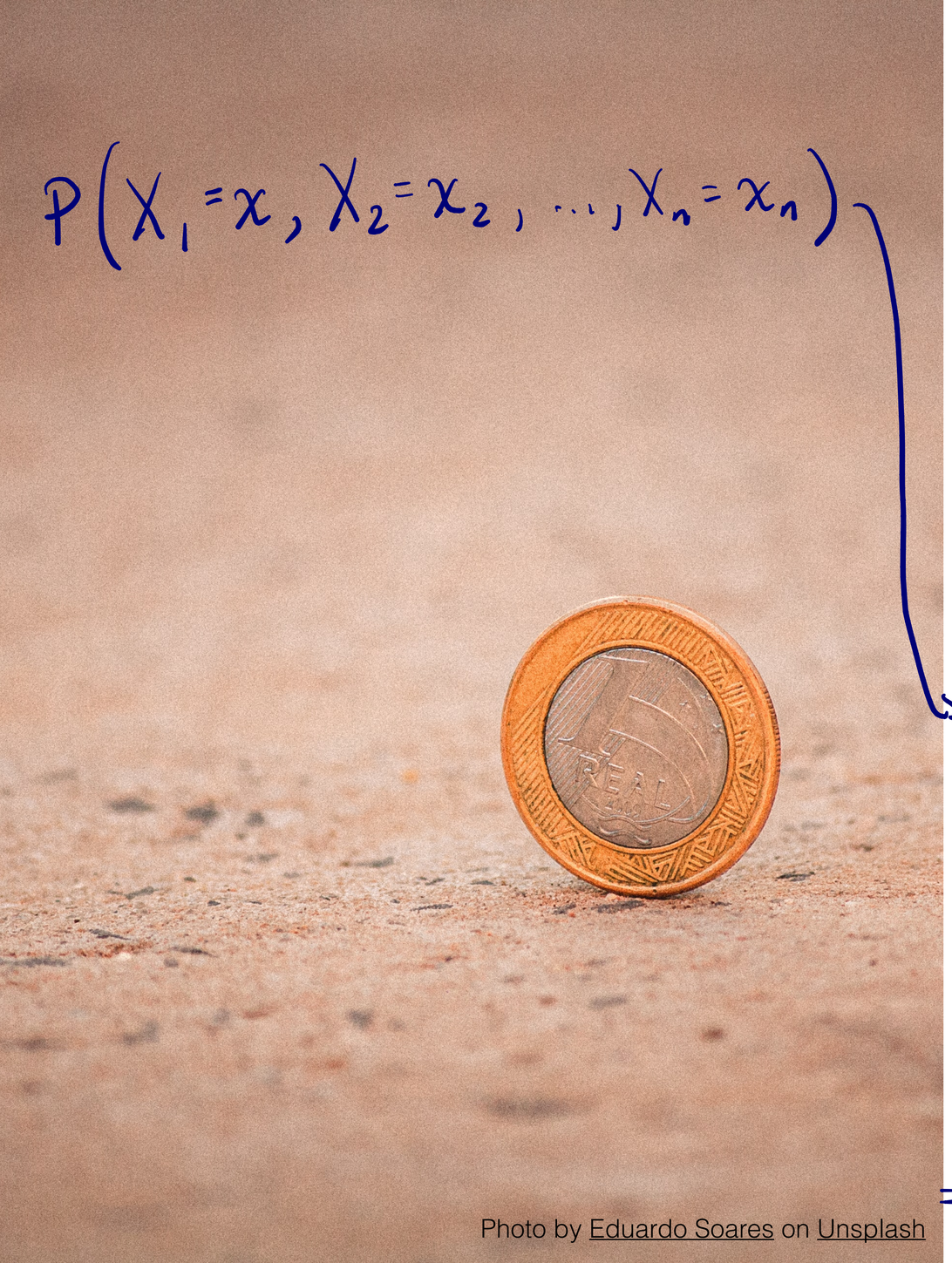


Photo by Eduardo Soares on Unsplash

Suppose we have a coin with unknown probability of heads, p, where $\{p : 0 \leq p \leq 1\}$ (this is called the parameter space of p). We flip the coin n times and record the number of heads. Find the MLE for p. $X_i \sim Bin(1, p)$

1.) pdf/pmf : $f(x_i; p) = P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$

2.) joint pmf/likelihood

$$f(\underline{x}; p) = \prod_{i=1}^{n} f(x_i; p) = p^{x_1} p^{x_2} \cdots p^{x_n} (1-p)^{1-x_1} \cdots (1-p)^{1-x_n}$$

$$= p^{\sum x_i} (1-p)^{n - \sum x_i} = L(p)$$

3.) Log-likelihood : $\ell(p) = \underbrace{\sum x_i}_{n\bar{x}} \log(p) + (n - \sum x_i) \log(1-p)$

4.) max : $\ell'(p) = \frac{n\bar{x}}{p} - \left(\frac{n - n\bar{x}}{1-p}\right) \overset{\text{set}}{=} 0$

$\Rightarrow n\bar{x}(1-p) - np + n\bar{x}p = 0 \Rightarrow n\bar{x} = np \Rightarrow$ MLE $\boxed{\hat{p} = \bar{X}}$

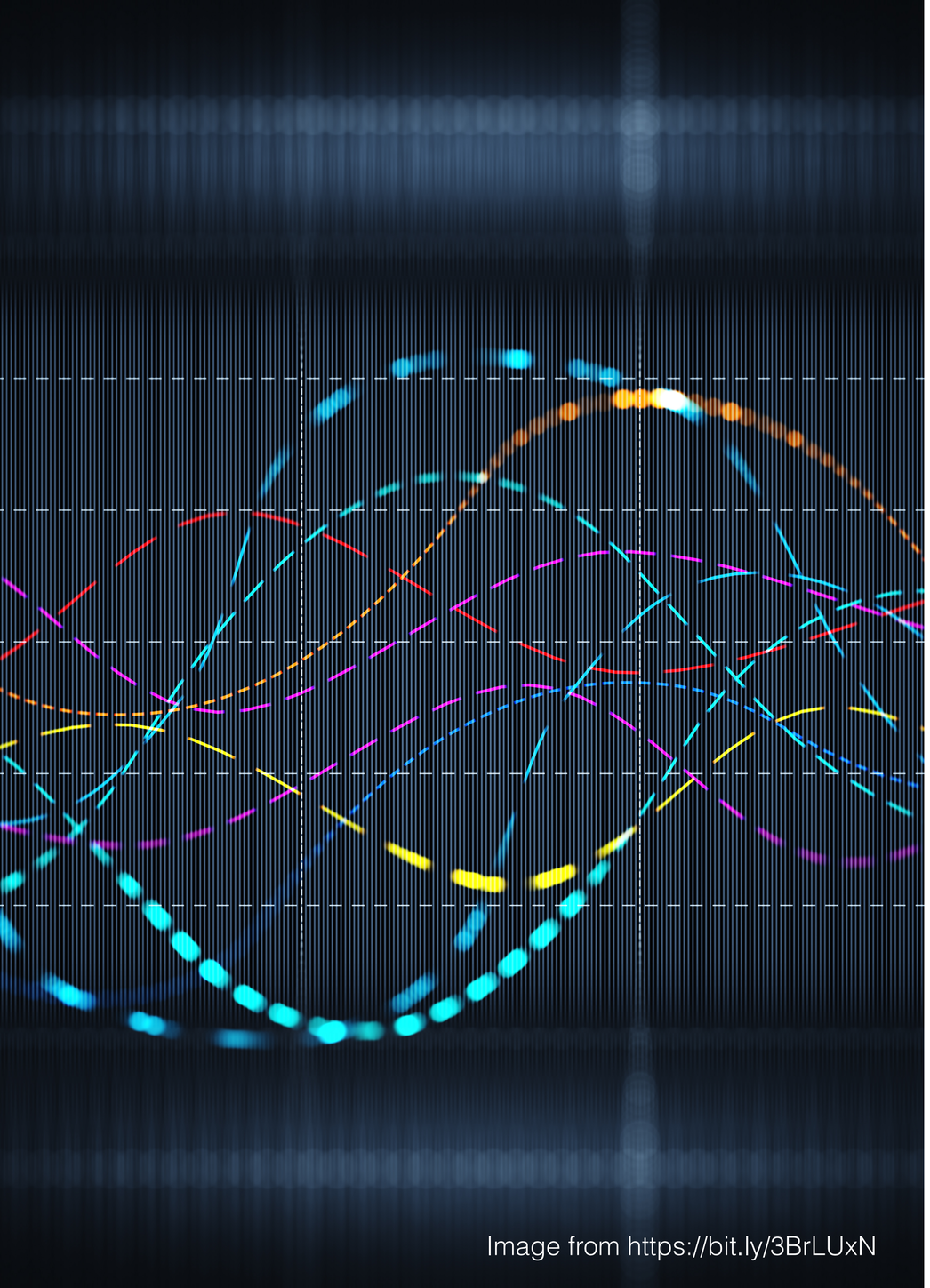


Image from https://bit.ly/3BrLUxN

Properties of the MLE. Let $\hat{\theta}_n$ be the MLE of θ.

1. **Consistency**. As $n \to \infty$,
$P\left(|\hat{\theta}_n - \theta| \leq \epsilon \right) = 1.$, $\epsilon > 0$

2. **Asymptotically unbiased**. As $n \to \infty$
$E\left(\hat{\theta}_n \right) = \theta.$

3. **Efficient**. As $n \to \infty$, $Var\left(\hat{\theta}_n \right)$ is as small as possible

4. **Asymptotic normality.** $\hat{\theta}_n \sim N\left(\theta, \sigma^2_{\hat{\theta}_n} \right)$

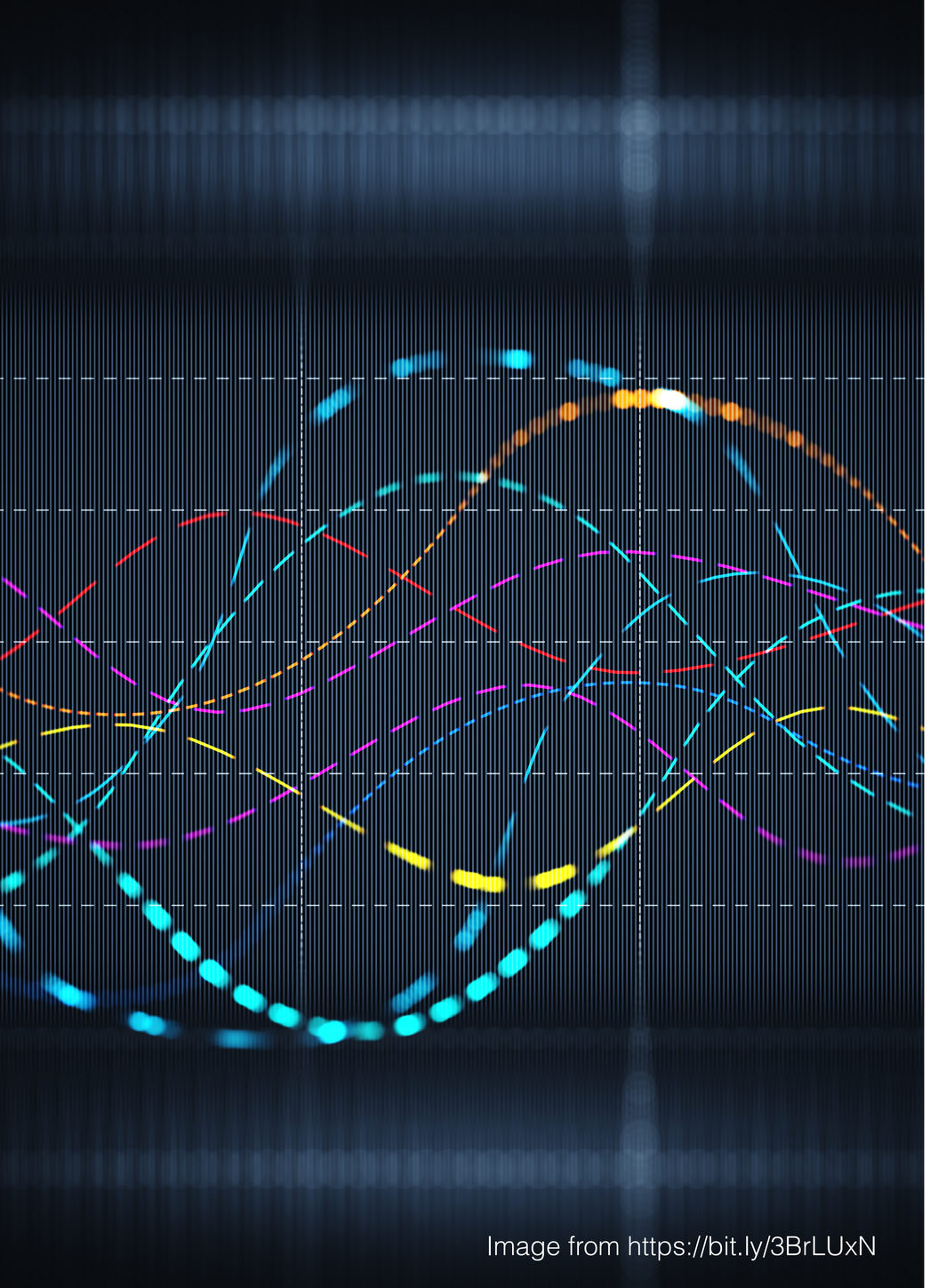


Image from https://bit.ly/3BrLUxN

$X_1, \ldots, X_n \sim Bin(1, p)$

Find the MLE of $\sigma^2 = Var(X_i) = p(1-p)$

We know $\hat{p}_{ML} = \bar{X}$. So, $\hat{\sigma}^2_{ML} = \hat{p}(1-\hat{p}) = \bar{X}(1-\bar{X})$

Invariance property of the MLE.

$\hat{\theta}_n = \hat{\theta}$

Let $\hat{\theta}_n$ be the MLE of θ and let $\tau(\theta)$ be some function of θ. Then the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$. The MLE of λ is $\hat{\lambda} = \frac{1}{\bar{X}}$. Find the MLE of $\mu = E(X_i)$. $= \frac{1}{\lambda} = \tau(\lambda)$

$\widehat{\tau(\lambda)} = \tau(\hat{\lambda}) = \frac{1}{\hat{\lambda}} = \frac{1}{1/\bar{x}} = \bar{X} = \hat{\mu}$

invariance

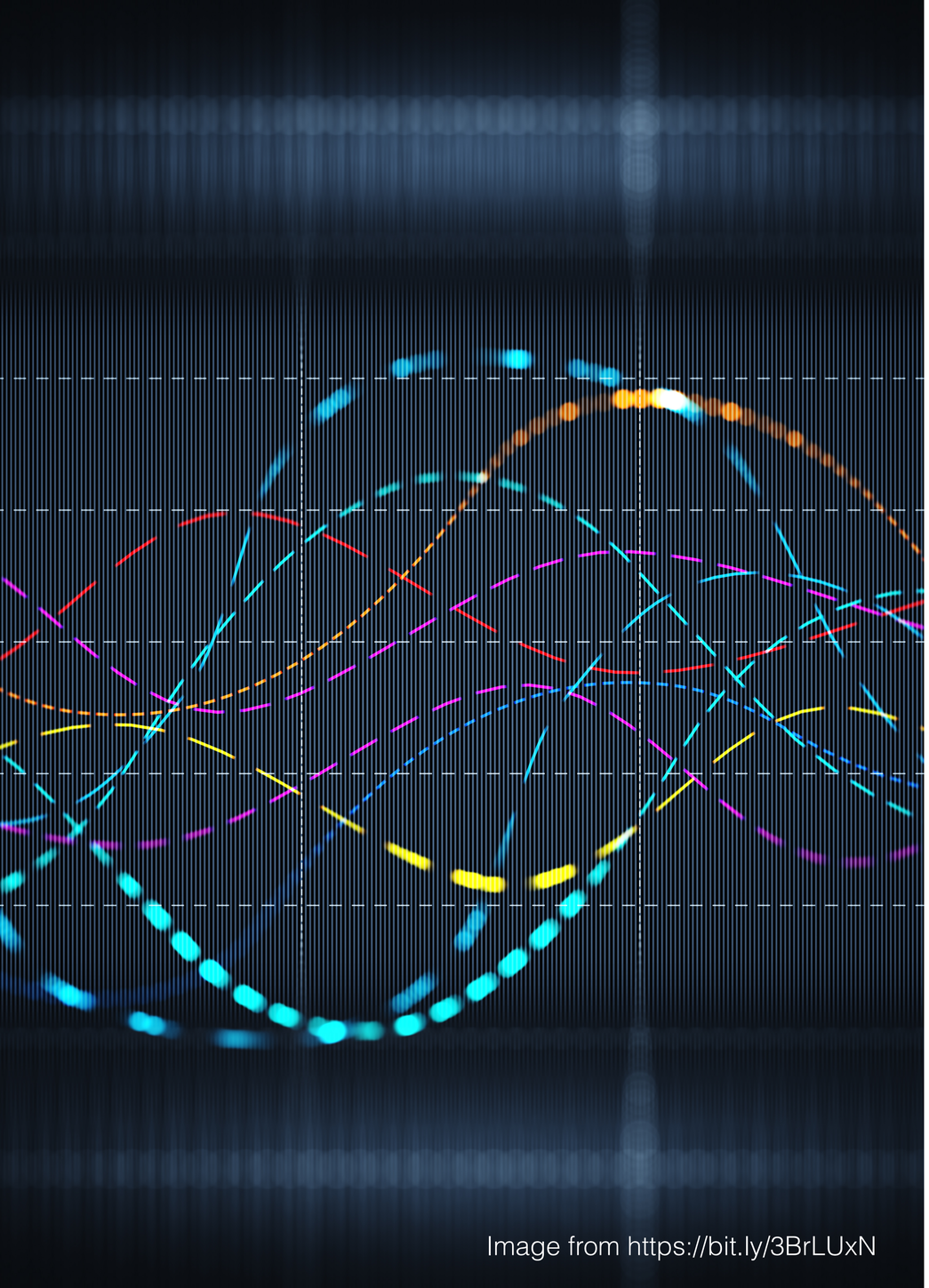


Image from https://bit.ly/3BrLUxN

All that to say, the MLE is pretty good!

Photo by Anne Nygård on Unsplash

Point estimators are nice, but they do not quantify uncertainty: how close is $\hat{\theta}$ to θ? It's not clear. All that we know is that the procedure used to produce $\hat{\theta}$ is "good" (e.g., unbiased, consistent, etc.).

An **interval estimator** is a range of values used to estimate θ.

"Intuitively, for a given confidence level $(1-\alpha)\times 100\,\%$, the shorter the confidence interval, the more informative it is for learning about the true value of the unknown parameter θ."

- Aris Spanos, Prob. Theory and Statistical Inference (501)

Photo by Anne Nygård on Unsplash

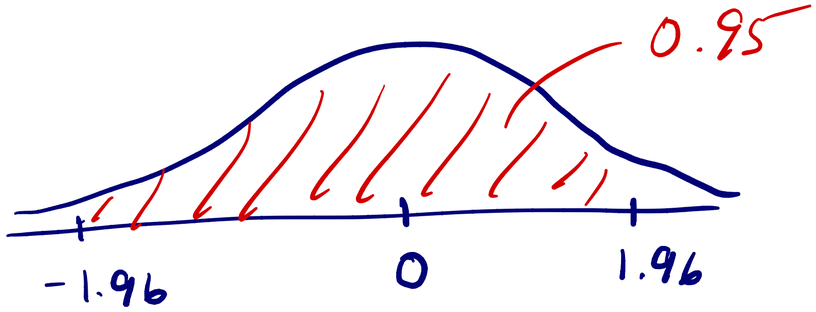
$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

Let's start with a simple example. Suppose that we have a simple random sample of n measurements from a normal population with unknown mean μ, and known standard deviation σ.

Standardizing the sample mean by first subtracting its expected value and then dividing by its standard deviation yields the standard normal variable:

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

Photo by Anne Nygård on Unsplash



CI at the 95% conf. level is $(L(X), U(X))$

Because the area under the standard normal curve between -1.96 and 1.96 is 0.95, we know:

$$P(-1.96 \le Z \le 1.96) = 0.95$$

$$\Rightarrow P\left(-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95$$

$$\Rightarrow P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Rightarrow P\left(\underbrace{\bar{X}-1.96\frac{\sigma}{\sqrt{n}}}_{L(X)} \le \mu \le \underbrace{\bar{X}+1.96\frac{\sigma}{\sqrt{n}}}_{U(X)}\right) = 0.95$$

Photo by Anne Nygård on Unsplash

The interval

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \left(L(X), U(X)\right)$$

Is called the $95\,\%$ confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies. So, the interval itself is a random interval.

Photo by Anne Nygård on Unsplash

The CI interval is centered at $\bar{X}$ and extends $1.96\frac{\sigma}{\sqrt{n}}$ to each side of $\bar{X}$.

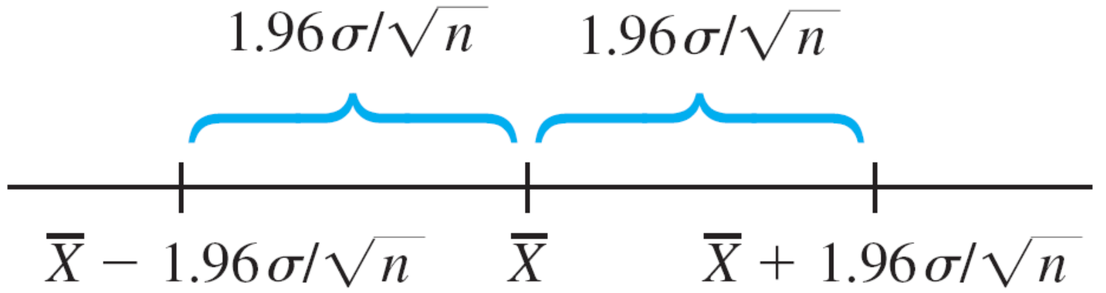


Photo by Anne Nygård on Unsplash

const

"We are 95 % confident that the true parameter is in this interval."

↑ random of many samples

What does that mean?

pre-sample: $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ covers μ w/ prob. 0.95

post-sample: $5 \pm 1.96/5$ ← fixed interval, so no probabalistic claims can be made


Photo by Luke Stackpoole on Unsplash

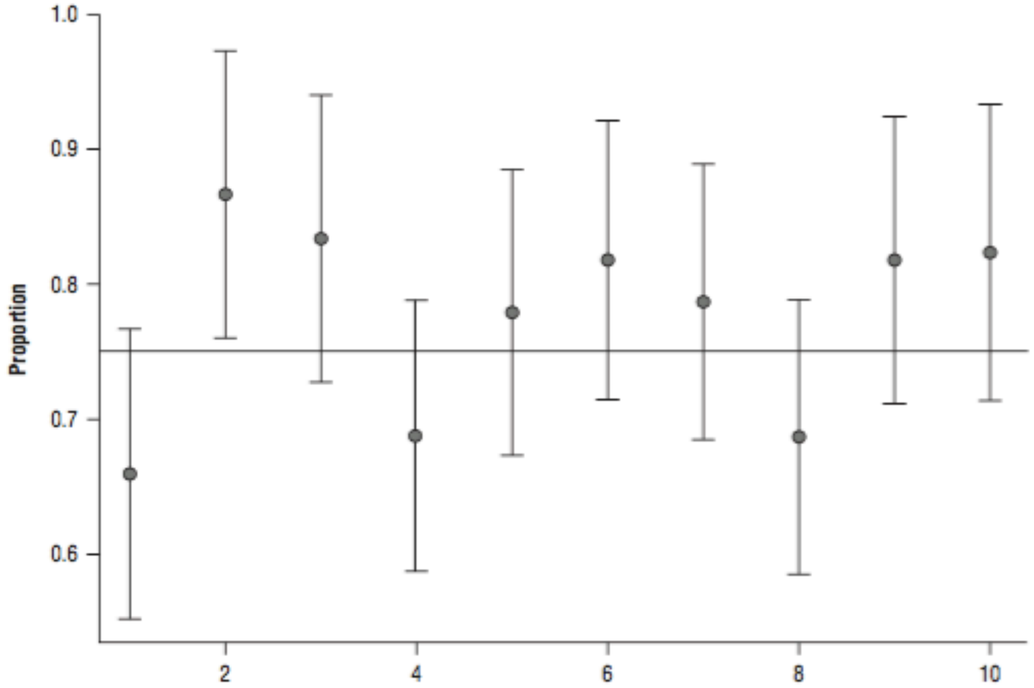
A correct interpretation of "$95\,\%$ confidence" relies on the long-run relative frequency interpretation of probability.

In repeated sampling, $95\,\%$ of the confidence intervals obtained from all samples will cover μ. The other $5\,\%$ of the intervals will not.

The confidence level ($95\,\%$) is not a statement about any particular interval. It is a statement about the reliability of the pre-data confidence interval formula.

Photo by Luke Stackpoole on Unsplash

Figure 1: Confidence Interval

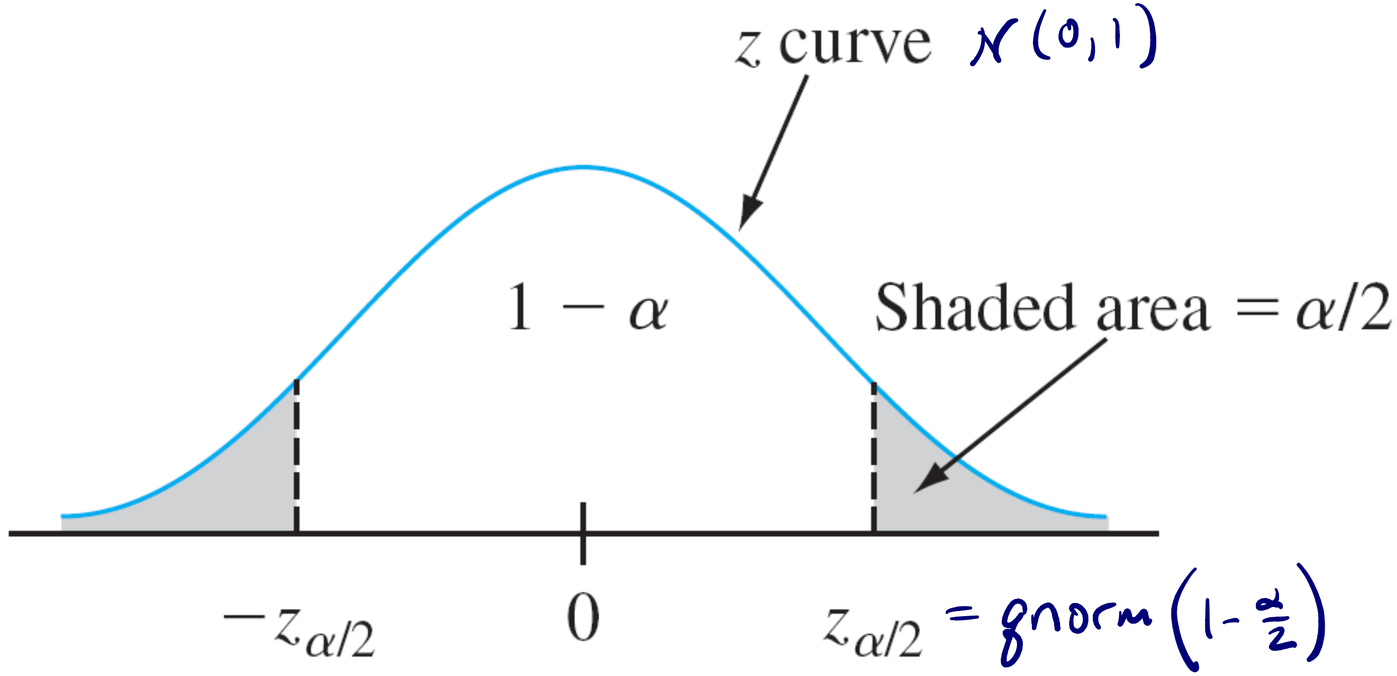


Note: Suppose that the true proportion of believers in climate change among French citizens is 0.75, as represented by the horizontal black line near the middle. This figure shows ten 90% confidence intervals used to estimate the proportion of believers in climate change among French citizens. Each circle represents a point estimate, p, calculated from a different sample of n French citizens; for each confidence interval, the length of the vertical line is twice the margin of error, E, for that interval. Notice that the second interval fails to cover the true proportion. For the 90% confidence interval procedure, it is expected that about one in every ten intervals will fail to cover the true proportion.

Photo by Luke Stackpoole on Unsplash

Robust misinterpretation of confidence intervals (https://bit.ly/30EEXft)

Photo by Anne Nygård on Unsplash



When $\alpha = 0.05$, $z_{\alpha/2} = 1.96$

Photo by Anne Nygård on Unsplash

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2)$$

σ^2 Known

A $(1-\alpha)\times 100\,\%$ **confidence interval** for the mean μ when the value of σ is known is given by:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Photo by Anne Nygård on Unsplash

Example: A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

(a) Calculate a confidence interval for true average hole diameter using a confidence level of $90\,\%$.

$\alpha = 0.1$, $z_{\alpha/2} = z_{0.05} = \text{qnorm}(0.95) \approx 1.64$

$\bar{x} \pm 1.64\left(\frac{\sigma}{\sqrt{n}}\right) = (5.400, 5.452)$

(b) What about the $99\,\%$ confidence interval?

$\alpha = 0.01 \Rightarrow z_{\alpha/2} = z_{0.005} \approx 2.54$

$CI_{99\%} = (5.385, 5.467)$

(c) What are the advantages and disadvantages to a wider confidence interval?

~~$P(\mu \in (5.400, 5.452)) = 0.9$~~

Photo by Anne Nygård on Unsplash

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with an unknown average response time of μ, and a known standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to use it to estimate μ. We can perform a task n times and measure the response times. What sample size n is necessary to ensure that the resulting $95\,\%$ CI has a width of (at most) 10 units? milliseconds

$$w = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 10 \iff 10 = 2(1.96)\left(\frac{25}{\sqrt{n}}\right)$$

$$\Rightarrow n \approx 96.04$$

$$\Rightarrow \text{take a sample of size } n = 97$$

Photo by Anne Nygård on Unsplash

When we don't know the population standard deviation, σ, we will need to work with the **sample standard deviation** s. Remember that s is calculated as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s = \sqrt{s^2}$$

With this, we instead work with the standardized random variable:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim ?$$

What is the distribution of this standardized random variable?

Photo by Anne Nygård on Unsplash

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z$$

Previously, there was randomness only in the numerator of Z, by virtue of the estimator $\bar{X}$.

In the new standardized variable, both $\bar{X}$ *and* s are random.

$$T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

When n is large, the extra variability does not matter much and can be ignored.

When n is smaller, the distribution of this new variable should be wider than the normal to reflect the extra uncertainty.

Photo by Anne Nygård on Unsplash

$$\tau = S(\mathbf{X}) = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim ?$$

When the sample size, n is…

"Large" (roughly: $n \geq 30$)

$s(\mathbf{X}) \overset{approx}{\sim} N(0,1)$

So, a "normal" confidence interval is appropriate!

Small (roughly: $n < 30$)

$s(\mathbf{X})$ is not "normal", so we need to figure out its distribution.

	n >= 30	n < 30
Underlying normal distribution	σ known ✓	σ known ✓
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known ✓	σ known
	σ unknown	σ unknown

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown ✓	σ unknown

Photo by Jonathan Farber on Unsplash

Let $\bar{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ (and unknown variance σ^2), the random variable

$$T = S(\underset{\sim}{X}) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

has a probability distribution called a ***t*-distribution** with $n - 1$ degrees of freedom (df).

t-Distribution

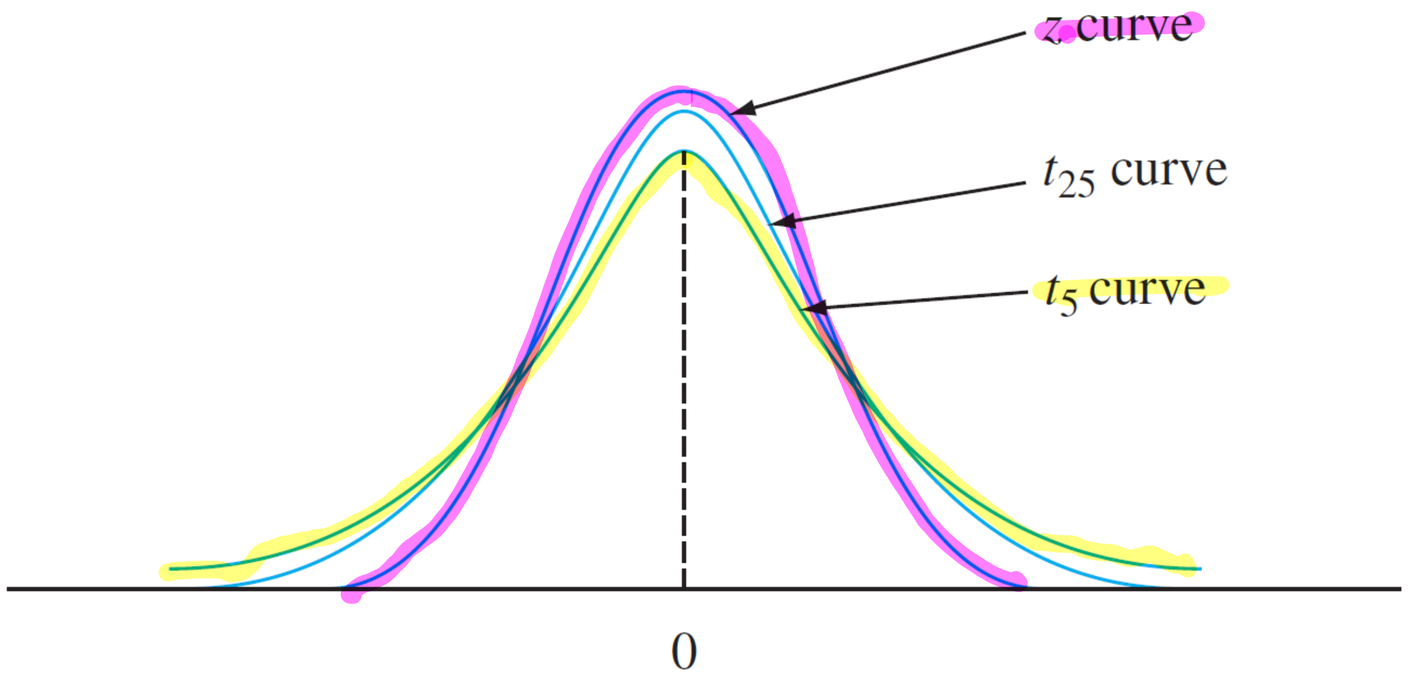


Photo by Jonathan Farber on Unsplash

Let t_ν denote the t-distribution with ν df.

1. Each t_ν curve is bell-shaped and centered at 0.
2. Each t_ν curve is more spread out than the standard normal (z) curve.
3. As ν increases, the spread of the corresponding t_ν curve decreases.
4. As $\nu \to \infty$, the sequence of t_ν curves approaches the standard normal curve.

Photo by Jonathan Farber on Unsplash

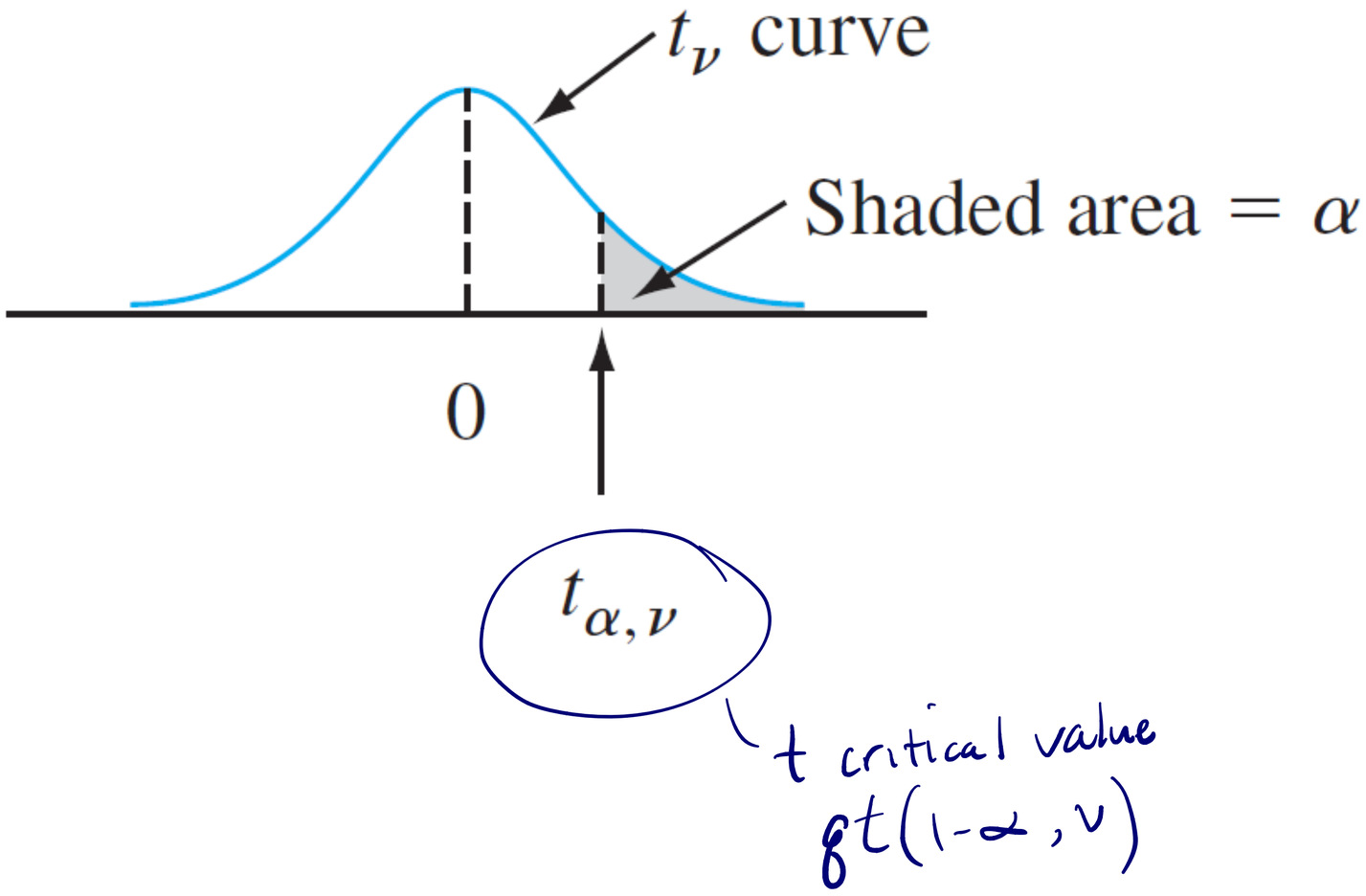


Photo by Jonathan Farber on Unsplash

The probabilities of t curves are found in a similar way as the normal curve.

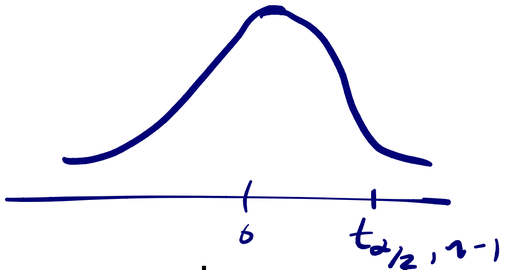
Example: obtain $t_{.05,15}$

$qt(1-0.05, 15) \approx 1.75$

($qnorm(1-0.05) \approx 1.64$)

Photo by Jonathan Farber on Unsplash

$$P\left(-t_{\alpha/2,n-1} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2,n-1}\right) = 1-\alpha$$



Let $\bar{X}$ and S^2 be the sample mean and sample ~~standard deviation~~ variance computed from the results of a random sample from a normal population with mean μ. Then a $(1-\alpha)\times 100\,\%$ t-confidence interval for the mean μ is:

$$\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \;,\; \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

Photo by Jonathan Farber on Unsplash

Example: Suppose that the GPA measurements for 23 students follow a normal distribution. The sample mean is 3.146. The sample standard deviation is 0.308. Calculate a $90\,\%$ confidence interval for the mean GPA.

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

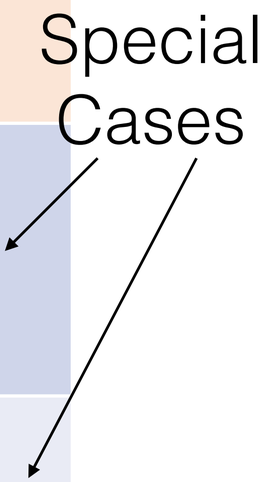


Photo by Anne Nygård on Unsplash

When $n < 30$ and the underlying distribution is unknown, we have to:

1. Make a specific assumption about the form of the population distribution and derive a CI based on that assumption.

2. Use other methods (such as bootstrapping) to make reasonable confidence intervals.