Math for ML

Week 6.1: Central Limit Theorem, Distributions, and the MLE

By: Samuel Deng

Logistics & Announcements

- PS4. due Tomorrow, Tuesday 11:59 PM.
- PS5. due next Tues. 11:59 PM.
- Project due next Mon. 11:59 PM.

That's it :)

☆ COURSE EVALUATIONS!

Lesson Overview

Gaussian Distribution. We define perhaps the most important "named" probability distribution, the Gaussian/"Normal" distribution, and go over some key properties.

Central Limit Theorem. We state and prove the central limit theorem, the statement that the sample average of *many* independent random variables converges in distribution to the Gaussian. It doesn't matter what distribution those random variables take!

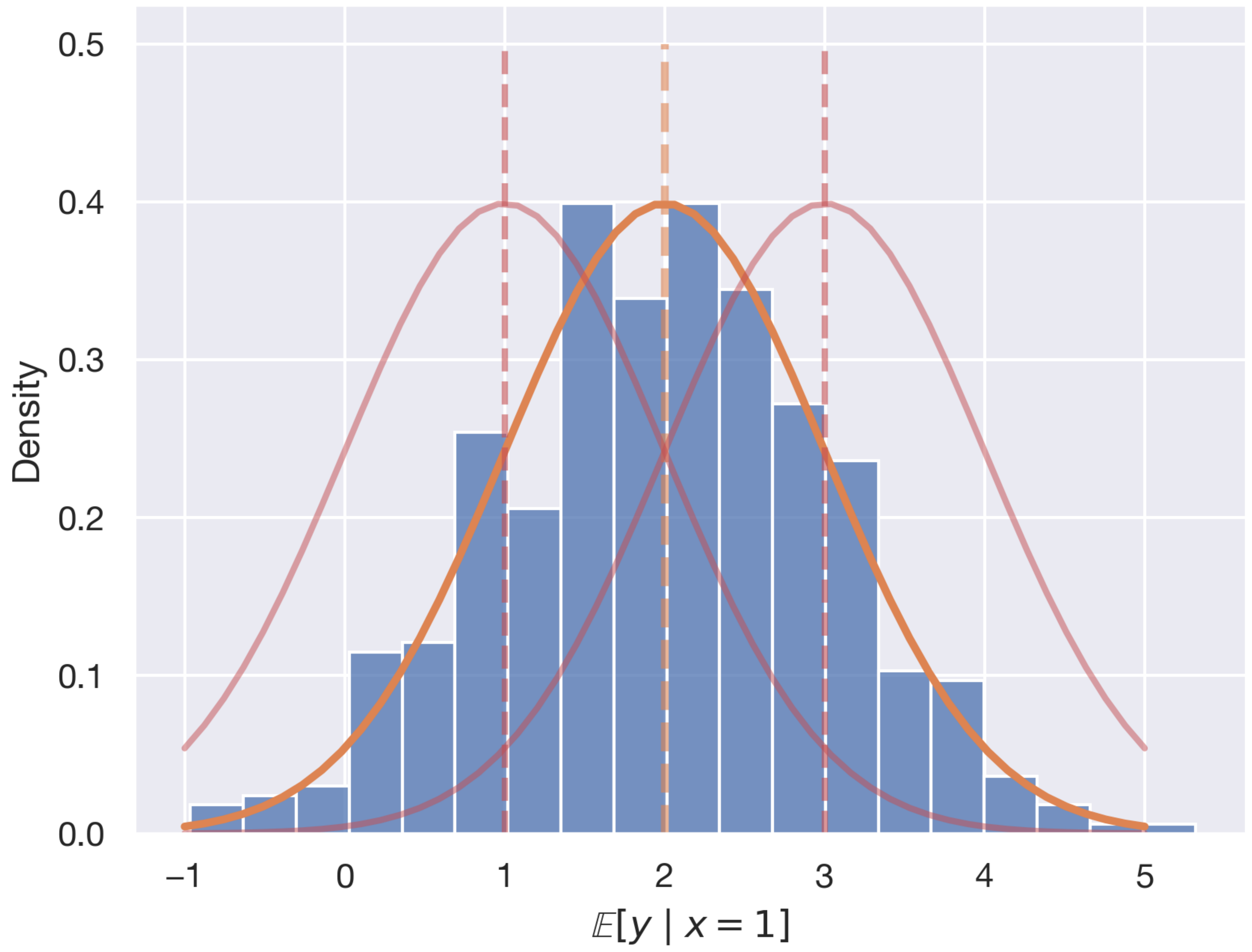
"Named" Distributions. We review other common "named" distributions for discrete and continuous random variables.

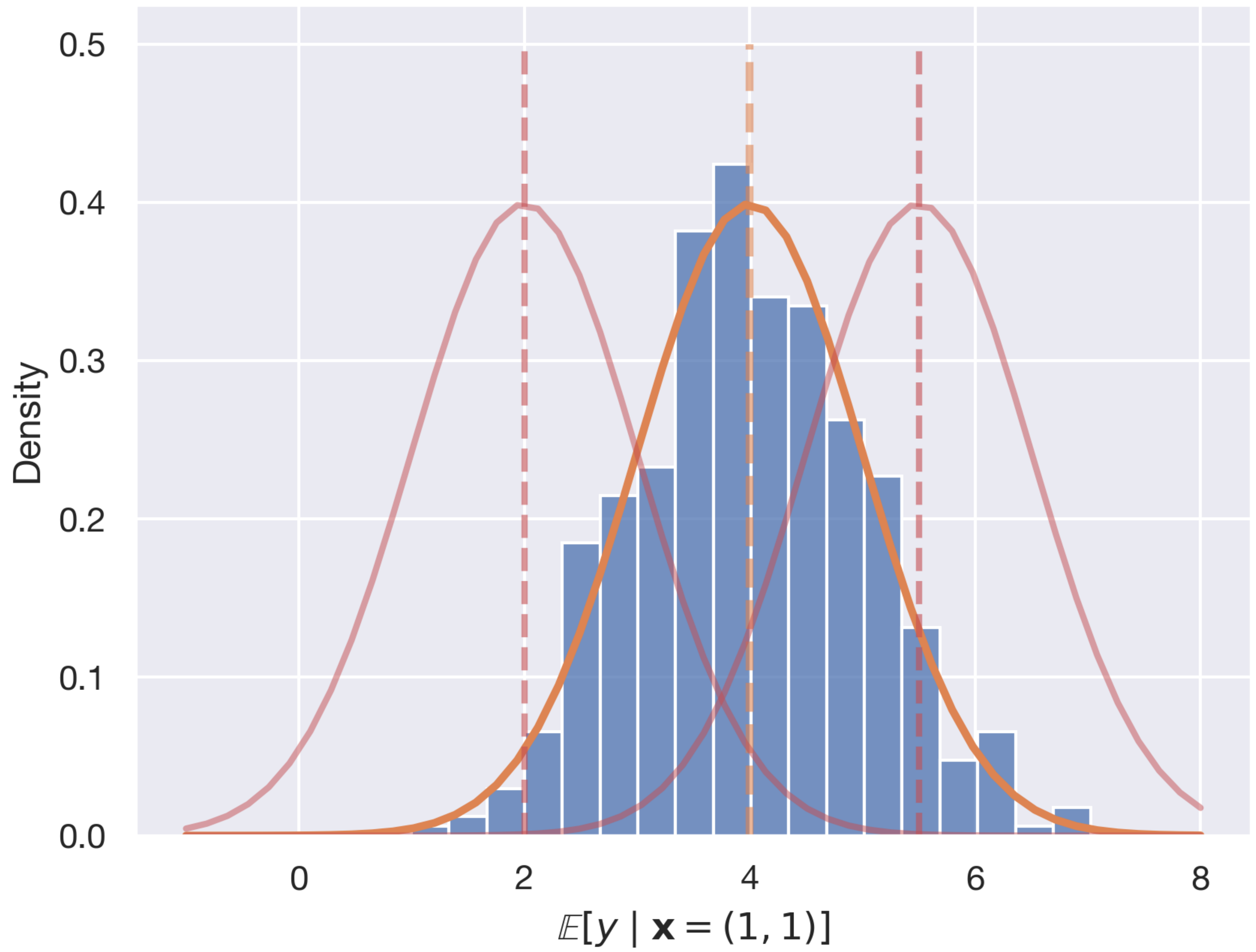
Maximum likelihood estimation. We define maximum likelihood estimation (MLE), a statistical/probabalistic perspective towards finding a well-generalizing model for data.

MLE and OLS. We explore the connection between MLE and OLS by defining the Gaussian error model. In this model, MLE and OLS correspond.

Lesson Overview

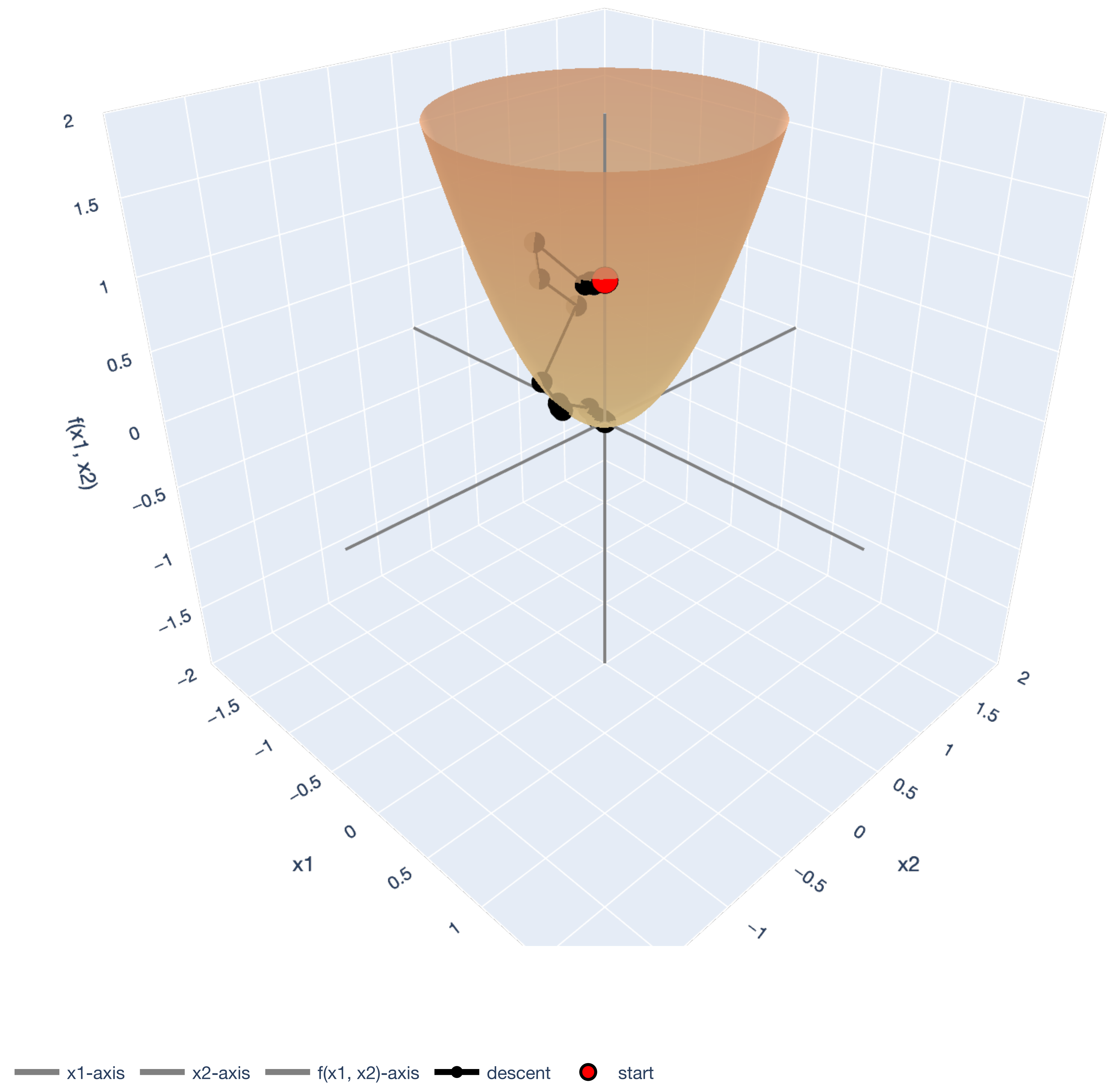
Big Picture: Least Squares

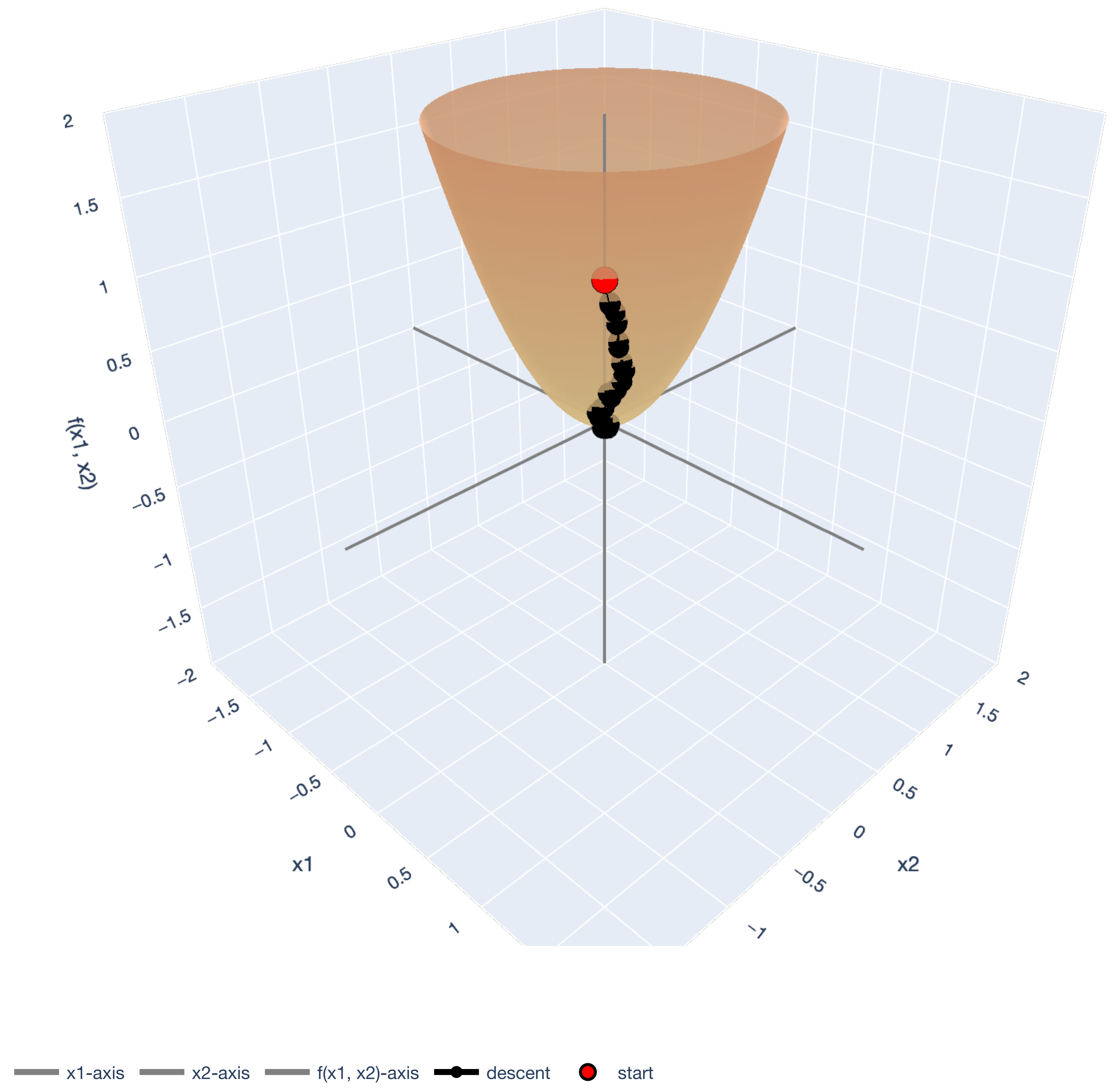




Lesson Overview

Big Picture: Gradient Descent



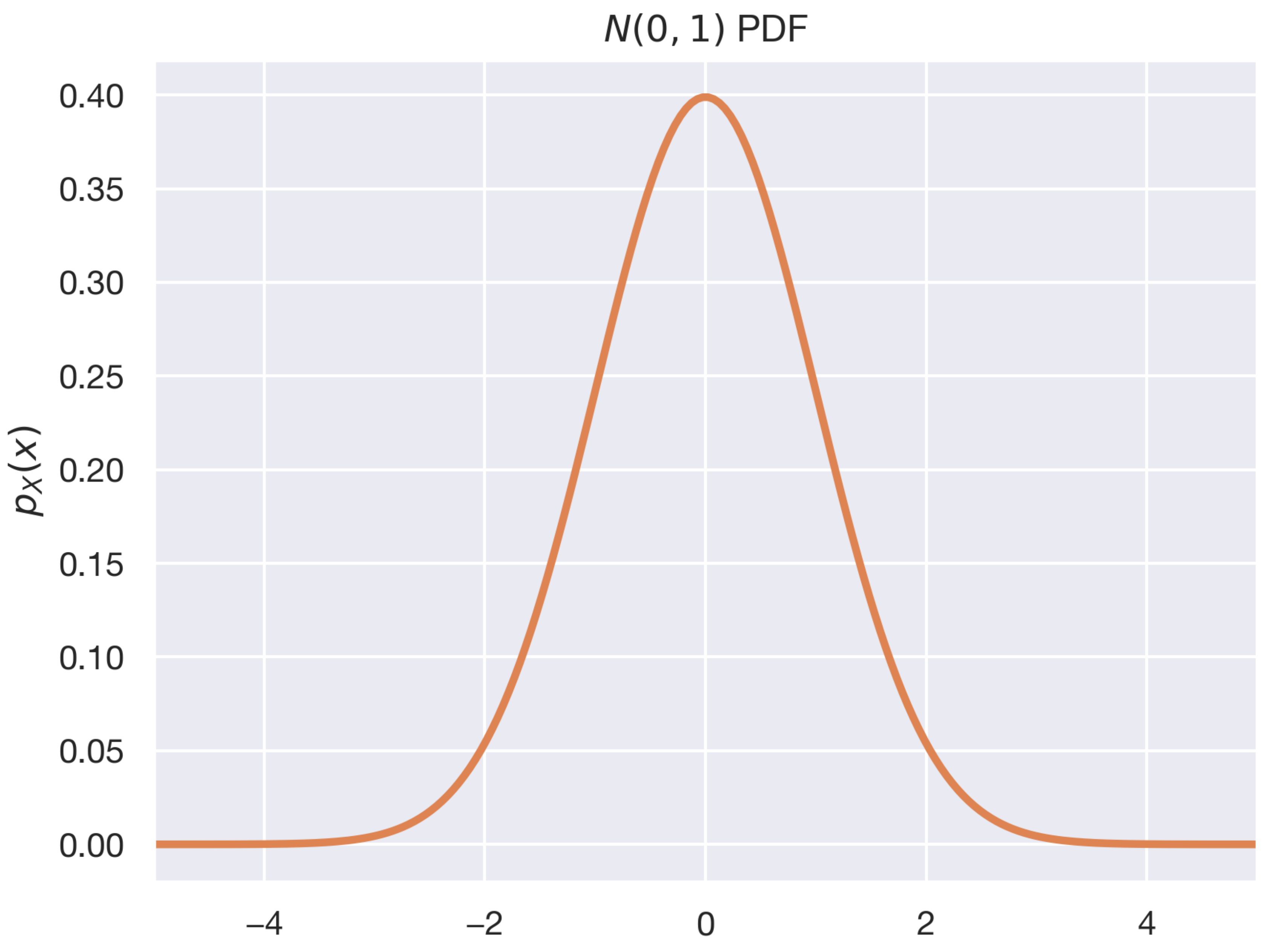


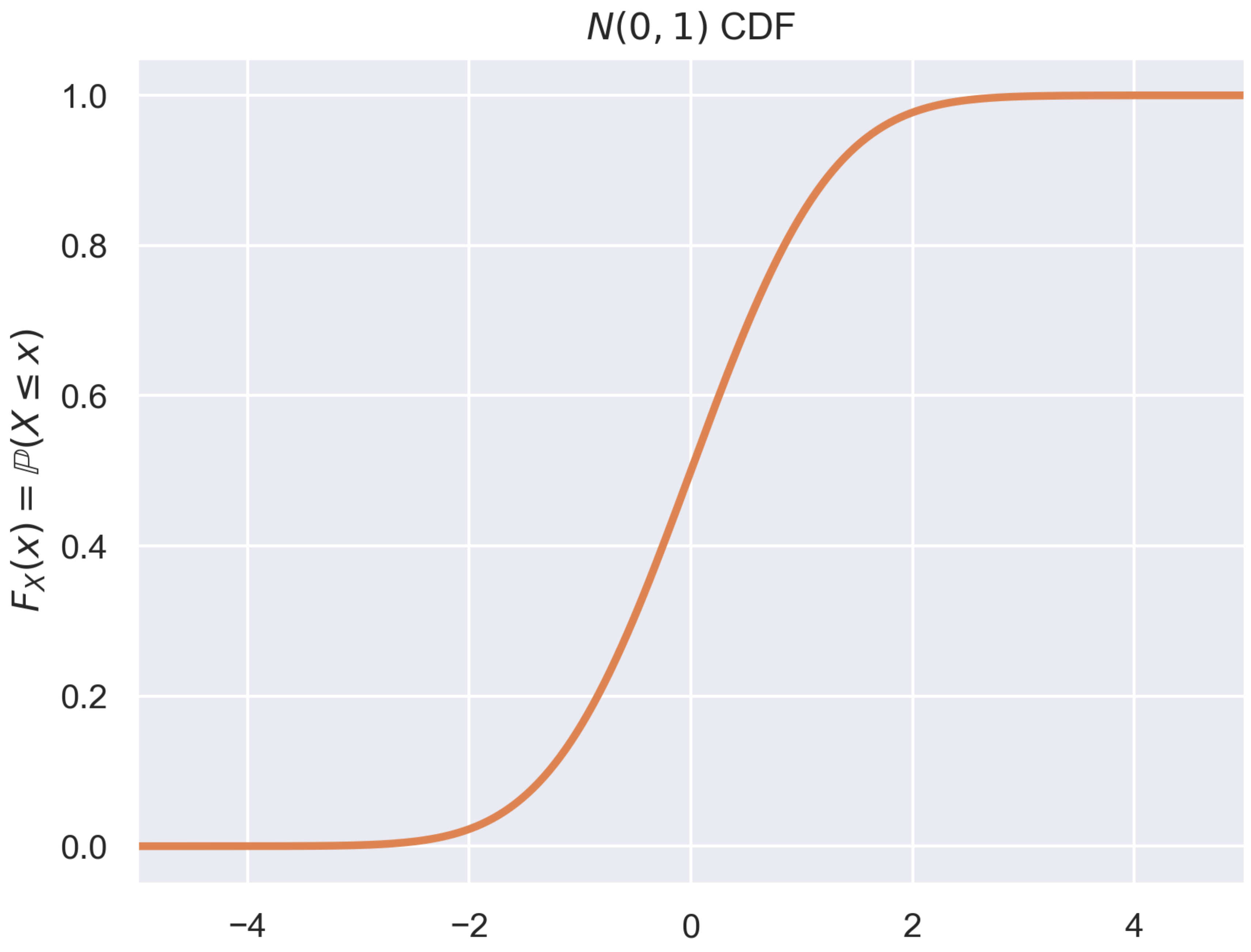
The Gaussian Distribution

Definition and Properties

The Gaussian Distribution
Intuition and Shape

The ***Gaussian/Normal*** distribution with parameters μ and σ has a “bell-shaped” PDF centered at μ and “spread” depending on the parameter σ.





The Gaussian Distribution

Standard Gaussian Definition

A random variable Z has a ***standard Gaussian/Normal distribution*** denoted $Z \sim N(0,1)$ if it has PDF:

$$p_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \text{ for all } z \in \mathbb{R}.$$

This random variable has mean $\mathbb{E}[Z] = 0$ and variance $\mathrm{Var}(Z) = 1$.

(traditionally, standard Gaussians are denoted with Z, PDF $\phi(z)$, and CDF $\Phi(z)$).

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

The Gaussian Distribution

General Definition

A random variable X has a ***Gaussian/Normal distribution*** with parameters μ and σ, denoted $X \sim N(\mu, \sigma^2)$ if it has PDF:

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \text{ for all } x \in \mathbb{R}.$$

This random variable has mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$.

The Gaussian Distribution

Properties of Gaussians

Standardization. If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0,1)$. As a result:

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Probability statements about general Gaussian ⇄ standard Gaussian.

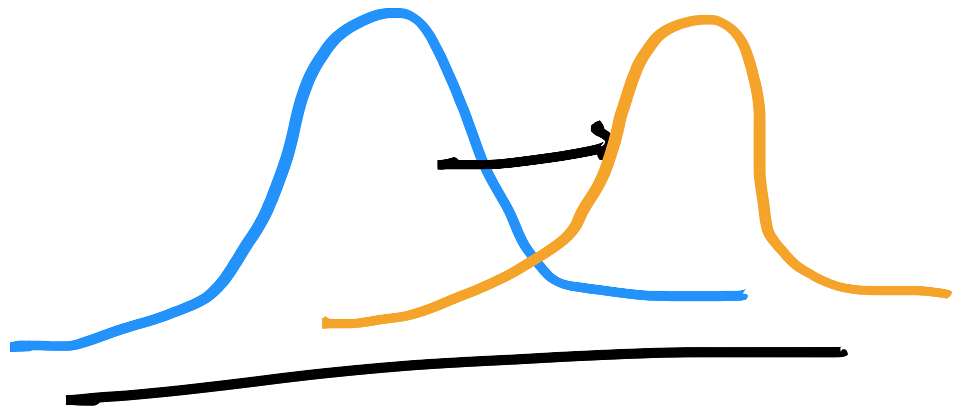
The Gaussian Distribution

Properties of Gaussians

Standardization. If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0,1)$. As a result:

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Standard to general. If $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.



The Gaussian Distribution

Properties of Gaussians

Standardization. If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0,1)$. As a result:

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Standard to general. If $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.

Sums of Gaussians. If $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1,\ldots,n$ are independent, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Central Limit Theorem

Intuition and Simulations

Statistical Estimation
Intuition

In ***probability theory***, we assumed we knew some data generating process (as a *distribution*) $\mathbb{P}_{\mathbf{x}}$, and we analyzed observed data under that process.

$$\mathbb{P}_{\mathbf{x}} \implies \mathbf{x}_1, \ldots, \mathbf{x}_n.$$

Statistics can be thought of as the "reverse process." We see some data and we try to make inferences about the process that generated the data.

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \implies \mathbb{P}_{\mathbf{x}}$$

Statistical Estimation

Intuition

In ***probability theory***, we assumed we knew some data generating process (as a *distribution*) $\mathbb{P}_{\mathbf{x}}$, and we analyzed observed data under that process.

$$\mathbb{P}_{\mathbf{x}} \implies \mathbf{x}_1, \ldots, \mathbf{x}_n.$$

Statistics can be thought of as the "reverse process." We see some data and we try to make inferences about the process that generated the data.

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \implies \mathbb{P}_{\mathbf{x}}$$

In order to do so, we need to formalize the notion that "collecting a lot of data" gives us a peek at the underlying process!

Law of Large Numbers
Theorem Statement

Theorem (Weak Law of Large Numbers). Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with finite mean $\mu := \mathbb{E}[X_i]$. Let their *sample average* be denoted as

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\overline{X}_n - \mu \quad < \epsilon \right) = 1.$$

This type of convergence is also called ***convergence in probability***.

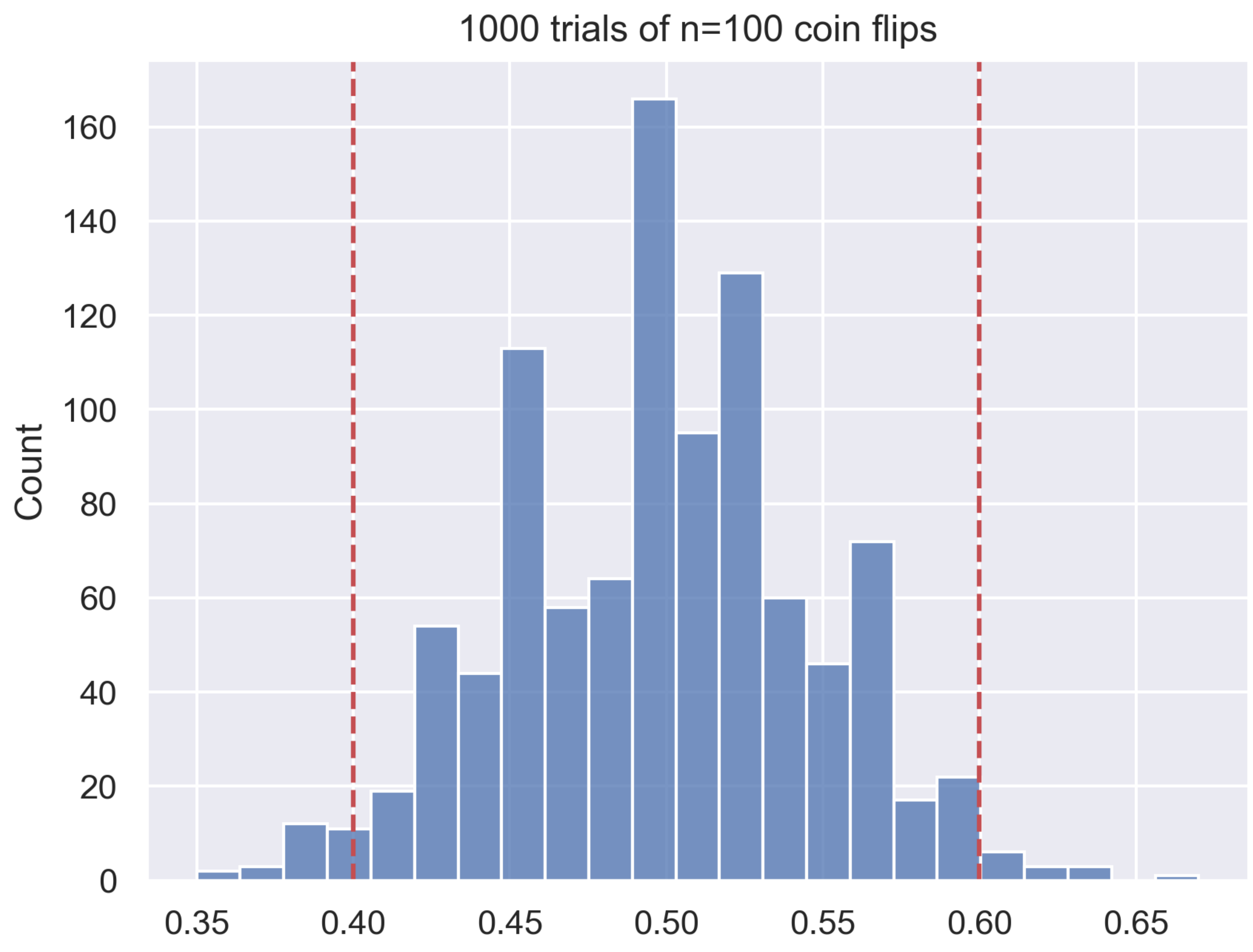
Law of Large Numbers

Example: Mean Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Law of large numbers states that for *any* $\epsilon > 0$*, no matter how small*:

$$\lim_{n \to \infty} \mathbb{P}(\left| \overline{X}_n - 1/2 \right| < \epsilon) = 1$$



Law of Large Numbers

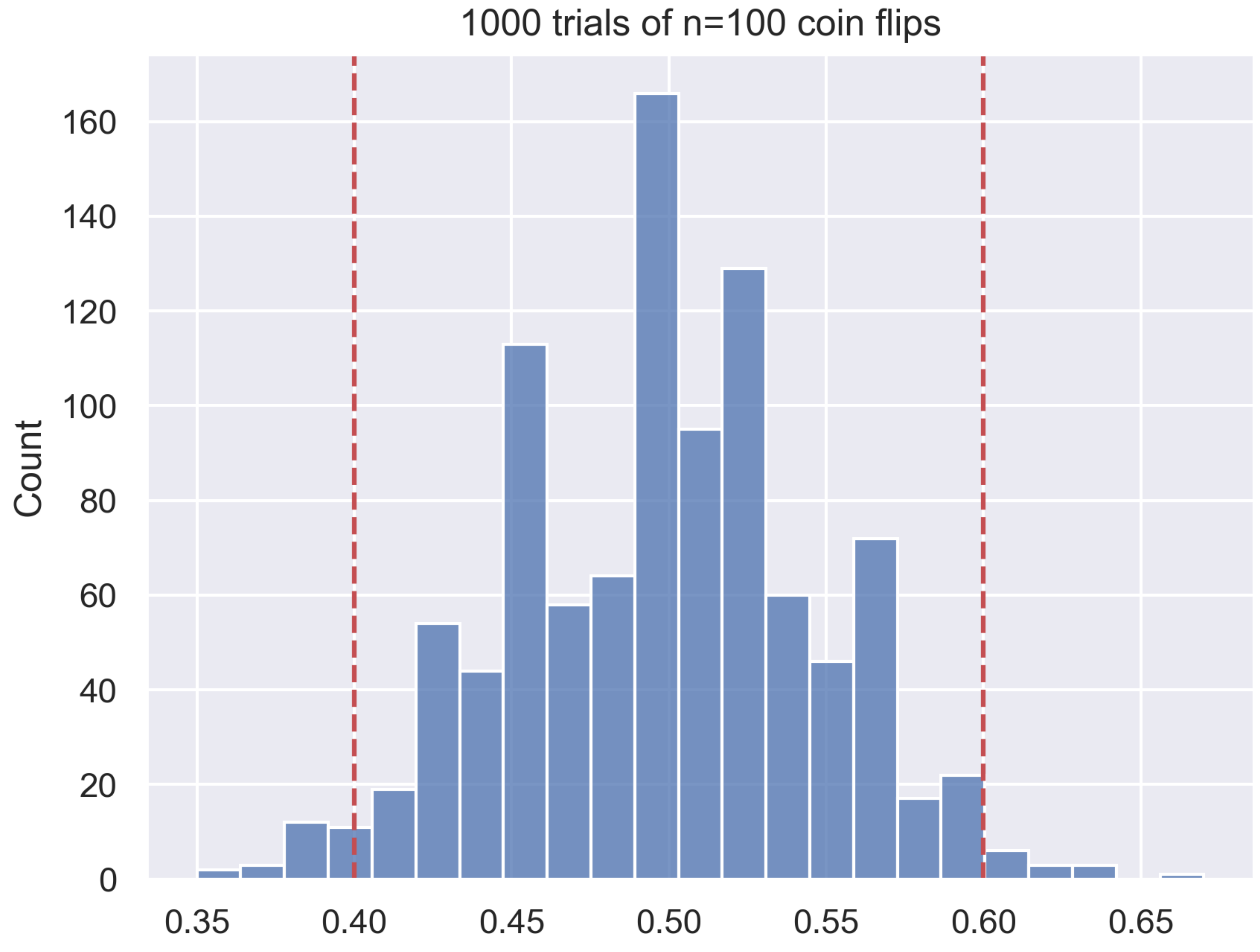
Example: Mean Estimator for Coins

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Law of large numbers states that for *any* $\epsilon > 0$, *no matter how small*:

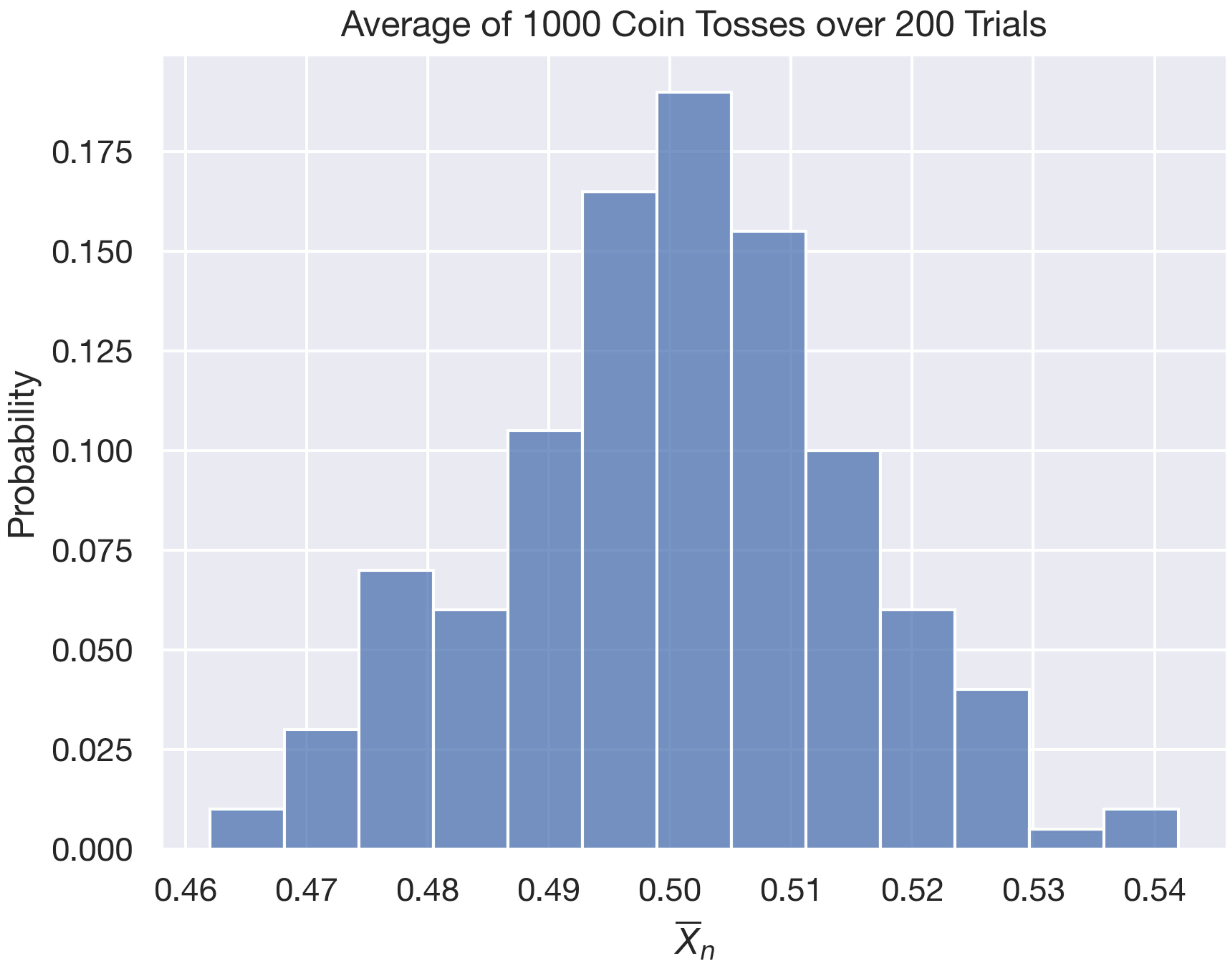
$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - 1/2\right| < \epsilon\right) = 1$$

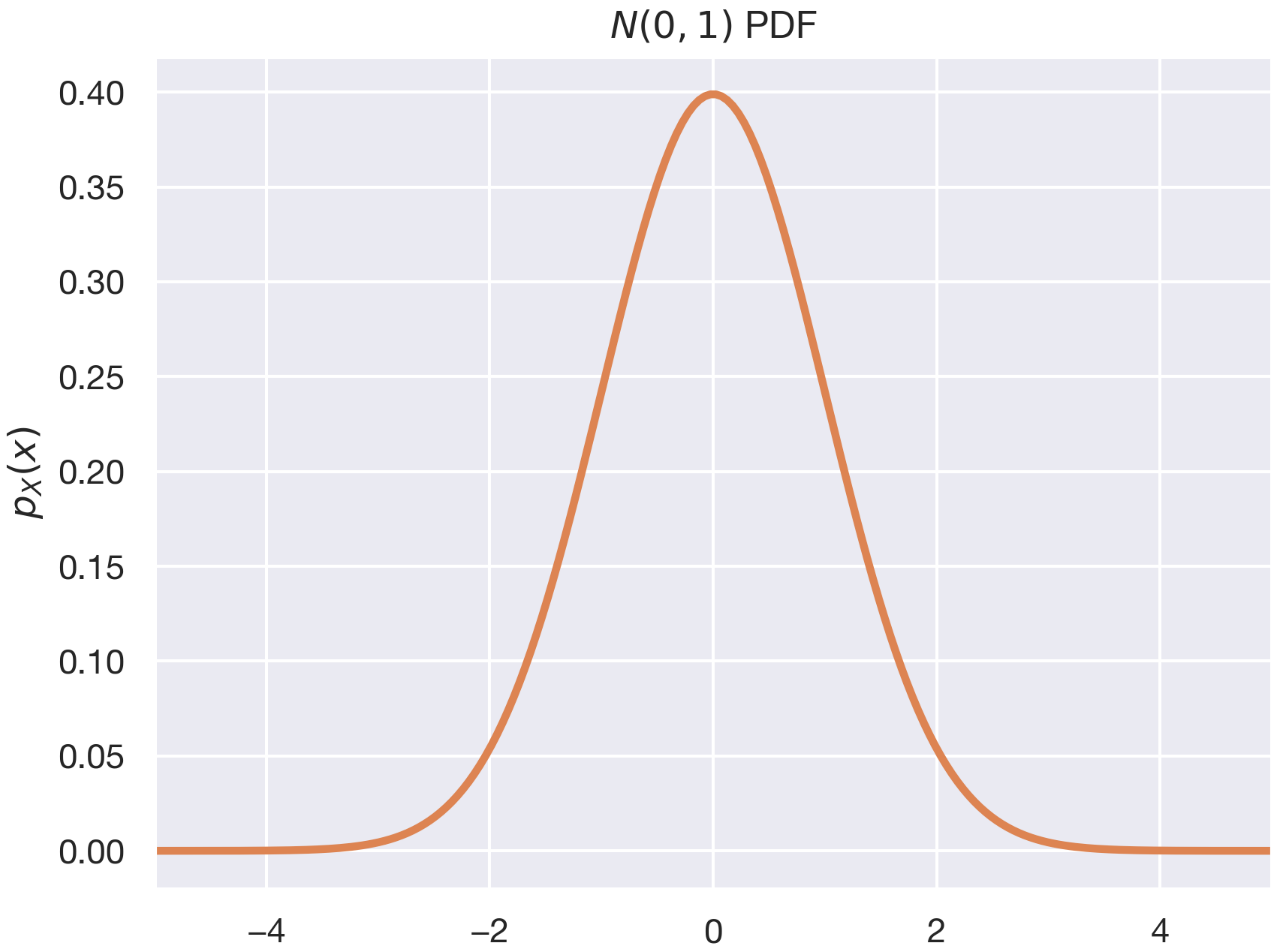


But can we say something more about the distribution of the random variable $\overline{X}_n$?

Central Limit Theorem

Intuition



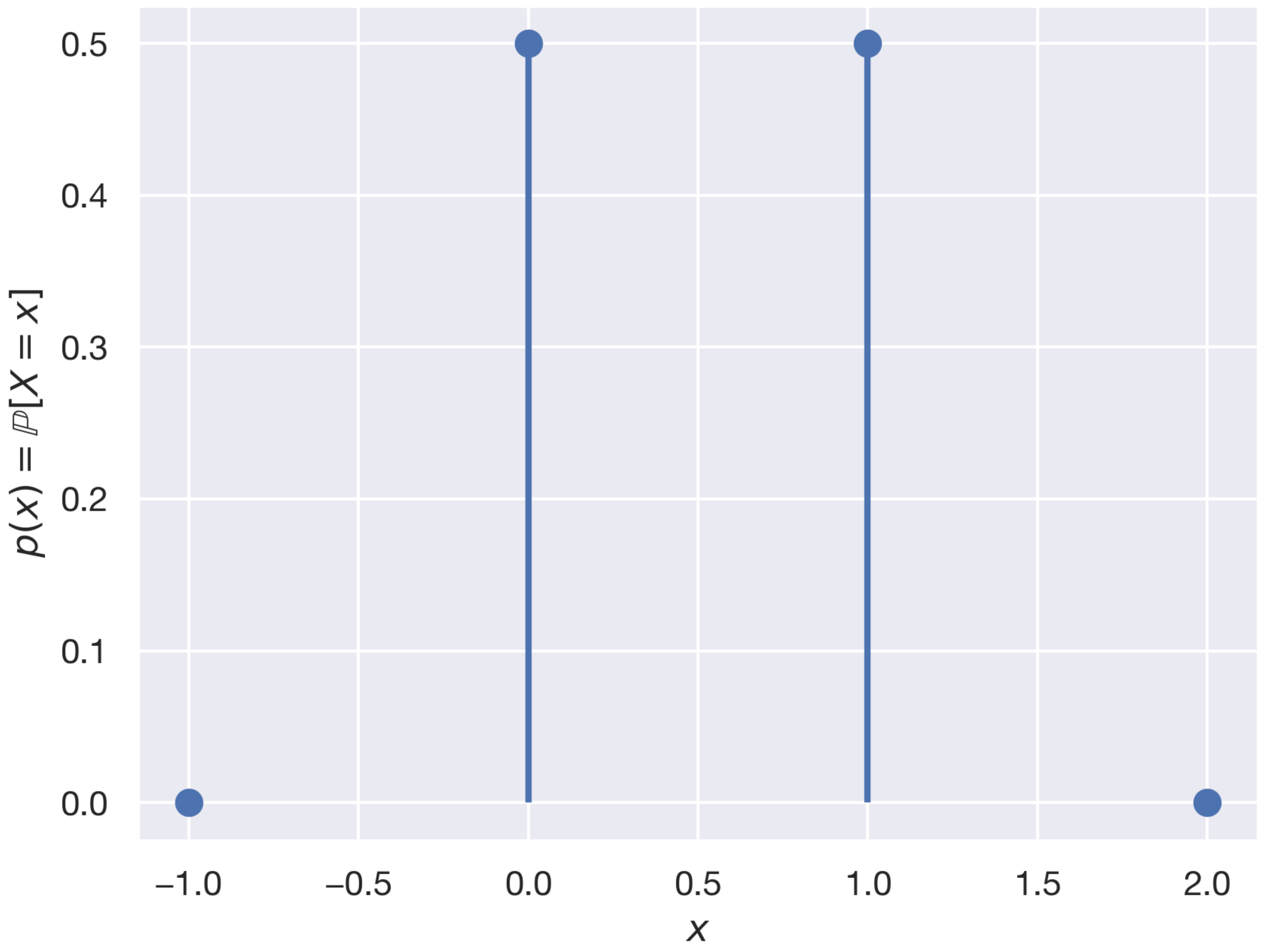


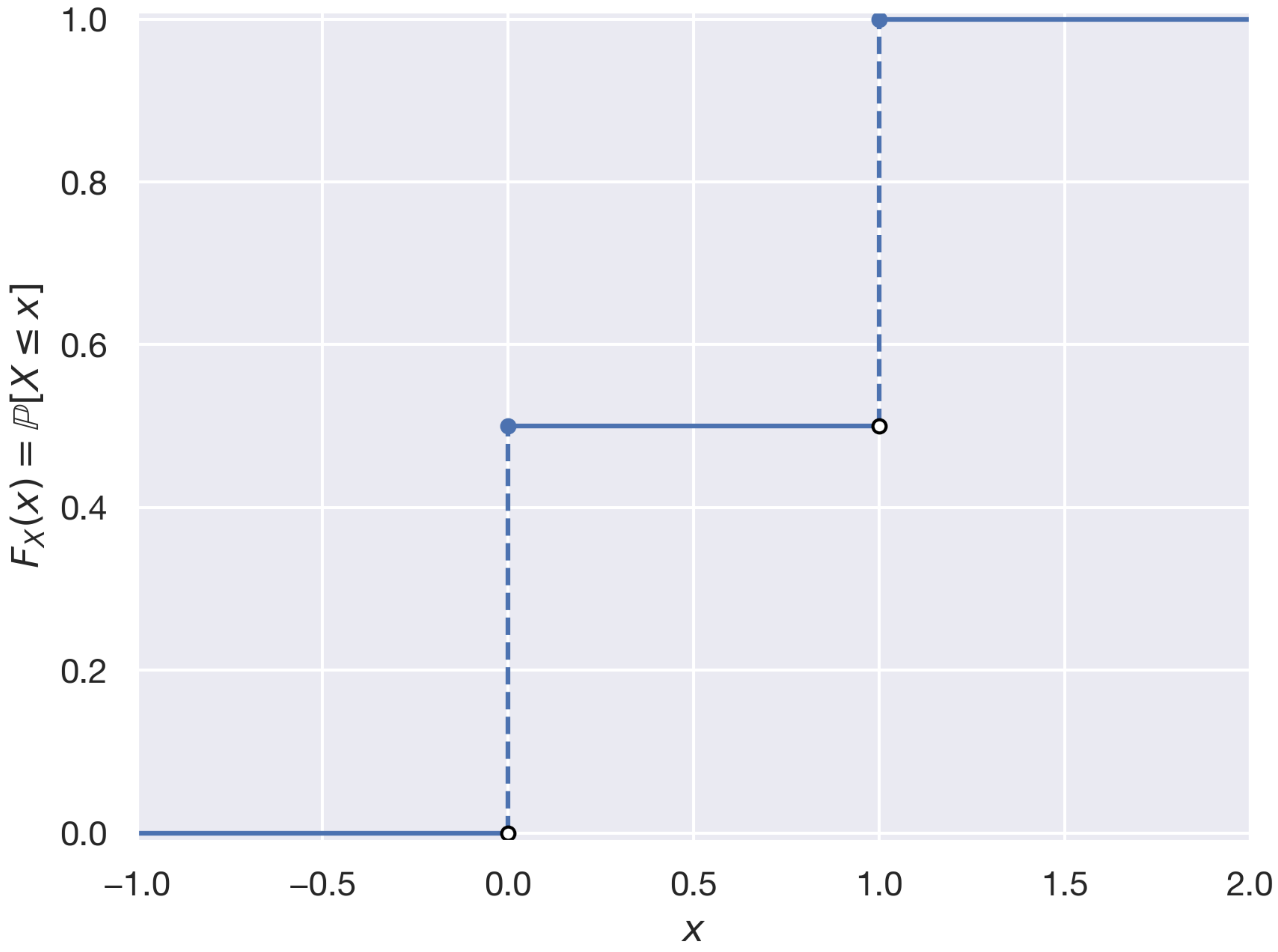
Central Limit Theorem

Experiment: Coin Tosses

$$\mathbb{P}[X = 0] = \mathbb{P}[X = 1] = 1/2$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$





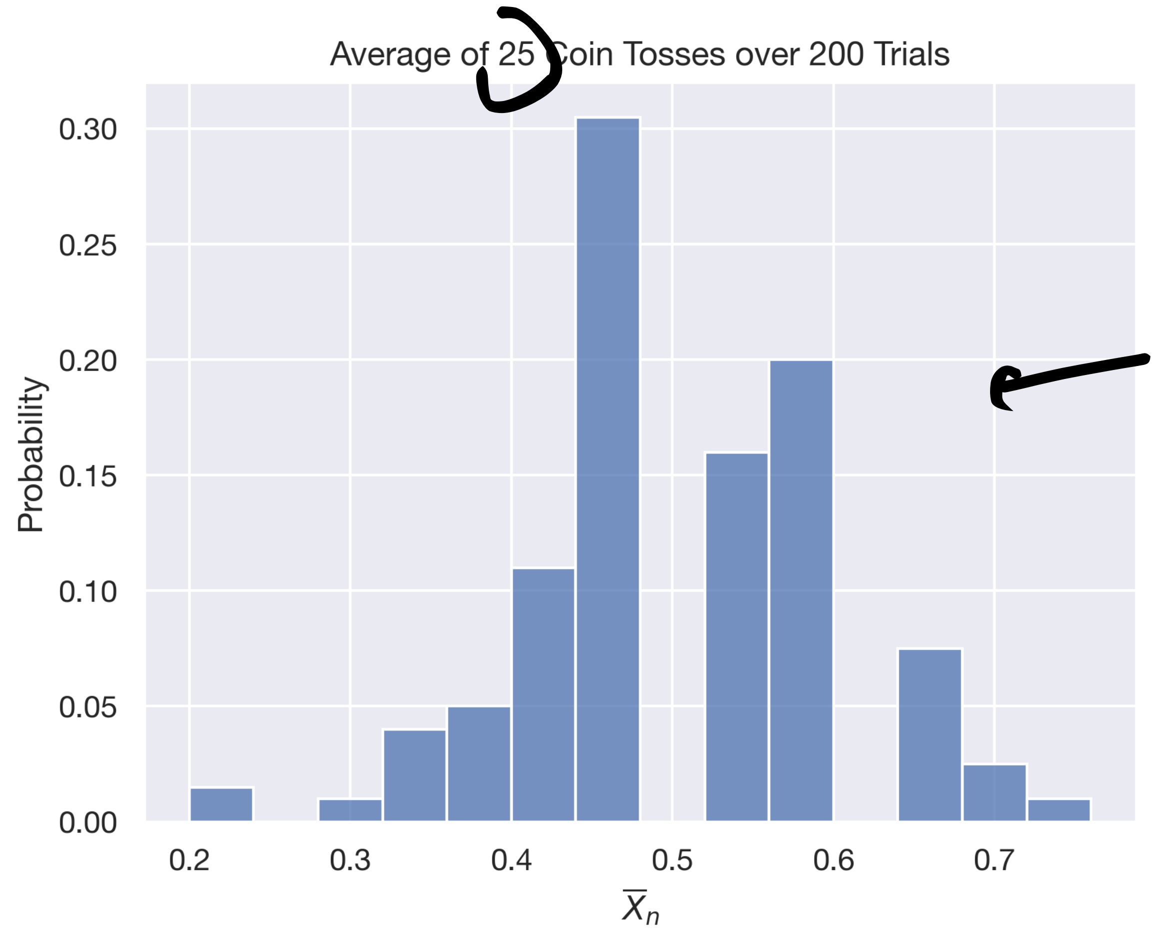
Central Limit Theorem

Experiment: Coin Tosses

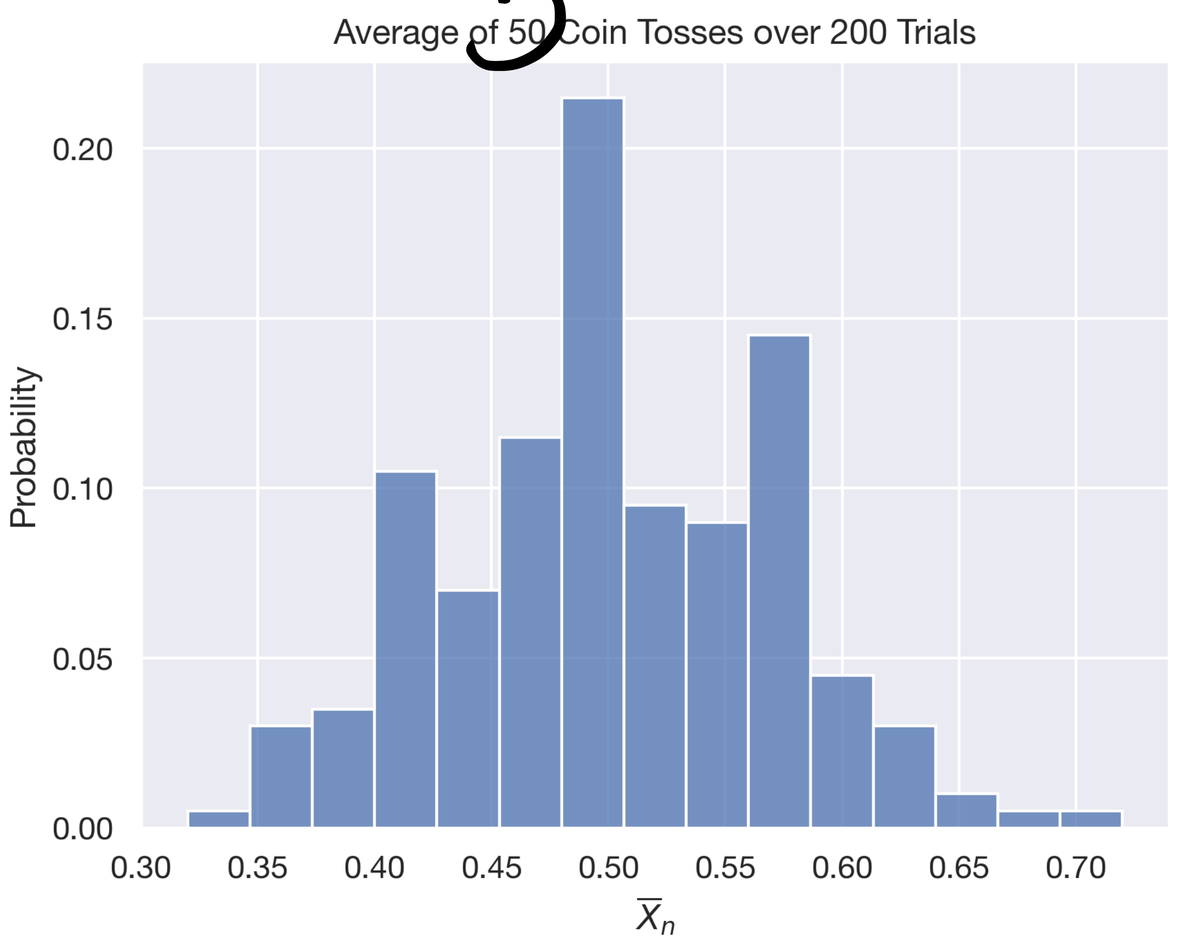
① Fix n for # of coin tosses.

② T trials of n coin tosses (200)

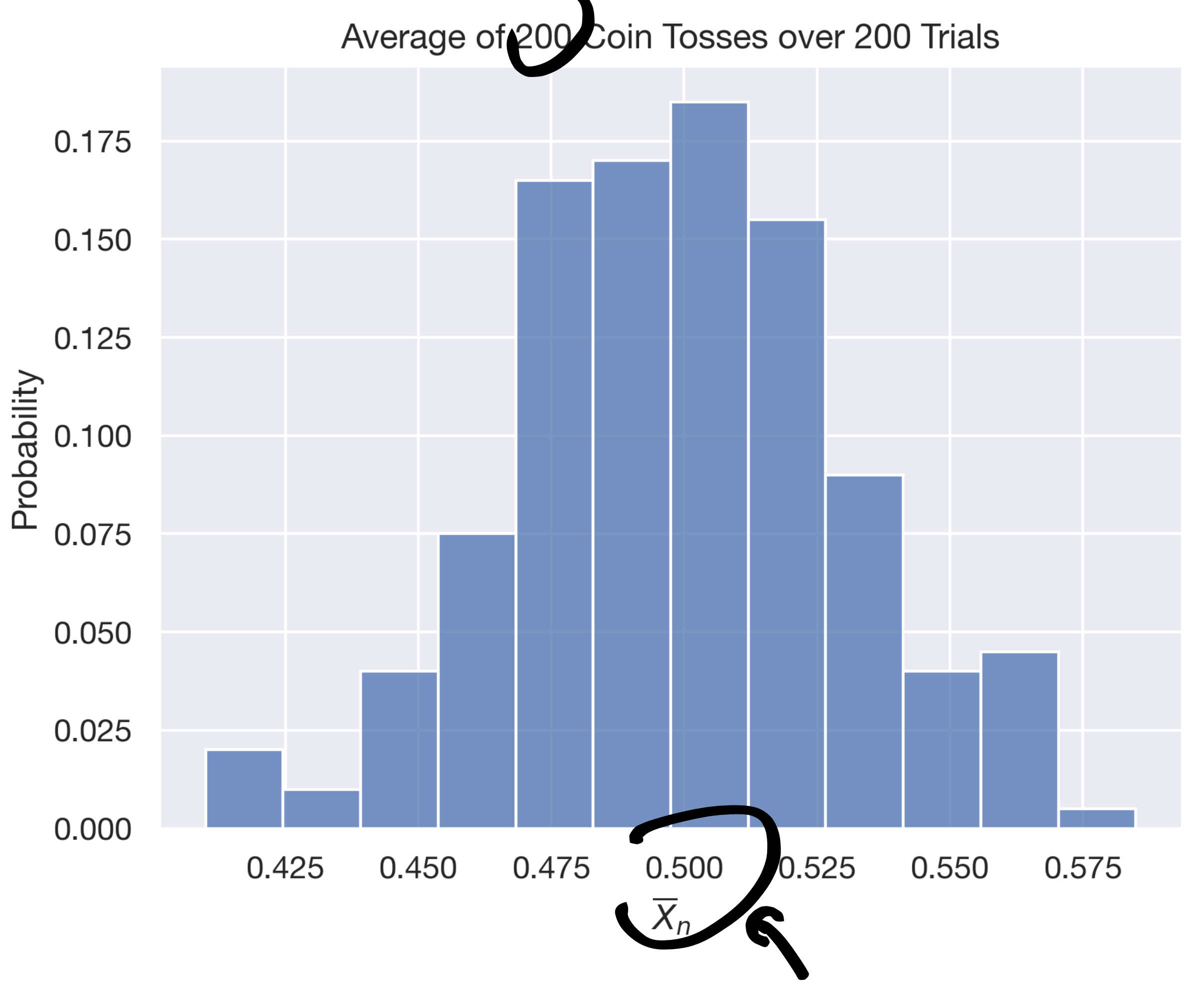
n = 25

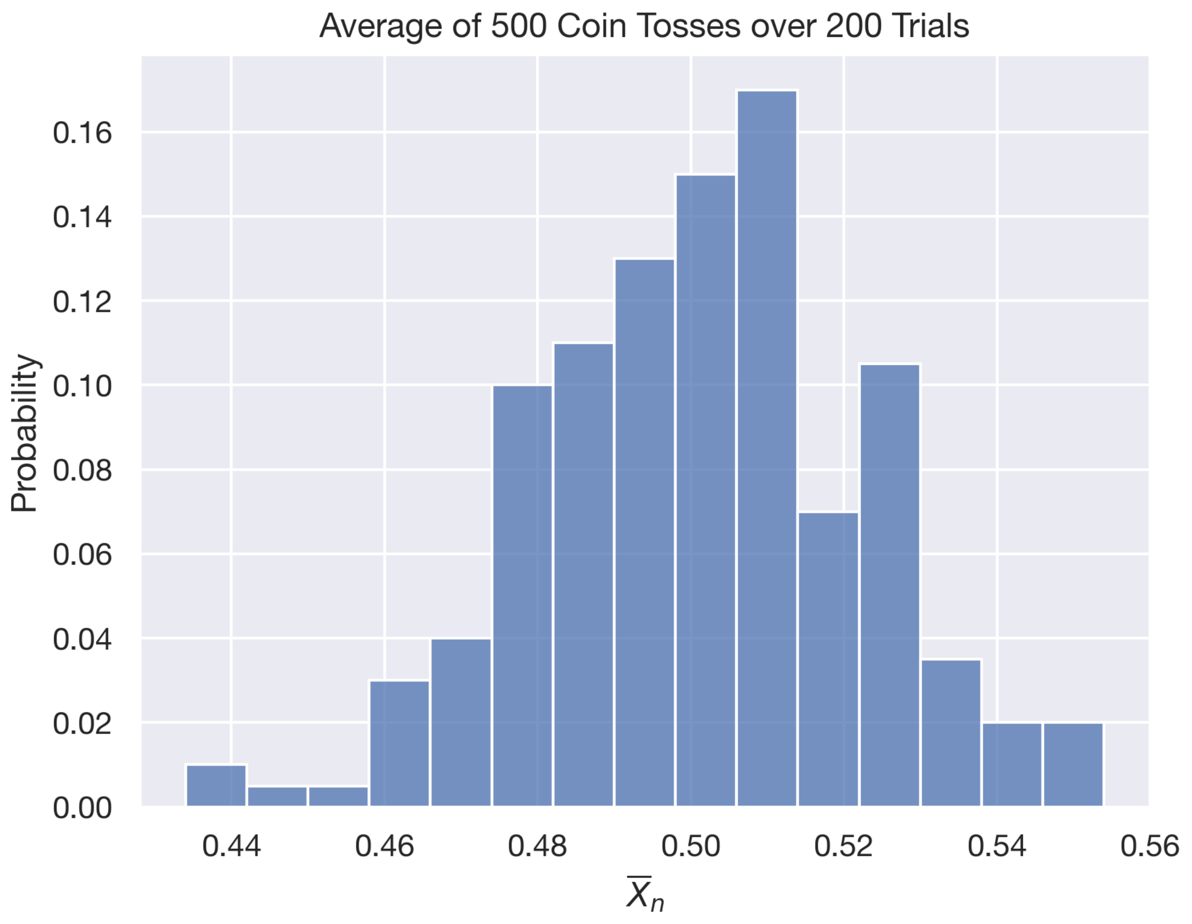


n = 50



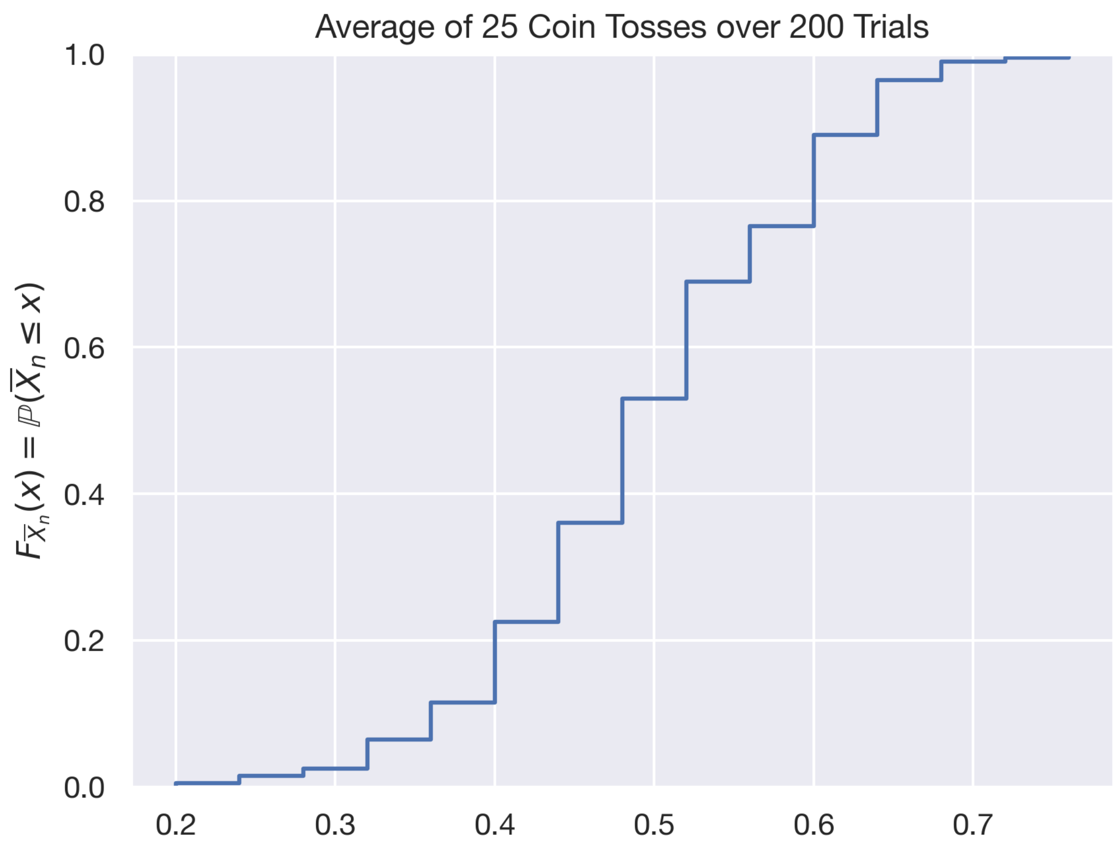
n = 200

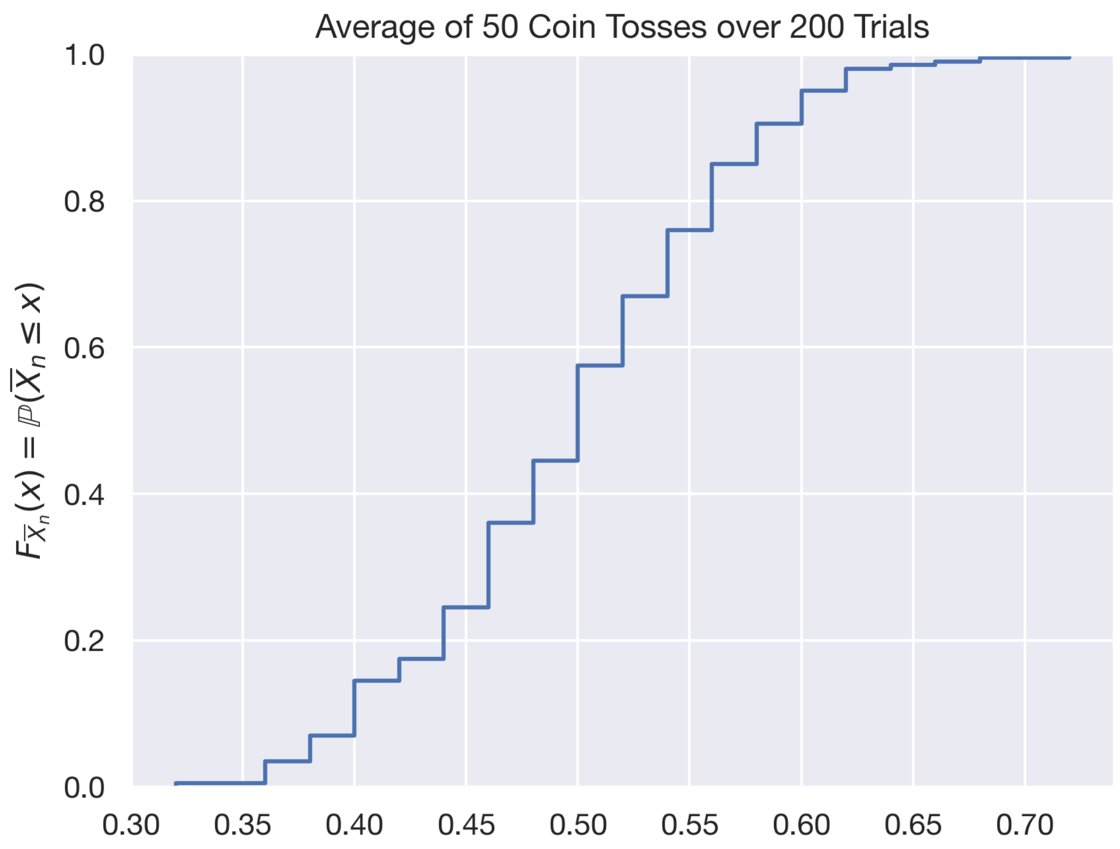


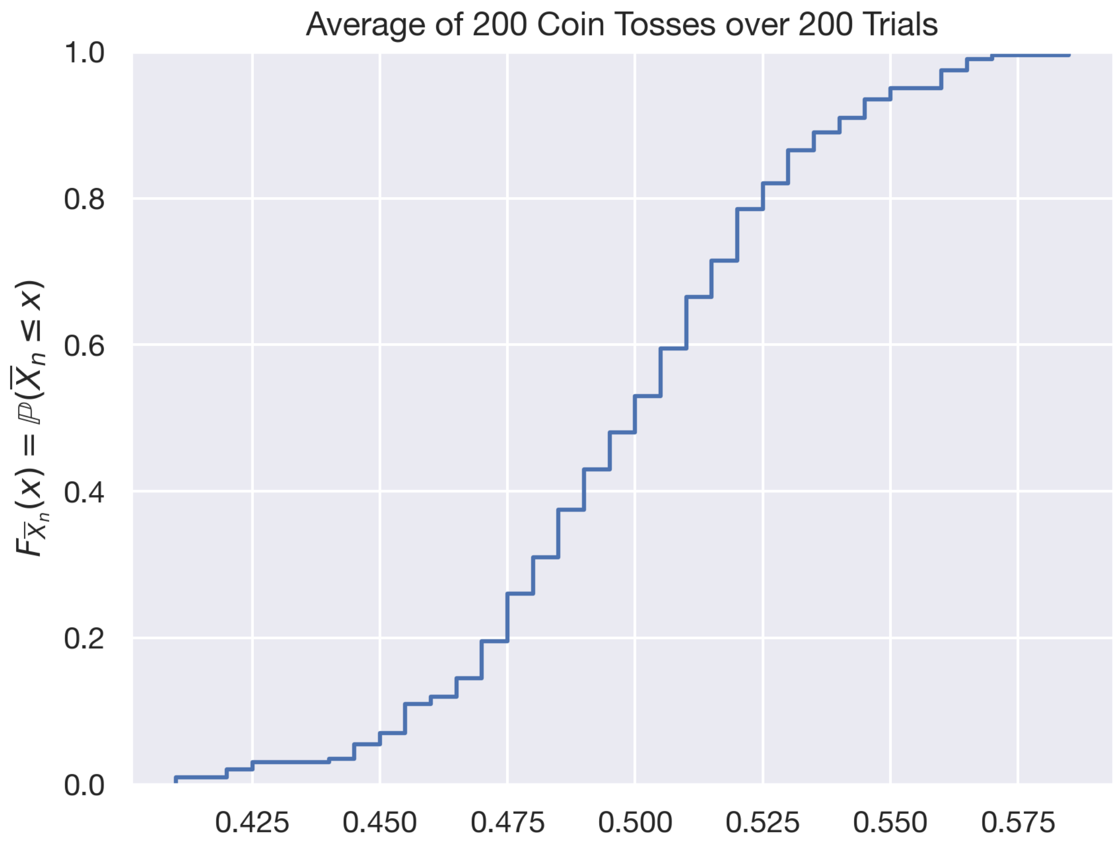


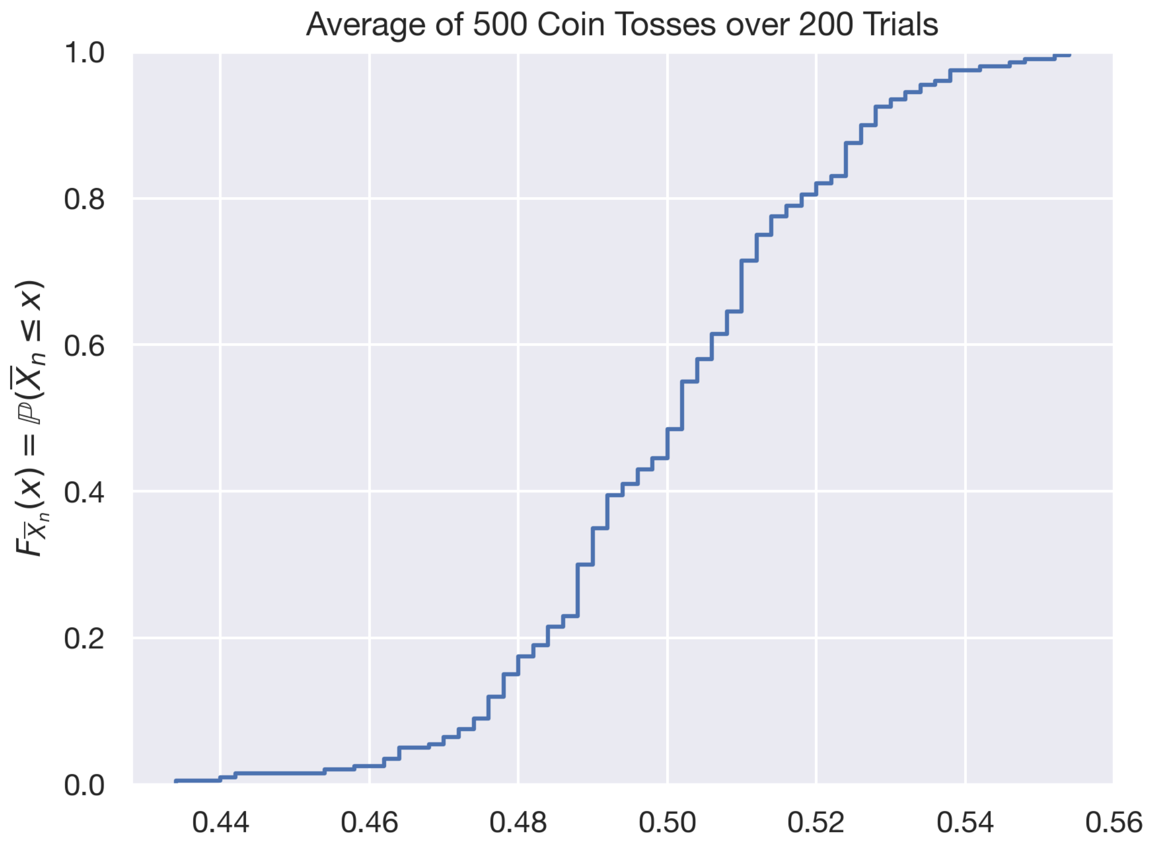
Central Limit Theorem

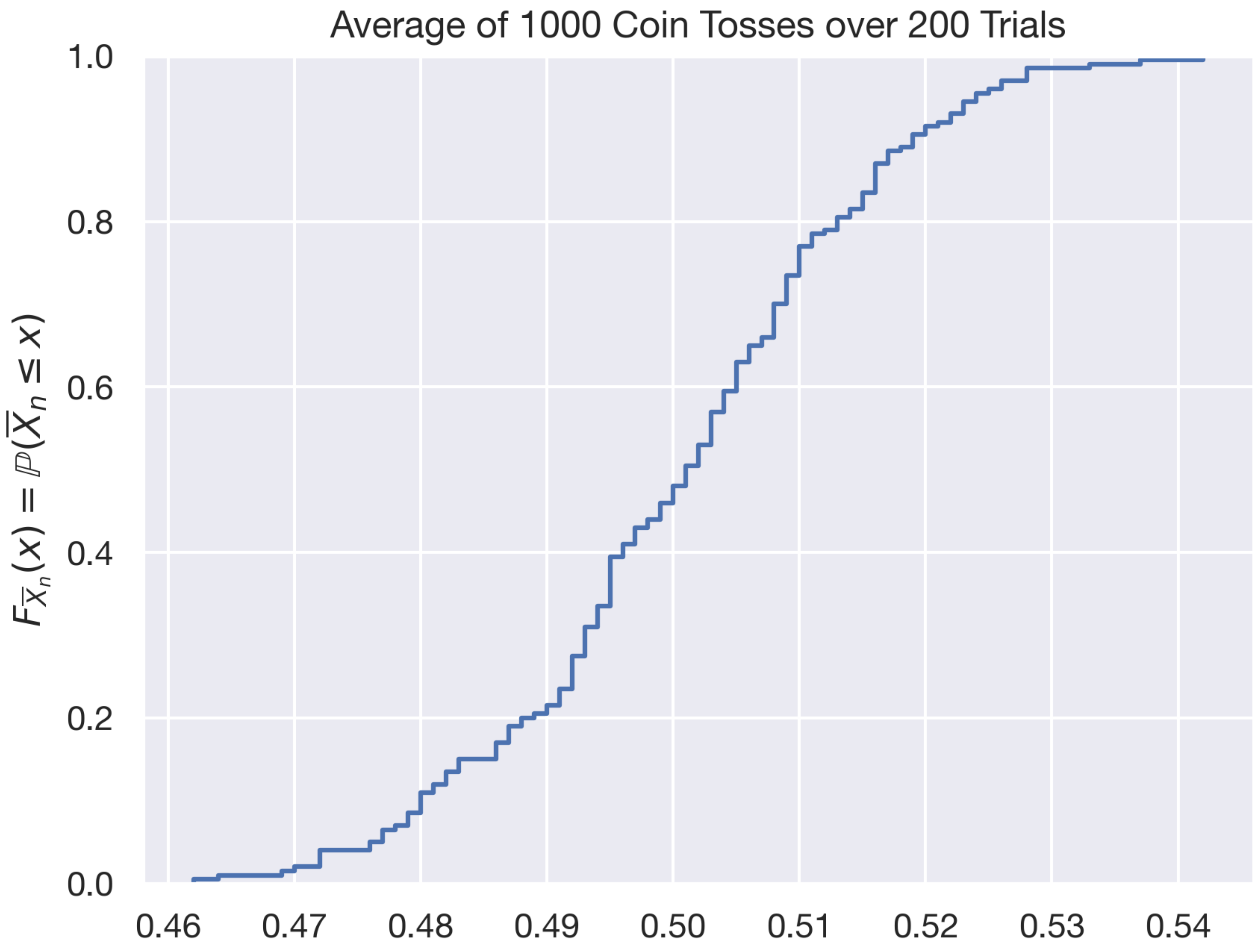
Experiment: Coin Tosses



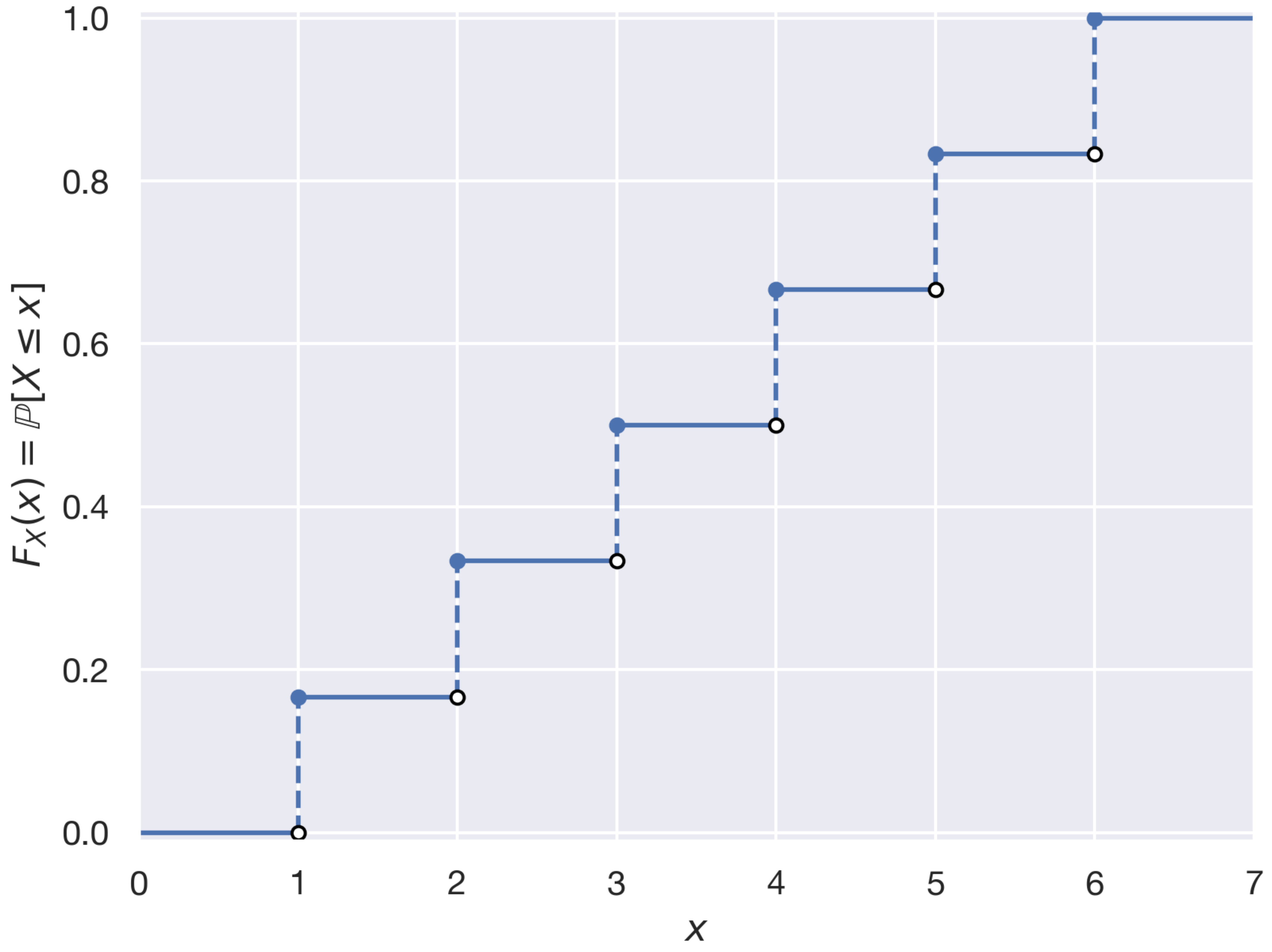


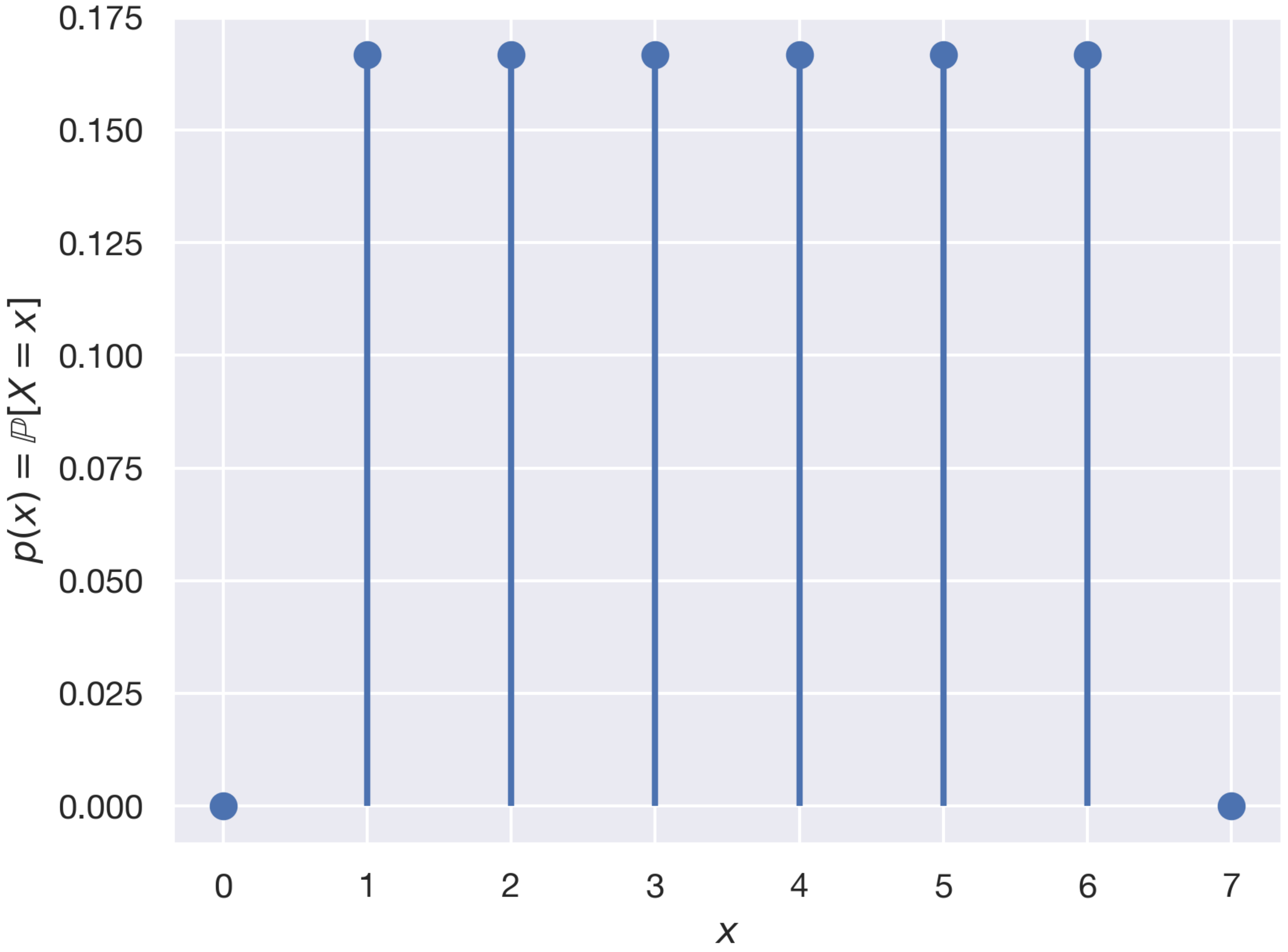






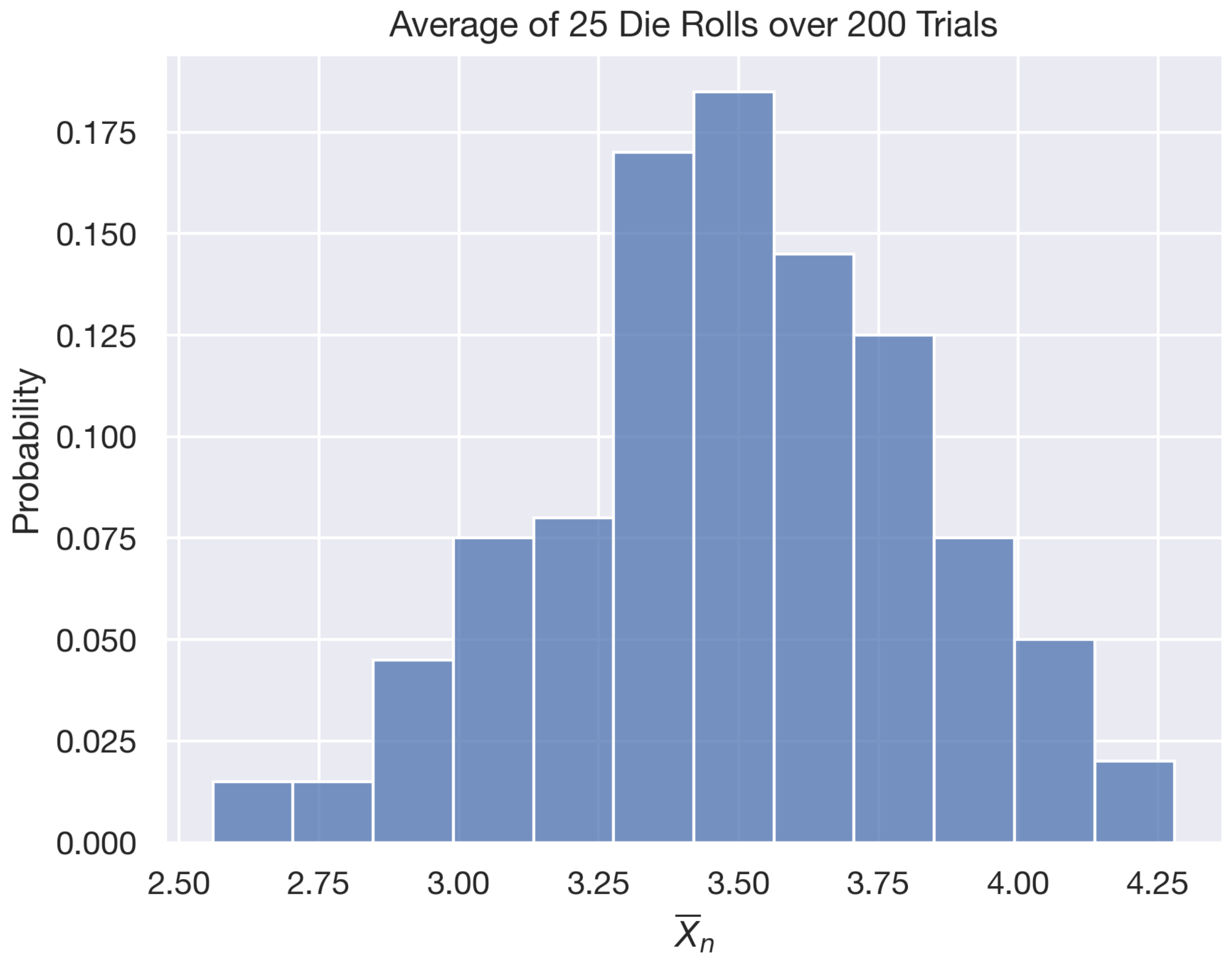
Central Limit Theorem
Experiment: Die Rolls

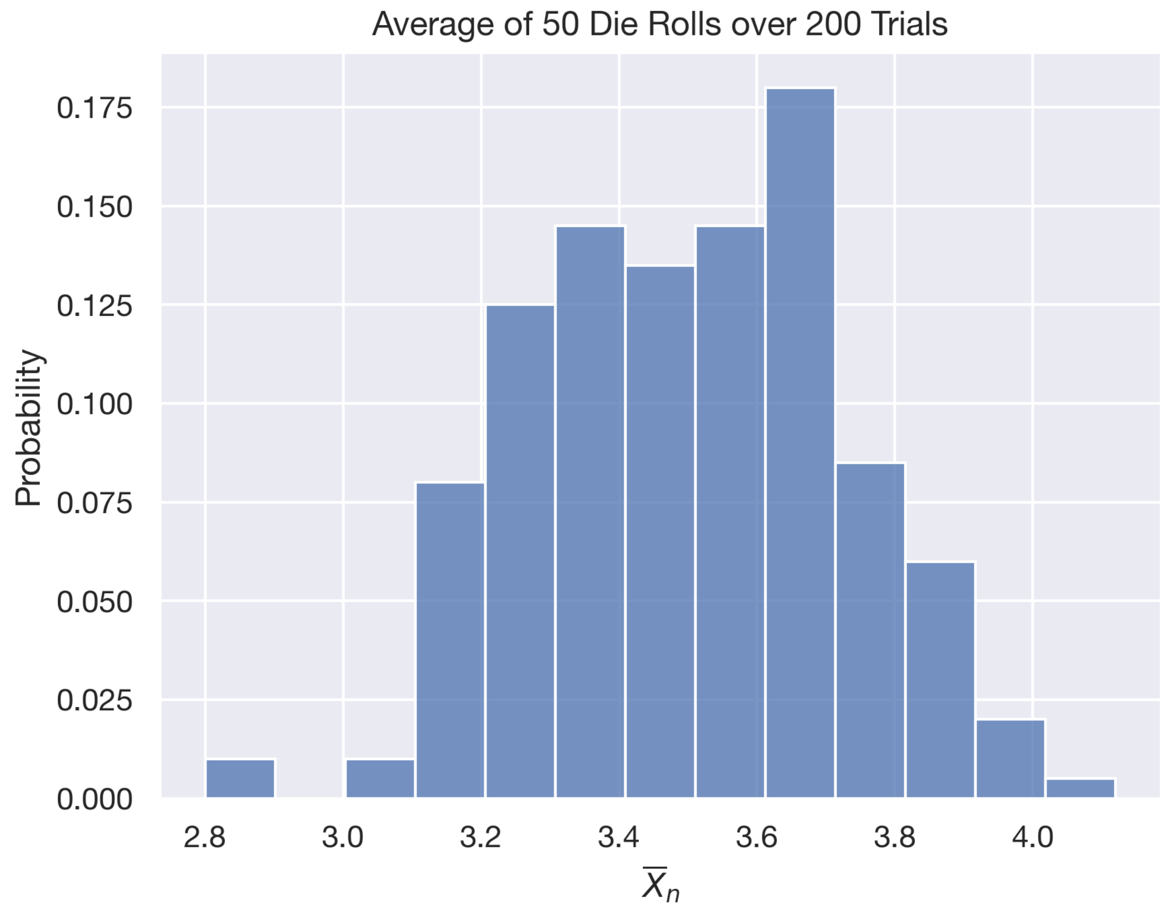


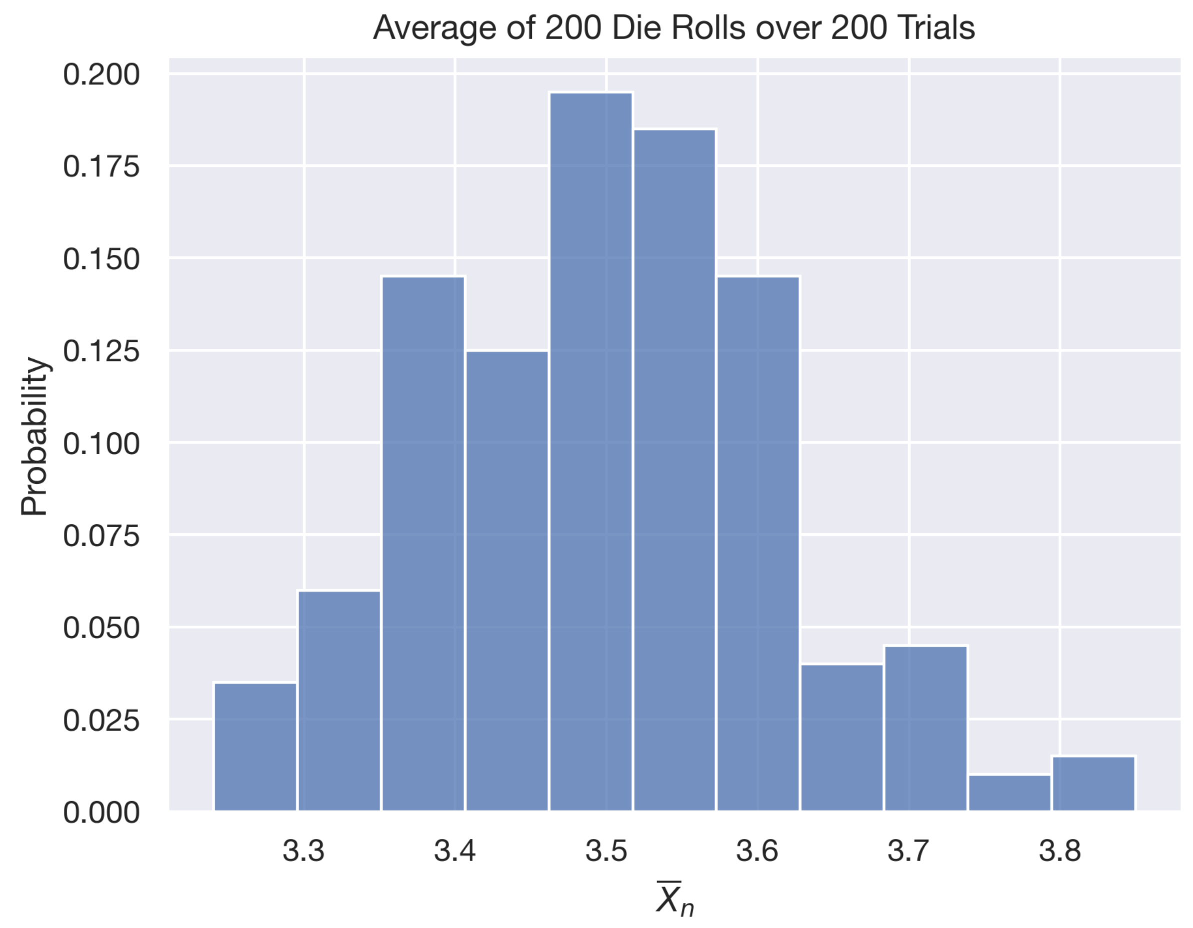


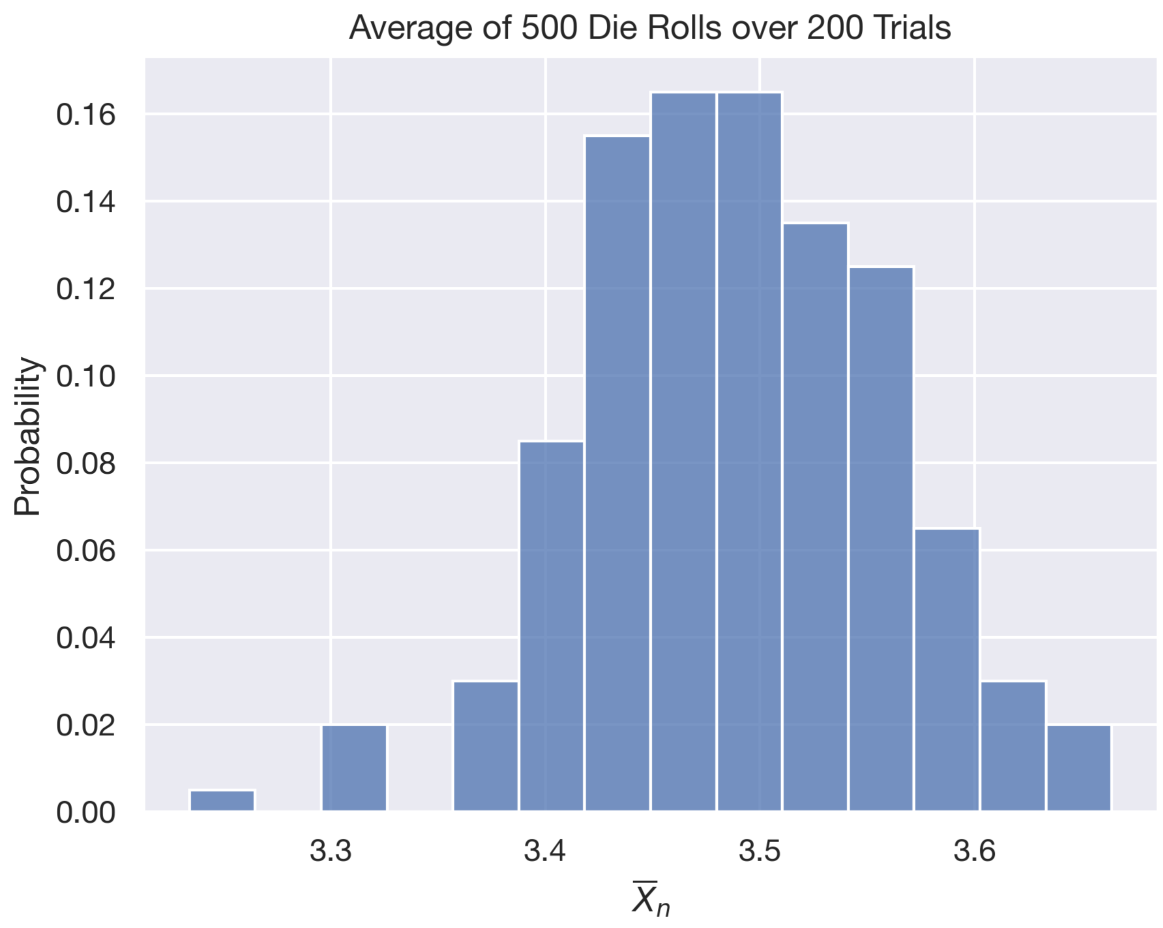
Central Limit Theorem

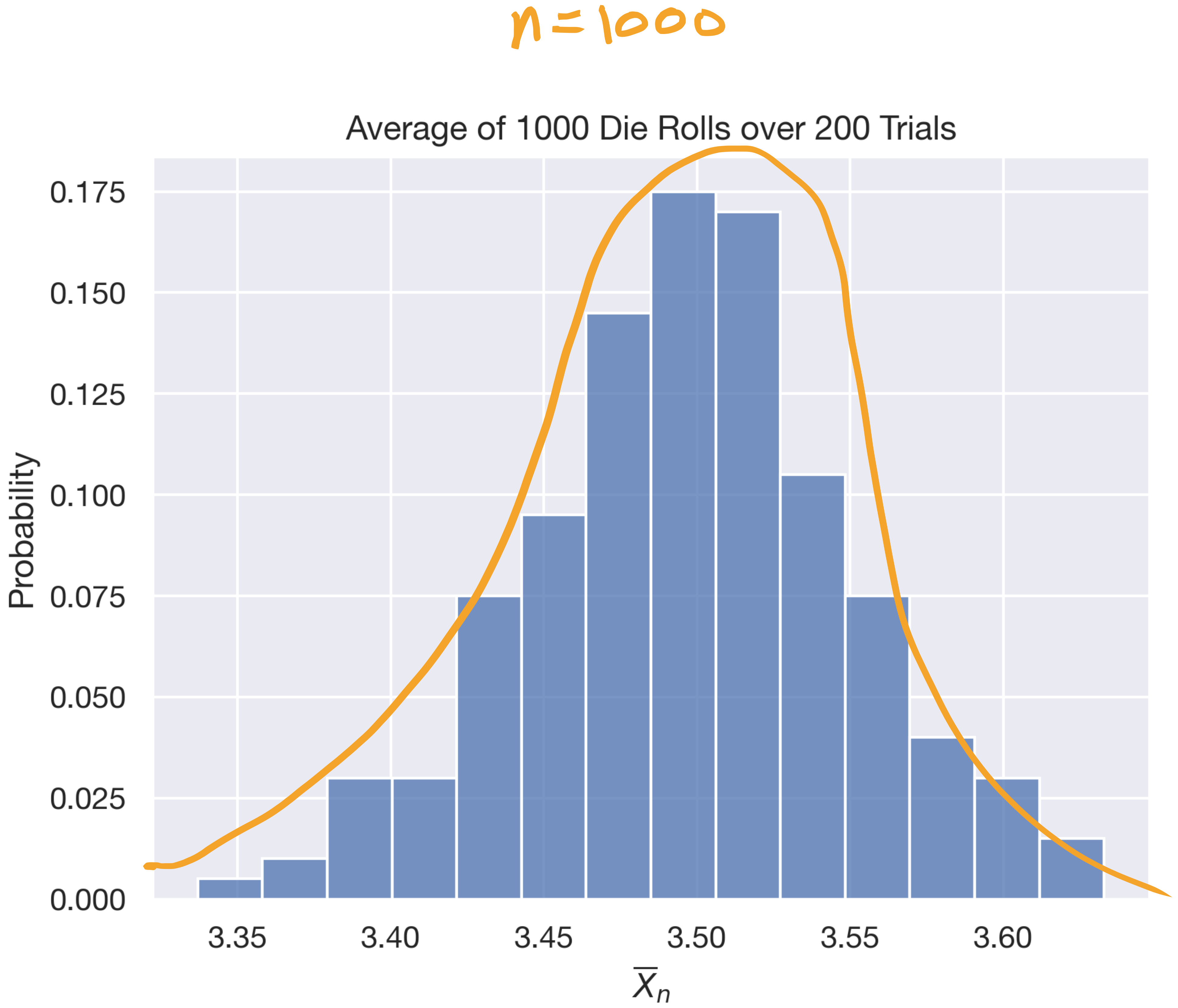
Experiment: Die Rolls





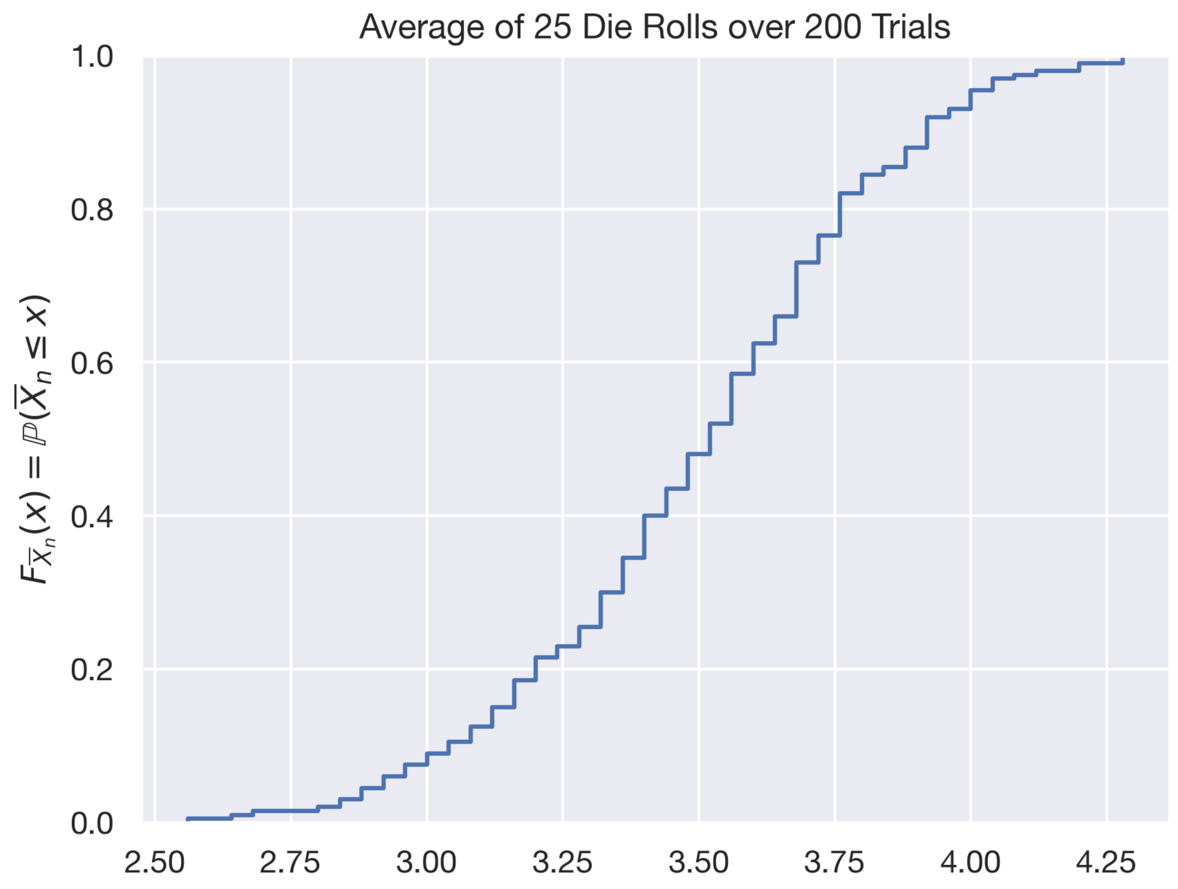


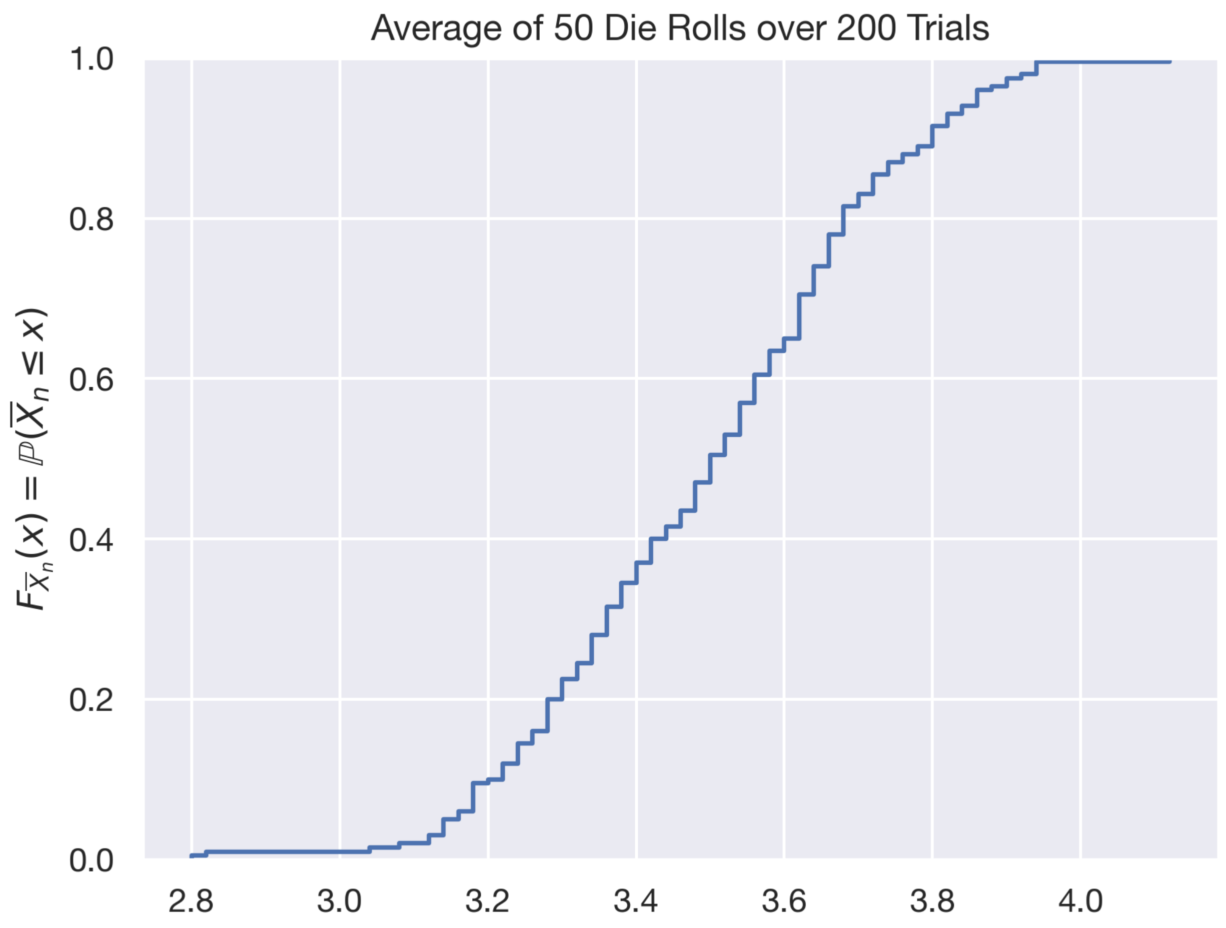


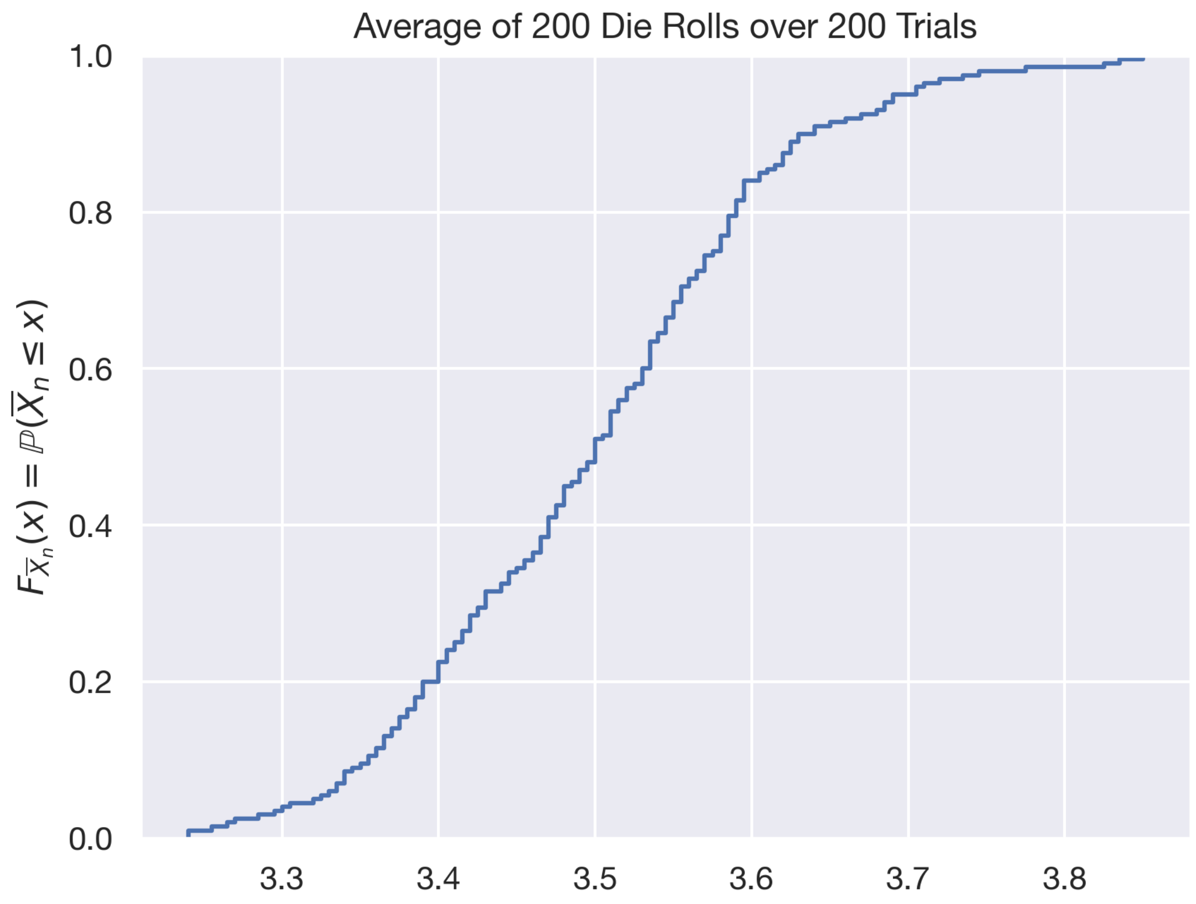


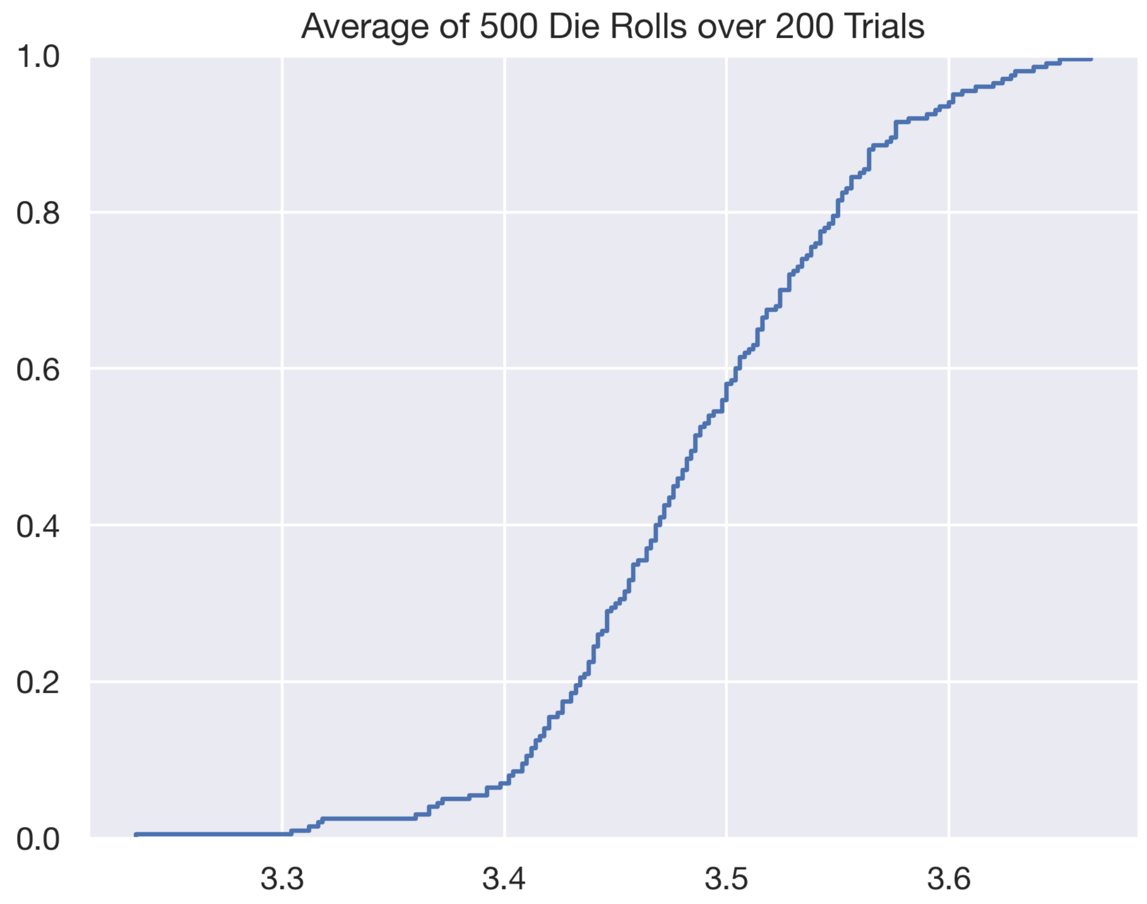
Central Limit Theorem

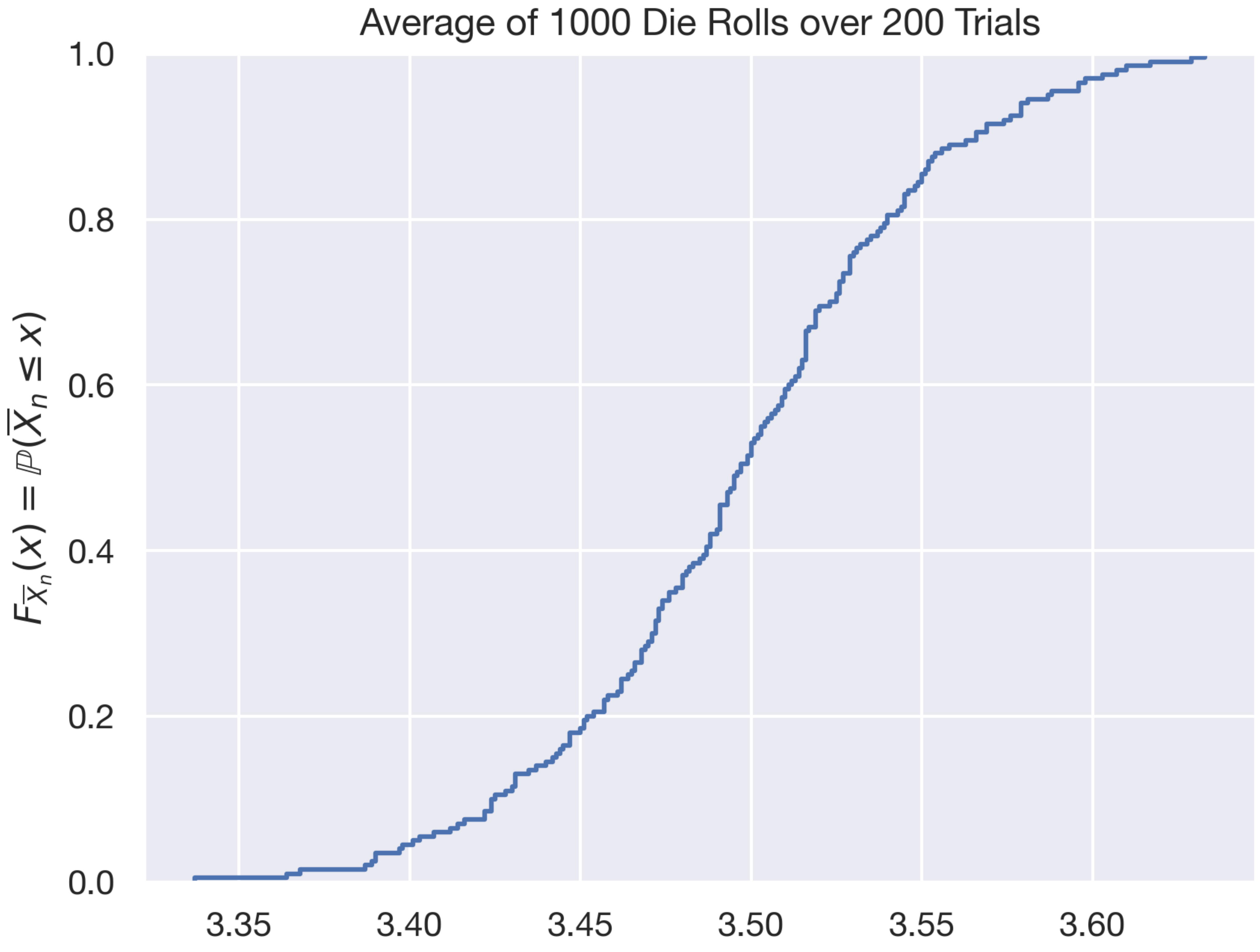
Experiment: Die Rolls







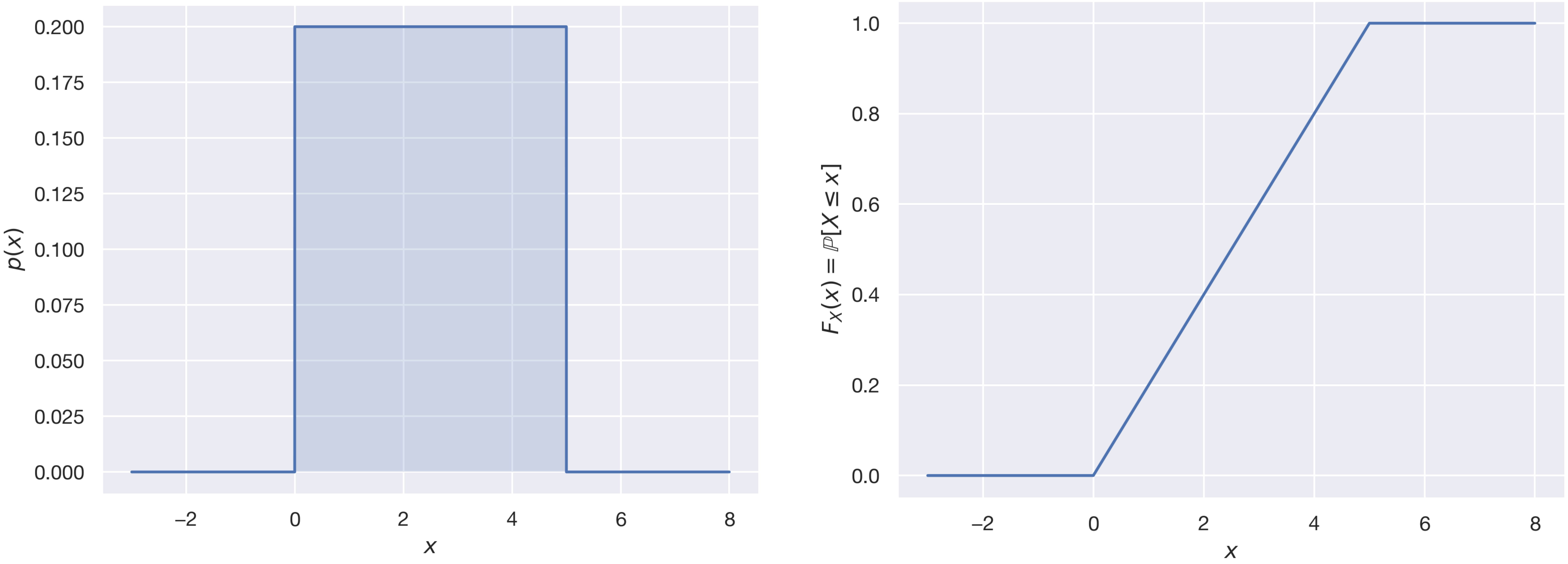




Central Limit Theorem

Experiment: Drawing uniform real value

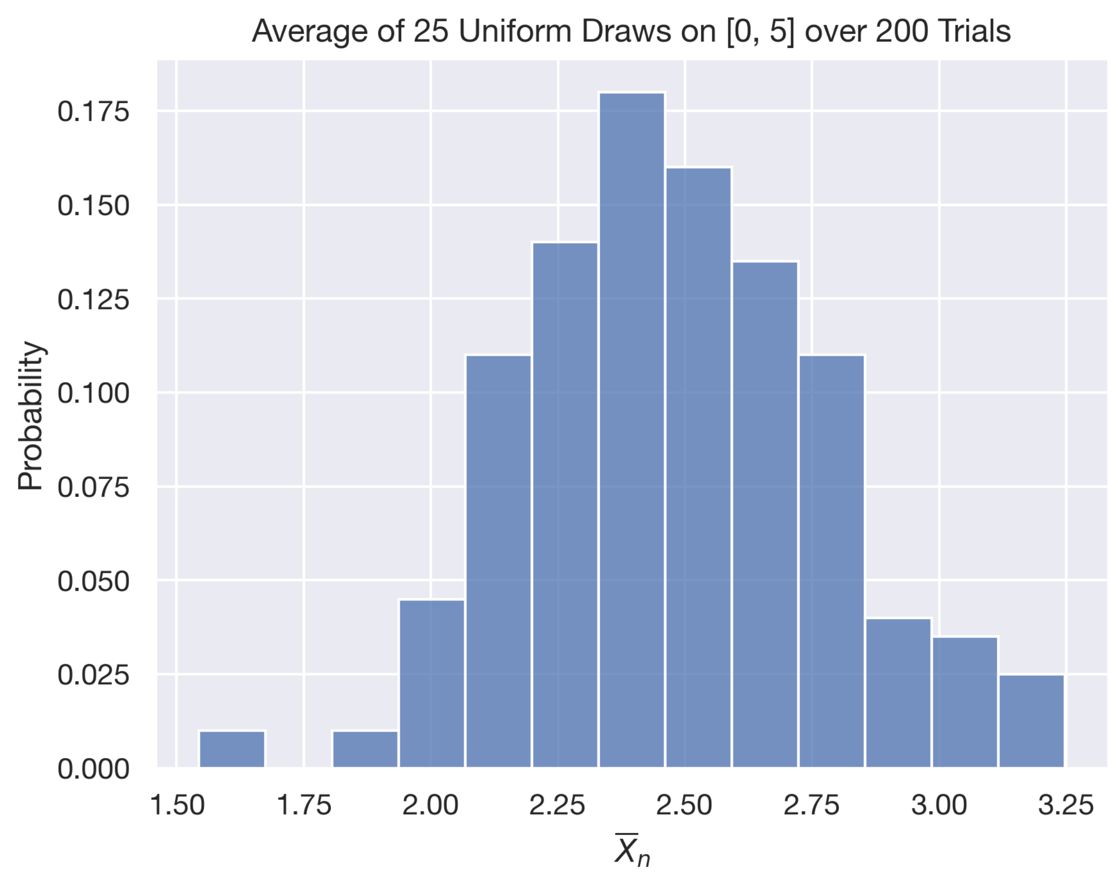
Unif([0, 5])

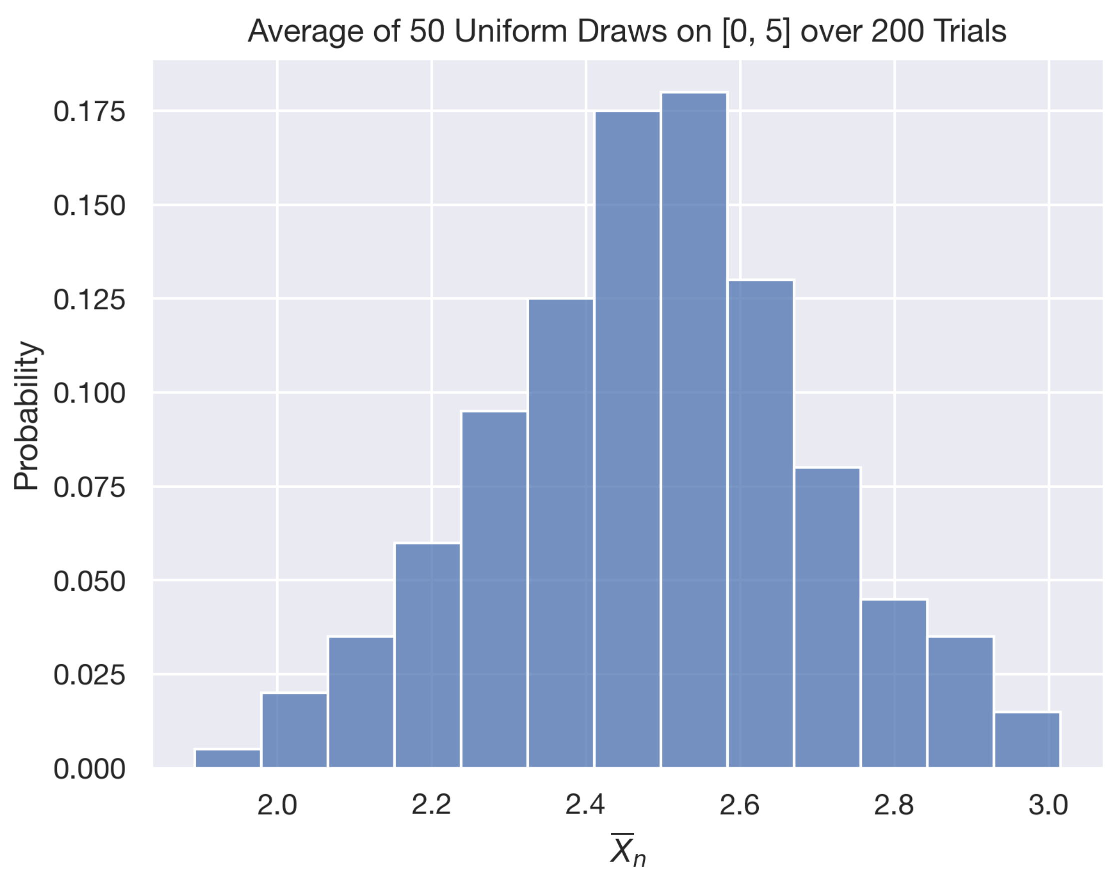


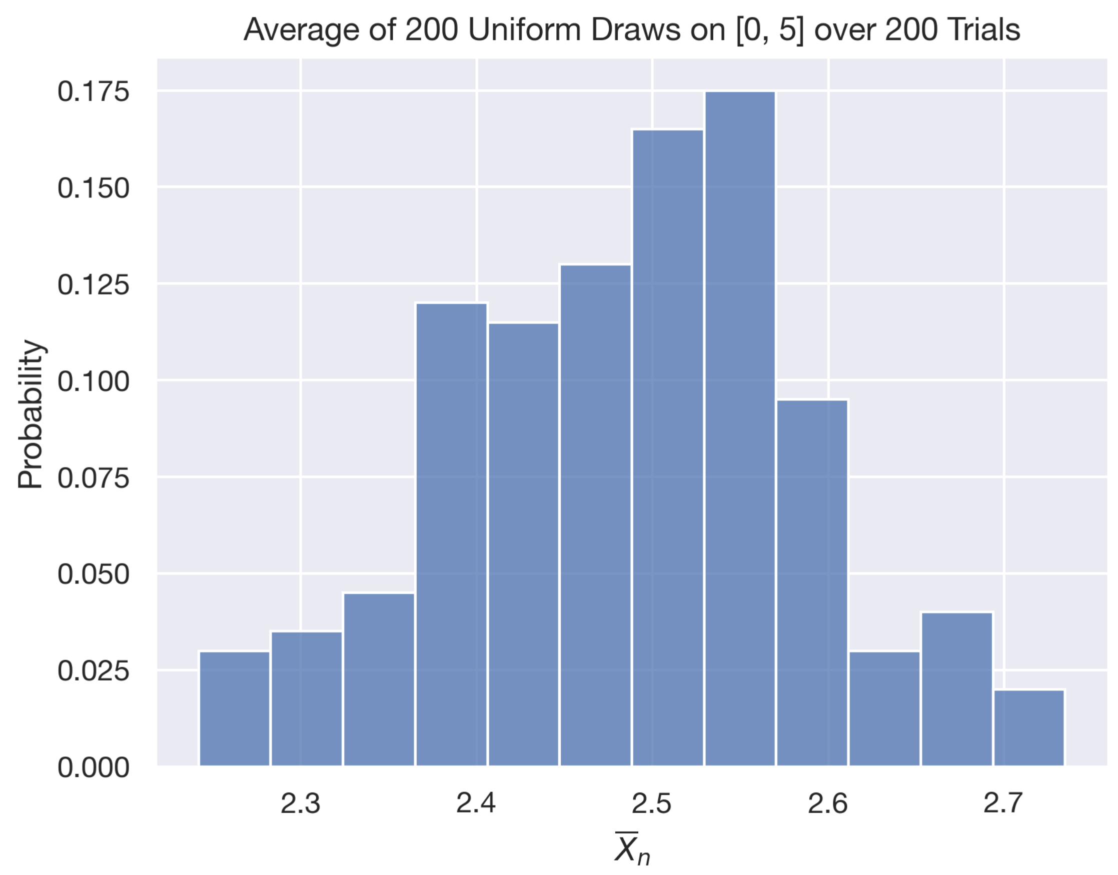
Central Limit Theorem

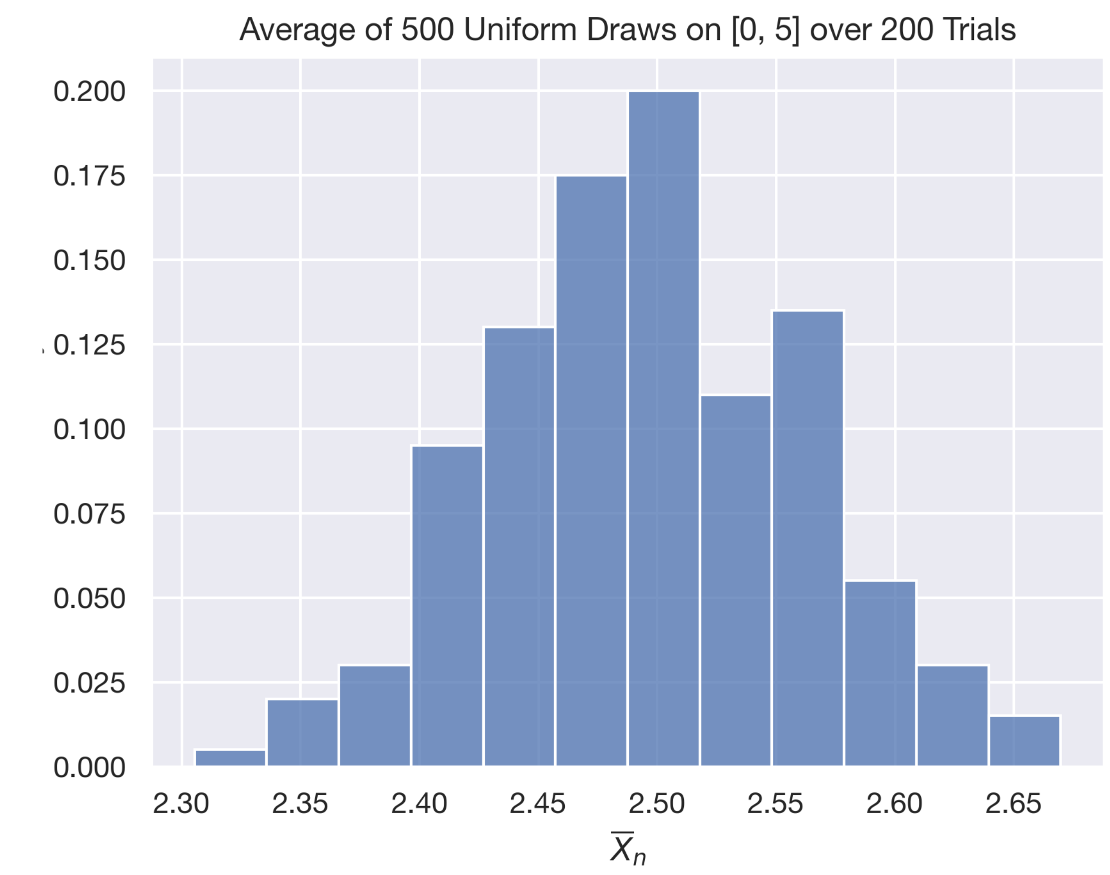
Experiment: Drawing uniform real value

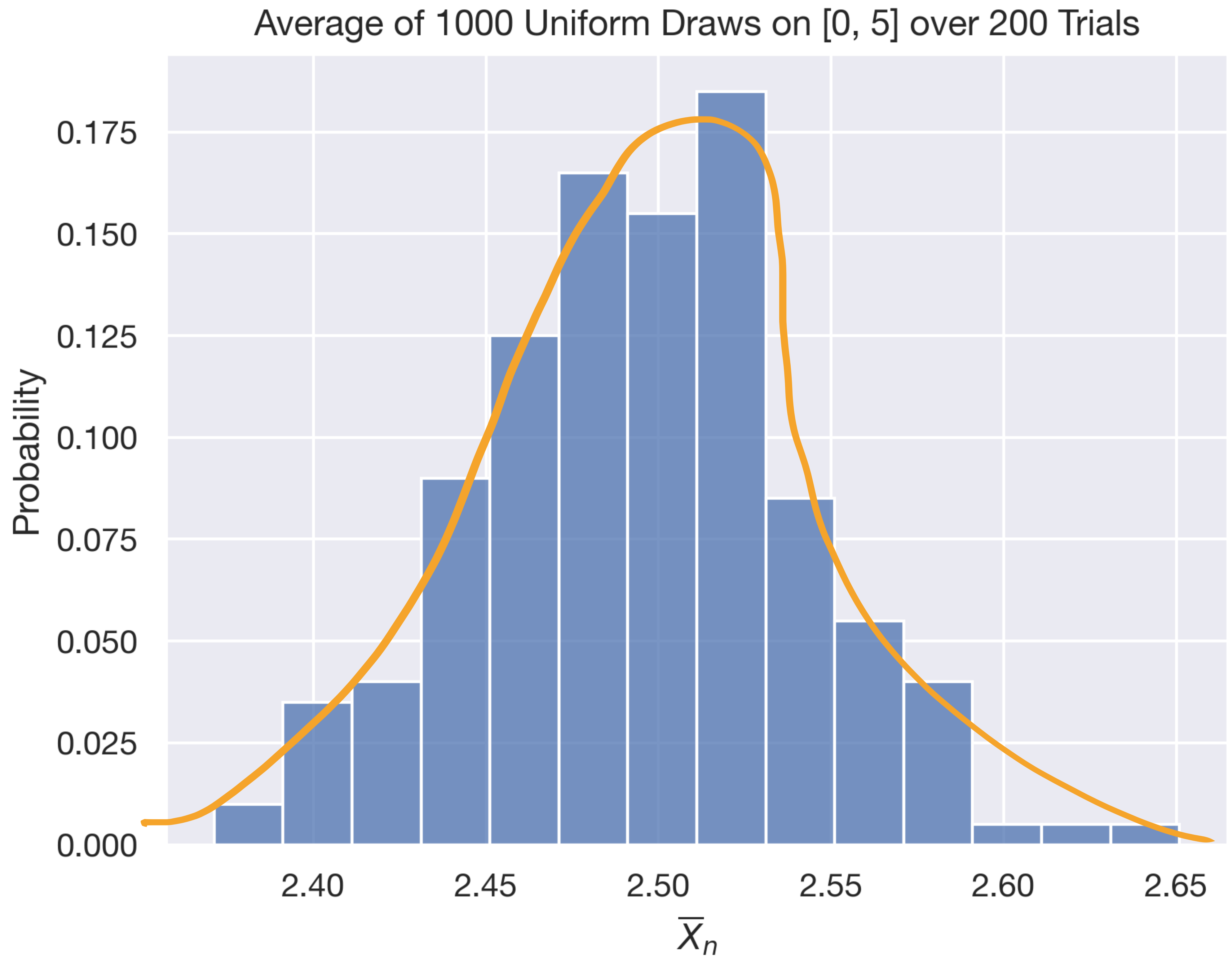
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$





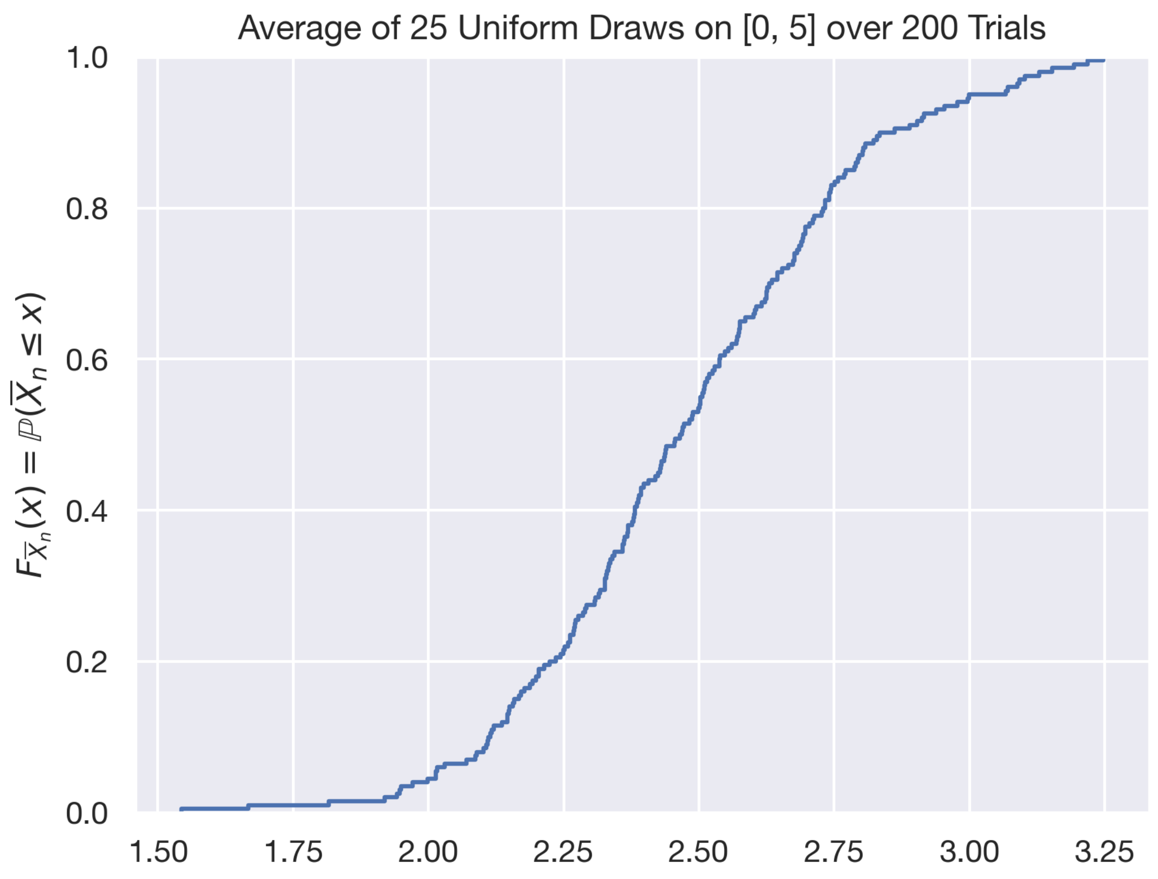


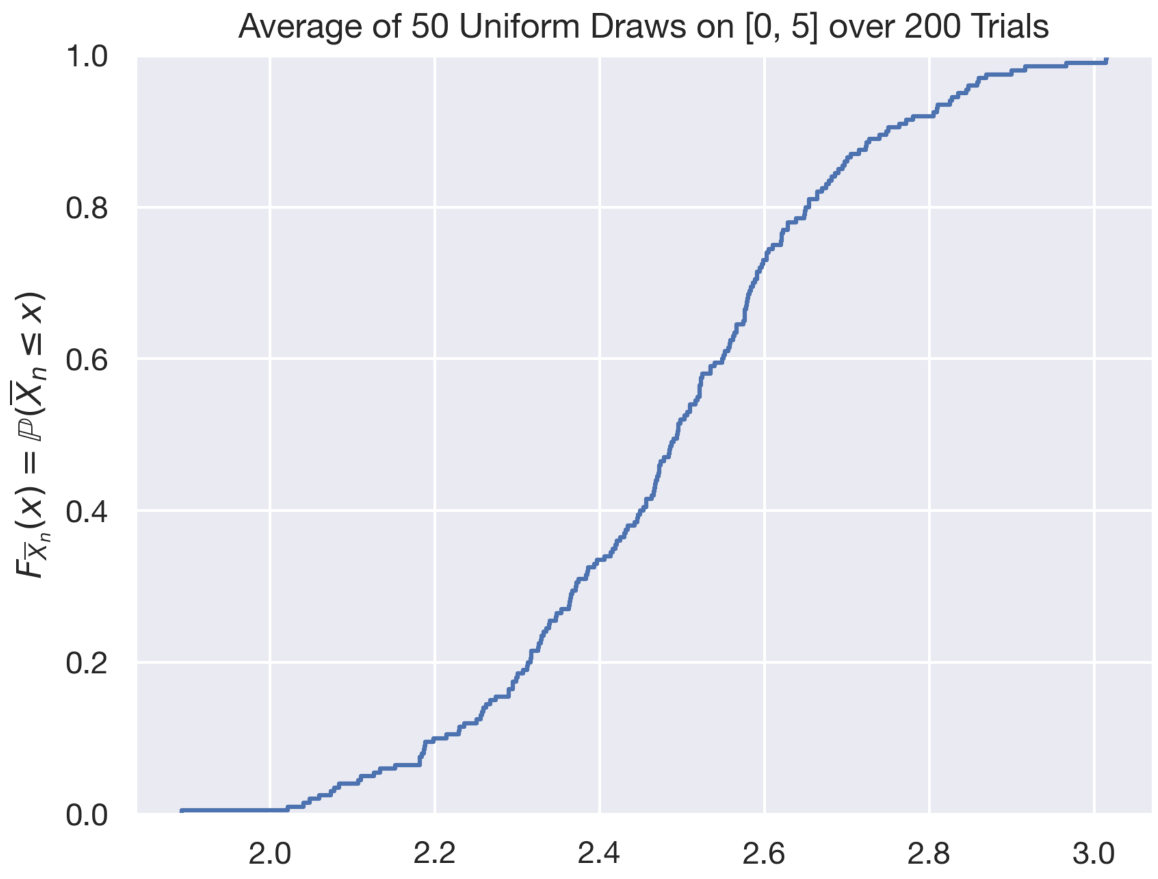


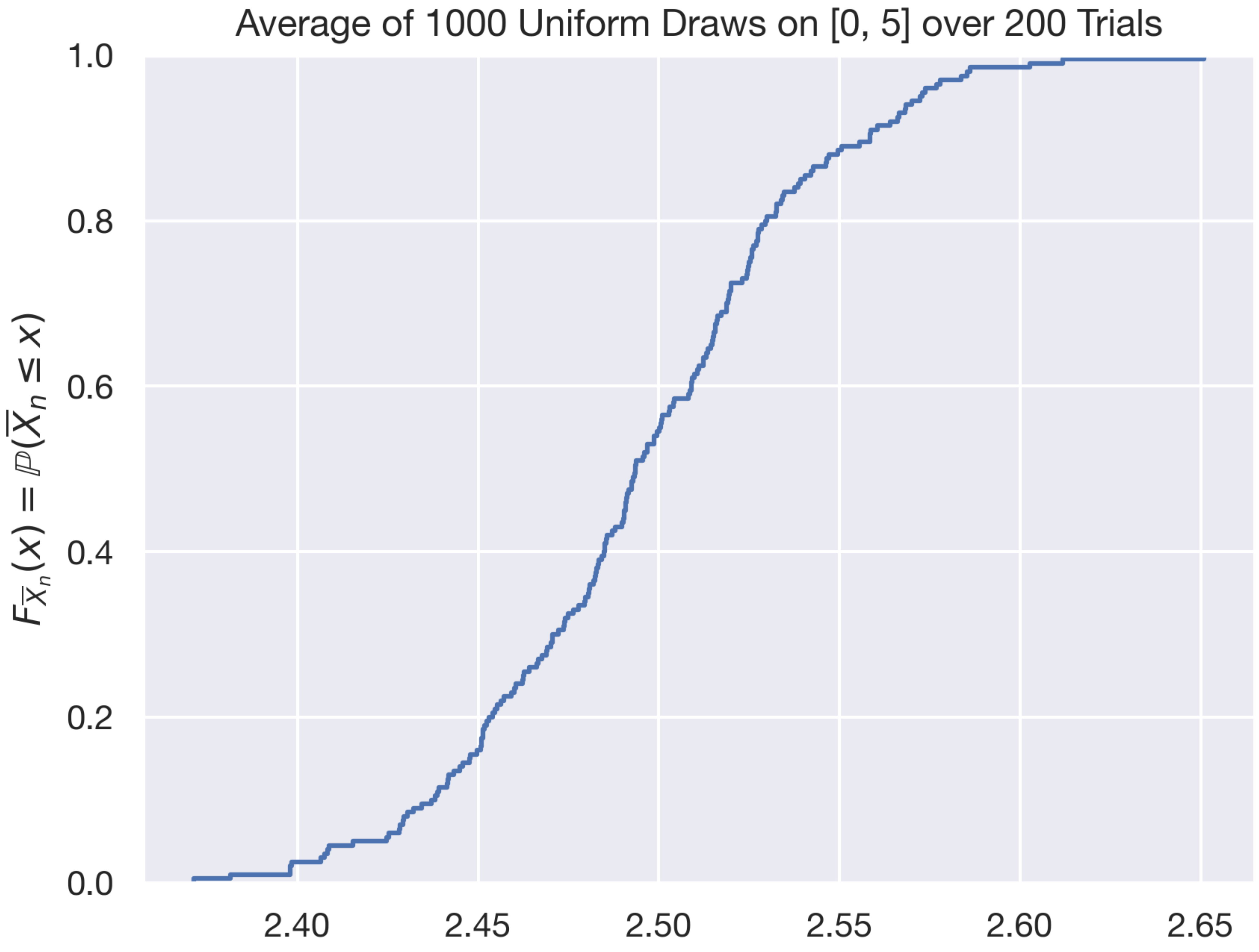


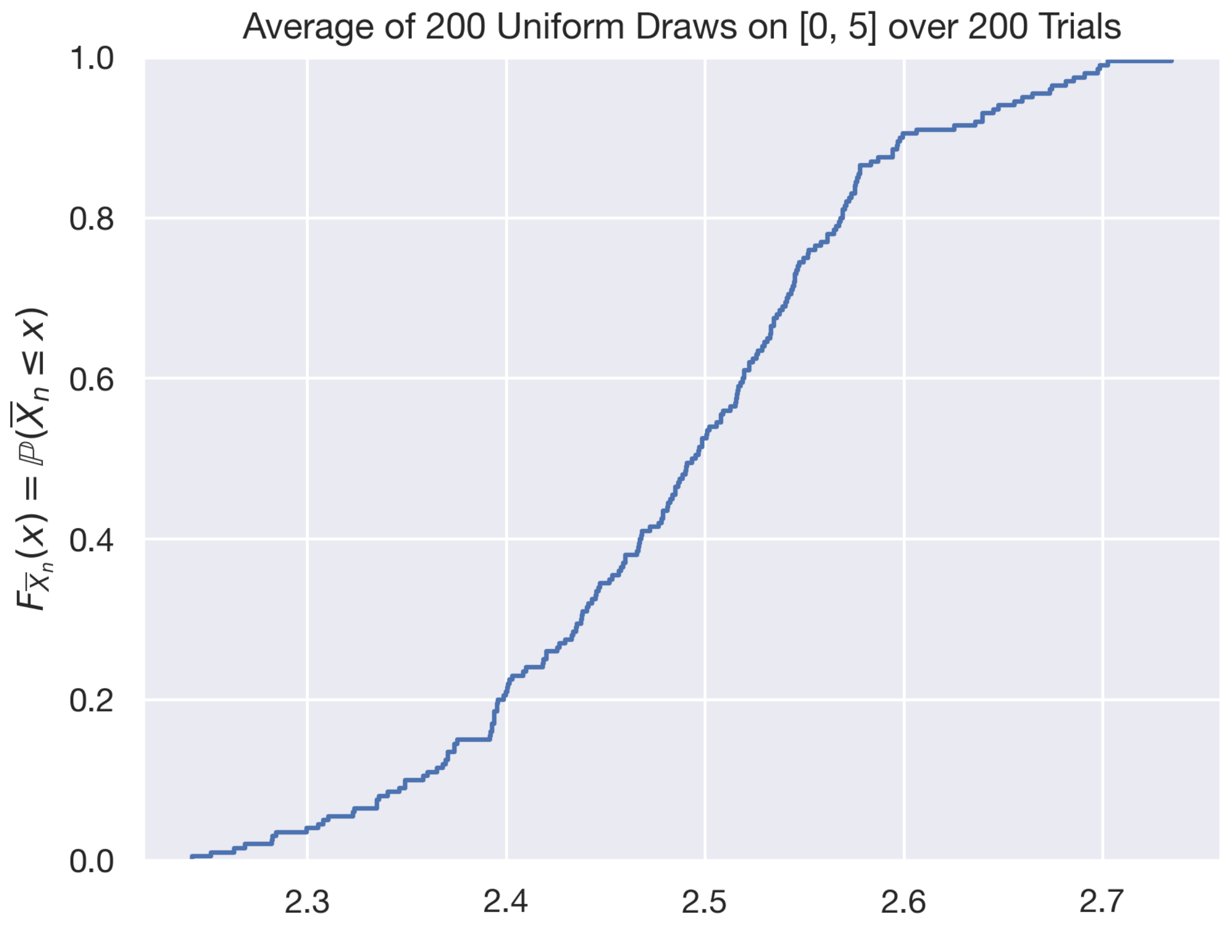
Central Limit Theorem

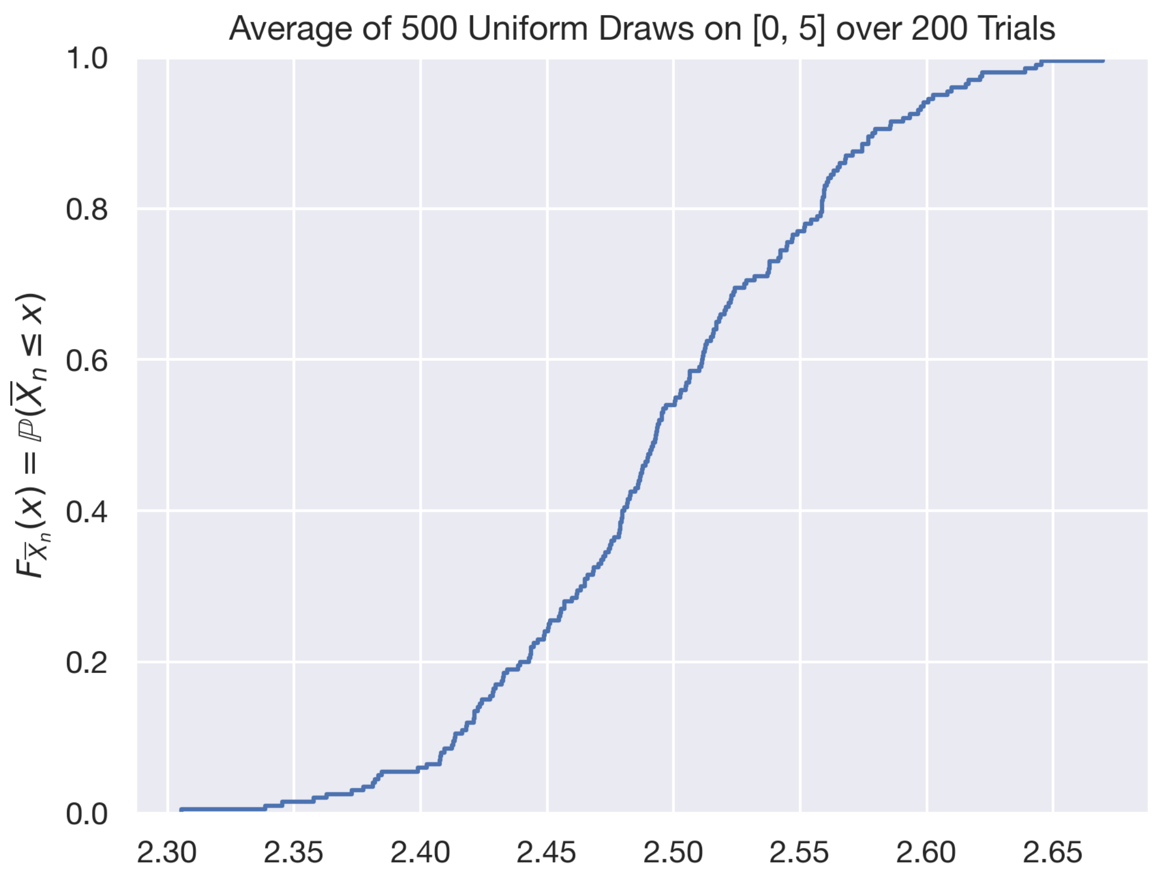
Experiment: Drawing uniform real value











Convergence and MGFs
Tools for CLT Proof

Convergence in Distribution

Intuition

A sequence of random variables $X_1, X_2, X_3, \ldots$ ***converges in distribution*** to another random variable X if:

For large enough n, the distribution of X_n starts looking indistinguishable from the distribution of X.

Convergence in Distribution

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables, and let X be another random variable. Let F_n be the CDF of X_n and let F_X be the CDF of X, so:

$$F_n(x) = \mathbb{P}[X_n \leq x] \text{ and } F_X(x) = \mathbb{P}[X \leq x].$$

Then the sequence (X_n) ***converges in distribution*** to X, written $X_n \to_D X$ if

$$\lim_{n\to\infty} F_n(t) = F_X(t) \text{ for all } t \text{ for which } F_X \text{ is continuous.}$$

Convergence in Distribution

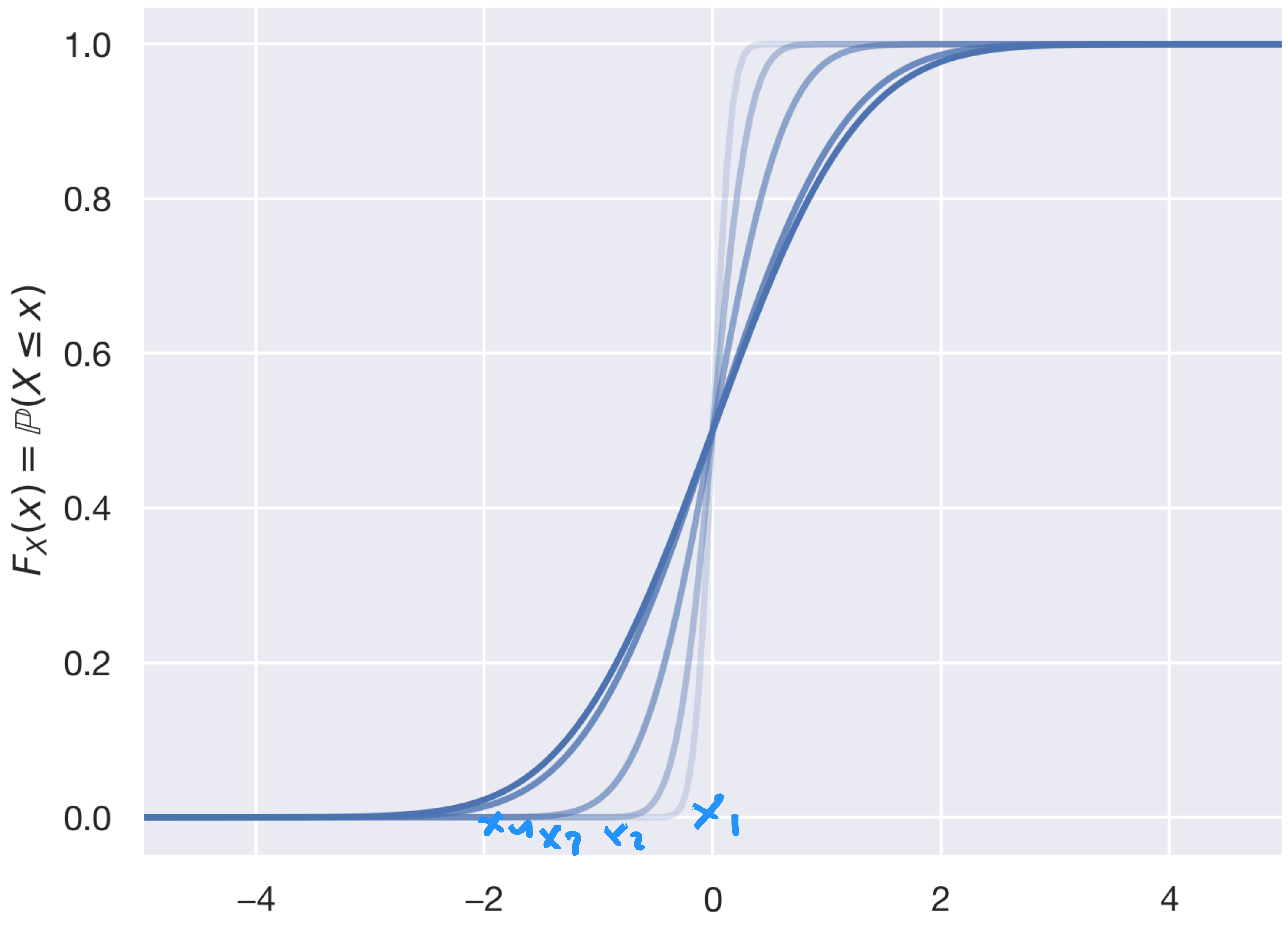
Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables, and let X be another random variable. Let F_n be the CDF of X_n and let F_X be the CDF of X, so:

$$F_n(x) = \mathbb{P}[X_n \leq x] \text{ and } F_X(x) = \mathbb{P}[X \leq x].$$

Then the sequence (X_n) ***converges in distribution*** to X, written $X_n \to_D X$ if

$$\lim_{n\to\infty} F_n(t) = F_X(t) \text{ for all } t \text{ for which } F_X \text{ is continuous.}$$



Moment Generating Function

Intuition

The ***moment generating function (MGF)*** packs all the "moment" information of a random variable X into the Taylor-expandable function e^{tX}.

$$e^X = 1 + X + \frac{X^2}{2} + \frac{X^3}{3!} + \dots$$

Taylor Expansion of e^X.

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2} + \frac{t^3X^3}{3!} + \dots$$

$$\mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + t^2\frac{\mathbb{E}[X^2]}{2} + t^3\frac{\mathbb{E}[X^3]}{3!} + \dots$$

* Technical axioms that allow us to take integrals of infinite sums

Moment Generating Function

Intuition

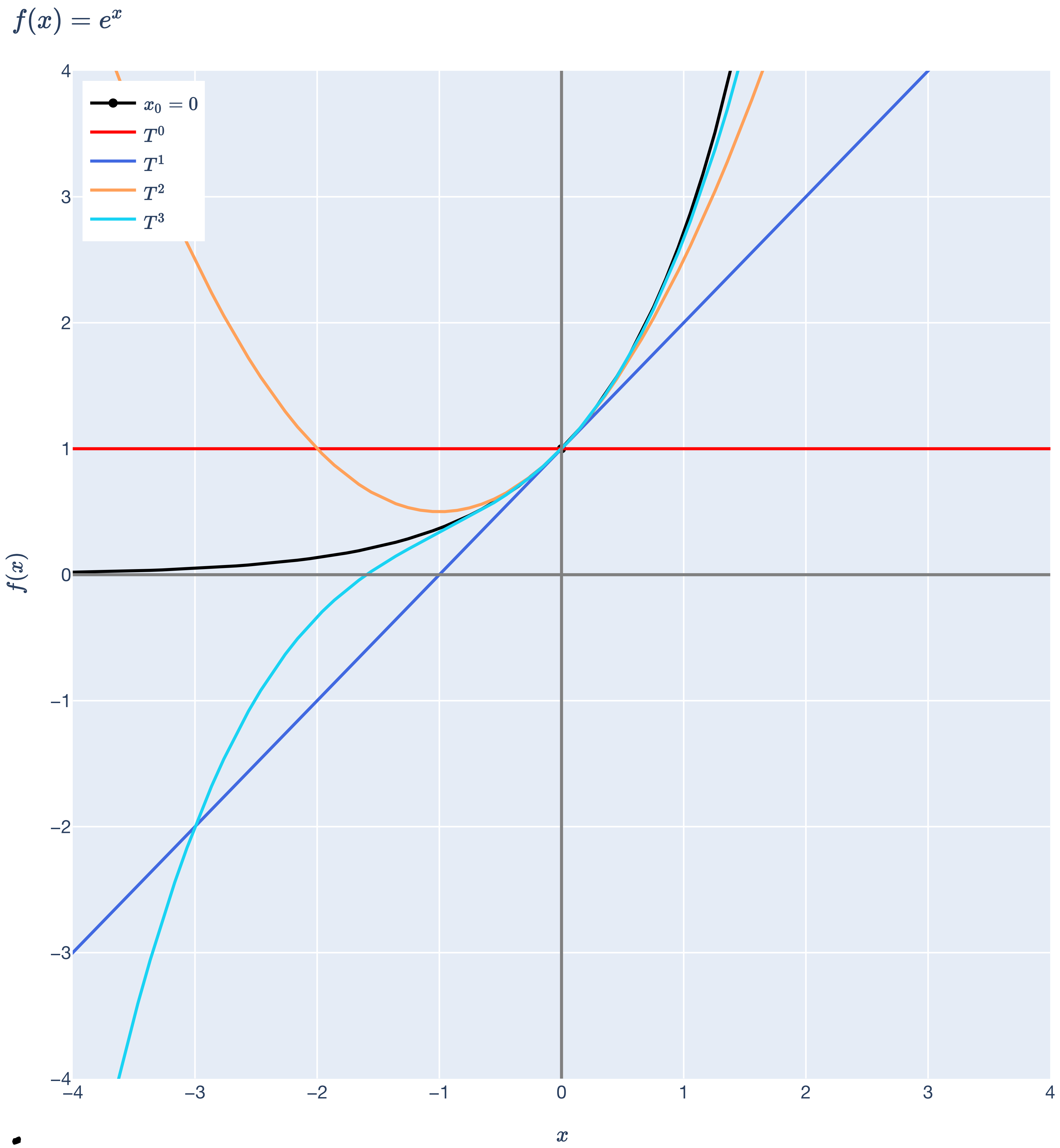
The ***moment generating function (MGF)*** packs all the “moment” information of a random variable X into the Taylor-expandable function e^{tX}.

$$e^X = 1 + X + \frac{X^2}{2} + \frac{X^3}{3!} + \dots$$

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2} + \frac{t^3X^3}{3!} + \dots$$

$$\mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + t^2\frac{\mathbb{E}[X^2]}{2} + t^3\frac{\mathbb{E}[X^3]}{3!} + \dots$$

$$\frac{d}{dt}\mathbb{E}[e^{tX}] = 0 + \mathbb{E}[X] + t\,\mathbb{E}[X^2] + \frac{t^2\,\mathbb{E}[X^3]}{2} + \dots$$



Moment Generating Function

Definition

$X \sim N(0,1)$

$M_X(t) = e^{t^2/2}$

$M_X'(t) = e^{t^2/2} \cdot t$

$M_X'(0) = 0$

The ***moment generating function (MGF)*** of a random variable X is the function $M_X : \mathbb{R} \to \mathbb{R}$ defined by:

$$M_X(t) := \mathbb{E}[e^{tX}] = \int e^{tx} dF_X(x).$$

$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \dots$

If M_X is well-defined in an interval around $t = 0$,

$$M_X'(0) = \left[\frac{d}{dt}\mathbb{E}[e^{tX}]\right]_{t=0} = \mathbb{E}\left[\frac{d}{dt}e^{tX}\right]_{t=0} = \mathbb{E}[Xe^{tX}]_{t=0} = \mathbb{E}[X].$$

Generally, the kth derivative at $t = 0$ gives the ***kth moment of X:***

$$M^{(k)}(0) = \mathbb{E}[X^k].$$

$\mathbb{E}[X^k]$ ← kth moment

Moment Generating Function

MGF characterizes the distribution

Theorem (MGF characterizes distributions). Let X and Y be random variables. If there exists some $\delta \in \mathbb{R}$ where $M_X(t) = M_Y(t)$ for all t in a neighborhood $B_\delta(0)$ around 0, then X and Y have the same distribution:

$$\mathbb{P}_X = \mathbb{P}_Y \text{ and } F_X(t) = F_Y(t) \text{ for their CDFs } F_X \text{ and } F_Y.$$

Moment Generating Function

MGF of Standard Normal

Theorem (MGF characterizes distributions). Let X and Y be random variables. If there exists some $\delta \in \mathbb{R}$ where $M_X(t) = M_Y(t)$ for all t in a neighborhood $B_\delta(0)$ around 0, then X and Y have the same distribution:

$$\mathbb{P}_X = \mathbb{P}_Y \text{ and } F_X(t) = F_Y(t) \text{ for their CDFs } F_X \text{ and } F_Y.$$

Theorem (MGF of Standard Normal). Let $Z \sim N(0,1)$. The MGF of Z exists and is given by:

$$M_Z(t) = e^{t^2/2}.$$

$\mathbb{E}[e^{tZ}] = \int e^{tZ} \, d\mathbb{P}_Z$

(complete the square)

Moment Generating Function

MGF of Standard Normal

Theorem (MGF characterizes distributions). Let X and Y be random variables. If there exists some $\delta \in \mathbb{R}$ where $M_X(t) = M_Y(t)$ for all t in a neighborhood $B_\delta(0)$ around 0, then X and Y have the same distribution:

$$\mathbb{P}_X = \mathbb{P}_Y \text{ and } F_X(t) = F_Y(t) \text{ for their CDFs } F_X \text{ and } F_Y.$$

Theorem (MGF of Standard Normal). Let $Z \sim N(0,1)$. The MGF of Z exists and is given by:

$$M_Z(t) = e^{t^2/2}.$$

Theorem (Sums of independent RVs). If $X_1, \ldots, X_n$ are independent random variables and $S = \sum_{i=1}^{n} X_i$, then $M_S(t) = \prod_{i=1}^{n} M_{X_i}(t)$ where $M_{X_i}(t)$ is the MGF of X_i.

Pf: $\mathbb{E}[e^{t\sum_{i=1}^{n} X_i}] = \mathbb{E}[e^{tX_1} \cdot e^{tX_2} \cdots e^{tX_n}]$

$= \mathbb{E}[e^{tX_1}] \cdot \mathbb{E}[e^{tX_2}] \cdots \mathbb{E}[e^{tX_n}]$

Product of individual MGFs of X_i.

Central Limit Theorem

Proof and Implications

Central Limit Theorem

Theorem Statement

Theorem (Central Limit Theorem). Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with finite mean $\mu := \mathbb{E}[X_i]$ and finite variance $\sigma^2 := \mathrm{Var}(X_i)$. Let their *sample average* be denoted as $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ and let their "standardized" average be:

$$Z_n := \frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$

$\sqrt{\mathrm{Var}(\overline{X}_n)} = \frac{\sigma}{\sqrt{n}}$

Then, Z_n converge to $Z \sim N(0,1)$ in distribution. That is, $Z_n \to_D Z$:

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \Phi(z) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \mathbb{P}(Z \le z).$$

Central Limit Theorem

Theorem Statement

Theorem (Central Limit Theorem). Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with finite mean $\mu := \mathbb{E}[X_i]$ and finite variance $\sigma^2 := \mathrm{Var}(X_i)$. Let their *sample average* be denoted as $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ and let their "standardized" average be:

$$Z_n := \frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

Then, Z_n converge to $Z \sim N(0,1)$ in distribution. That is, $Z_n \to_D Z$:

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \mathbb{P}(Z \leq z).$$

Probability statements about $\overline{X}_n$ can be approximated using a Gaussian distribution!

Central Limit Theorem

Proof of CLT

Without loss of generality, assume $\mu = 0$.

Goal: Show the MGF of $Z_n := \sqrt{n}\overline{X}_n/\sigma$ approaches $M_Z(t) = e^{t^2/2}$.

Central Limit Theorem

Proof of CLT

Without loss of generality, assume $\mu = 0$.

Goal: Show the MGF of $Z_n := \sqrt{n}\overline{X}_n/\sigma$ approaches $M_Z(t) = e^{t^2/2}$.

Step 1: Use MGF property on sums of independent random variables.

Let $S_n := \sum_{i=1}^{n} X_i$, so $Z_n = \dfrac{S_n}{\sigma\sqrt{n}}$. Because $X_1, \ldots, X_n$ are independent, the MGF follows:

$$M_{S_n}(t) = \left(M_{X_i}(t)\right)^n,$$ where M_{X_i} is the MGF of any X_i.

Therefore,

$$M_{Z_n}(t) = \left(M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

Central Limit Theorem

Proof of CLT

Without loss of generality, assume $\mu = 0$.

Goal: Show the MGF of $Z_n := \sqrt{n}\overline{X}_n/\sigma$ approaches $M_Z(t) = e^{t^2/2}$.

Step 2: Use Taylor expansion and (Peano's) Taylor's Theorem on $M_X(s)$ for some s.

From Step 1,

$s = \frac{t}{\sigma\sqrt{n}}$

$$M_{Z_n}(t) = \left(M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

Now, expand the Taylor series of $M_X(s)$ around $s = 0$:

$$M_X(s) = M_X(0) + sM_X'(0) + \frac{1}{2}s^2M_X''(0) + R(s),$$ where $R(s)$ is a remainder.

By Peano's form of Taylor's Theorem, $R(s)/s^2 \to 0$ as $s \to 0$.

Central Limit Theorem

Proof of CLT

Without loss of generality, assume $\mu = 0$.

Goal: Show the MGF of $Z_n := \sqrt{n}\overline{X}_n/\sigma$ approaches $M_Z(t) = e^{t^2/2}$.

Step 3: Plug in the moments $M_X(0)$, $M_X'(0)$ and $M_X''(0)$.

We know that $M_X(0) = 1$ (from def of MGF), $M_X'(0) = \mu = 0$ (from assumption), and $M_X''(0) = \sigma^2$ (from definition of $\mathrm{Var}(X)$). Plug these in:

$$M_X(s) = M_X(0) + sM_X'(0) + \frac{1}{2}s^2M_X''(0) + R(s)$$

$$\implies M_X(s) = 1 + \frac{\sigma^2}{2}s^2 + R(s).$$

$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$M_X'(0) = \mathbb{E}[X] = 0.$

$M_X''(0) = \mathbb{E}[X^2] = \mathrm{Var}(X) = \sigma^2.$

$M_X(0) = \mathbb{E}[e^0] = 1.$

Central Limit Theorem

Proof of CLT

Without loss of generality, assume $\mu = 0$.

Goal: Show the MGF of $Z_n := \sqrt{n}\overline{X}_n/\sigma$ approaches $M_Z(t) = e^{t^2/2}$.

Step 4: Replace $s = t/\sigma\sqrt{n}$ and get back to the MGF of interest, $M_{Z_n}(t)$.

Let $s = \dfrac{t}{\sigma\sqrt{n}}$, so $s^2 = \dfrac{t^2}{\sigma^2 n}$. From Step 3,

$$M_X(s) = 1 + \frac{\sigma^2}{2}s^2 + R(s)$$
$$\implies M_X\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + R(s).$$

From Step 1, we have found:

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + R(s)\right)^n.$$

Central Limit Theorem

Proof of CLT

Without loss of generality, assume $\mu = 0$.

Goal: Show the MGF of $Z_n := \sqrt{n}\overline{X}_n/\sigma$ approaches $M_Z(t) = e^{t^2/2}$.

Step 5: Send $n \to \infty$ and exploit Peano's form of Taylor's Theorem to conclude.

Let $s = \dfrac{t}{\sigma\sqrt{n}}$, so $s^2 = \dfrac{t^2}{\sigma^2 n}$. From Step 4,

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + R(s)\right)^n.$$

0

$s \to 0 \Rightarrow R(s) \to 0$
$n \to \infty$

As $n \to \infty$, $s^2 \to 0$ and $R(s)/s^2 \to 0$. By definition of e^a for some $a \in \mathbb{R}$,

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a \text{ if } a_n \to a \Longrightarrow \lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2} = M_Z(t).$$

The Gaussian Distribution

Properties of Gaussians

Standardization. If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0,1)$. As a result:

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Standard to general. If $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.

Sums of Gaussians. If $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ are independent, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Central Limit Theorem

Equivalent Approximations

$\approx$:= limiting distribution as $n \to \infty$.

For i.i.d. random variables $X_1, \ldots, X_n$, let:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$Z_n := \frac{\bar{X}_n - \mu}{\sqrt{\text{Var}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

$$\Longrightarrow$$

For large enough n, the CLT statement allows the equivalent approximations…

$$Z_n \approx N(0,1)$$

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

$$\bar{X}_n - \mu \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2)$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx N(0,1)$$

Central Limit Theorem

Equivalent Approximations

For i.i.d. random variables $X_1, \ldots, X_n$, let:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Z_n := \frac{\bar{X}_n - \mu}{\sqrt{\mathrm{Var}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

$$\Longrightarrow$$

For large enough n, the CLT statement allows the equivalent approximations…

$$Z_n \approx N(0,1)$$

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

$$\bar{X}_n - \mu \approx N\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \approx N(0,\sigma^2)$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx N(0,1)$$

Central Limit Theorem

Two Implications

SAMPLING DISTRIBUTION := distribution of $\bar{X}_n$.

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ for large enough } n.$$

This says two things:

1. The mass of $\bar{X}_n$ centers to μ, the true mean of the i.i.d. random variables.

2. The spread of draws from $\bar{X}_n$ gets smaller and smaller as n grows.

Central Limit Theorem

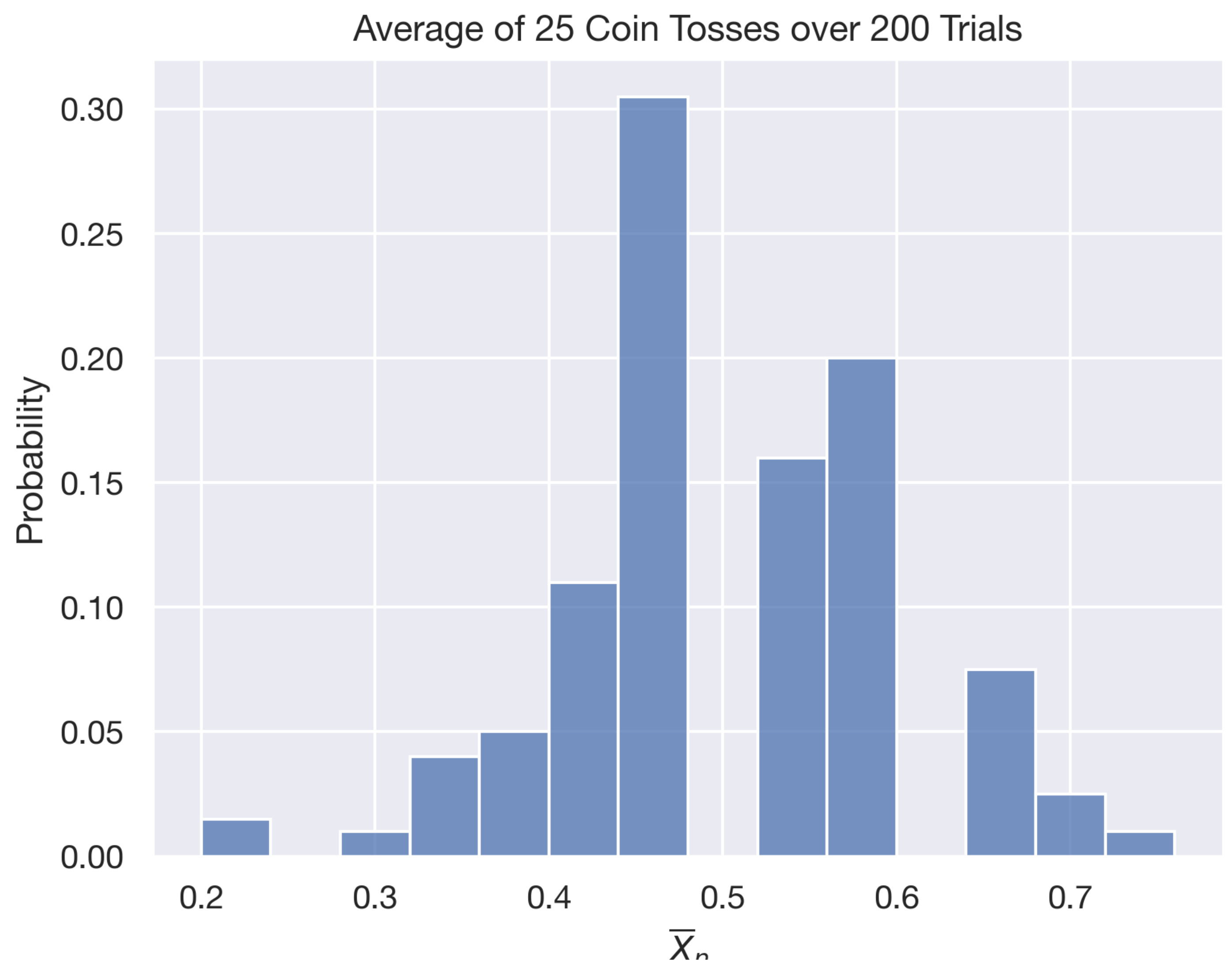
Two Implications

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$ for large enough n.

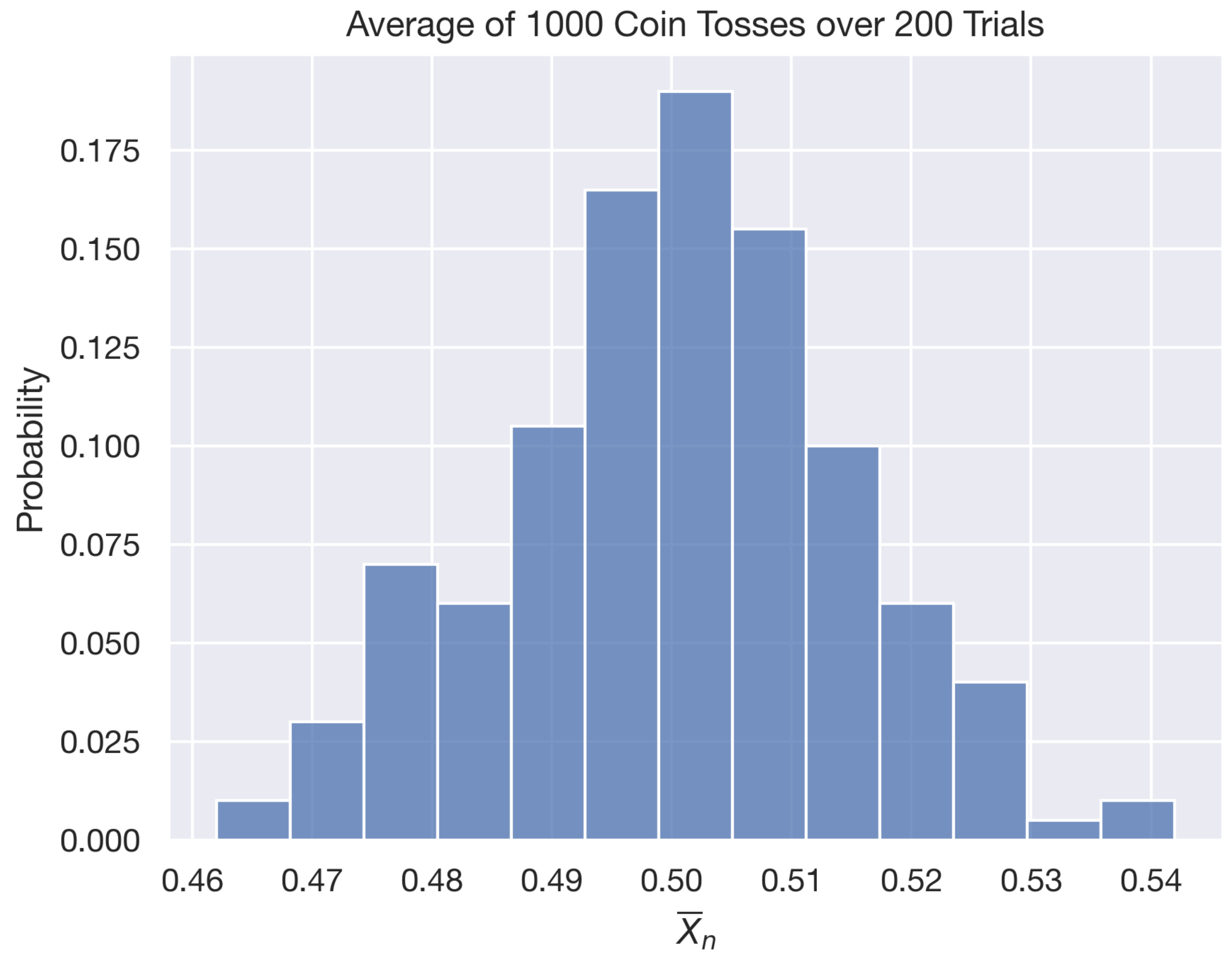
This says two things:

1. The mass of $\overline{X}_n$ centers to μ, the true mean of the i.i.d. random variables.

2. The spread of draws from $\overline{X}_n$ gets smaller and smaller as n grows.



n = 25



n = 1000

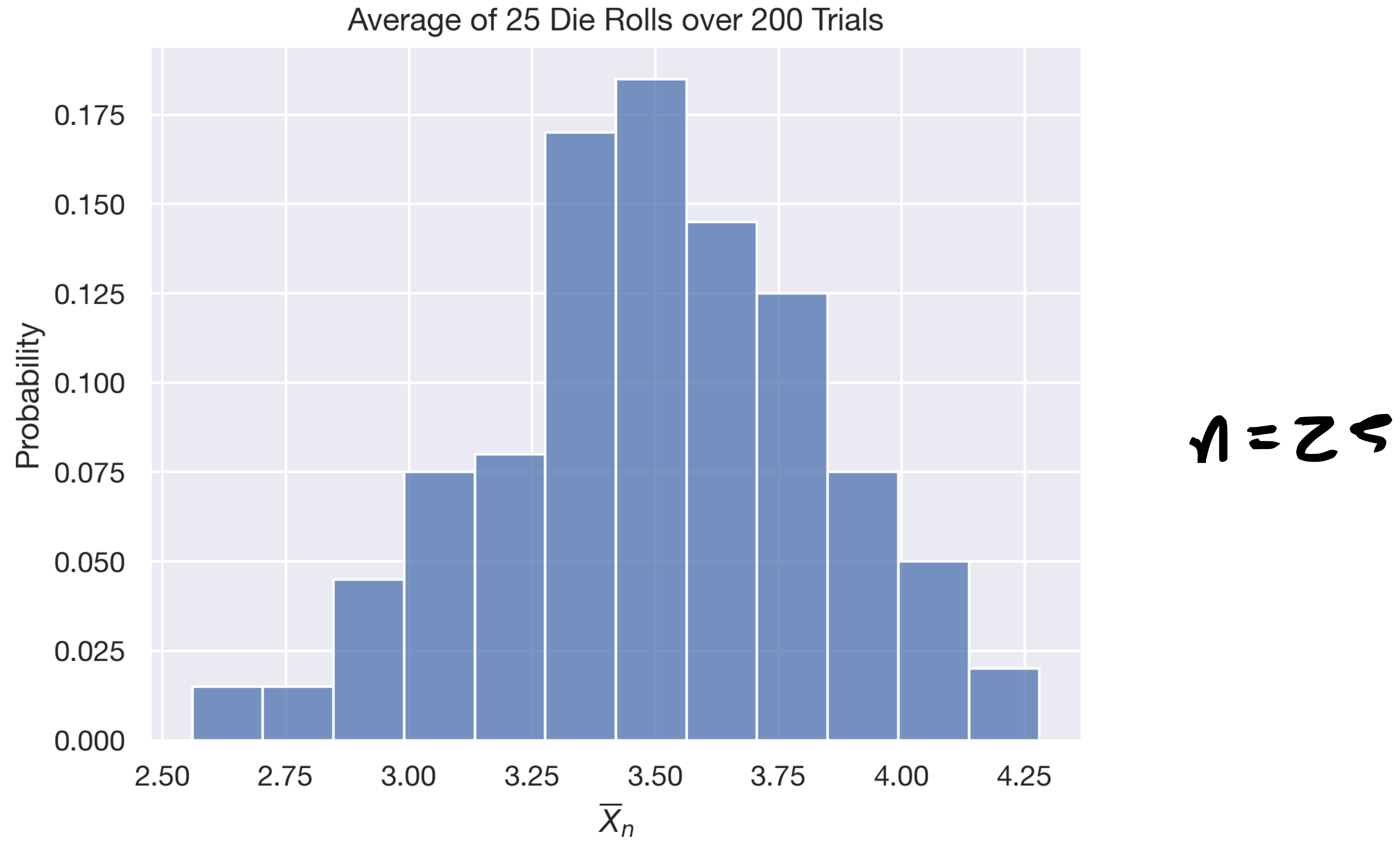
Central Limit Theorem
Two Implications

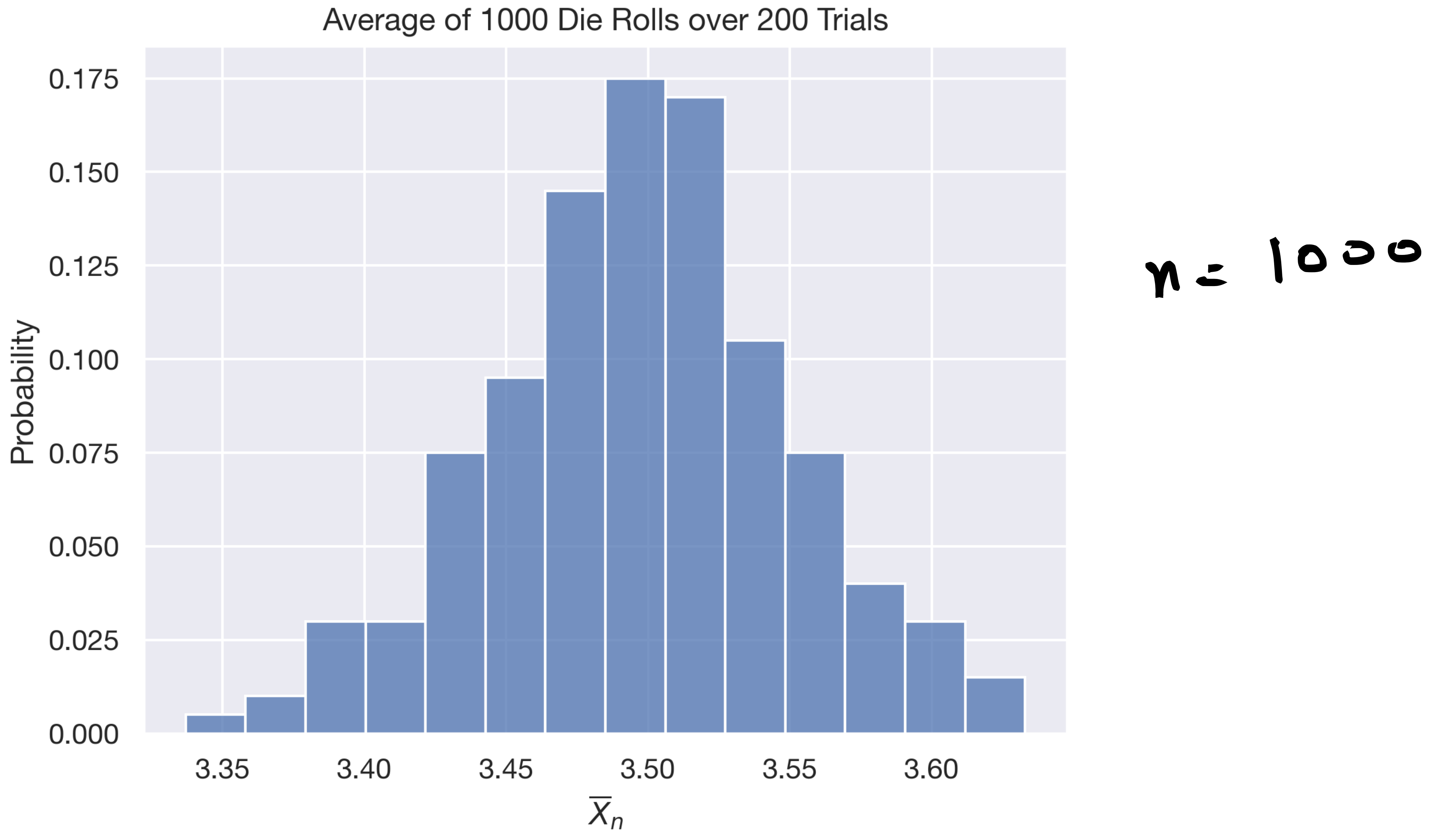
$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$ for large enough n.

This says two things:

1. The mass of $\overline{X}_n$ centers to μ, the true mean of the i.i.d. random variables.

2. The spread of draws from $\overline{X}_n$ gets smaller and smaller as n grows.





Central Limit Theorem

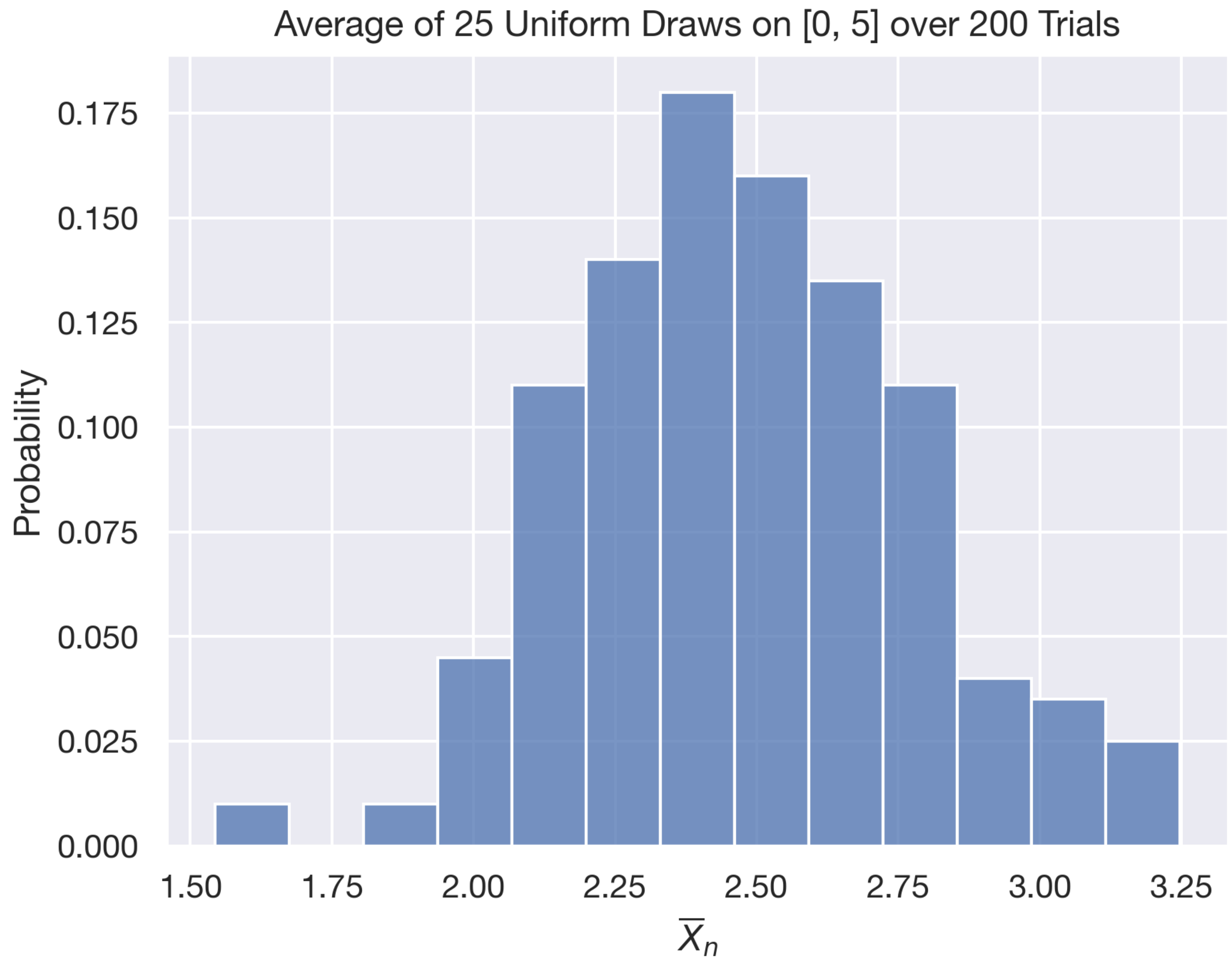
Two Implications

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$ for large enough n.

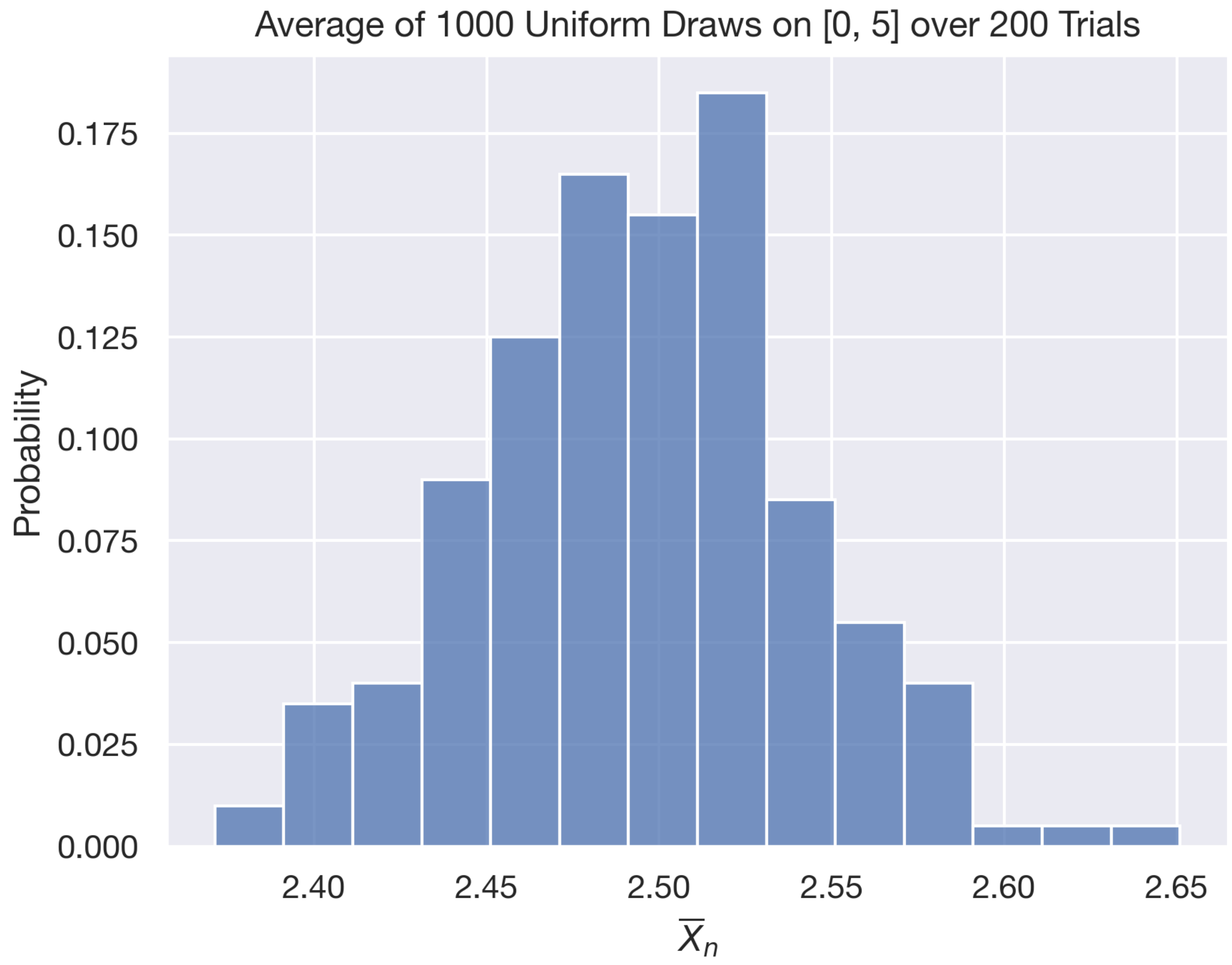
This says two things:

1. The mass of $\overline{X}_n$ centers to μ, the true mean of the i.i.d. random variables.

2. The spread of draws from $\overline{X}_n$ gets smaller and smaller as n grows.



$n = 25$



$n = 1000$

“Named” Distributions

Discrete Examples

Discrete Distributions

Discrete Random Variables

A ***discrete random variable*** X takes on a finite or countably infinite number of values.

CDF. $F_X(x) := \mathbb{P}(X \leq x)$.

A = event (subset of $\mathbb{R}$).

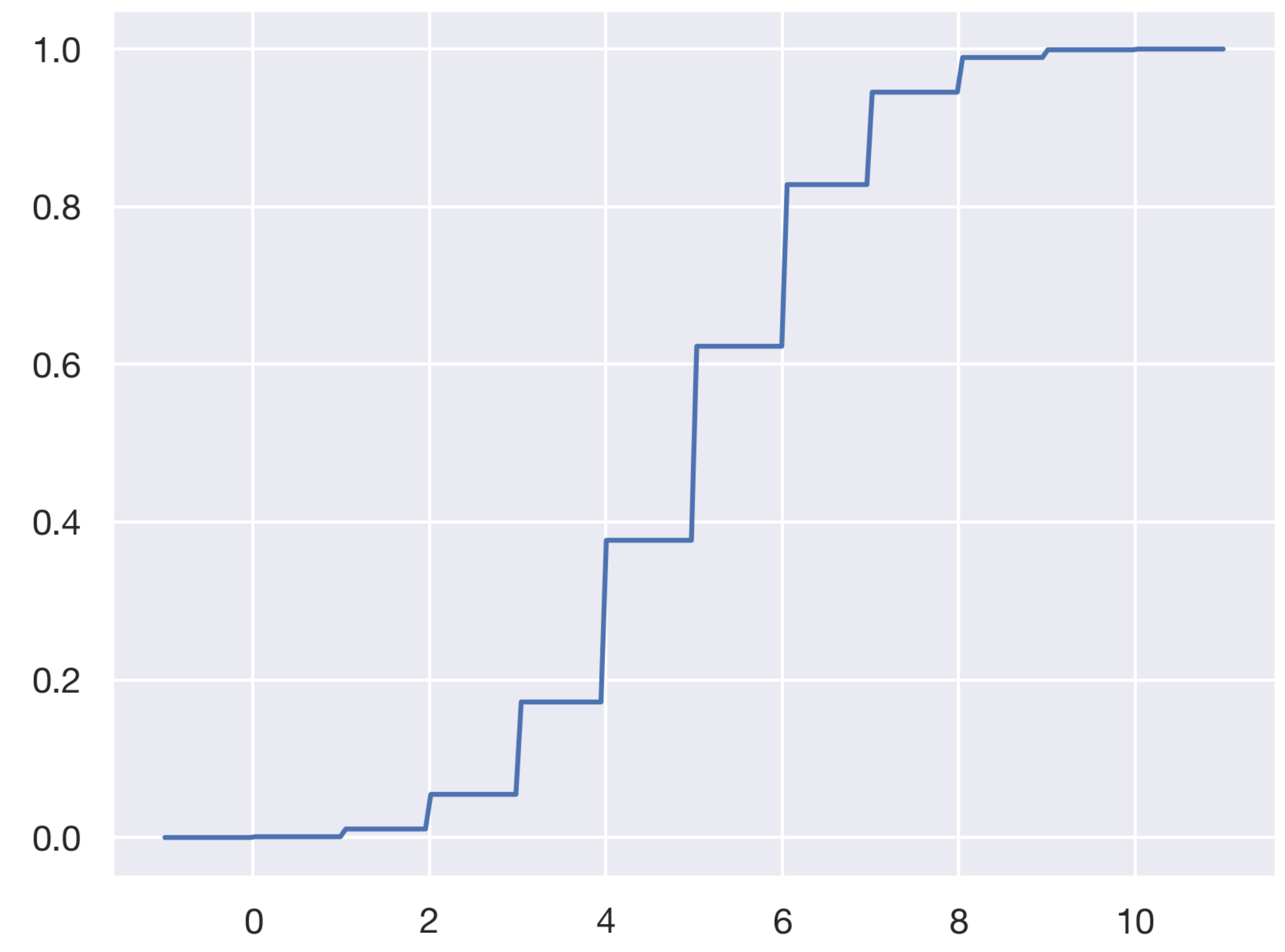
PMF. $p_X(x) = \mathbb{P}(X = x)$ and $\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x)$.

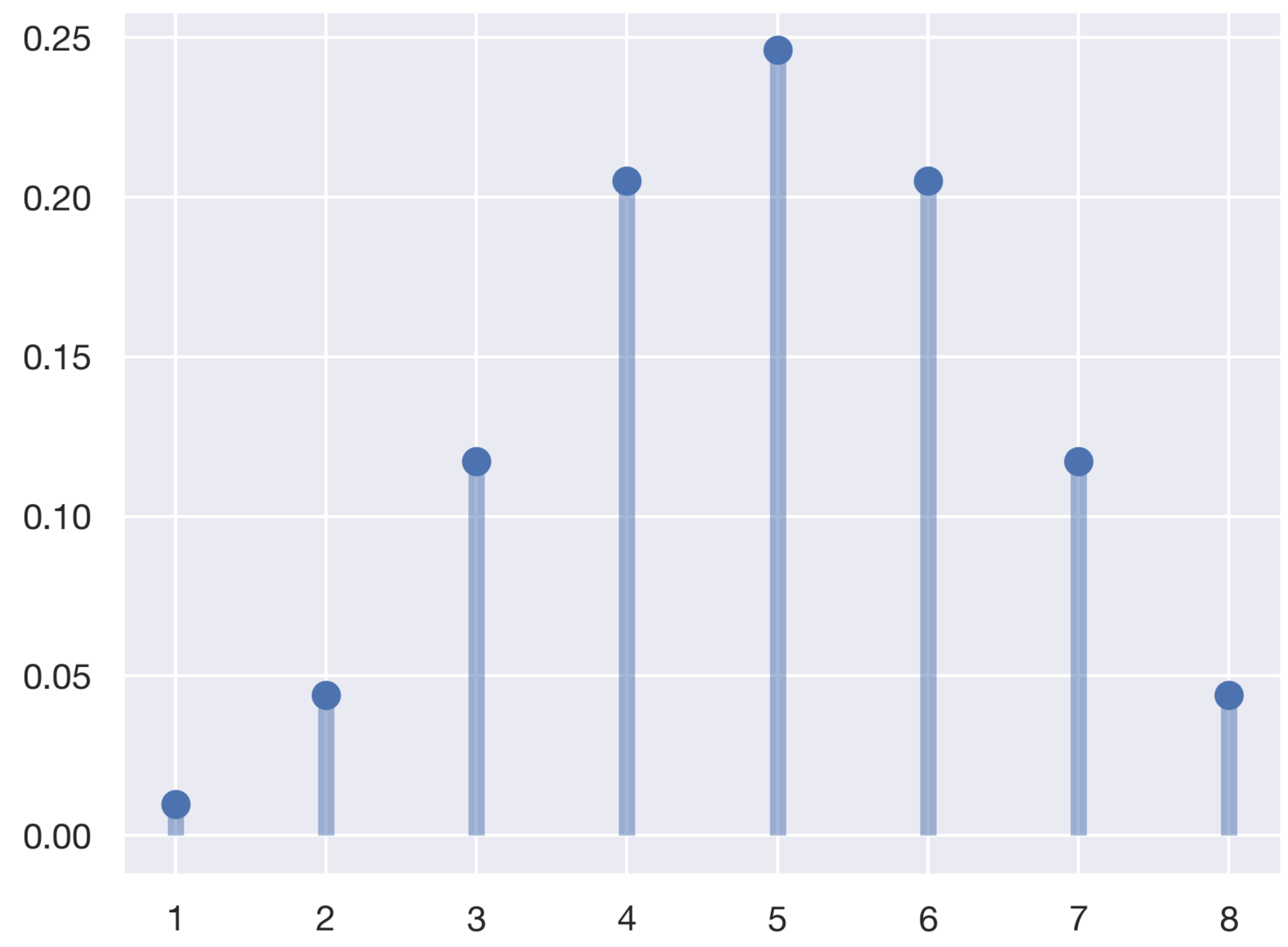
PMF is the height of the "jump" of F_X at x.

PMF is nonnegative.

PMF sums to 1.

Expectation. $\mathbb{E}(X) = \sum_x x p_X(x) = \sum_x x \mathbb{P}(X = x)$.





The Point Mass Distribution

"Story" of the Distribution

A single point $a \in \mathbb{R}$ has all the probability mass, every other point has zero mass.

Example. Let X be a random variable putting all its mass on $a = 1$.

The Point Mass Distribution

Properties

$X \sim \delta_a$ ← kronecker delta

Parameters: $a \in \mathbb{R}$, the point mass.

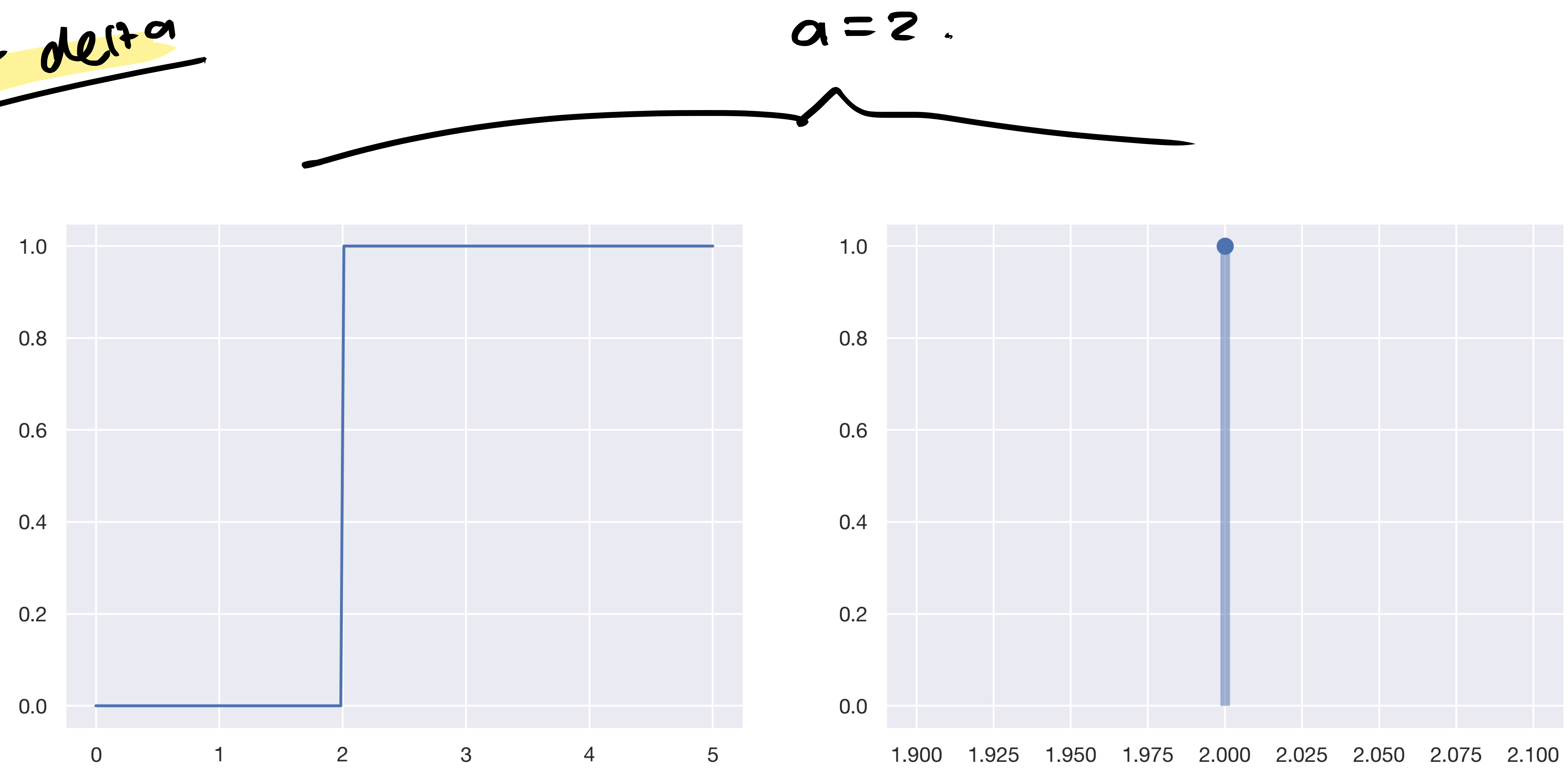
CDF: $F_X(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a \end{cases}$

PMF: $p_X(x) = \begin{cases} 1 & x = a \\ 0 & x \neq a \end{cases}$

Mean: $\mathbb{E}[X] = a$.

Variance: $\mathrm{Var}(X) = 0$.

MGF: $M_X(t) = e^{ta}$. $\mathbb{E}[e^{tX}] = e^{ta}$



The Discrete Uniform Distribution

“Story” of the Distribution

Randomly choose an element in a finite set S, with equal probability for each element.

Example. Let X be the number on the roll of a fair, six-sided die.

The Discrete Uniform Distribution

Properties

$\lfloor x \rfloor$ = "Round x down."

$$X \sim \mathrm{DUnif}(k)$$

k = 5

Parameters: $k \in \mathbb{N}$, the number of possible states, denoted $\{1, 2, \ldots, k\}$.

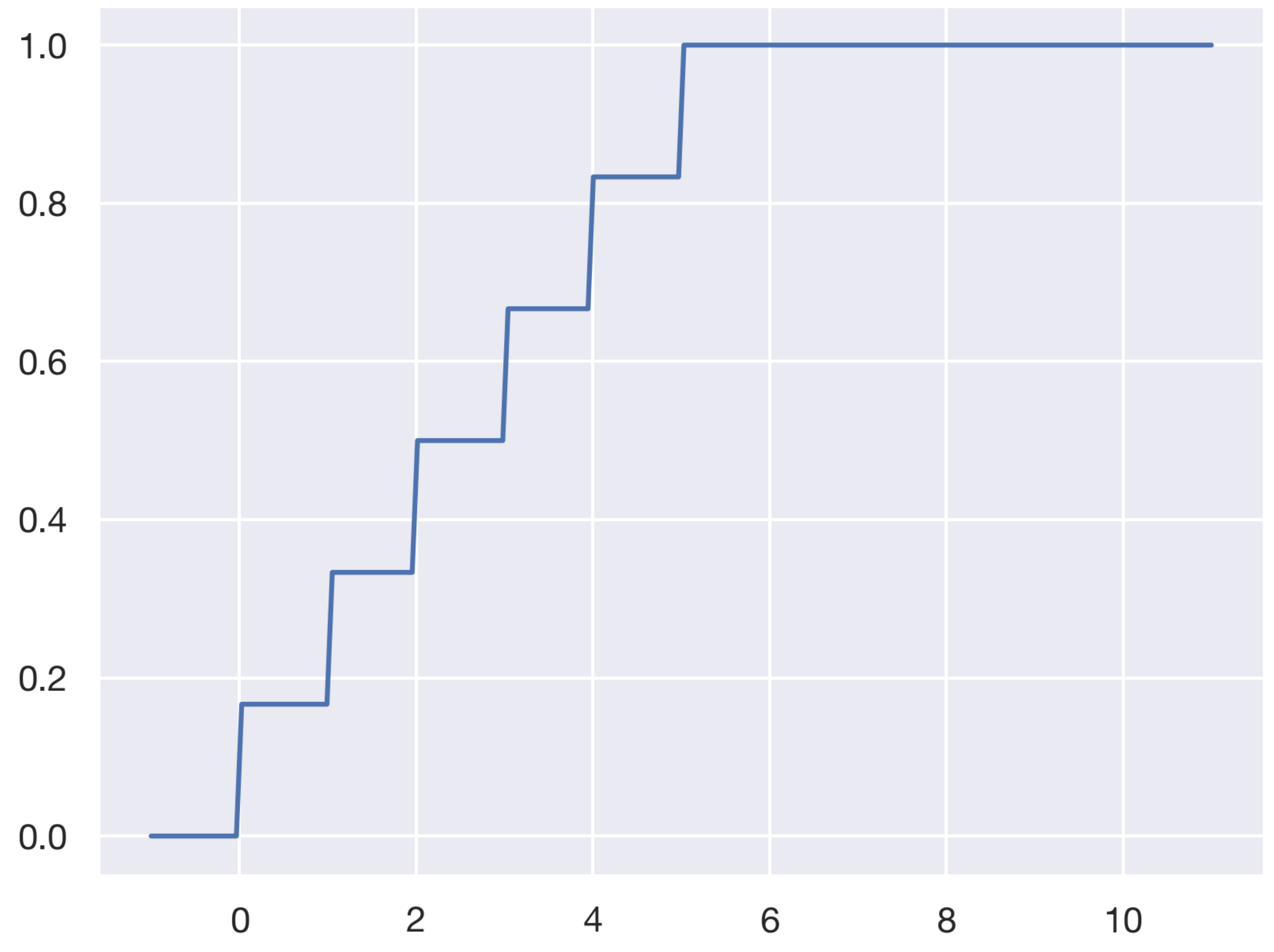
CDF: $F_X(x) = \dfrac{\lfloor k \rfloor}{n}$

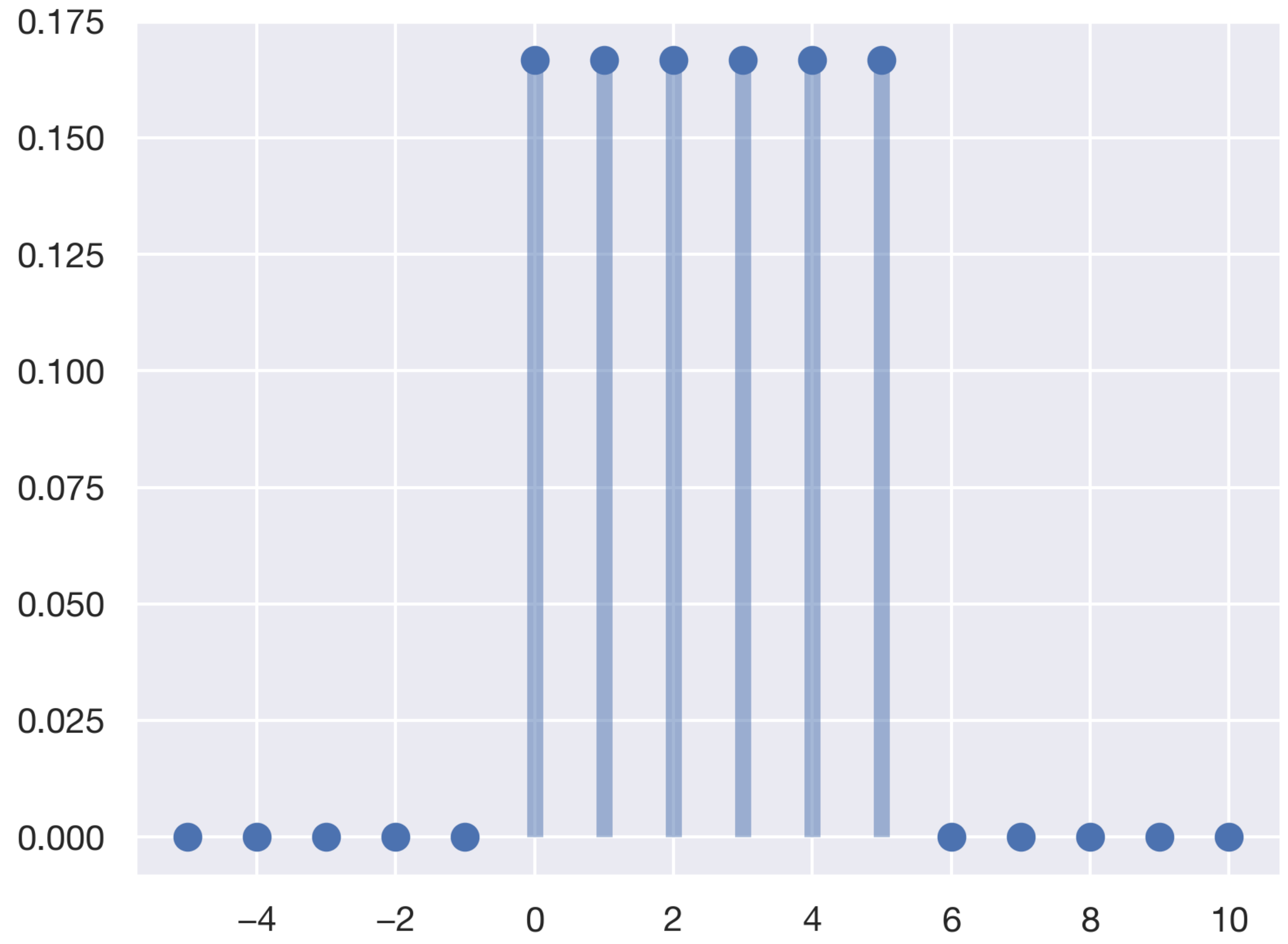
PMF: $p_X(x) = \begin{cases} 1/k & x = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}$

Mean: $\mathbb{E}[X] = \dfrac{k+1}{2}$.

Variance: $\mathrm{Var}(X) = \dfrac{k^2 - 1}{12}$.

MGF: $M_X(t) = \dfrac{e^t(1 - e^{kt})}{k(1 - e^t)}$.





The Bernoulli Distribution

"Story" of the Distribution

Flip a coin that lands heads with probability p and tails with probability $1 - p$.

Example. Let X denote the outcome of a presidential election with two candidates and a tie-breaking mechanism, with 1 indicating Candidate A and 0 indicating Candidate B.

The Bernoulli Distribution

Properties

Indicator RVs : $\mathbb{1}\{\text{Event}\}$

$$X \sim \text{Ber}(p)$$

Parameters: $p \in [0,1]$, the success probability.

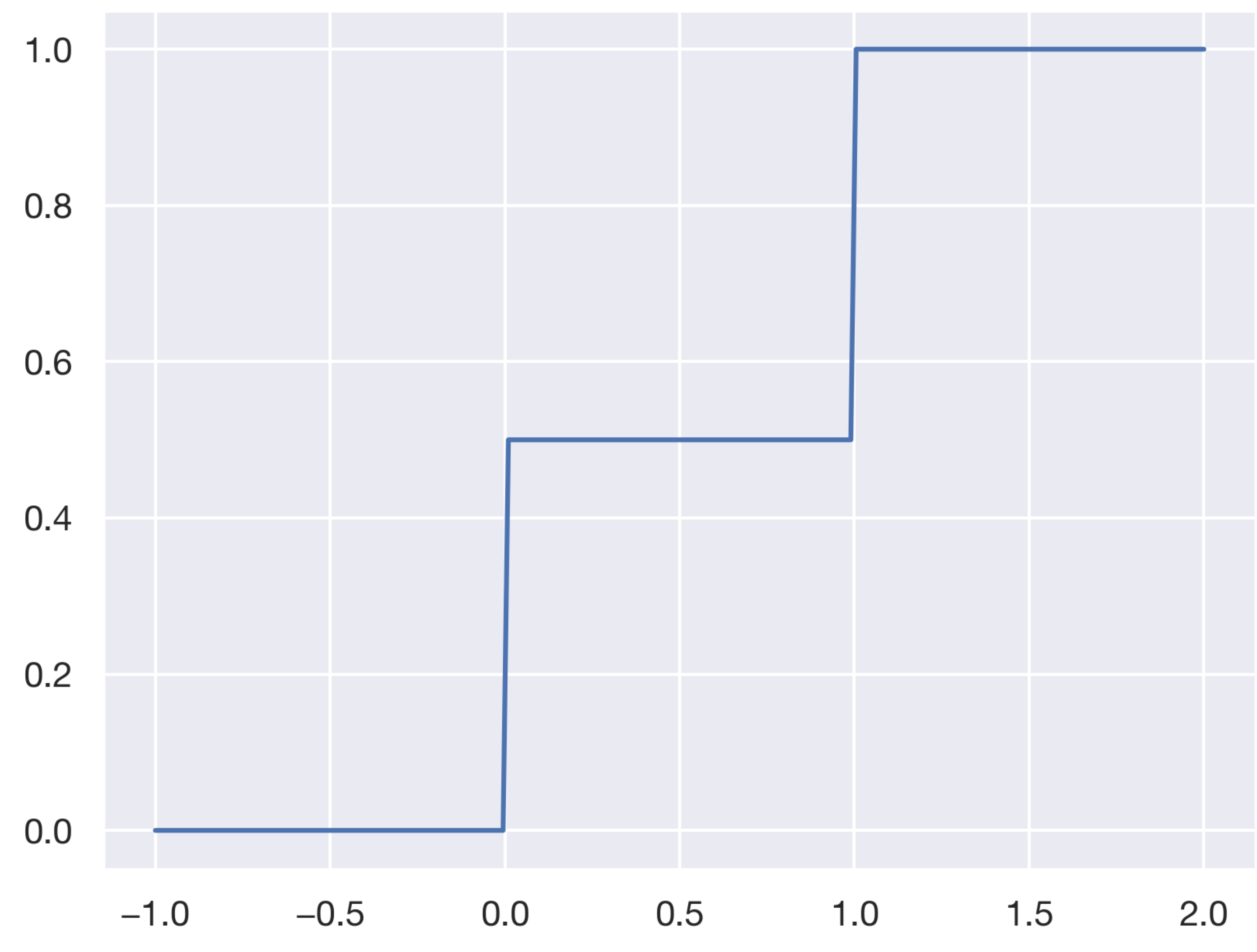
CDF: $F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$

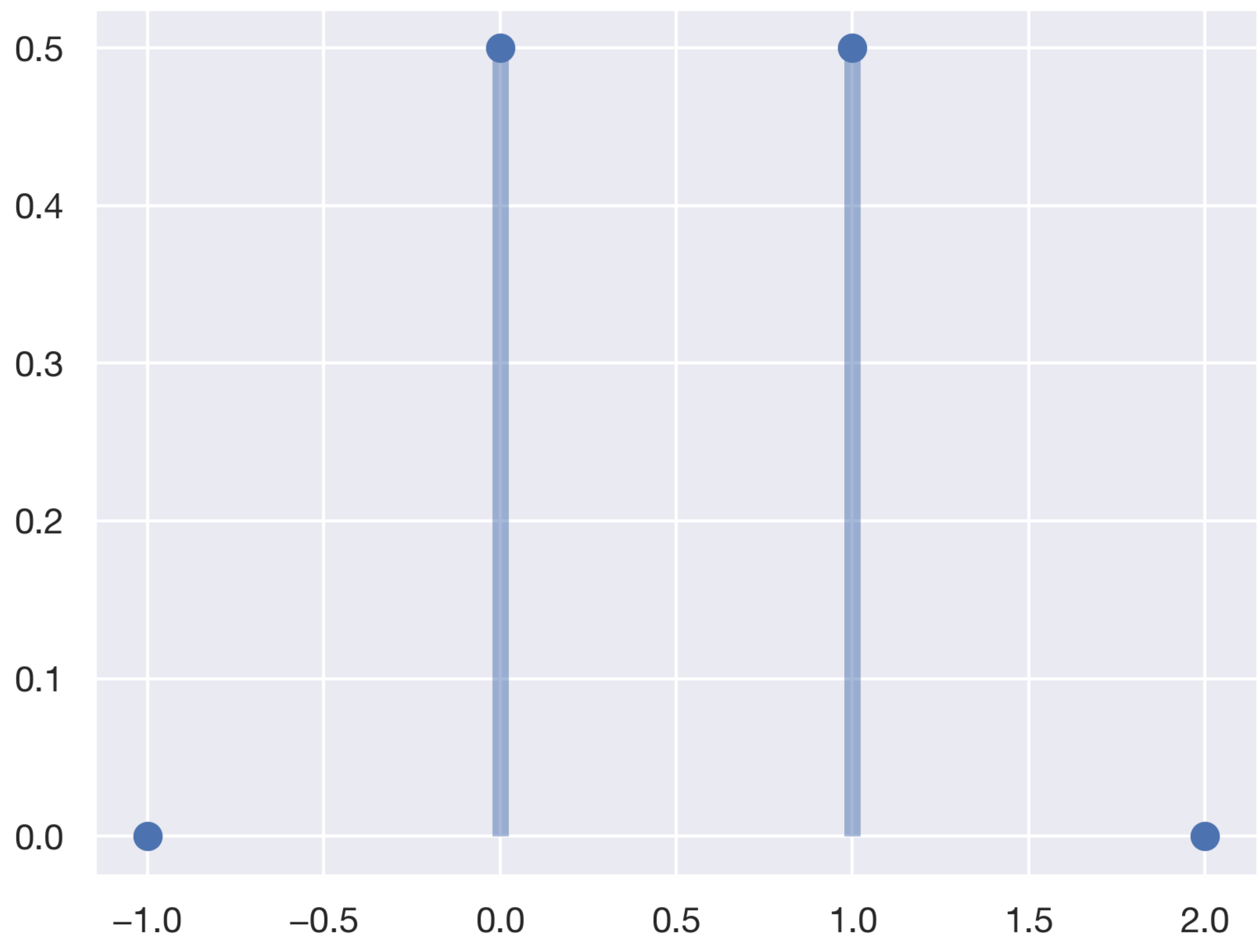
PMF: $p_X(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \\ 0 & \text{otherwise} \end{cases}$

Mean: $\mathbb{E}[X] = p$.

Variance: $\text{Var}(X) = p(1-p)$.

MGF: $M_X(t) = 1 - p + pe^t$.





The Binomial Distribution

"Story" of the Distribution

★ Sum of n independent Bernoullis

$\sum_{i=1}^{n} X_i$

Flip n independent coins, each landing heads with probability p and tails with probability $1-p$, and count the number of heads.

Example. Consider an urn with 7 orange balls and 3 green balls. Let X count the total number of orange balls drawn after drawing $n = 10$ balls *with replacement* from the urn.

The Binomial Distribution

Properties

$$X \sim \text{Bin}(n, p)$$

→ Also: Sum of n Bernoulli's.

order doesn't matter.

Parameters: $n \in \{0,1,2,\ldots\}$, the number of trials.
$p \in [0,1]$, the success probability.

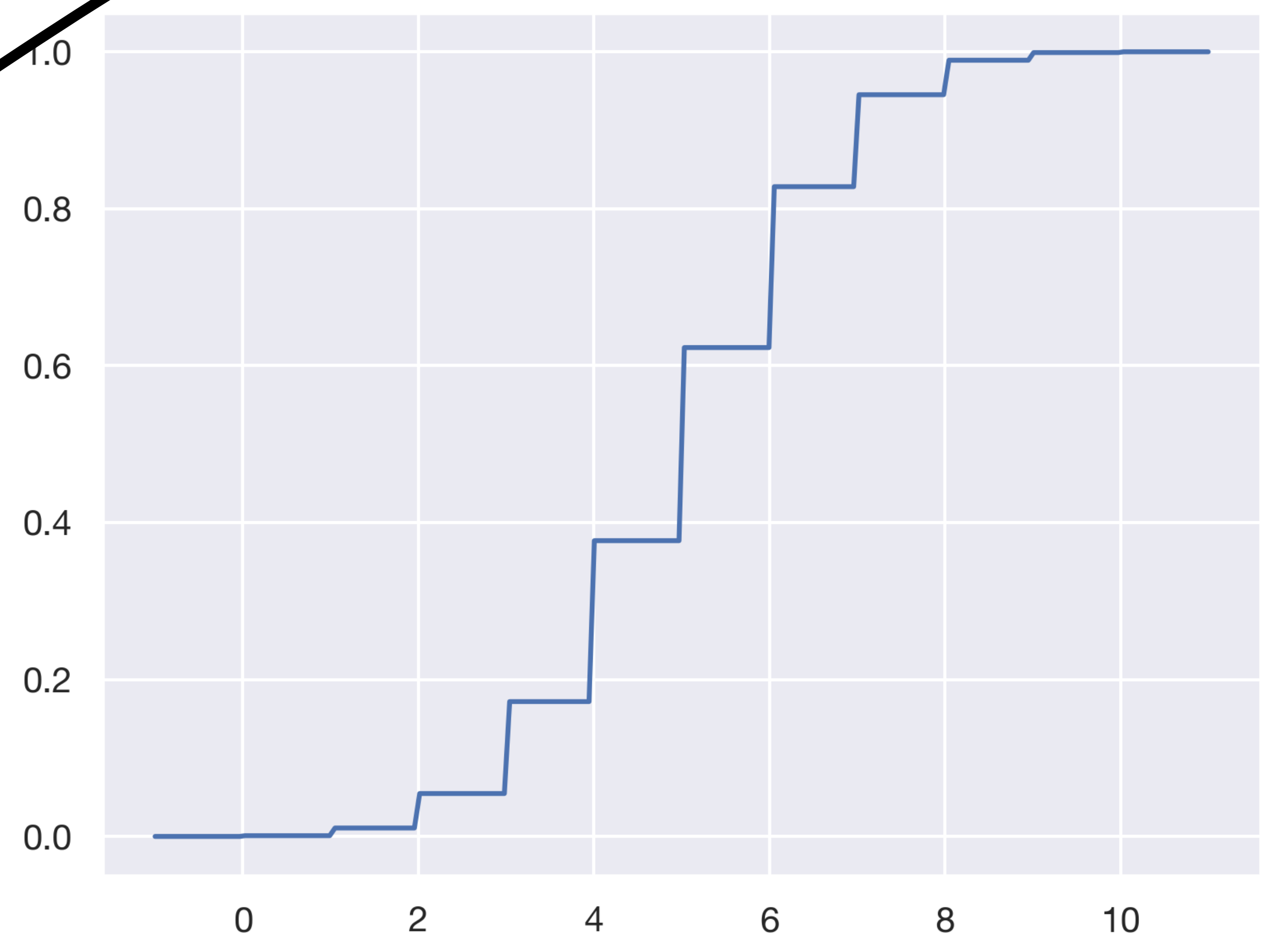
CDF: $F_X(x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$

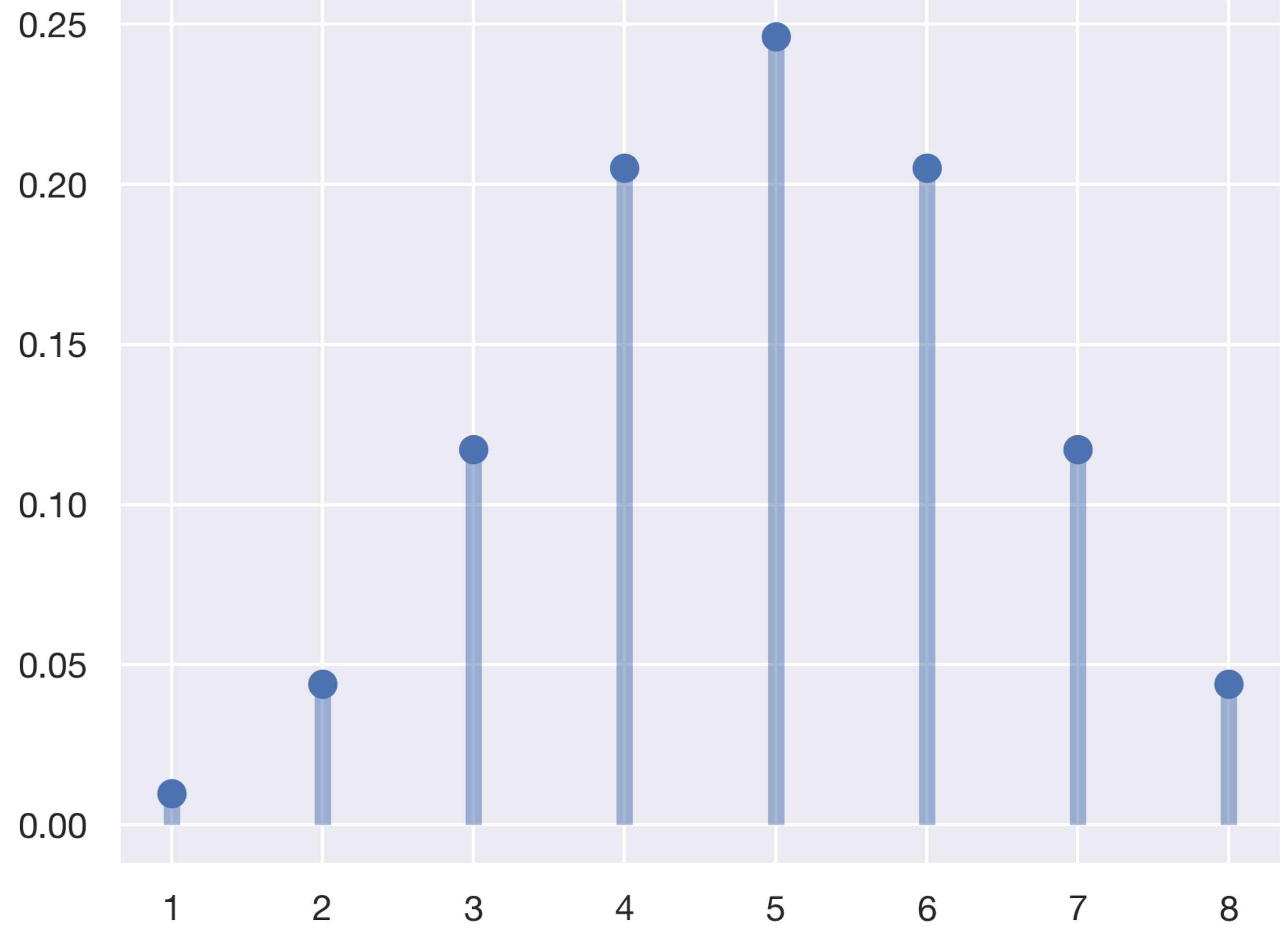
PMF: $p_X(x) = \binom{n}{x} p^k (1-p)^{n-x}$

Mean: $\mathbb{E}[X] = np$.

Variance: $\text{Var}(X) = np(1-p)$.

MGF: $M_X(t) = (1 - p + pe^t)^n$.





The Geometric Distribution

“Story” of the Distribution

Flip coins, each landing heads with probability p and tails with probability $1 - p$, until you see your first head. How many trials occurred?

Example. Let X be the number of rolls needed from repeatedly rolling a fair, six-sided die until 3 shows up.

The Geometric Distribution
Properties

SUPPORT: points that have nonzero mass

$$X \sim \text{Geom}(p)$$

Parameters: $p \in [0,1]$, the success probability.

CDF: $F_X(x) = 1 - (1-p)^{\lfloor x \rfloor}$ if $x \geq 1$, 0 otherwise

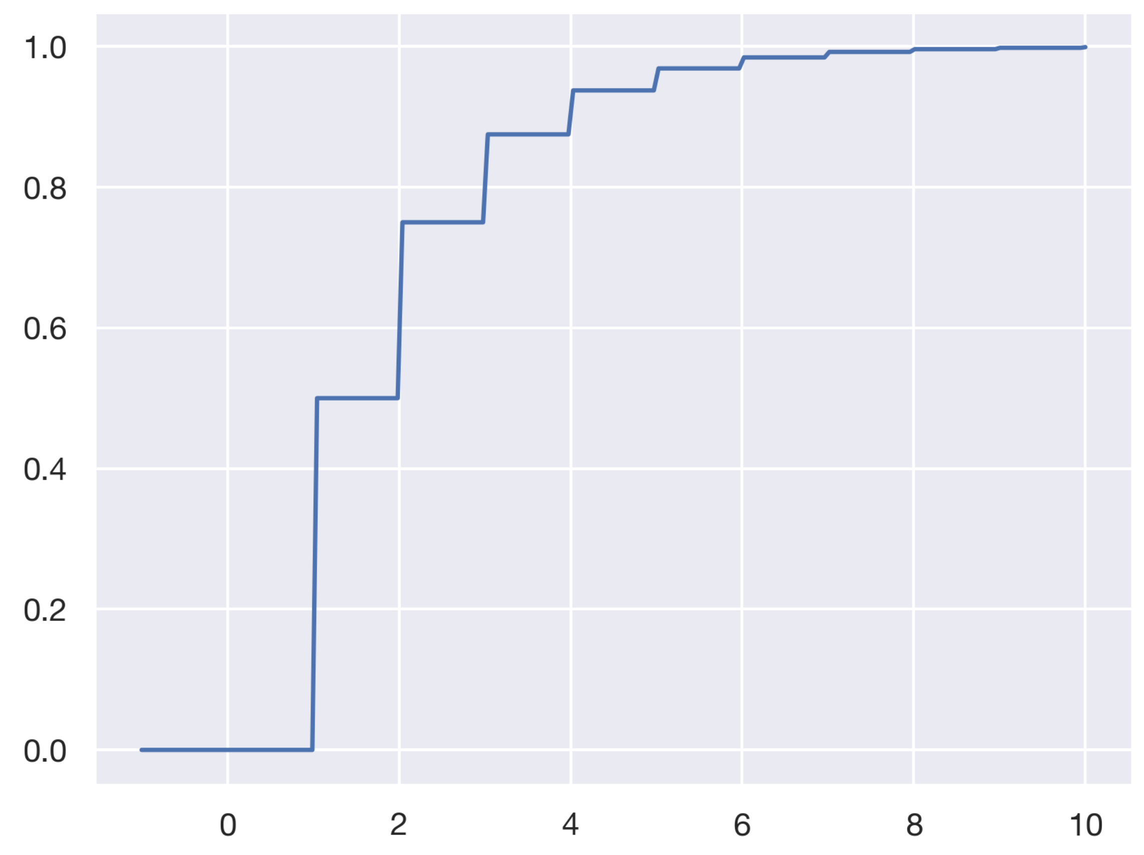
PMF: $p_X(x) = \begin{cases} (1-p)^{x-1}p & x \in \{1,2,3,\ldots,\} \\ 0 & \text{otherwise} \end{cases}$

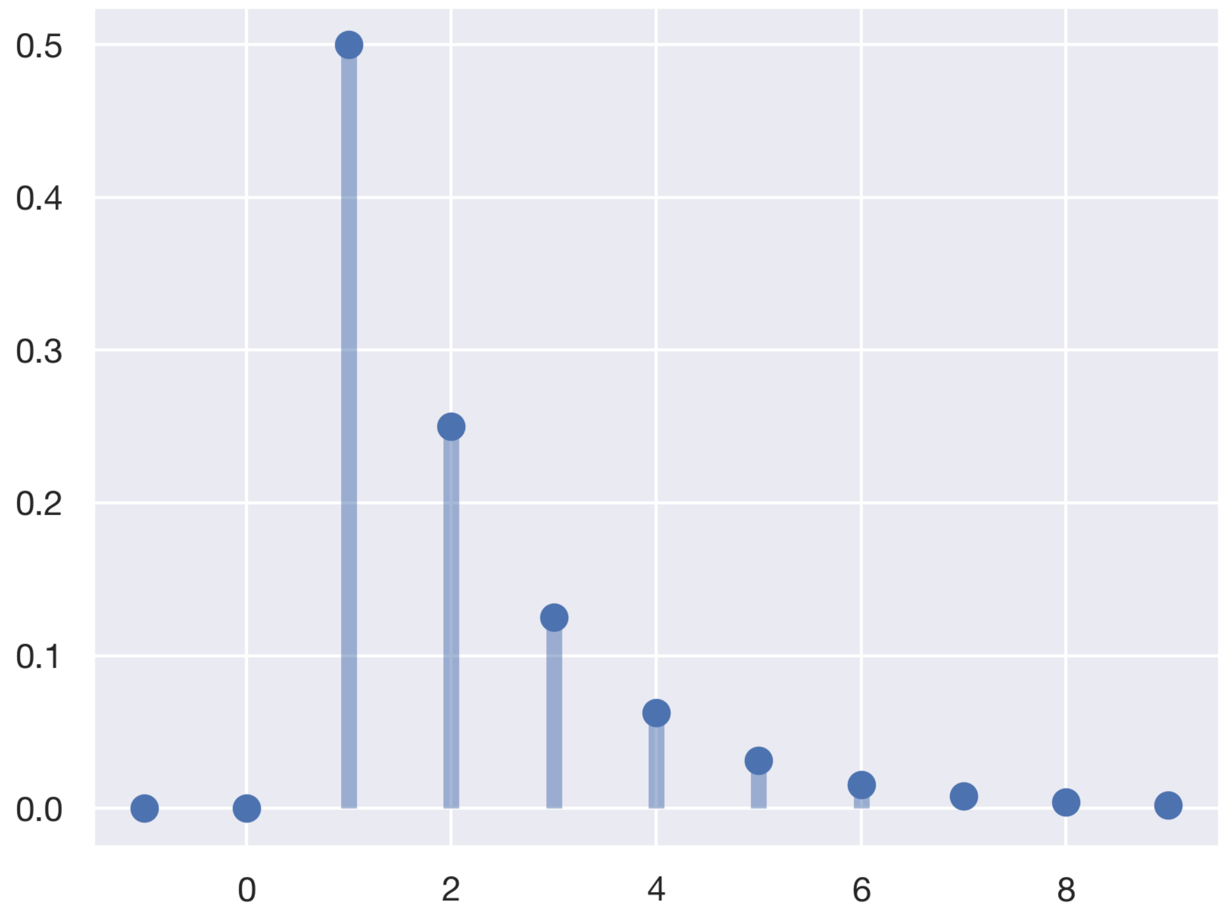
Mean: $\mathbb{E}[X] = 1/p$.

Variance: $\text{Var}(X) = \dfrac{1-p}{p^2}$.

MGF: $M_X(t) = \dfrac{pe^t}{1-(1-p)e^2}$ for $t < -\ln(1-p)$.

NOTE: Infinite support!





The Poisson Distribution

“Story” of the Distribution

RATE

Count the number of rare events in a fixed time interval, if the average number of events in that interval is λ.

Example. Let X be the number of text messages you receive in a given hour if you receive an average of $\lambda = 3$ messages per hour.

The Poisson Distribution

"Story" of the Distribution

Count the number of rare events in a fixed time interval, if the average number of events in that interval is λ.

Example. Let X be the number of text messages you receive in a given hour, if you receive an average of $\lambda = 3$ messages per hour.

Example. Let X count the number of times a raindrop hits a specific square inch in a minute, if that square inch receives an average of $\lambda = 10$ drops per minute.

The Poisson Distribution
Properties

$$X \sim \text{Pois}(\lambda)$$

Parameters: $\lambda \in (0,\infty)$, the success rate.

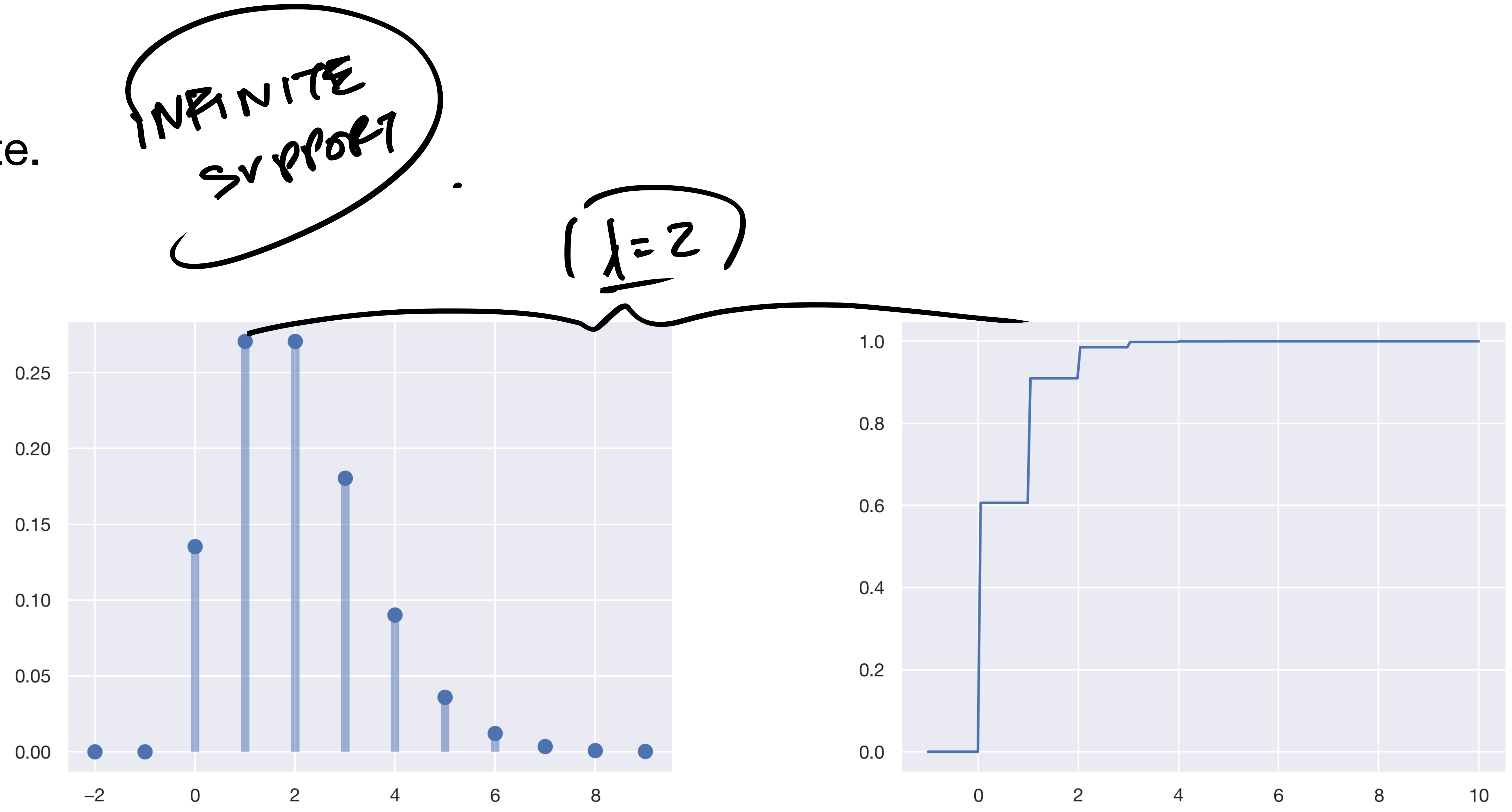
CDF: $F_X(x) = e^{-\lambda} \sum_{j=0}^{\lfloor x \rfloor} \frac{\lambda^j}{j!}$

PMF: $p_X(x) = \frac{\lambda^k e^{-\lambda}}{k!}$

Mean: $\mathbb{E}[X] = \lambda.$

Variance: $\text{Var}(X) = \lambda.$

MGF: $M_X(t) = \exp(\lambda(e^t - 1)).$



“Named” Distributions

Continuous Examples

Continuous Distributions

Continuous Random Variables

A ***continuous random variable*** X takes on an uncountably infinite number of values. The probability at any point x is 0.

CDF. $F_X(x) := \mathbb{P}(X \leq x)$.

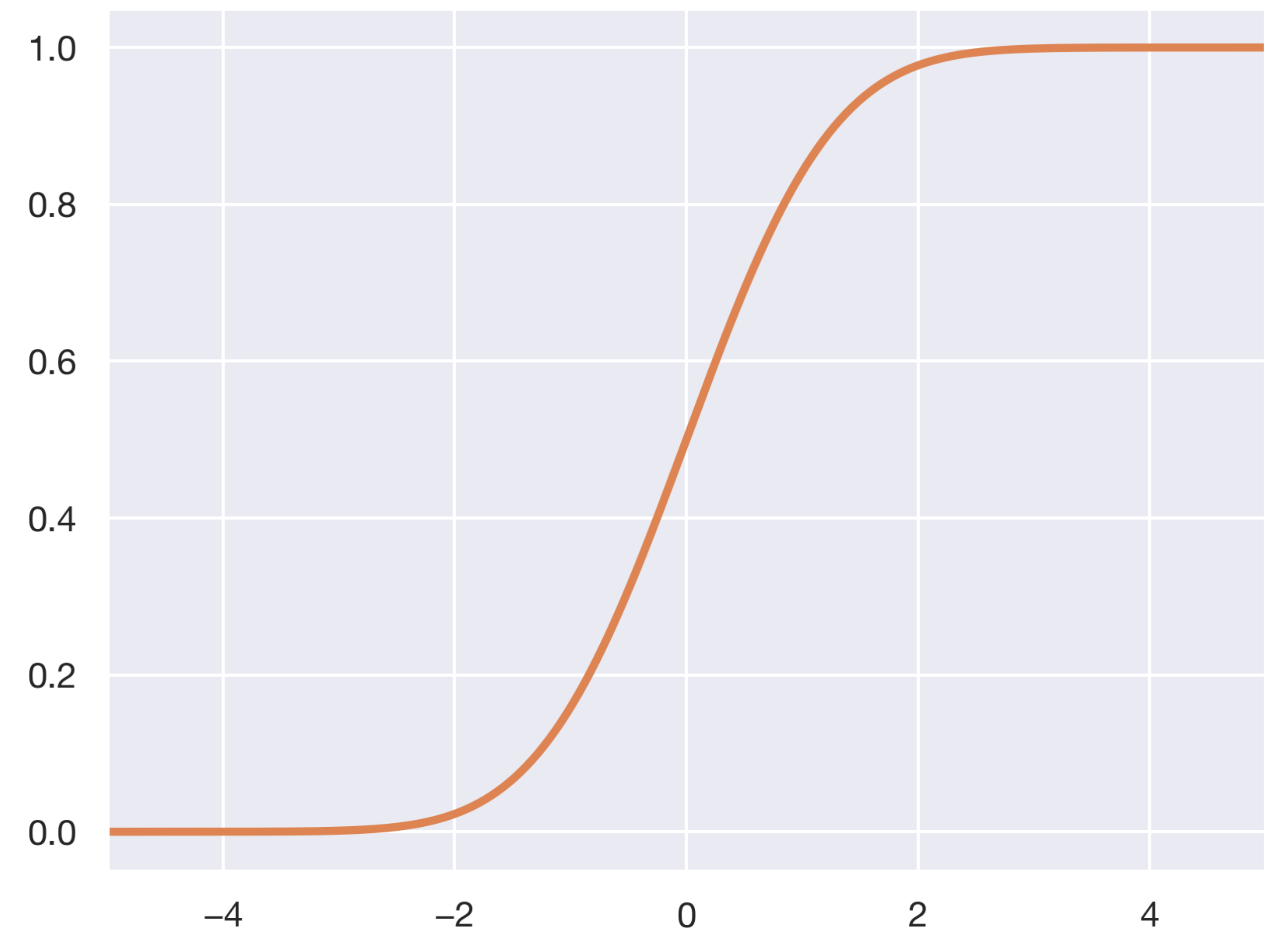
PDF. $p_X(x) = F'(x)$ and $\mathbb{P}(X \in A) = \int_A p_X(x)dx$.

PDF is the derivative of F.

PDF is nonnegative and integrates to 1.

PDF *does not* give probabilities at points.

Expectation. $\mathbb{E}(X) = \int_{-\infty}^{\infty} xp_X(x)dx$.



The Uniform Distribution

“Story” of the Distribution

Draw a completely random number in the continuous interval from a to b.

Example. Let X be where you randomly break a stick of length $b = 20$ inches.

$[0, 20]$.

The Uniform Distribution
Properties

$$X \sim \text{Unif}(a, b)$$

Parameters: $-\infty < a < b < \infty$, the interval boundaries.

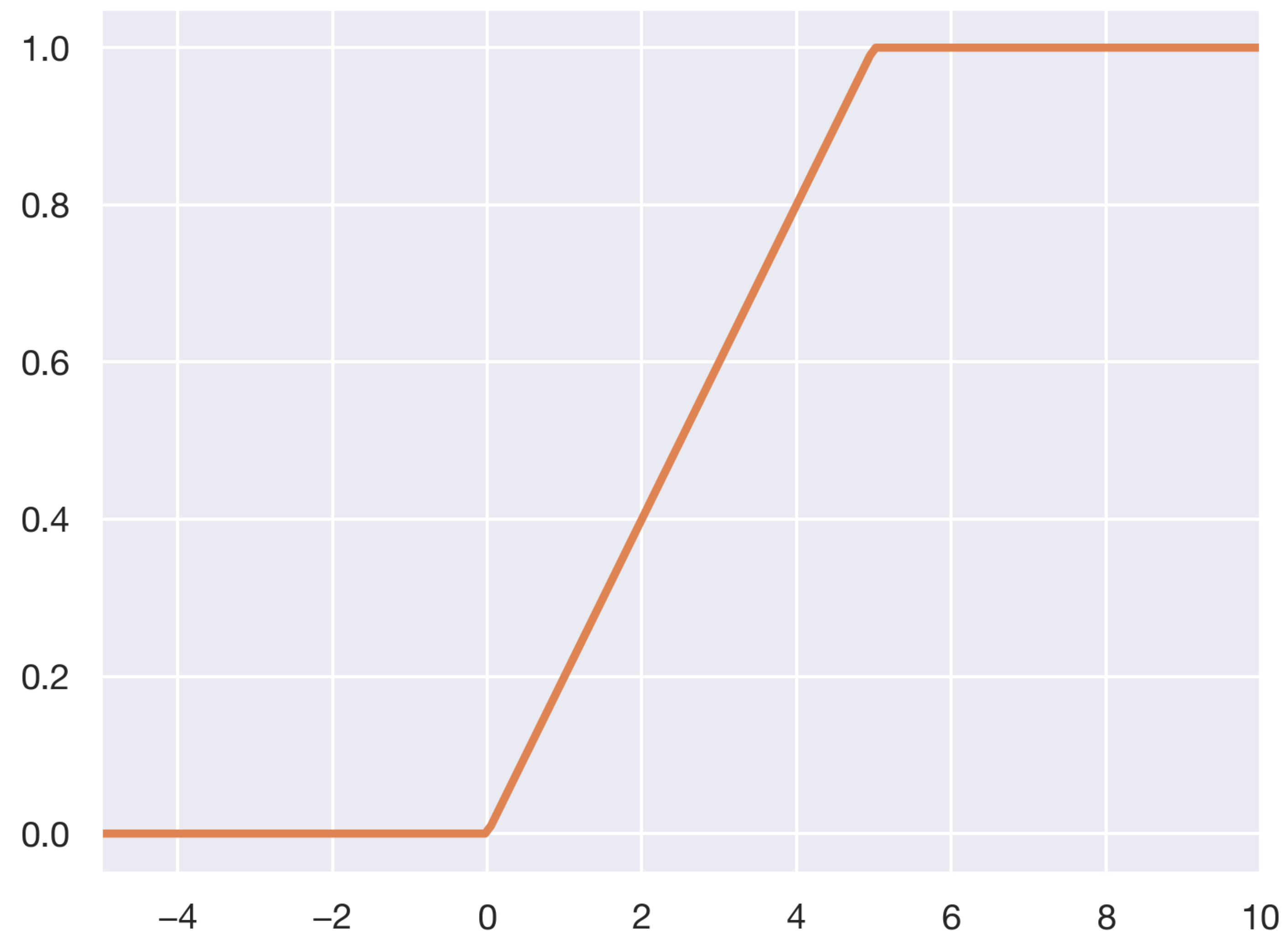
CDF: $F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \end{cases}$

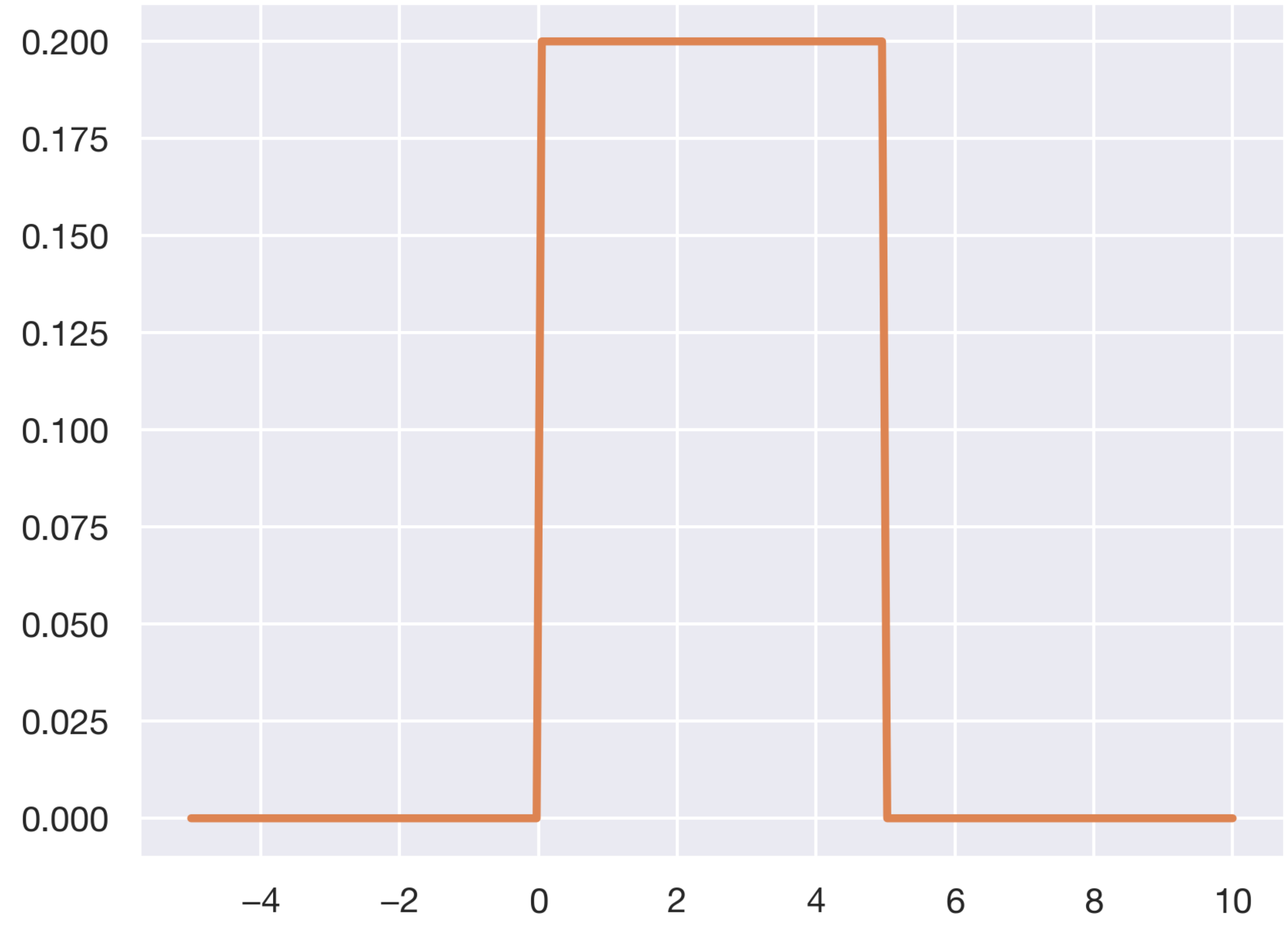
PDF: $p_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

Mean: $\mathbb{E}[X] = \frac{1}{2}(a + b)$.

Variance: $\text{Var}(X) = \frac{1}{12}(b - a)^2$.

MGF: $M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$ for $t \neq 0$ and $M_X(0) = 1$.





The Gaussian Distribution

"Story" of the Distribution

Draw a random number with probability distributed according to a "bell-shaped" curve.

Example. Let X be the height of a human male.

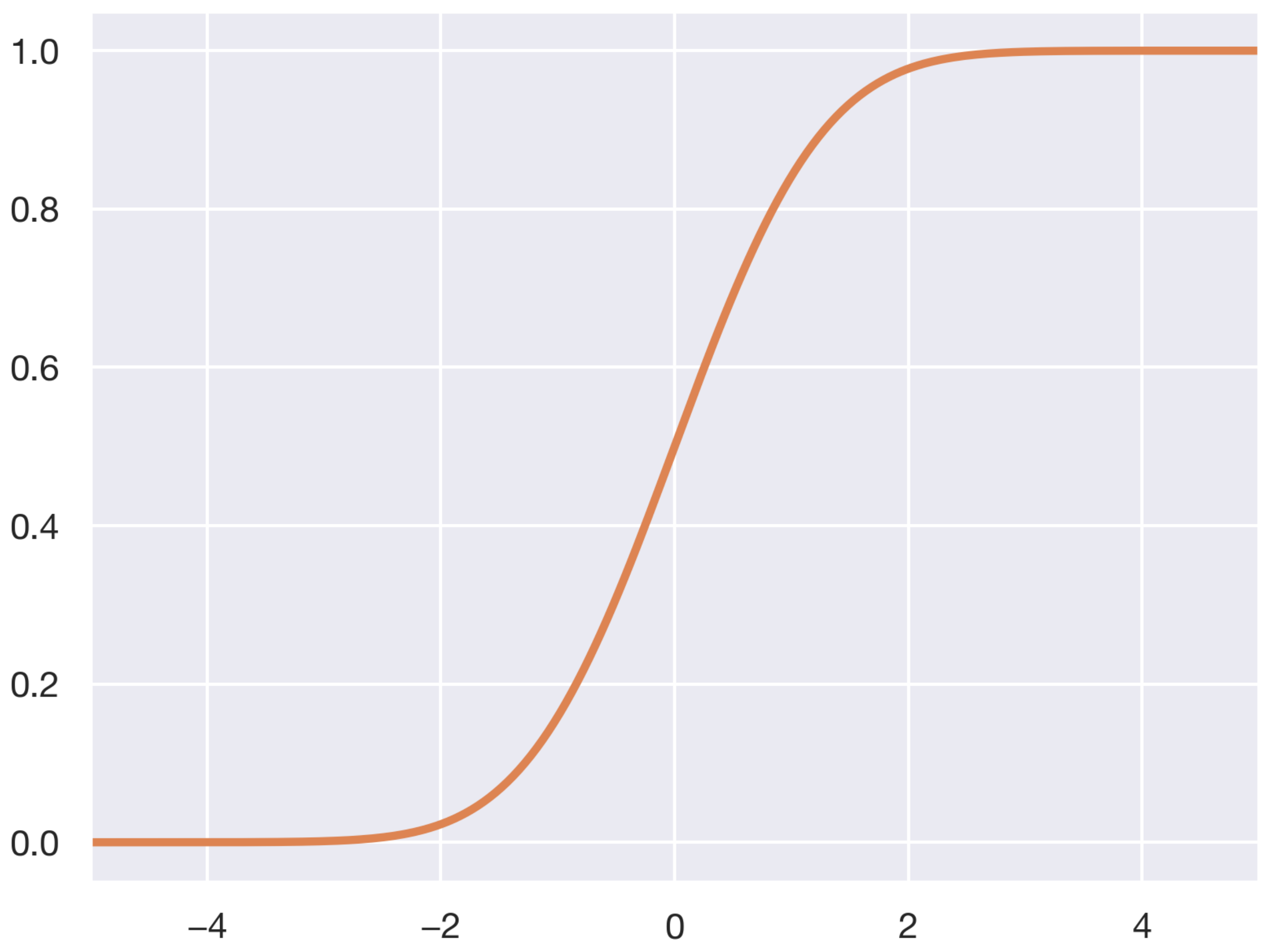
The Gaussian Distribution
Properties

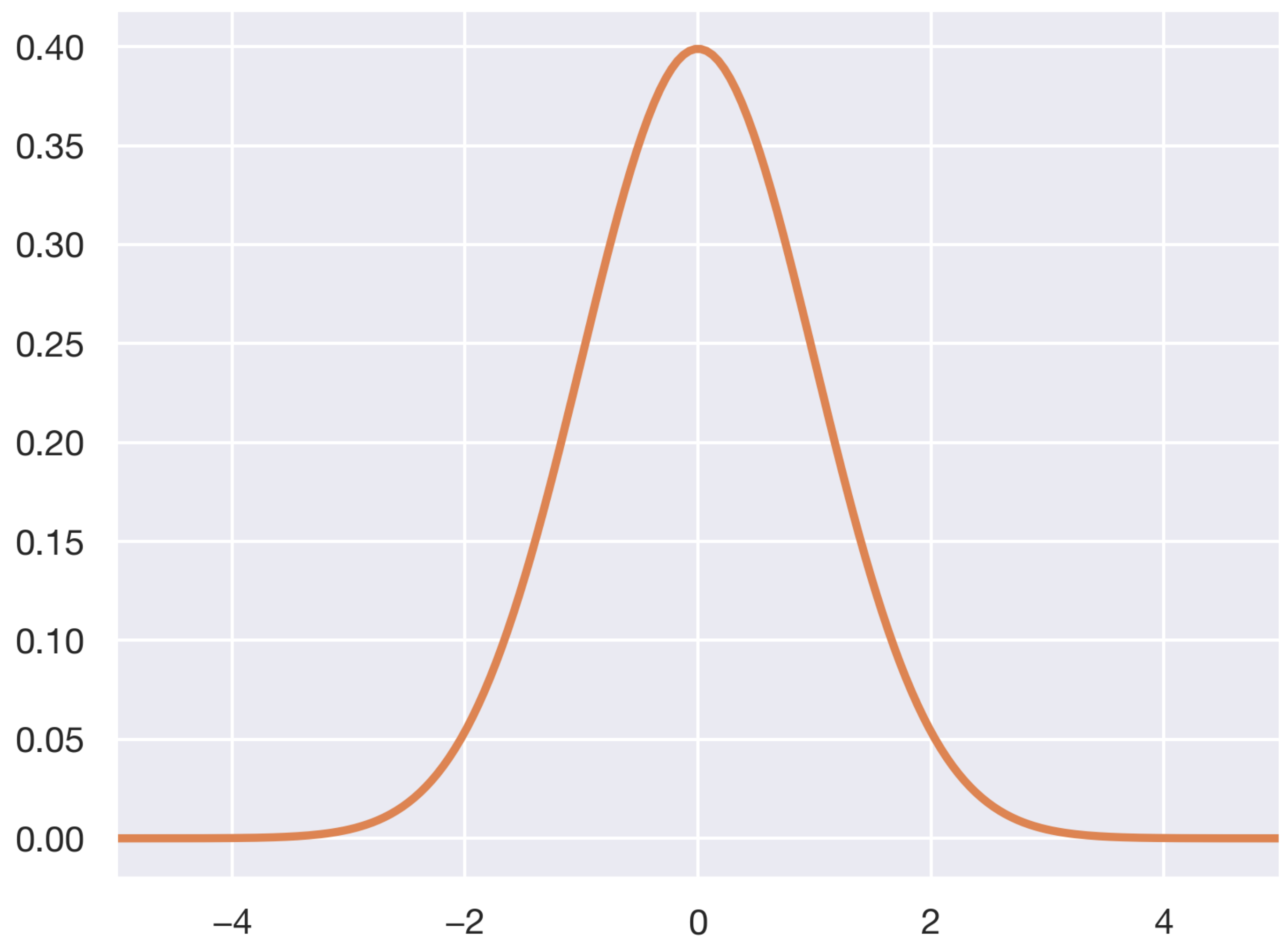
$\Phi :=$ CDF of standard $N(0,1)$.

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

Parameters: $\mu \in \mathbb{R}$, the mean and $\sigma^2 \in \mathbb{R}_{>0}$, the variance.

CDF: $F_X(x) = \int_{-\infty}^{x} p_X(x)dx = \Phi\left(\frac{x-\mu}{\sigma}\right)$
(no closed form)





PDF: $p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Mean: $\mathbb{E}[X] = \mu$.

Variance: $\mathrm{Var}(X) = \sigma^2$.

MGF: $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$.

$e^{t^2/2}$

(Proof: complete the square).

The Chi-squared Distribution

"Story" of the Distribution

$z_1 \sim N(0,1)$
$\vdots$
$z_k \sim N(0,1)$

$\Rightarrow \boxed{\sum_{i=1}^{k} z_i^2}$

Add up k independent, squared standard Gaussian random variables.

Example. Let $\mathbf{z} = (z_1, z_2)$ be a random vector with independent entries $z_1 \sim N(0,1)$ and $z_2 \sim N(0,1)$. Then, $X = ||\mathbf{z}||^2$ is a Chi-squared random variable with $k = 2$.

The Chi-squared Distribution

Properties

$$X \sim \chi^2(k)$$

Parameters: k, the "degrees of freedom."

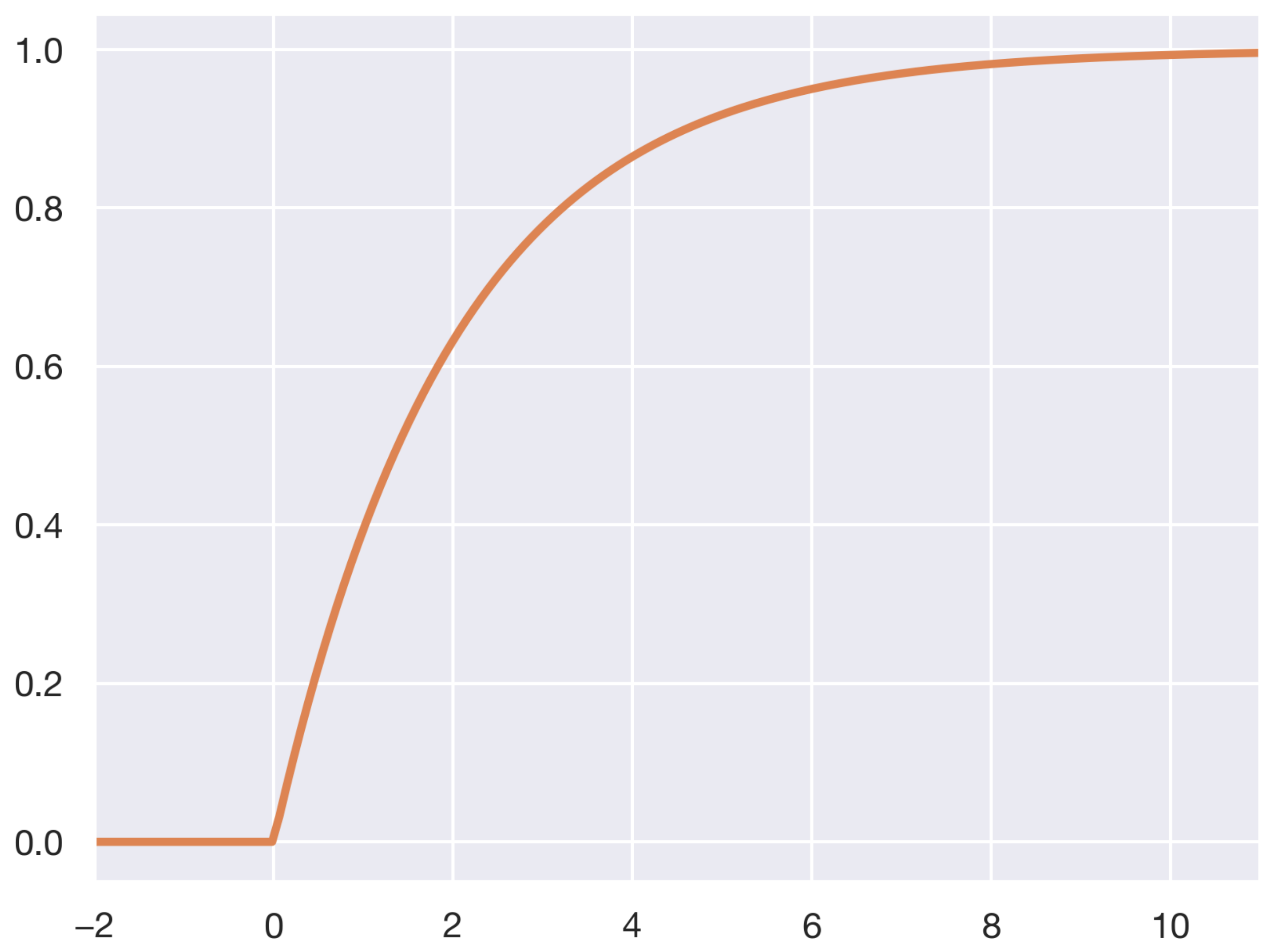
CDF: $F_X(x; 2) = 1 - e^{-x/2}$ (more complicated for $k \neq 2$)

PDF: $p_X(x) = \begin{cases} \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\int_0^\infty t^{k-1}e^{-t}dt} & x > 0 \\ 0 & \text{otherwise} \end{cases}$

Mean: $\mathbb{E}[X] = k$.

Variance: $\text{Var}(X) = 2k$.

MGF: $M_X(t) = (1-2t)^{-k/2}$ for $t < 1/2$.



The Exponential Distribution

"Story" of the Distribution

The waiting time for a success in continuous time, where λ is the rate at which successes arrive.

Example. Let X be the time between receiving one text message and the next, where λ is the rate of text messages per unit time.

The Exponential Distribution
PDF, CDF, and MGF

$$X \sim \text{Expo}(\lambda)$$

$F_X(x) = \mathbb{P}[X \leq x]$

Parameters: $\lambda > 0$, the success rate.

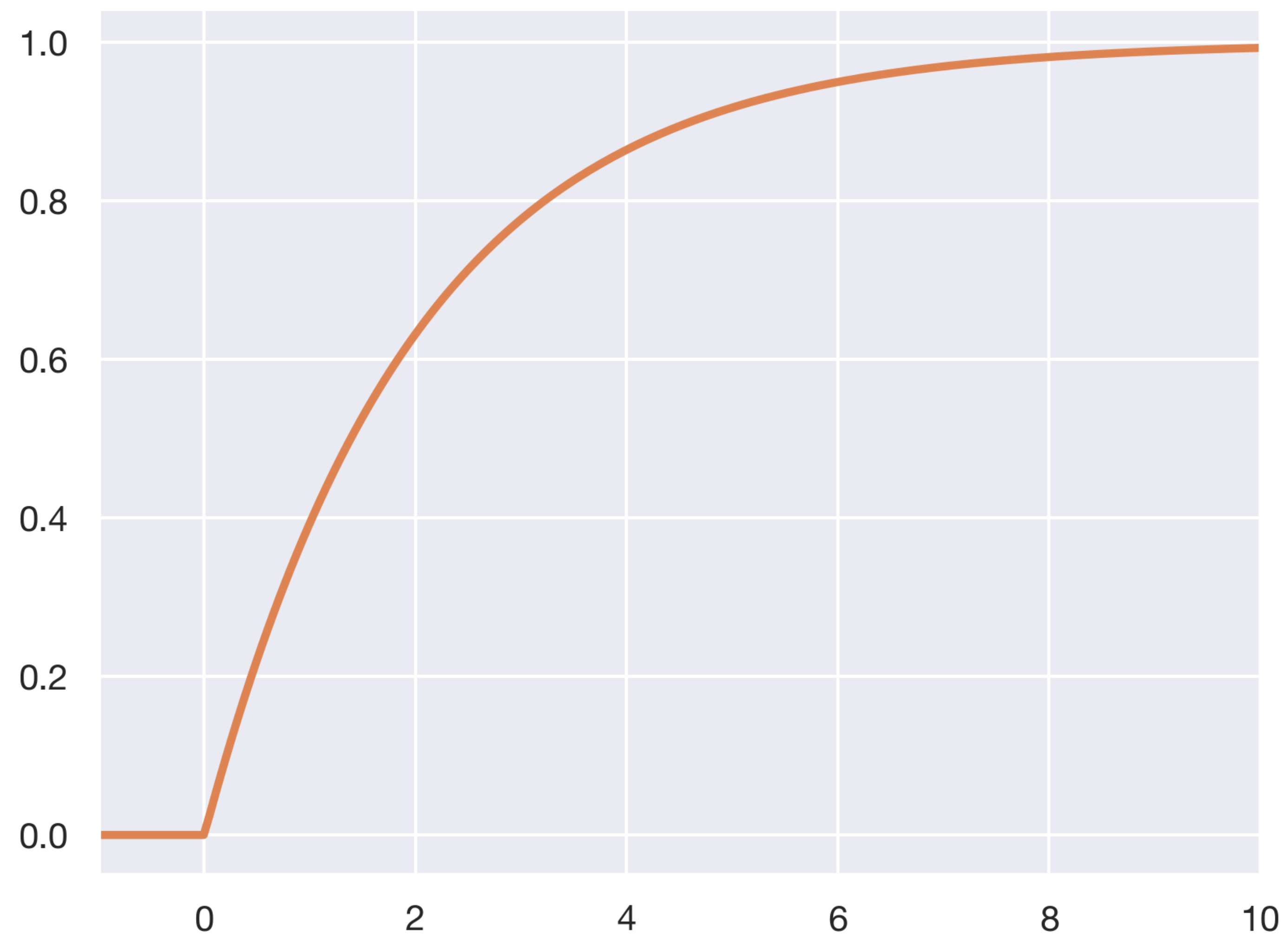
CDF: $F_X(x) = 1 - e^{-\lambda x}$

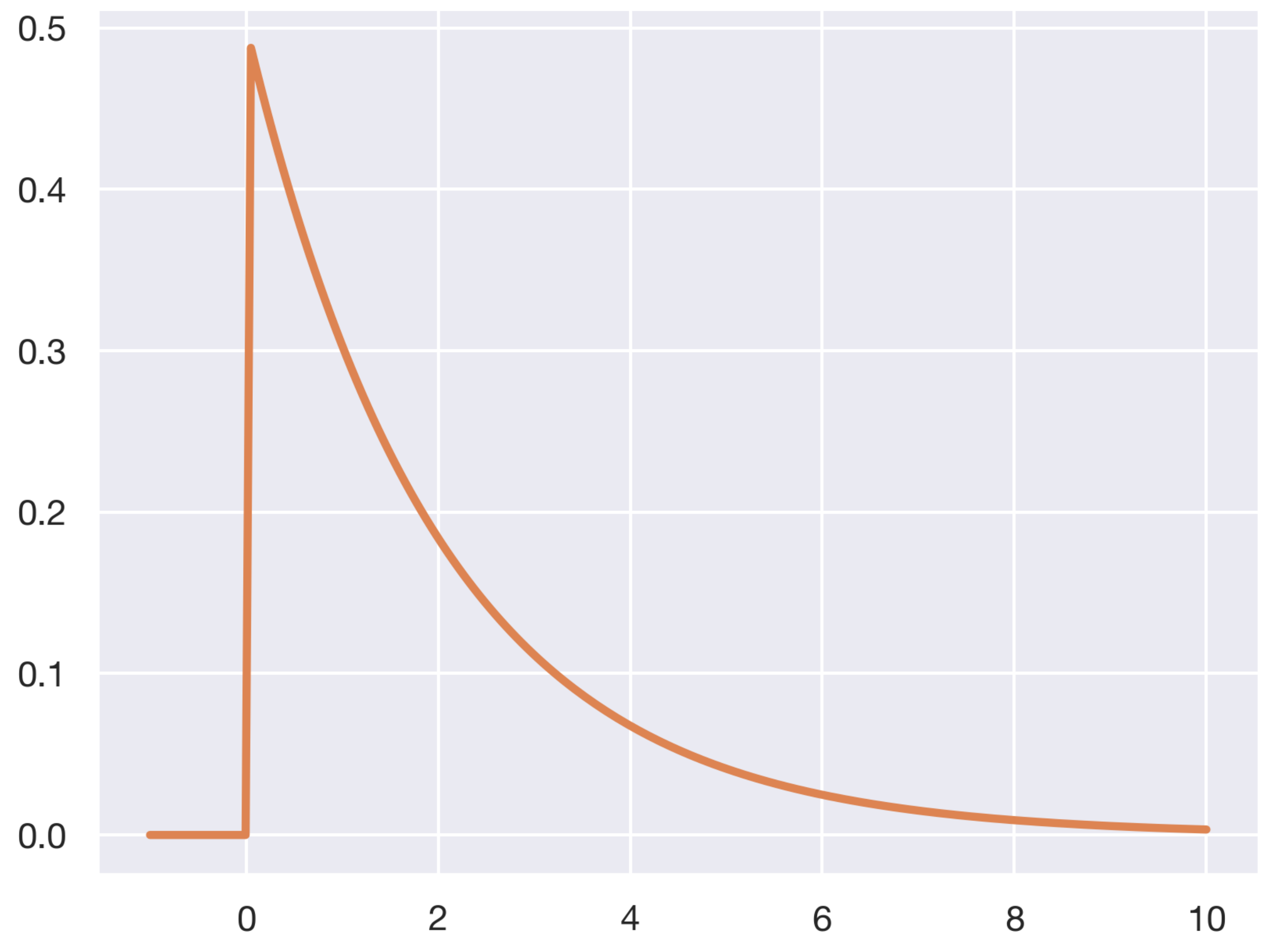
PDF: $p_X(x) = \lambda e^{-\lambda x}$

Mean: $\mathbb{E}[X] = 1/\lambda$.

Variance: $\text{Var}(X) = 1/\lambda^2$.

MGF: $M_X(t) = \dfrac{\lambda}{\lambda - t}$ for $t < \lambda$.





Maximum Likelihood Estimation

Intuition and Definition

Statistical Estimator

Intuition

θ = param.

$\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} x_i$.

A ***(statistical) estimator*** is a "best guess" at some (unknown) quantity of interest (the ***estimand***) using observed data.

We will only concern ourselves with ***point estimation***, where we want to estimate a single, fixed quantity of interest (as opposed to, say, an interval).

The quantity doesn't have to be a single number; it could be, for example, a fixed vector, matrix, or function.

Statistical Estimator

Definition

Let $X_1, \ldots, X_n$ be n i.i.d. random variables drawn from some distribution $\mathbb{P}_X$. An ***estimator*** $\hat{\theta}_n$ of some fixed, unknown parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n).$$

Defined similarly for random vectors.

Empirical Risk Minimization (ERM)

What we've been doing

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists an **unknown** joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y) \leftarrow \mathbb{E}_{x_0,y_0}[(y_0 - f(x_0))^2].$$

($d\mathbb{P}(\mathbf{x}, y)$: $p(x,y)\,dx\,dy$)

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2. = \frac{1}{n}\|Xw - y\|^2$$

if $f(x) = \hat{w}^\top x$.

Parametric Estimation vs. ERM

A different approach

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists an **unknown** joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We then went on to minimize the ***empirical risk*** to get our model $f : \mathbb{R}^d \to \mathbb{R}$.

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2.$$

This uses no information about the distribution of the data!

Parametric Estimation

Intuition

Suppose we have a good guess at the underlying distribution generating some i.i.d. data $X_1, \ldots, X_n$.

"My data is probably generated from a Poisson distribution."

Then, we can restrict our attention to estimating a ***parametric model***, a function $p(x; \theta)$ that depends on parameters $\theta = (\theta_1, \ldots, \theta_k)$ belonging to some parameter space $\Theta \subseteq \mathbb{R}^k$.

"Let's estimate $\lambda \in \mathbb{R}$ in the PMF $p(x; \lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!}$."

Parametric Estimation

Intuition

Suppose we have a good guess at the underlying distribution generating some i.i.d. data $X_1, \ldots, X_n$.

"My data is probably generated from a Poisson distribution."

Then, we can restrict our attention to estimating a ***parametric model***, a function $p(x; \theta)$ that depends on parameters $\theta = (\theta_1, \ldots, \theta_k)$ belonging to some parameter space $\Theta \subseteq \mathbb{R}^k$.

"Let's estimate $\lambda \in \mathbb{R}$ in the PMF $p(x; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$."

If our assumption is good, then a good estimate $\hat{\theta}_n$ of θ might tell us everything we need to know about our data!

Parametric Estimation

Definition

A ***parametric model*** is a class of functions of the form:

$$\mathscr{F} := \left\{ f(x;\theta) : \theta \in \Theta \right\},$$

where $\Theta \subseteq \mathbb{R}^k$ is the ***parameter space*** and $\theta = (\theta_1, \ldots, \theta_k)$ are the ***model parameters***.

Example. The parameter space for the Gaussian distribution $N(\mu, \sigma^2)$ is

$$\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}.$$

Example. The parameter space for the Bernoulli distribution $\mathrm{Ber}(p)$ is

$$\Theta = \{p : 0 \leq p \leq 1\}.$$

Maximum Likelihood Estimation

Intuition

A common way to do *parametric estimation* given i.i.d. data $X_1, \ldots, X_n$ is ***maximum likelihood estimation.***

We assume that $X_1, \ldots, X_n$ came from a distribution with PDF $p(x; \theta)$ and parameter space $\Theta \subseteq \mathbb{R}^k$.

"Assume that the data come from a Gaussian with

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

We consider the ***likelihood function*** which maps from parameters Θ to some positive number: the "likelihood" of those parameters explaining the data.

Maximum Likelihood Estimation

Intuition

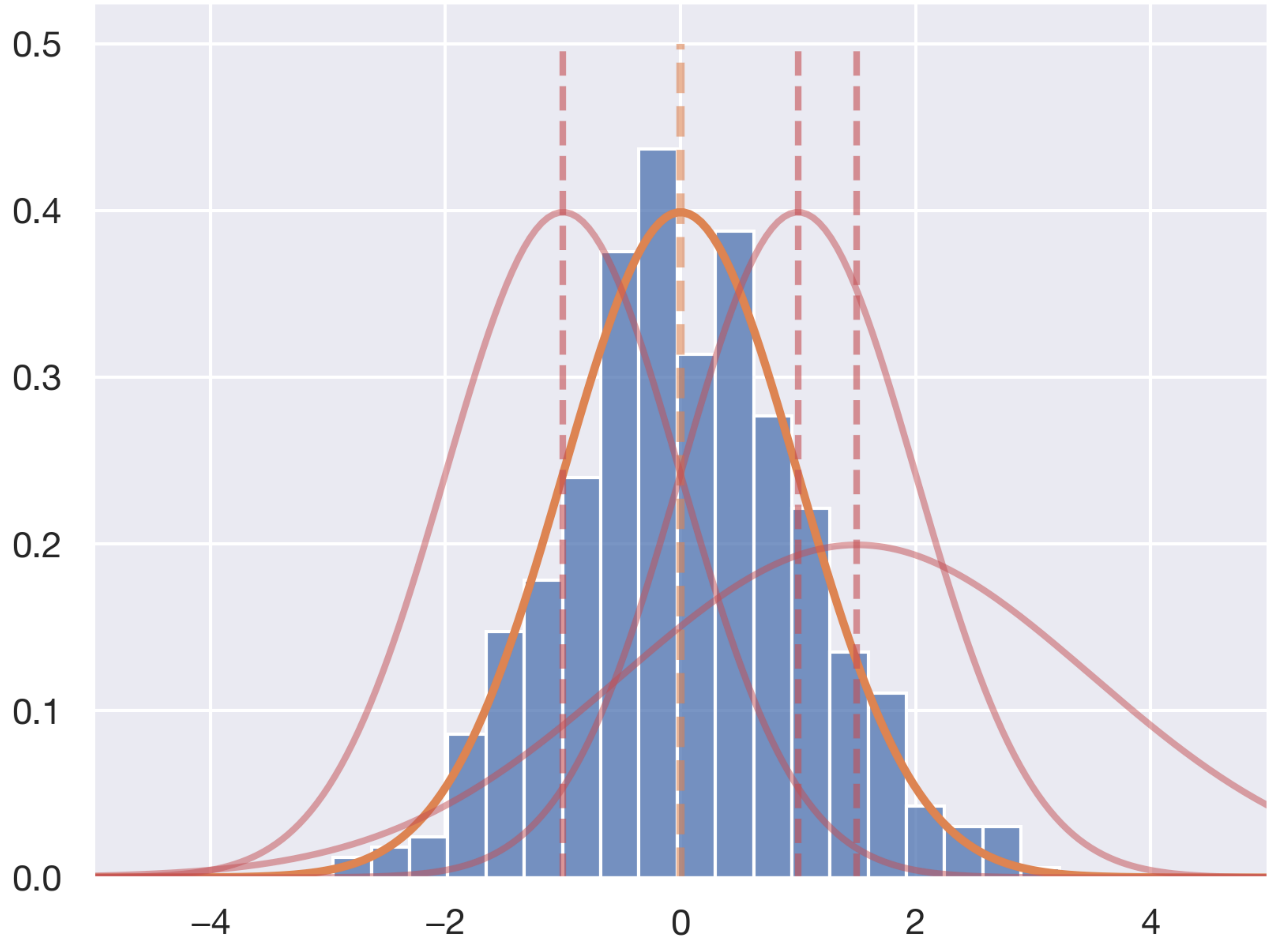
A common way to do *parametric estimation* given i.i.d. data $X_1, \ldots, X_n$ is ***maximum likelihood estimation.***

We assume that $X_1, \ldots, X_n$ came from a distribution with PDF $p(x; \theta)$ and parameter space $\Theta \subseteq \mathbb{R}^k$.

"Assume that the data come from a Gaussian with

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\text{"}$$

We consider the ***likelihood function*** which maps from parameters Θ to some positive number: the "likelihood" of those parameters explaining the data.



Maximum Likelihood Estimation

Intuition

A common way to do *parametric estimation* given i.i.d. data $X_1, \ldots, X_n$ is ***maximum likelihood estimation.***

We assume that $X_1, \ldots, X_n$ came from a distribution with PDF $p(x; \theta)$ and parameter space $\Theta \subseteq \mathbb{R}^k$.

"Assume that the data come are Poisson with $p_X(x) = \dfrac{\lambda^k e^{-\lambda}}{k!}$*"*

We consider the ***likelihood function*** which maps from parameters Θ to some positive number: the "likelihood" of those parameters explaining the data.

Maximum Likelihood Estimation

Definition

Consider the parametric model

$$\mathscr{F} := \left\{ f(x;\theta) : \theta \in \Theta \right\}.$$

Let $X_1, \ldots, X_n$ be i.i.d. random variables (or random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$). The ***likelihood function*** is the function $L_n : \Theta \to [0,\infty)$ defined by:

parameters

"score" likelihood.

$$L_n(\theta) := \prod_{i=1}^{n} f(X_i;\theta).$$

Joint Distribution of $X_1, \ldots, X_n$

$P(x_1, \ldots, x_n)$

$= p(x_1)\, P(x_2) \ldots P(x_n)$

$= \prod_{i=1}^{n} P(x_i).$

Note that $X_1, \ldots, X_n$ are fixed here, so this is *just* a function of θ.

"How well does θ describe my data $X_1, \ldots, X_n$?"

Maximum Likelihood Estimation

The Log-Likelihood

Consider the parametric model

$$\mathcal{F} := \left\{ f(x;\theta) : \theta \in \Theta \right\}.$$

Let $X_1, \ldots, X_n$ be i.i.d. random variables (or random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$). The ***likelihood function*** is the function $L_n : \Theta \to [0,\infty)$ defined by:

$$L_n(\theta) := \prod_{i=1}^{n} f(X_i;\theta).$$

$$\log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

The ***log-likelihood function*** is the function defined by:

$$\mathcal{L}_n(\theta) := \log L_n(\theta).$$

Maximum Likelihood Estimation

The Maximum Likelihood Estimator

Consider the parametric model

$$\mathscr{F} := \left\{ f(x;\theta) : \theta \in \Theta \right\}.$$

Let $X_1, \ldots, X_n$ be i.i.d. random variables (or random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$). The ***likelihood function*** is the function $L_n : \Theta \to [0,\infty)$ defined by:

$$L_n(\theta) := \prod_{i=1}^{n} f(X_i;\theta).$$

The ***log-likelihood function*** is the function defined by:

$$\mathscr{L}_n(\theta) := \log L_n(\theta).$$

The ***maximum likelihood estimator*** $\hat{\theta}_{MLE}$ is the value of θ that maximizes $L_n(\theta)$.

Maximum Likelihood Estimation

Why log-likelihood?

The ***log-likelihood function*** is the function defined by:

$$\mathcal{L}_n(\theta) := \log L_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta).$$

The ***maximum likelihood estimator*** $\hat{\theta}_{MLE}$ is the value of θ that maximizes $L_n(\theta)$.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta) = \arg\max_{\theta} \mathcal{L}_n(\theta).$$

$\log(\,\cdot\,)$ is a *monotonic* function, so the maximizer of $\log f$ corresponds to the maximizer of f.

Maximum Likelihood Estimation

Why log-likelihood?

The ***log-likelihood function*** is the function defined by:

$$\mathscr{L}_n(\theta) := \log L_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta).$$

The ***maximum likelihood estimator*** $\hat{\theta}_{MLE}$ is the value of θ that *minimizes* $-L_n(\theta)$.

$$\hat{\theta}_{MLE} = \arg\min_{\theta} \; -L_n(\theta) = \arg\min_{\theta} \; -\mathscr{L}_n(\theta).$$

$\log(\,\cdot\,)$ is a *monotonic* function, so the maximizer of $\log f$ corresponds to the maximizer of f.

Maximum Likelihood Estimation

Example: Bernoulli

$x_1 = 0$
$x_2 = 1$
$x_3 = 0$
$\vdots$
$x_n = 1$

Example. Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, so our parametric model is:

$$\mathscr{F} = \left\{ f(x; p) = p^x (1-p)^{1-x} : p \in [0,1] \right\}$$

$$\Theta = \{ p : 0 \leq p \leq 1 \}$$

The unknown parameter θ is p.

Maximum Likelihood Estimation

Example: Bernoulli

Example. Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, so our parametric model is:

$$\mathscr{F} = \left\{ f(x; p) = p^x (1-p)^{1-x} : p \in [0,1] \right\}$$

$$\Theta = \{ p : 0 \leq p \leq 1 \}$$

The unknown parameter θ is p.

Likelihood function. The likelihood function is

JOINT PMF

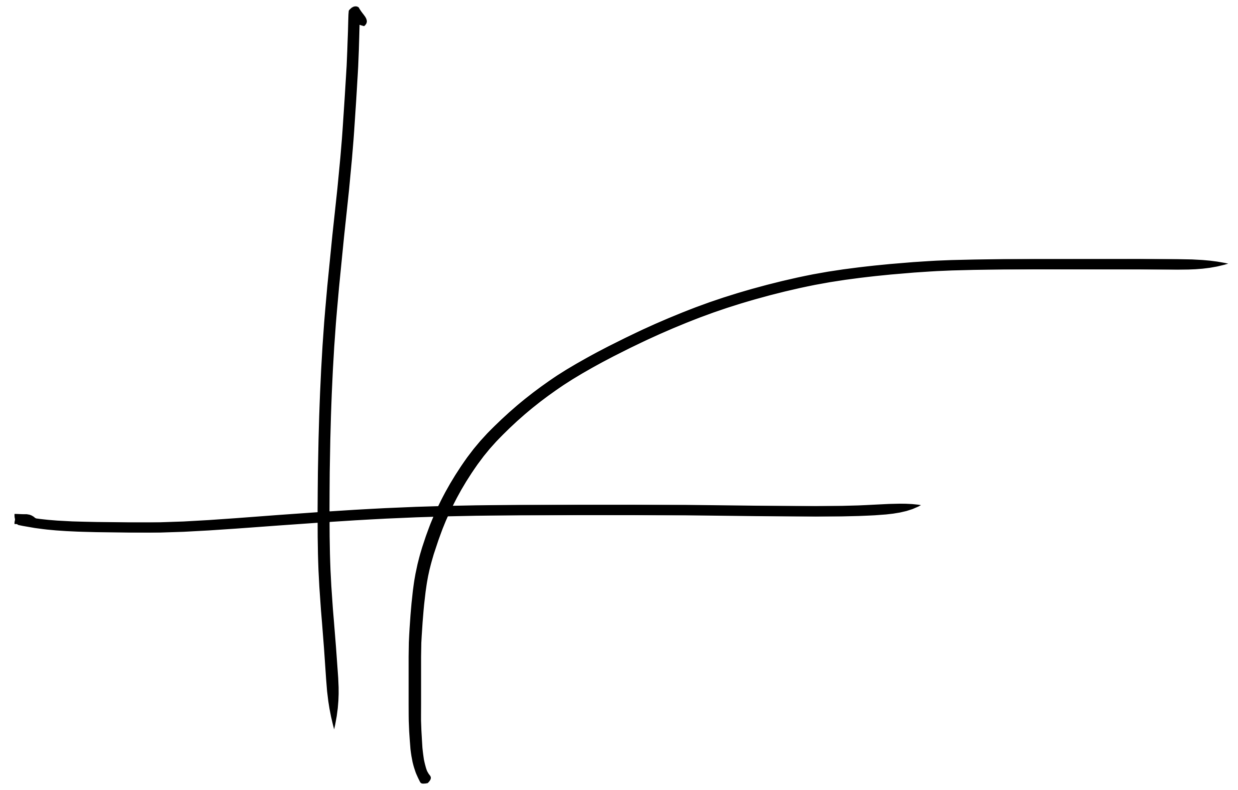
$$L_n(\theta) = L_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

Denote $S := \sum_{i=1}^{n} X_i$, and the likelihood function is:

$$L_n(p) = p^S (1-p)^{n-S}$$

Maximum Likelihood Estimation

Example: Bernoulli



Example. Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, so our parametric model is:

$$\mathscr{F} = \left\{ f(x;p) = p^x(1-p)^{1-x} : p \in [0,1] \right\}$$

$$\Theta = \{p : 0 \le p \le 1\}$$

The unknown parameter θ is p.

Likelihood function. The likelihood function is

$$L_n(\theta) = L_n(p) = \prod_{i=1}^{n} f(X_i;p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

Denote $S := \sum_{i=1}^{n} X_i$, and the likelihood function is:

$$L_n(p) = p^S(1-p)^{n-S}$$

S is fixed!

Log-likelihood function. The log-likelihood is

$\mathscr{L}_n(p) = S \log p + (n-S)\log(1-p)$. Now optimize this with respect to p!

Maximum Likelihood Estimation

Example: Bernoulli

Example. Suppose $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$, so our parametric model is:

$$\mathscr{F} = \left\{ f(x; p) = p^x (1-p)^{1-x} : p \in [0,1] \right\}$$

$$\Theta = \{ p : 0 \leq p \leq 1 \}$$

The unknown parameter θ is p.

Optimizing the negative log-likelihood. We need to solve the optimization problem:

$$\underset{p \in [0,1]}{\text{minimize}} \; -\mathscr{L}_n(p) = -S \log p + (S-n) \log(1-p).$$

Through first-order condition:

$$\nabla_p \mathscr{L}_n(p) = -\frac{S}{p} - \frac{S-n}{1-p} = 0.$$

Solving for p, we get:

$$\hat{p}_{MLE} = \frac{S}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Maximum Likelihood Estimation

Example: Bernoulli

Example. Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, so our parametric model is:

$$\mathcal{F} = \left\{ f(x;p) = p^x (1-p)^{1-x} : p \in [0,1] \right\}$$

$$\Theta = \{ p : 0 \leq p \leq 1 \}$$

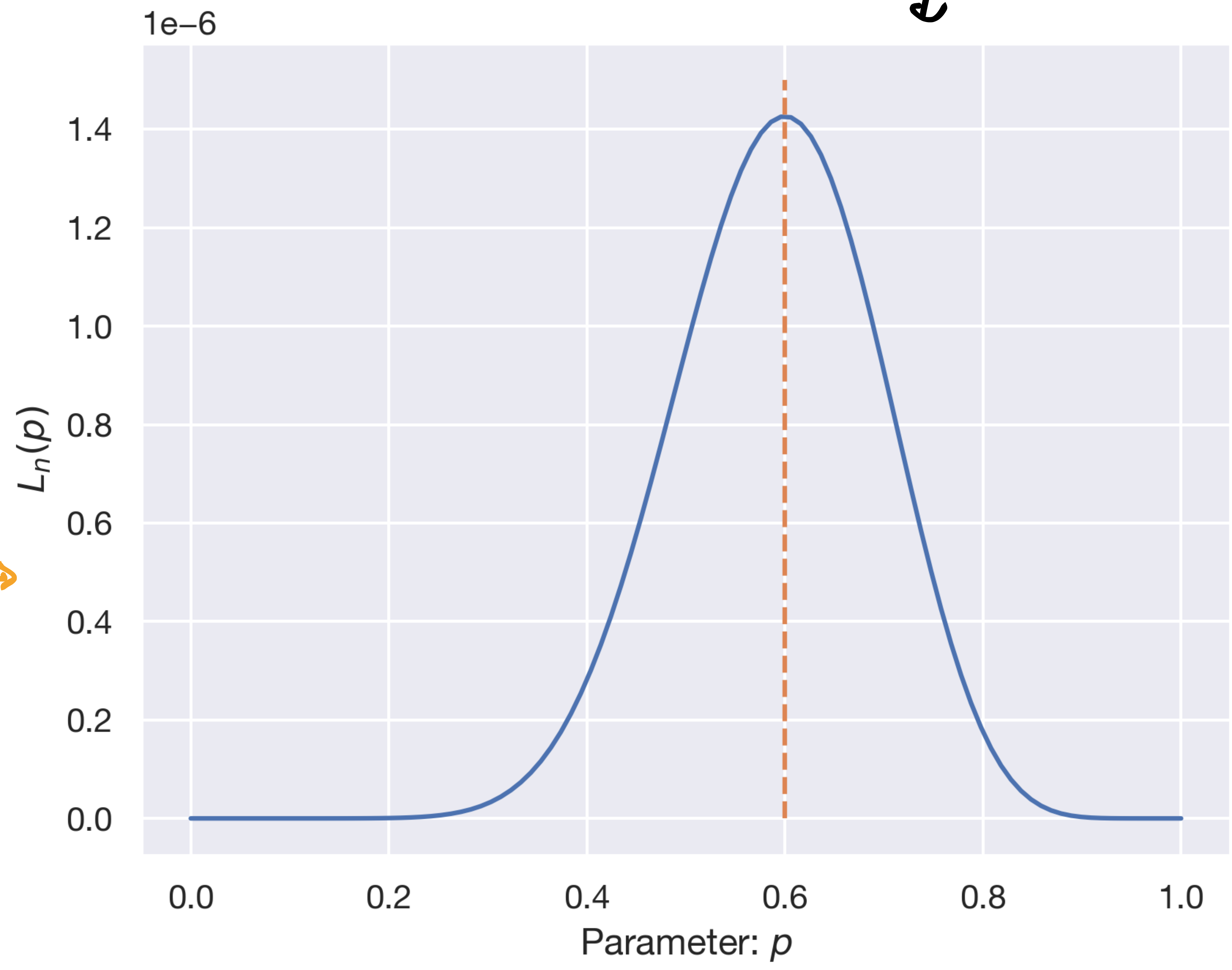
The unknown parameter θ is p.

The *likelihood function* is:

$$L_n(p) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}$$

The *maximum likelihood estimator* of the estimand p is:

$$\hat{p}_{MLE} = \frac{S}{n} = \frac{1}{n} \sum_{i=1}^n X_i.$$



Maximum Likelihood Estimation

Properties of the MLE

Under conditions on the *statistical model* with true parameter θ, the MLE is…

Consistent. As $n \to \infty$, the MLE $\hat{\theta}_{MLE}$ satisfies $\mathbb{P}[\,|\hat{\theta}_{MLE} - \theta| > \epsilon] \to 0$.

Equivariant. If $\hat{\theta}_{MLE}$ is the MLE of θ, then $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.

Asymptotically Normal. The random variable $(\hat{\theta} - \theta)/\sqrt{\hat{SE}} \to_D N(0,1)$, where $\hat{SE}$ is an estimate of the standard error.

Asymptotically optimal. Among all well-behaved estimators, the MLE has the smallest variance when $n \to \infty$.

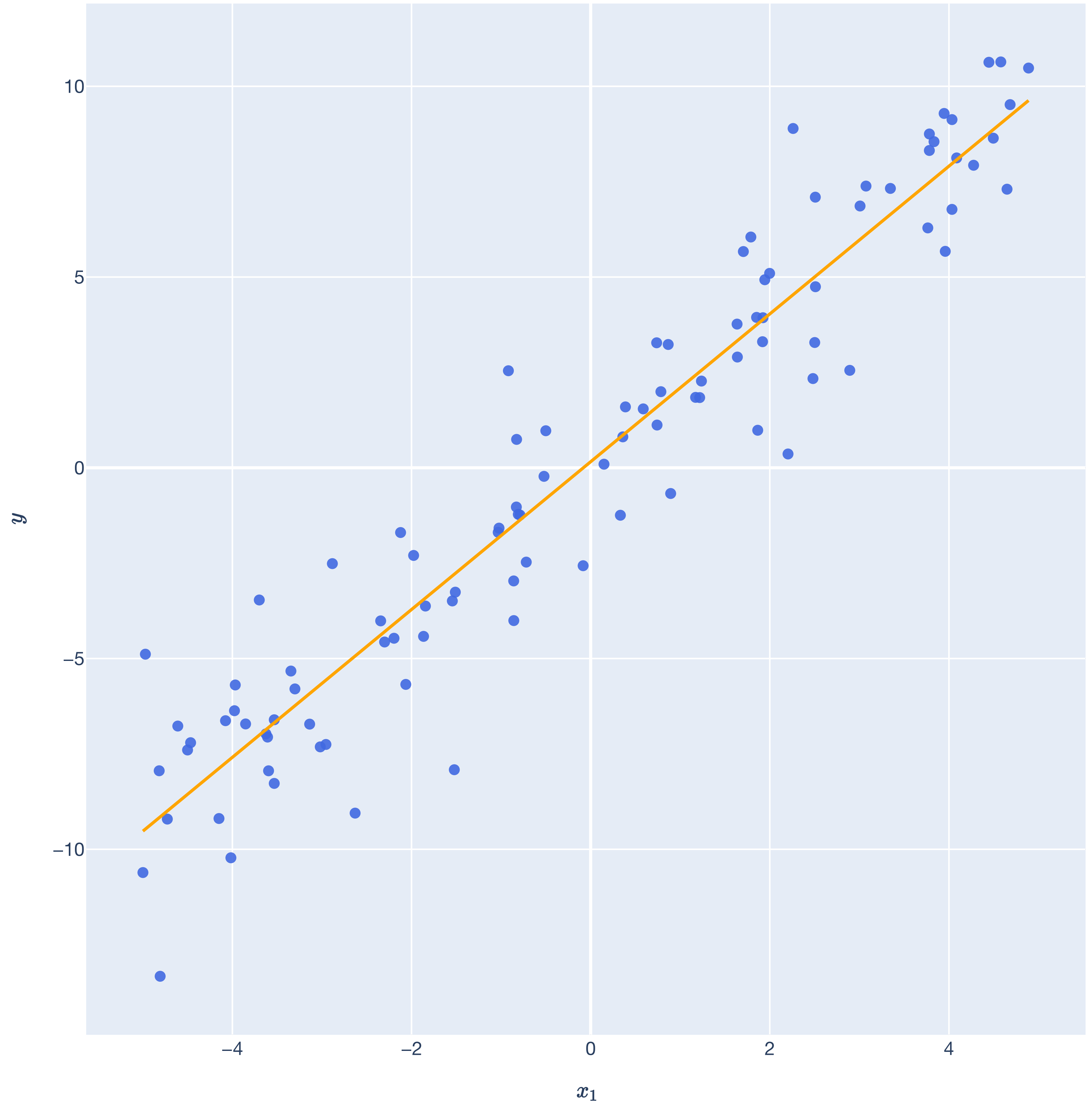
Gaussian Error Model

Further assumption on regression model

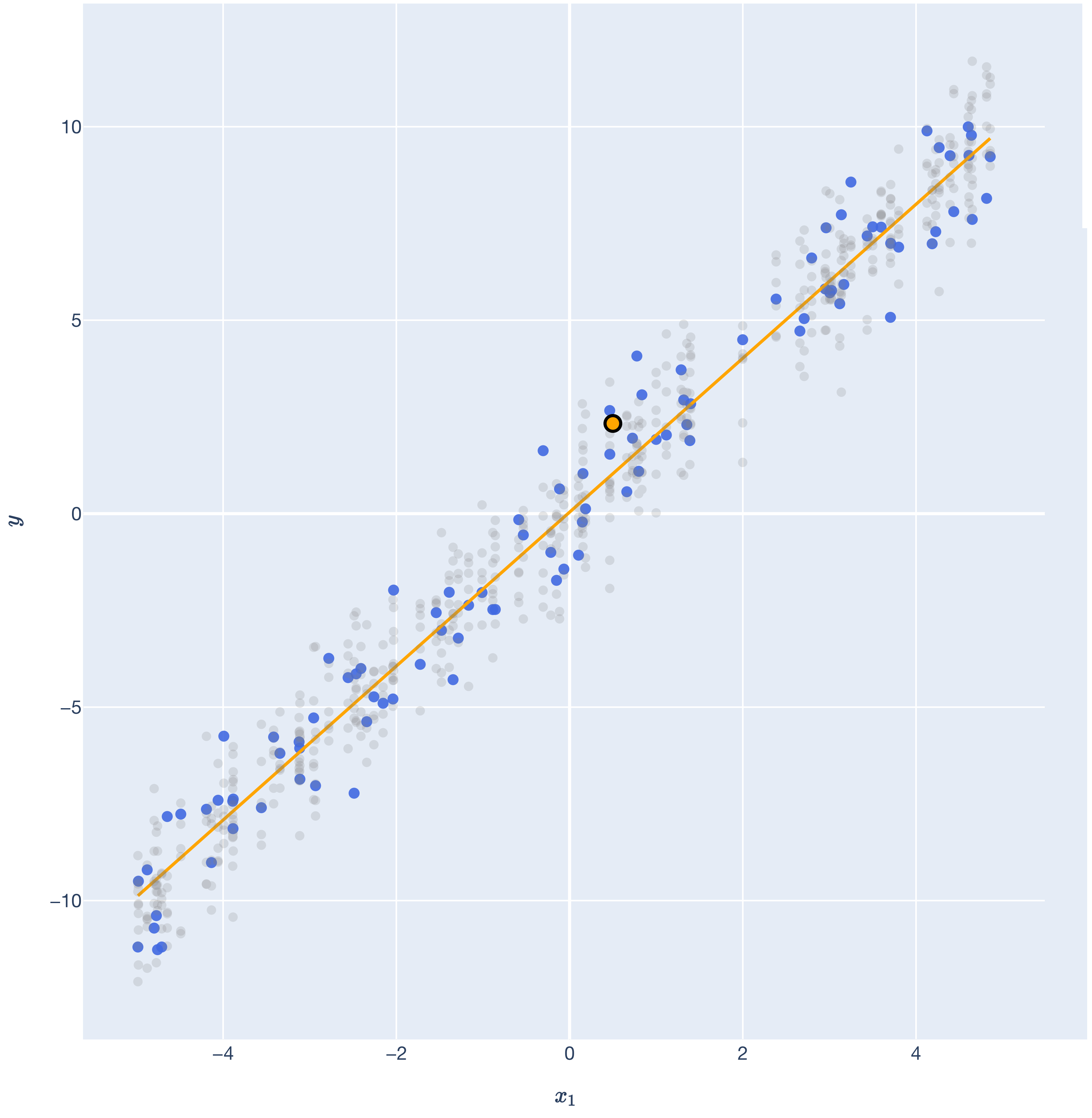
Regression

Setup

Collect labeled training data $\Longrightarrow$ Fit the model $\hat{\mathbf{w}}$ $\Longrightarrow$ Generalize on new $\mathbf{x}_0$



(ERM so far!)



Regression with randomness

Setup

(How we've been doing ERM).

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2.$$

Statistics of OLS

Theorem

Theorem (Statistical properties of OLS). Let $\mathbb{P}_{\mathbf{x},y}$ be a joint distribution $\mathbb{R}^d \times \mathbb{R}$ defined by the error model:

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where $\mathbf{w}^* \in \mathbb{R}^d$ and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$. Suppose we construct a random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and random vector $\mathbf{y} \in \mathbb{R}^n$ by drawing n random examples $(\mathbf{x}_i, y_i)$ from $\mathbb{P}_{\mathbf{x},y}$. Then, the OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the following statistical properties:

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$.

Variance: $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$.

Error Model

Statement of Error Model

Let $\mathbb{P}_{\mathbf{x},y}$ be a joint distribution $\mathbb{R}^d \times \mathbb{R}$ defined by the error model:

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where $\mathbf{w}^* \in \mathbb{R}^d$ and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

In matrix-vector form:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon,$$

where $\epsilon \in \mathbb{R}^d$ is a random vector with covariance matrix $\mathrm{Var}(\epsilon) = \sigma^2 \mathbf{I}$.

Gaussian Error Model

Motivation

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

We can think of ϵ as the randomness from the "unexplained" errors in modeling the relationship of y to $\mathbf{x}$ with a linear model $\mathbf{w}^* \in \mathbb{R}^d$. Possibly very complex!

$\Rightarrow$ Little unexplained errors cancel each other out on avg!

Gaussian Error Model

Motivation

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

We can think of ϵ as the randomness from the "unexplained" errors in modeling the relationship of y to $\mathbf{x}$ with a linear model $\mathbf{w}^* \in \mathbb{R}^d$. Possibly very complex!

From CLT: The distribution of the average of many random variables eventually looks Gaussian. Observable processes in Nature often arise from the sum of many "small contributions."

Gaussian Error Model

Definition

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

← Var(εᵢ) = σᵢ²

$\epsilon \sim N(0,\sigma^2)$ with $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

For realizations $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, each ϵ_i is i.i.d.

The constant variance assumption is known as ***homoskedasticity.***

heteroskedasticity = Var(εᵢ) depends on xᵢ.

Gaussian Error Model

Definition

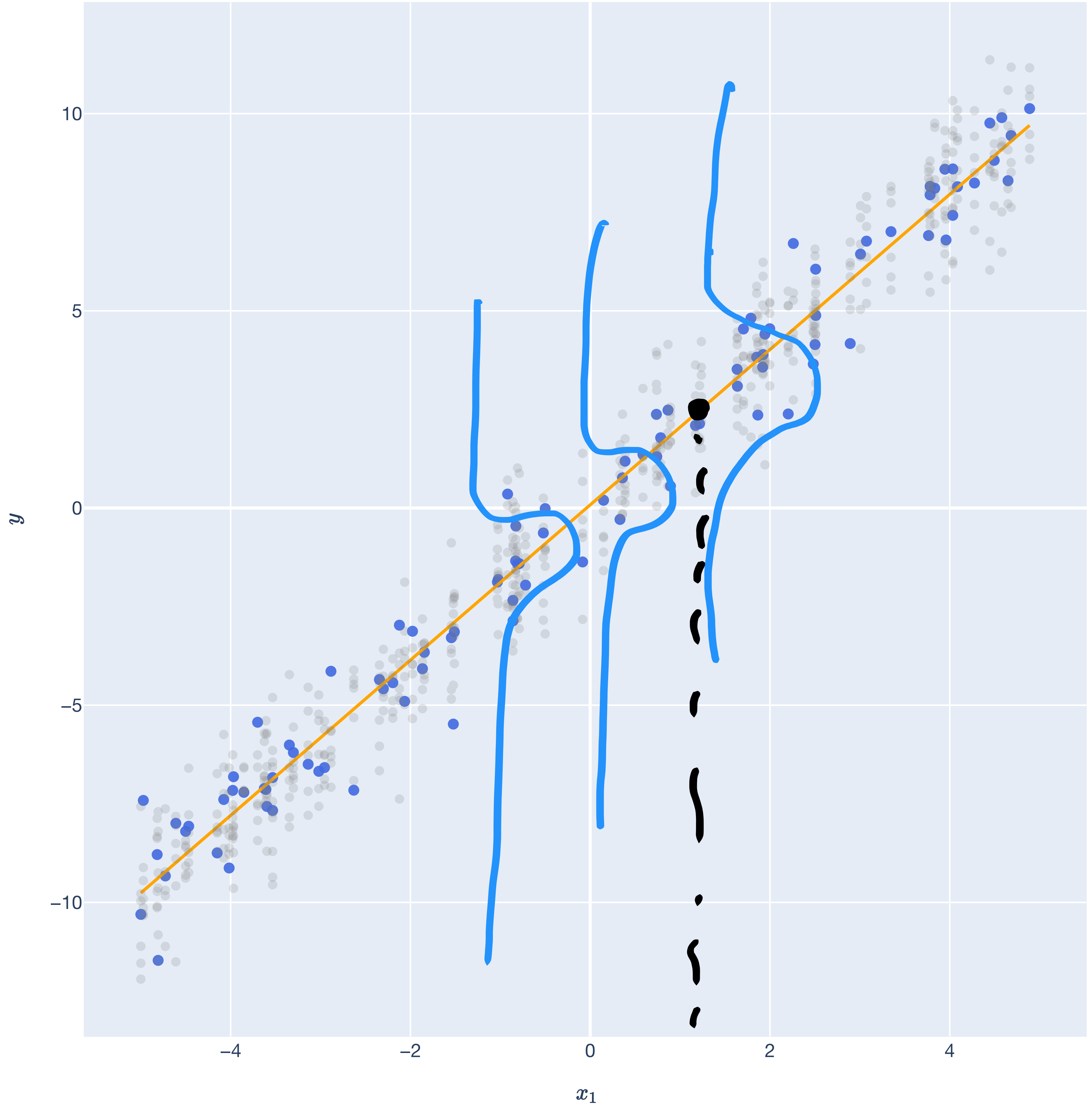
$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\epsilon \sim N(0,\sigma^2)$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

For realizations $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, each ϵ_i is i.i.d.

The constant variance assumption is known as ***homoskedasticity.***

$$\mathbb{E}[y_i \mid x_i] = \mathbb{E}[x_i^T w^*] + \mathbb{E}[\varepsilon_i \mid x_i] = x_i^T w^*$$



Gaussian Error Model

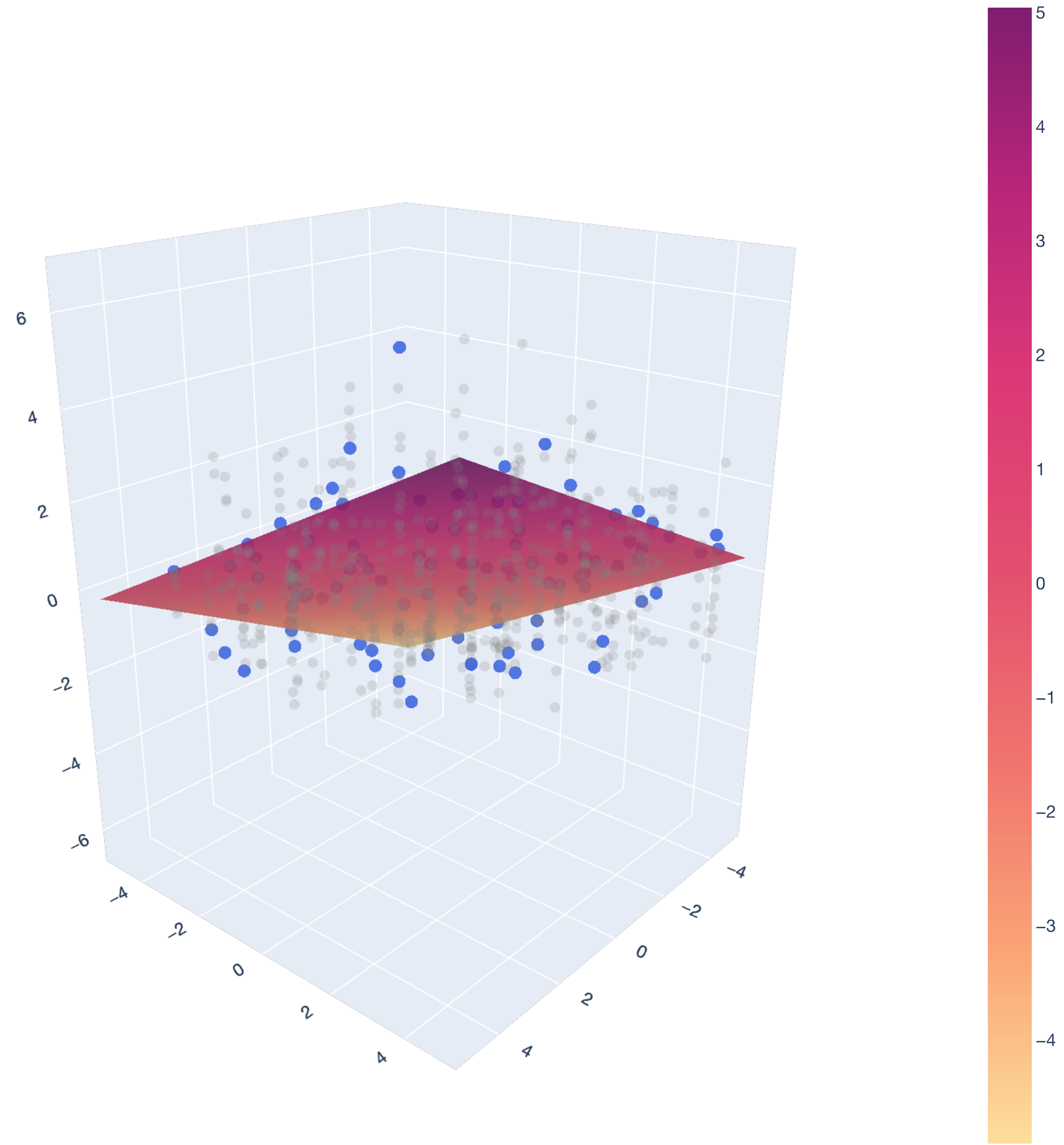
Definition

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\epsilon \sim N(0, \sigma^2)$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

For realizations $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, each ϵ_i is i.i.d.

The constant variance assumption is known as ***homoskedasticity.***



OLS and MLE

Equivalence under Gaussian errors

Problem Setup

Parametric Model

Assume we are in the *Gaussian error model*:

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\epsilon \sim N(0, \sigma^2)$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

For realizations $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, each ϵ_i is i.i.d.

Problem Setup

Parametric Model

$\mu = \mathbb{E}[y|x] = x^T w^*$

$\sigma^2 = \mathrm{Var}(\varepsilon | x) = \mathrm{Var}(\varepsilon)$

Assume we are in the *Gaussian error model*:

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\epsilon \sim N(0, \sigma^2)$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$.

For realizations $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, each ϵ_i is i.i.d.

This defines a *parametric model* on the conditional distribution $\mathbb{P}_{y|\mathbf{x}}$, with parameters $\theta = (\mathbf{w}^*, \sigma)$, with PDF:

$$p(y \mid \mathbf{x}; \mathbf{w}^*, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(y - \mathbf{x}^\top \mathbf{w}^*)^2 / 2\sigma^2\right\}.$$

Problem Setup

Log-Likelihood Function

Parametric model with parameters $\mathbf{w}^*$ and σ:

$$p(y \mid \mathbf{x}; \mathbf{w}^*, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(y - \mathbf{x}^\top \mathbf{w}^*)^2 / 2\sigma^2\right\}.$$

Given i.i.d. data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, the ***likelihood function*** is given by:

$$L_n(\mathbf{w}^*, \sigma) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i; \mathbf{w}^*, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left\{-(y_i - \mathbf{x}_i^\top \mathbf{w}^*)^2 / 2\sigma^2\right\}$$

Problem Setup

Log-Likelihood Function

Parametric model with parameters $\mathbf{w}^*$ and σ:

$$p(y \mid \mathbf{x}; \mathbf{w}^*, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(y - \mathbf{x}^\top \mathbf{w}^*)^2/2\sigma^2\right\}.$$

Given i.i.d. data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, the ***likelihood function*** is given by:

$$L_n(\mathbf{w}^*, \sigma) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i; \mathbf{w}^*, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left\{-(y_i - \mathbf{x}_i^\top \mathbf{w}^*)^2/2\sigma^2\right\}$$

The ***log-likelihood function*** is given by:

$$\mathscr{L}_n(\mathbf{w}^*, \sigma) = \log L_n(\mathbf{w}^*, \sigma) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w}^*)^2}{2\sigma^2}$$

Problem Setup

Log-Likelihood Function

Parametric model with parameters $\mathbf{w}^*$ and σ:

$$p(y \mid \mathbf{x}; \mathbf{w}^*, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(y - \mathbf{x}^\top \mathbf{w}^*)^2 / 2\sigma^2\right\}.$$

Given i.i.d. data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, the ***likelihood function*** is given by:

$$L_n(\mathbf{w}^*, \sigma) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i; \mathbf{w}^*, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left\{-(y_i - \mathbf{x}_i^\top \mathbf{w}^*)^2 / 2\sigma^2\right\}$$

The ***log-likelihood function*** is given by:

$$\mathcal{L}_n(\mathbf{w}^*, \sigma) = \log L_n(\mathbf{w}^*, \sigma) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w}^*)^2}{2\sigma^2}$$

Let's optimize and solve this for $\mathbf{w}^*$!

Finding the MLE

Solving the MLE Optimization Problem

The ***log-likelihood function*** is given by:

$$\mathscr{L}_n(\mathbf{w}, \sigma) = \log L_n(\mathbf{w}, \sigma) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}$$

We want to optimize and solve this for the estimand $\mathbf{w}$ (we don't care about estimating σ). To get $\hat{\mathbf{w}}_{MLE}$, we solve the optimization problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{maximize}}\ \mathscr{L}_n(\mathbf{w}) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}$$

Finding the MLE

Solving the MLE Optimization Problem

The ***log-likelihood function*** is given by:

$$\mathcal{L}_n(\mathbf{w}, \sigma) = \log L_n(\mathbf{w}, \sigma) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}$$

We want to optimize and solve this for the estimand $\mathbf{w}$ (we don't care about estimating σ). To get $\hat{\mathbf{w}}_{MLE}$, we solve the optimization problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \; -\mathcal{L}_n(\mathbf{w}) = -n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}$$

Finding the MLE

Solving the MLE Optimization Problem

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

In matrix-vector form, this is the same as the optimization problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2\sigma^2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Finding the MLE

Solving the MLE Optimization Problem

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

In matrix-vector form, this is the same as the optimization problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \; \frac{1}{2\sigma^2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

But $1/2\sigma^2$ is just a constant, so this is equivalent to OLS!

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \; \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

OLS and MLE

Theorem Statement

squared loss: $\ell(f,(x,y)) := (f(x)-y)^2 \longrightarrow \sum_{i=1}^{n} \ell(f,(x_i,y_i))$

$|f(x)-y|$

0-1 loss: $\ell(f,(x,y)) := \mathbb{1}\{f(x) \neq y\}$

Theorem (OLS and MLE). Suppose that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d. samples in $\mathbb{R}^d \times \mathbb{R}$ with conditional distribution $\mathbb{P}_{y|\mathbf{x}}$ defined by:

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon,$$

where $\epsilon_i \sim N(0,\sigma^2)$ and each ϵ_i is independent. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$ contain all the i.i.d. samples. Then, the maximum likelihood estimate (MLE) $\hat{\mathbf{w}}_{MLE}$ of the parameter $\mathbf{w}^*$ is given by the OLS estimator:

$$\hat{\mathbf{w}}_{MLE} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

OLS and MLE

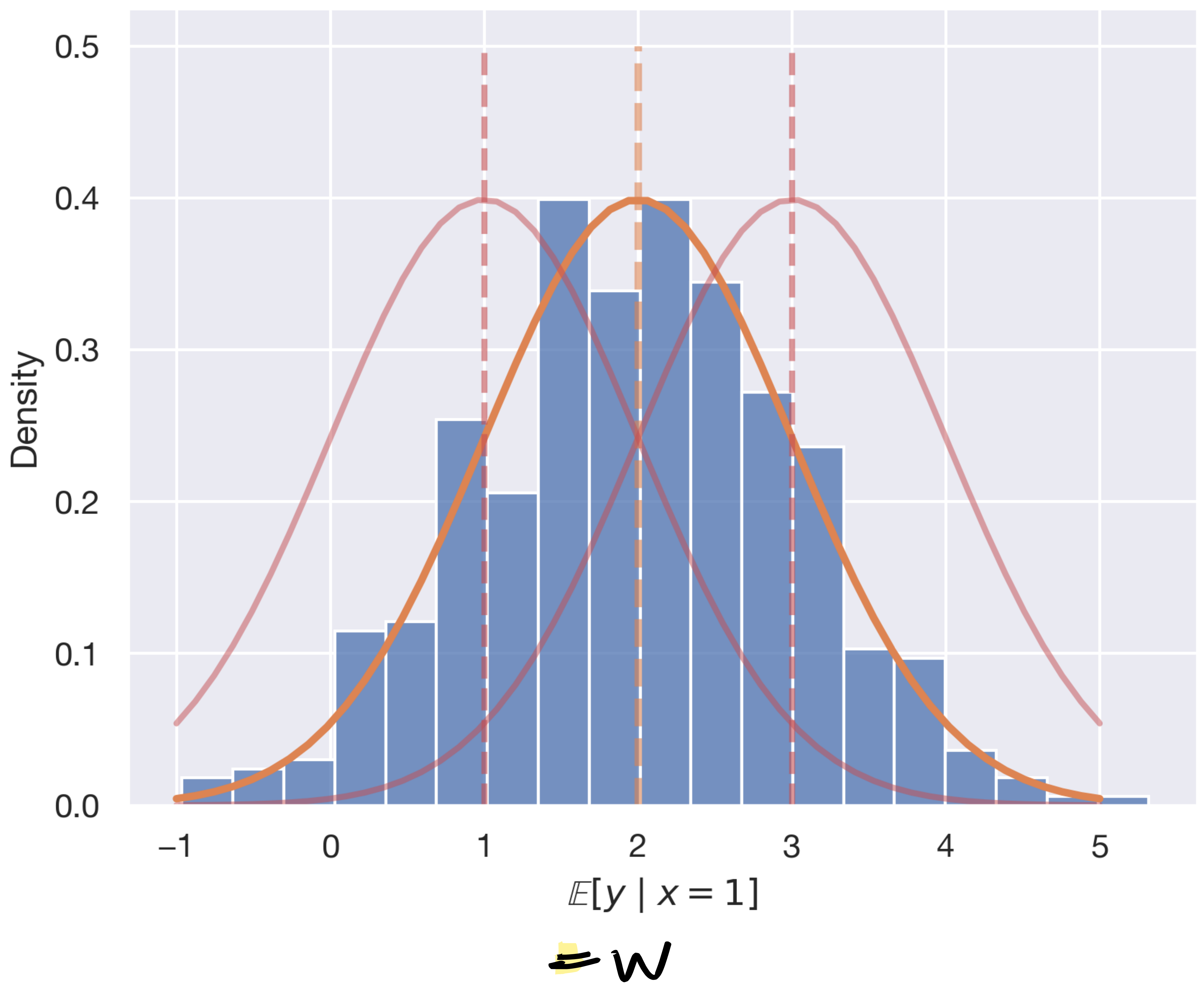
Theorem Statement

Theorem (OLS and MLE). Suppose that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d. samples in $\mathbb{R}^d \times \mathbb{R}$ with conditional distribution $\mathbb{P}_{y|\mathbf{x}}$ defined by:

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon,$$

where $\epsilon_i \sim N(0, \sigma^2)$ and each ϵ_i is independent. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ contain all the i.i.d. samples. Then, the maximum likelihood estimate (MLE) $\hat{\mathbf{w}}_{MLE}$ of the parameter $\mathbf{w}^*$ is given by the OLS estimator:

$$\hat{\mathbf{w}}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$



OLS and MLE

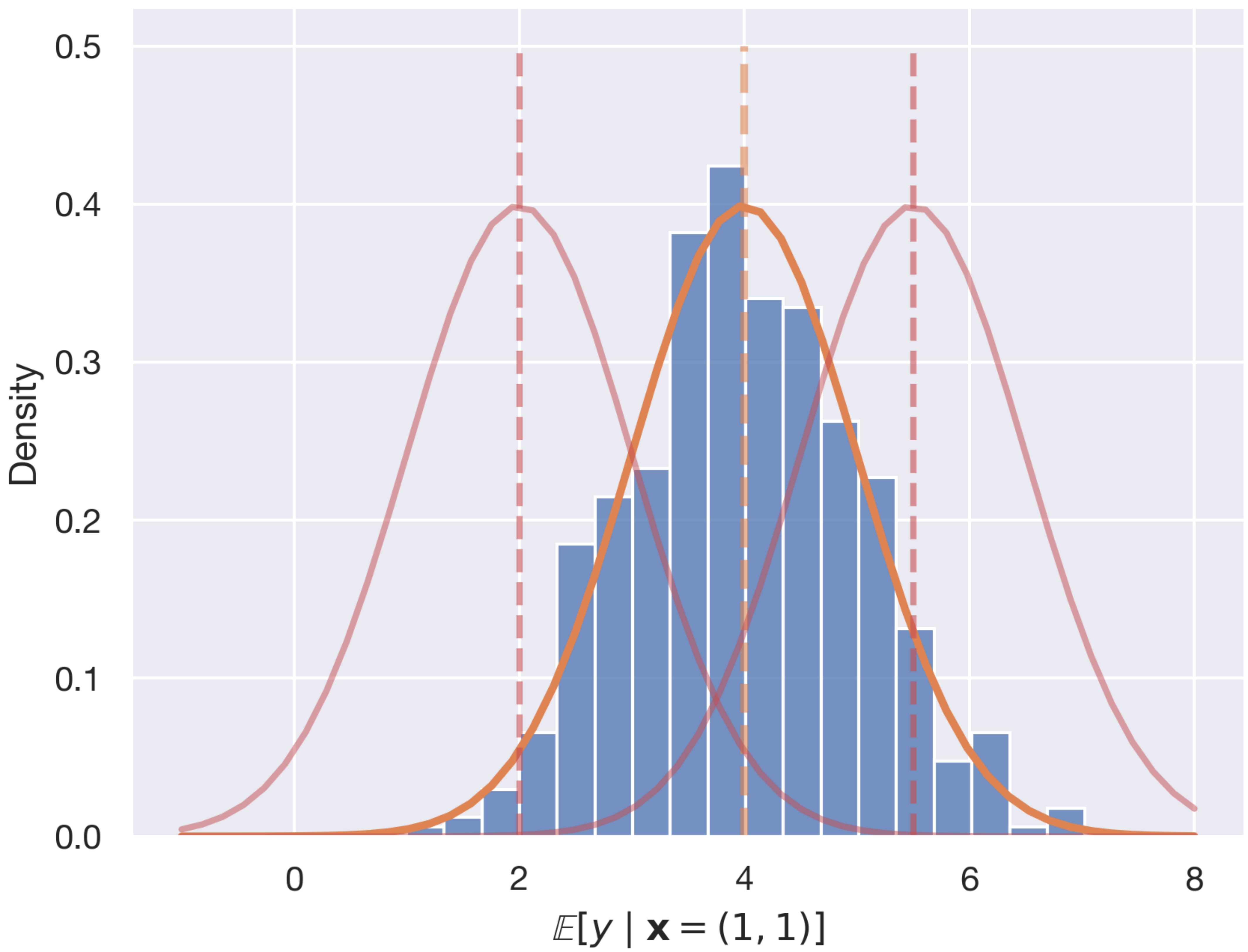
Theorem Statement

Theorem (OLS and MLE). Suppose that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d. samples in $\mathbb{R}^d \times \mathbb{R}$ with conditional distribution $\mathbb{P}_{y|\mathbf{x}}$ defined by:

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon,$$

where $\epsilon_i \sim N(0, \sigma^2)$ and each ϵ_i is independent. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ contain all the i.i.d. samples. Then, the maximum likelihood estimate (MLE) $\hat{\mathbf{w}}_{MLE}$ of the parameter $\mathbf{w}^*$ is given by the OLS estimator:

$$\hat{\mathbf{w}}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$



Recap

Lesson Overview

Gaussian Distribution. We define perhaps the most important "named" probability distribution, the Gaussian/"Normal" distribution, and go over some key properties.

Central Limit Theorem. We state and prove the central limit theorem, the statement that the sample average of *many* independent random variables converges in distribution to the Gaussian. It doesn't matter what distribution those random variables take!

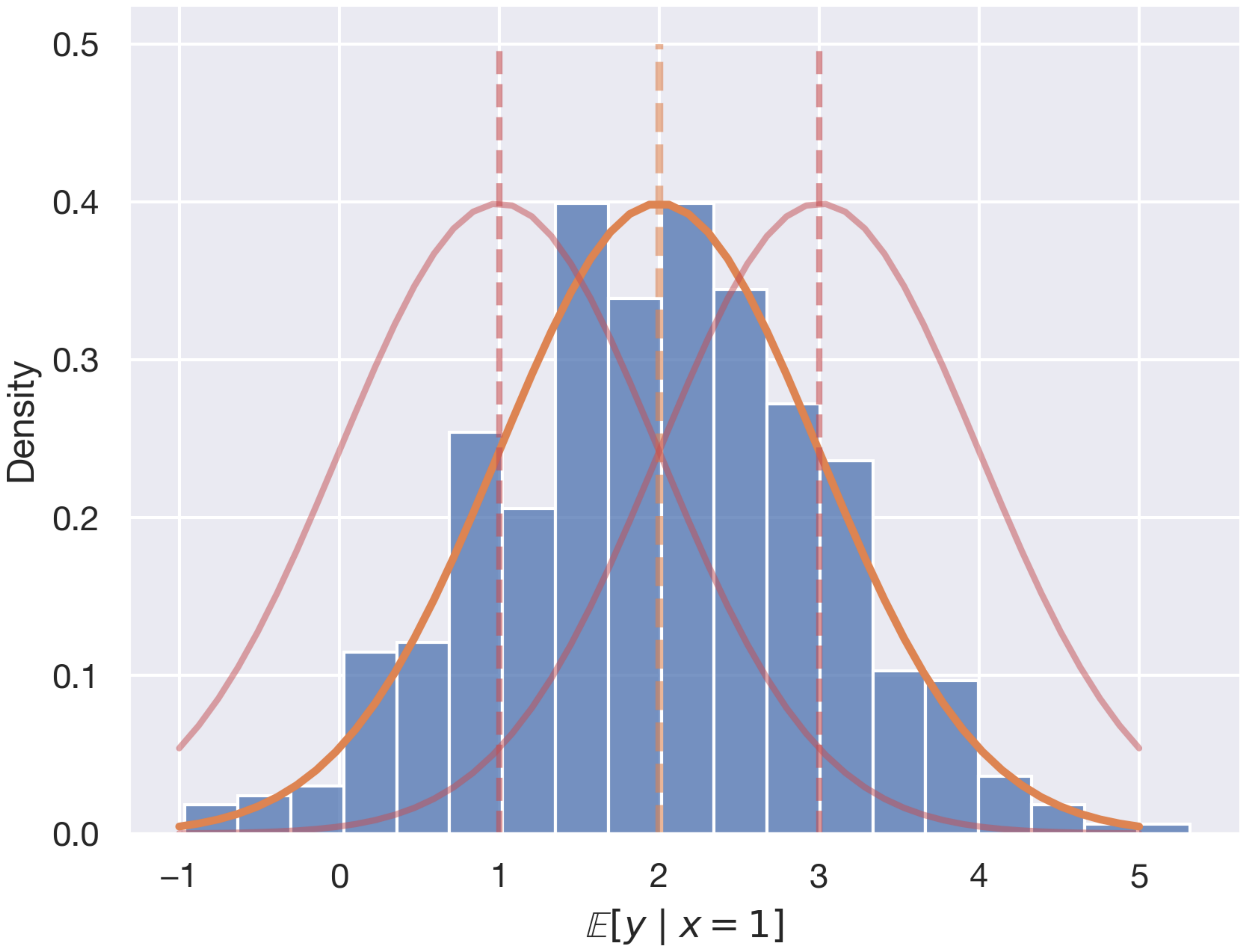
"Named" Distributions. We review other common "named" distributions for discrete and continuous random variables.

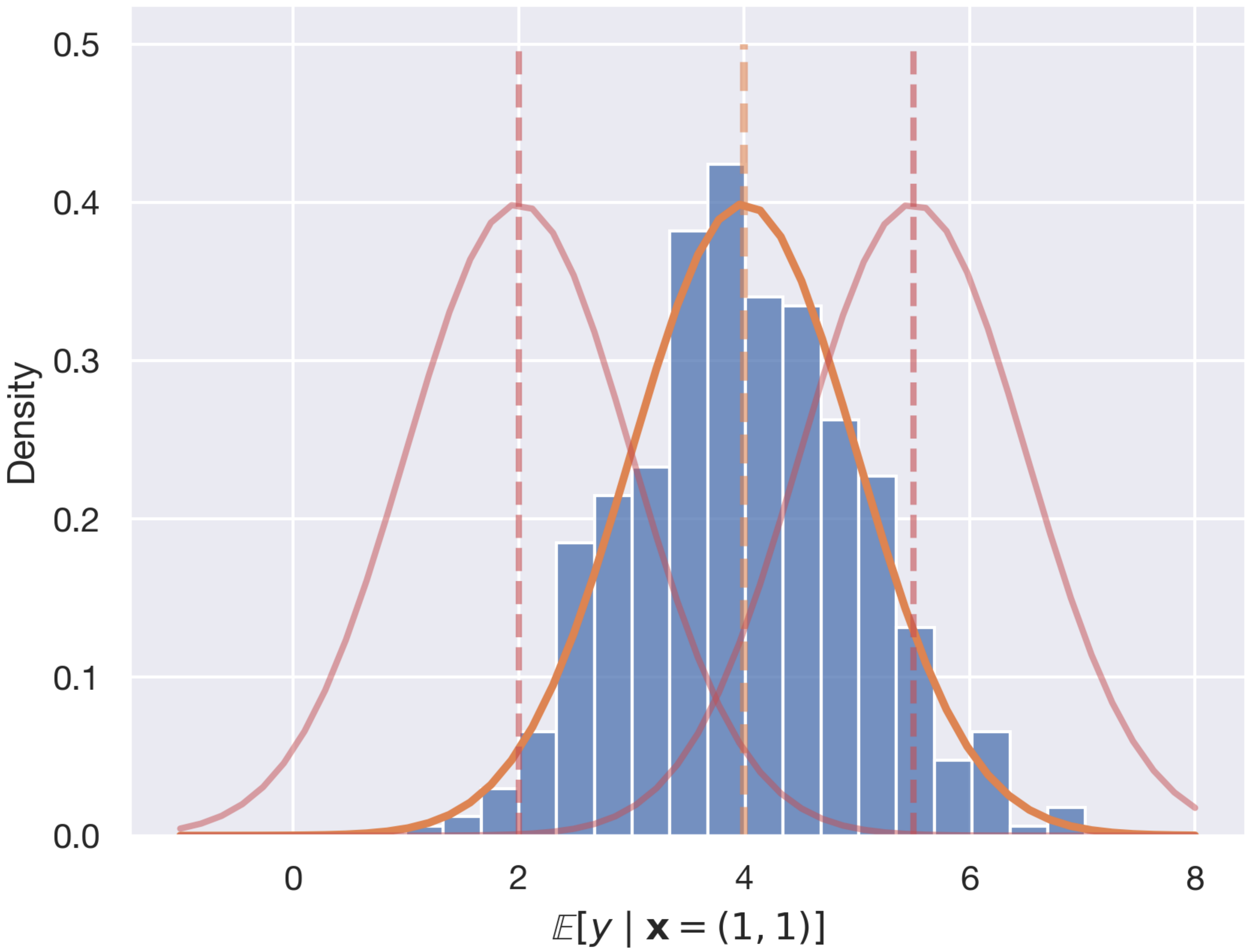
Maximum likelihood estimation. We define maximum likelihood estimation (MLE), a statistical/probabalistic perspective towards finding a well-generalizing model for data.

MLE and OLS. We explore the connection between MLE and OLS by defining the Gaussian error model. In this model, MLE and OLS correspond.

Lesson Overview

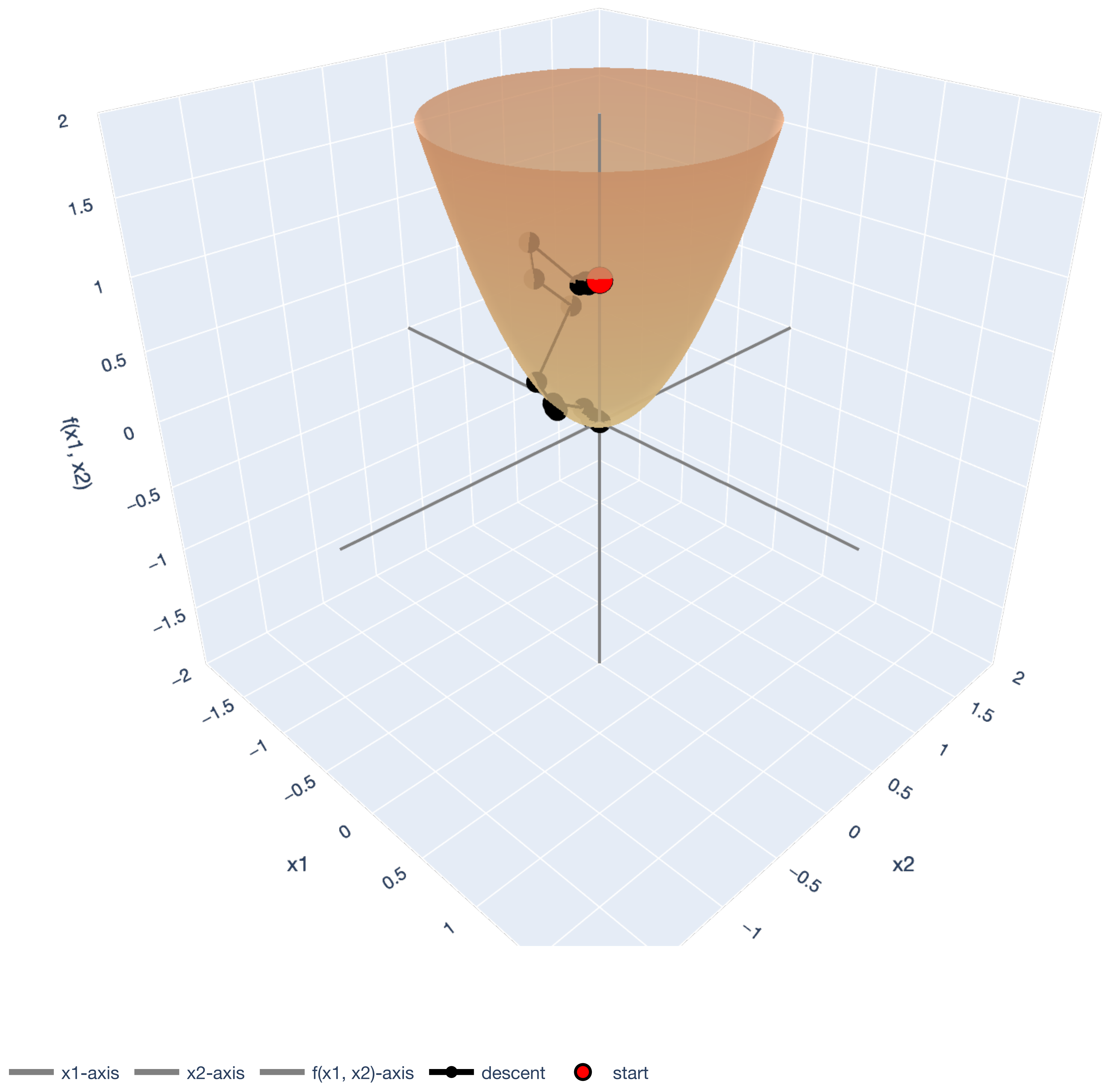
Big Picture: Least Squares

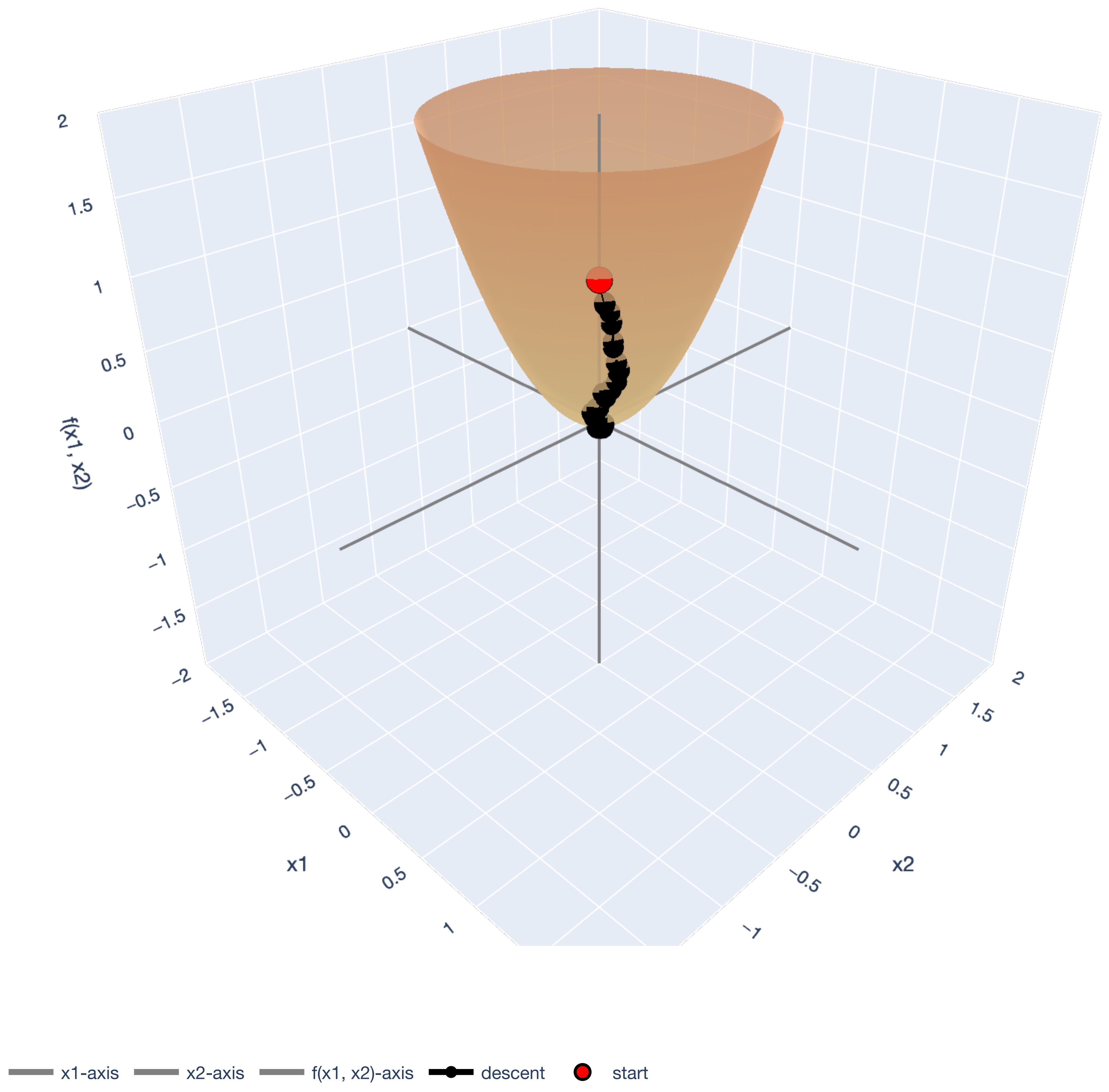




Lesson Overview

Big Picture: Gradient Descent





References

Mathematics for Machine Learning. Marc Pieter Deisenroth, A. Aldo Faisal, Cheng Soon Ong.

Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach. John H. Hubbard and Barbara Burke Hubbard.

"Gradient Descent." Daniel Hsu. Lecture notes from COMS4771 Machine Learning, Fall 2023.

"Lecture 2: Local Theory of Optimization." Santiago Balserio and Ciamac Moallemi. Lecture notes from B9118 Foundations of Optimization, Fall 2023.