etherpad.wikimedia.org/p/312 for (anonymous) questions/comments!

Central Limit Theorem

CSE 312 24Su

Lecture 15

Logistics

- Reminder about concept checks 12, 13, and 14, late due date tonight
- Midterm grades released on ~Wednesday
- Updated lecture notes for last Wed and Fri lecture on website

Normal Distributions

A **normal random variable** $X \sim \mathcal{N}(\mu, \sigma^2)$ has two parameters:

- $\mu = \mathbb{E}[X]$ is the mean
- $\sigma^2 = \mathrm{Var}(X)$ is the variance ($\sigma = \sqrt{Var(X)}$ is *standard deviation*)

and follows this ***probability density function*** (a bell curve!):

$$f_X(k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}}$$

Normal Distributions

The *CDF* has no closed form, so instead, we have a table containing values of the CDF for a standard normal random variable $\mathcal{N}(0,1)$.

To find the probability of a normal RV $X \sim \mathcal{N}(\mu, \sigma^2)$ being in some range…

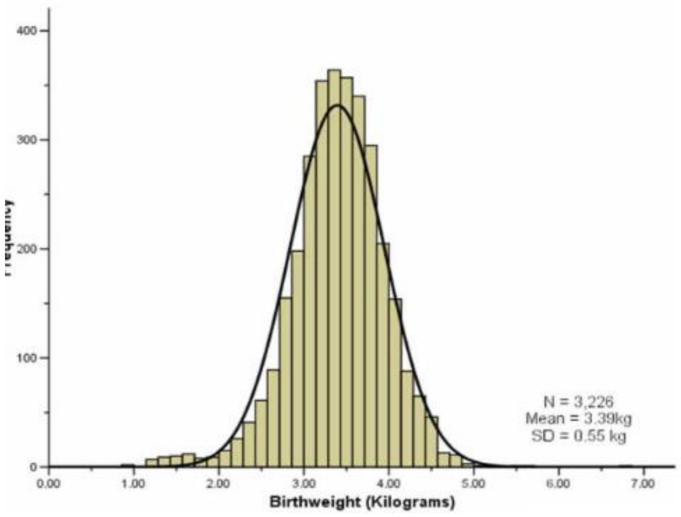
1. **Standardize** the normal random variable: $Z = \frac{X-\mu}{\sigma}$

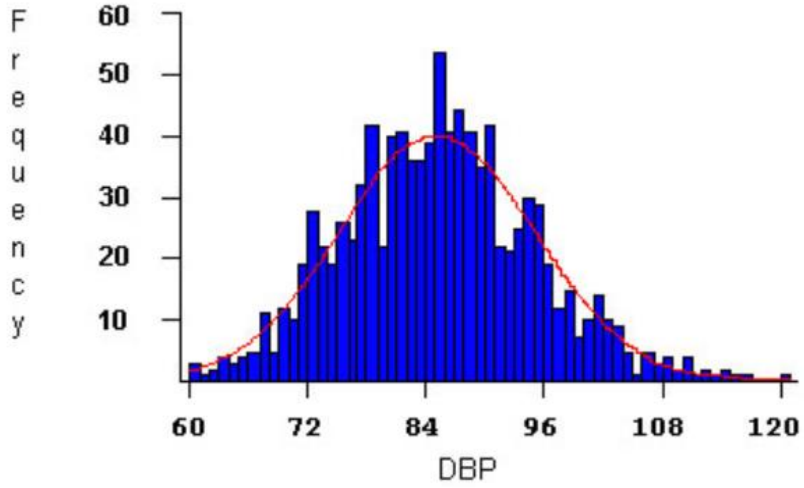
   note: when we standardize, the numbers left are called z-scores (the number of standard deviations away from the mean (e.g., $\mathbb{P}(Z \geq 2)$ *means we're finding probability of being more than 2 standard deviations away from the mean)*

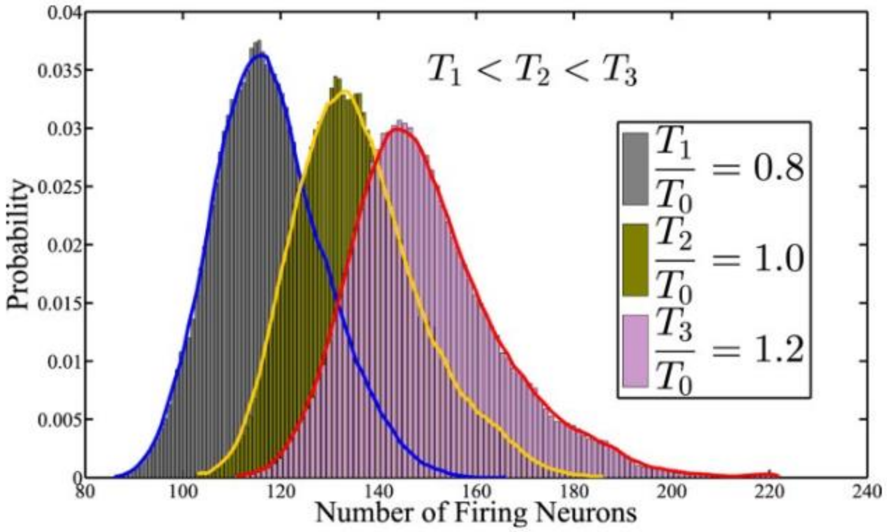
2. **Write probability expression in terms of** $\boldsymbol{\Phi}(\mathbf{z}) = \mathbb{P}(Z \leq z)$

3. **Look up the value(s)** in the table

Normal distributions show up everywhere!







But…why?

This is because of what we call, the *central limit theorem*!

"The sum of **any** independent random variables **approaches** a normal distribution. It becomes closer to normal as we sum more RVs together."

More formally, the Central Limit Theorem!

This is because of what we call, the *central limit theorem*!

"The sum of **any** independent random variables **approaches** a normal distribution. It becomes closer to normal as we sum more RVs together."

Central Limit Theorem

If $X_1, X_2, \dots, X_n$ are i.i.d. random variables, each with mean μ and variance σ^2

Let $Y_n = X_1 + X_2 + \cdots + X_n$

As $n \to \infty$, Y_n approaches a normal distribution $\mathcal{N}(n \cdot \mu, n \cdot \sigma^2)$

(i.e., CDF of Y_n converges to the CDF of $\mathcal{N}(n \cdot \mu, n \cdot \sigma^2)$)

What does i.i.d mean?

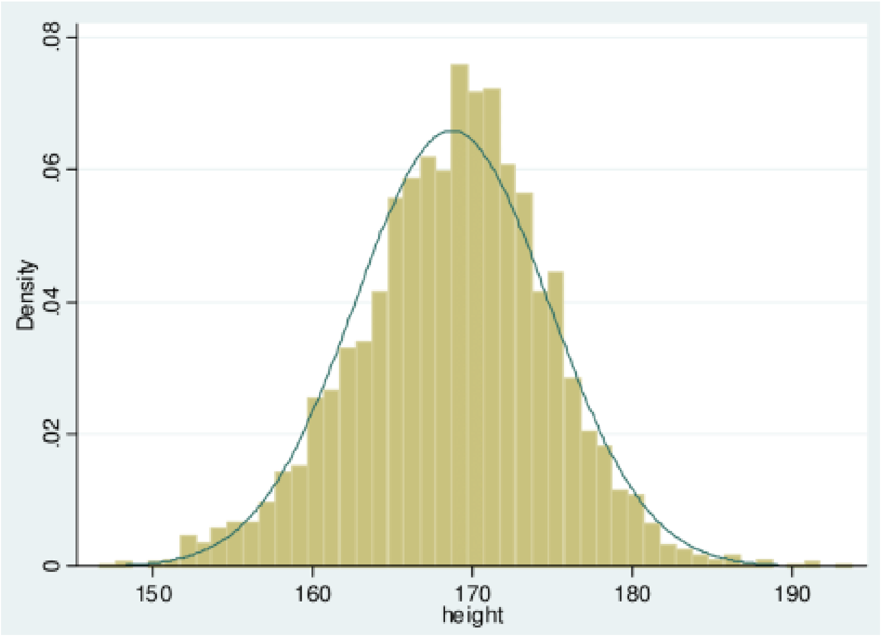
Independent and Identically Distributed (i.i.d)

For random variables $X_1, X_2, \ldots, X_n$ to be i.i.d., they must

- Be <u>mutually independent</u>
  "knowing value of one random variable doesn't give us info about the value of others"
- All have the <u>same PMF</u> (if discrete) or <u>PDF</u> (if continuous)
  "they follow the same probability distribution"

CLT with ***RVs that are NOT i.i.d***

(R.A. Fisher (1918))



Height may be expressed as the ***sum of many independent, random factors***

How much milk you drank per day as a child

Which variant of GH1 (a growth hormone) you have

How much protein/calcium/vitamins/minerals you had as a child

How many hours of sleep you averaged

How many hours of physical activity you averaged

$H = H_1 + H_2 +,,, + H_n$ -> $H \sim N(\ldots)$

A version of CLT does work here! But it's outside the scope of this class.

CLT with ***RVs that are i.i.d***

"number of firing neurons" – sum of indicator random variables for whether each neuron fired

Assume: each neuron is **independent**, and has the **same probability**

"number of people who voted for someone" – sum of indicator random variables for (assume people are independent)

Assume: each person makes **independent**, and has the **same probability**

"total amount invested in a year" – sum of random variables four amount invested each day (assume each investment is independent)

Assume: each day's investment is **independent** and follows the **same distribution**

We will use CLT in this class on problems like this.

A Sum of i.i.d Random Variables

If we have $X_1, X_2, \dots, X_n$ as i.i.d RVs each with mean μ and variance σ^2

$S_n = X_1 + X_2 + \cdots + X_n$ is the sum of those RVs. Then…

- **Expectation.** *by linearity of expectation…*

$$\mathbb{E}[S_n] = \mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] = n\mu$$

- **Variance**. *by linearity of variance <u>because of independence</u>*

$$Var(S_n) = Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(n\sigma^2$$

Proof of the CLT?

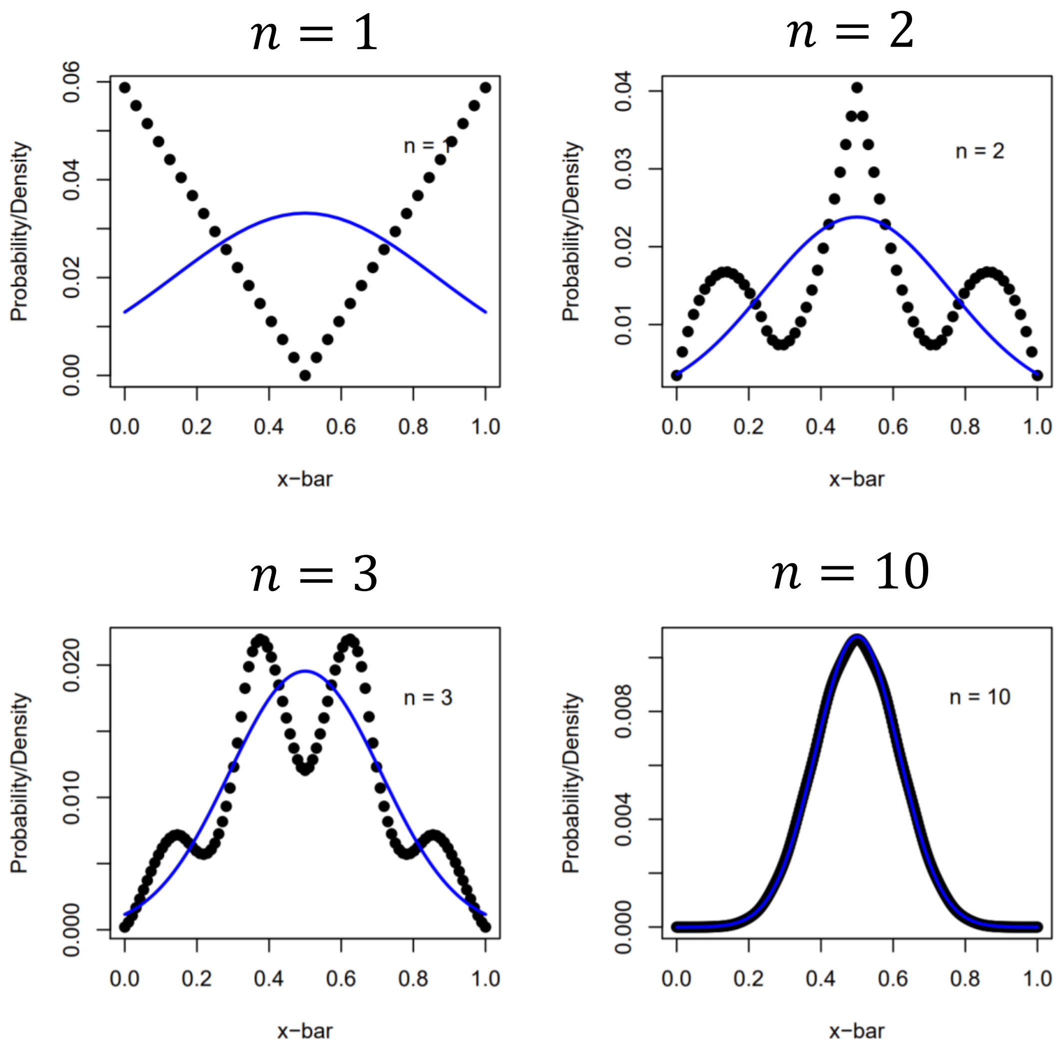
We're not going to cover the proof here.

How is the proof done?

Step 1: Prove that for all positive integers k, $[(Y_n)^k] \to \mathbb{E}[Z^k]$

Step 2: Prove that if $\mathbb{E}[(Y_n)^k] = \mathbb{E}[Z^k]$ for all k then $F_{Y_n}(z) = F_Z(z)$

"Proof by example"



CLT says a sum of n i.i.d RVs approaches a normal distribution as n gets larger

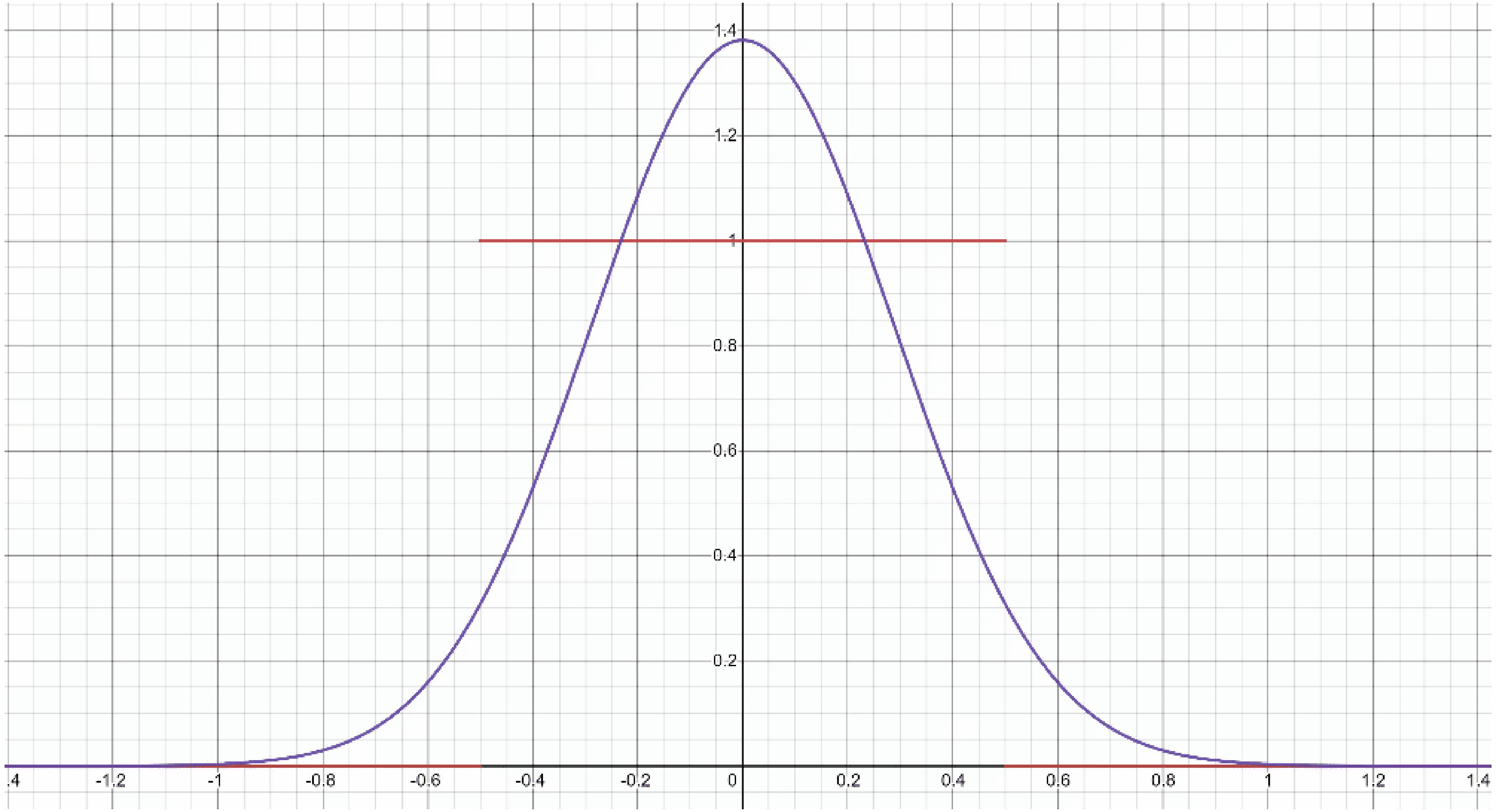
n is the number of i.i.d RVs summed

The **dotted lines** show an "empirical PMF" – a PMF estimated by running the experiment a large number of times.

The blue line is the normal RV that the CLT predicts.

Shown are $n = 1,2,3,10$

“Proof by example” -- uniform

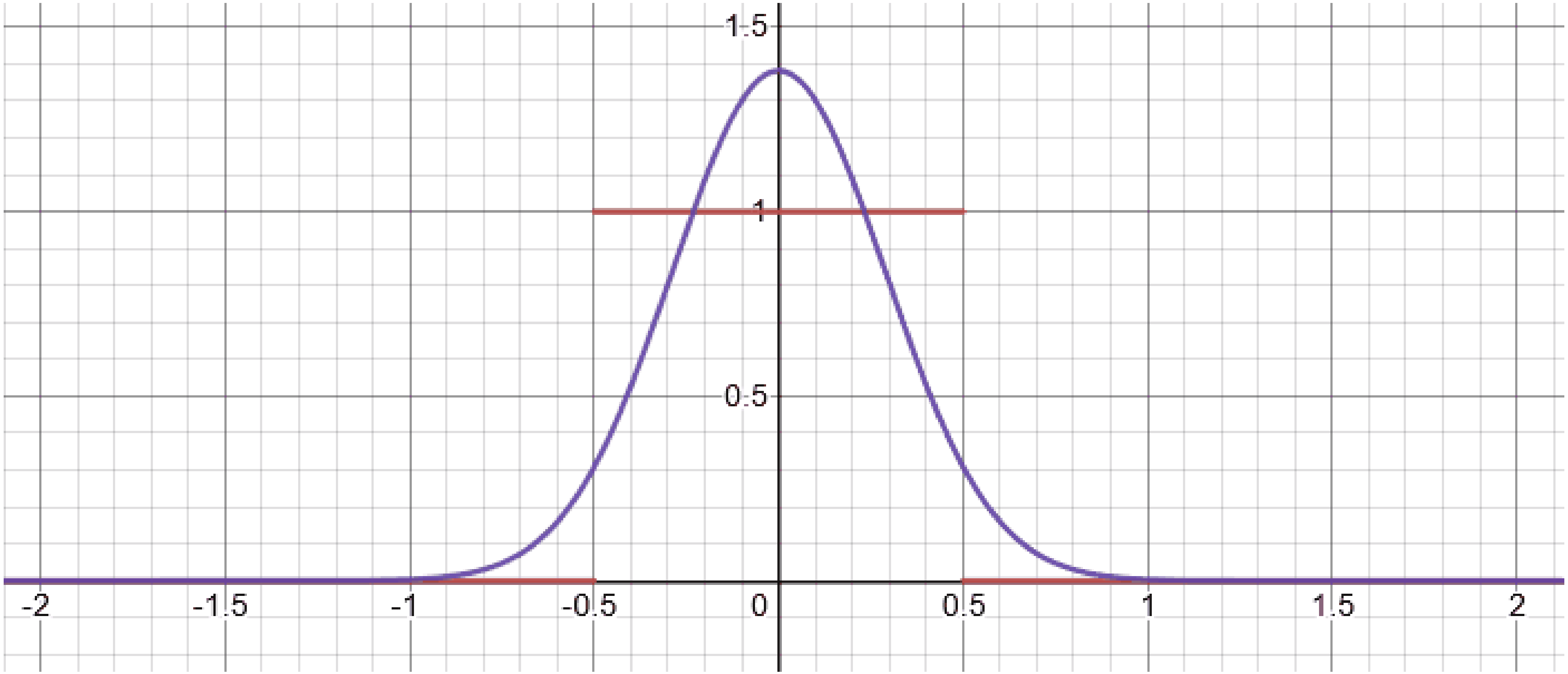


https://www.desmos.com/calculator/2n2m05a9km

“Proof by example” -- uniform

$$\boldsymbol{X = X_1}$$

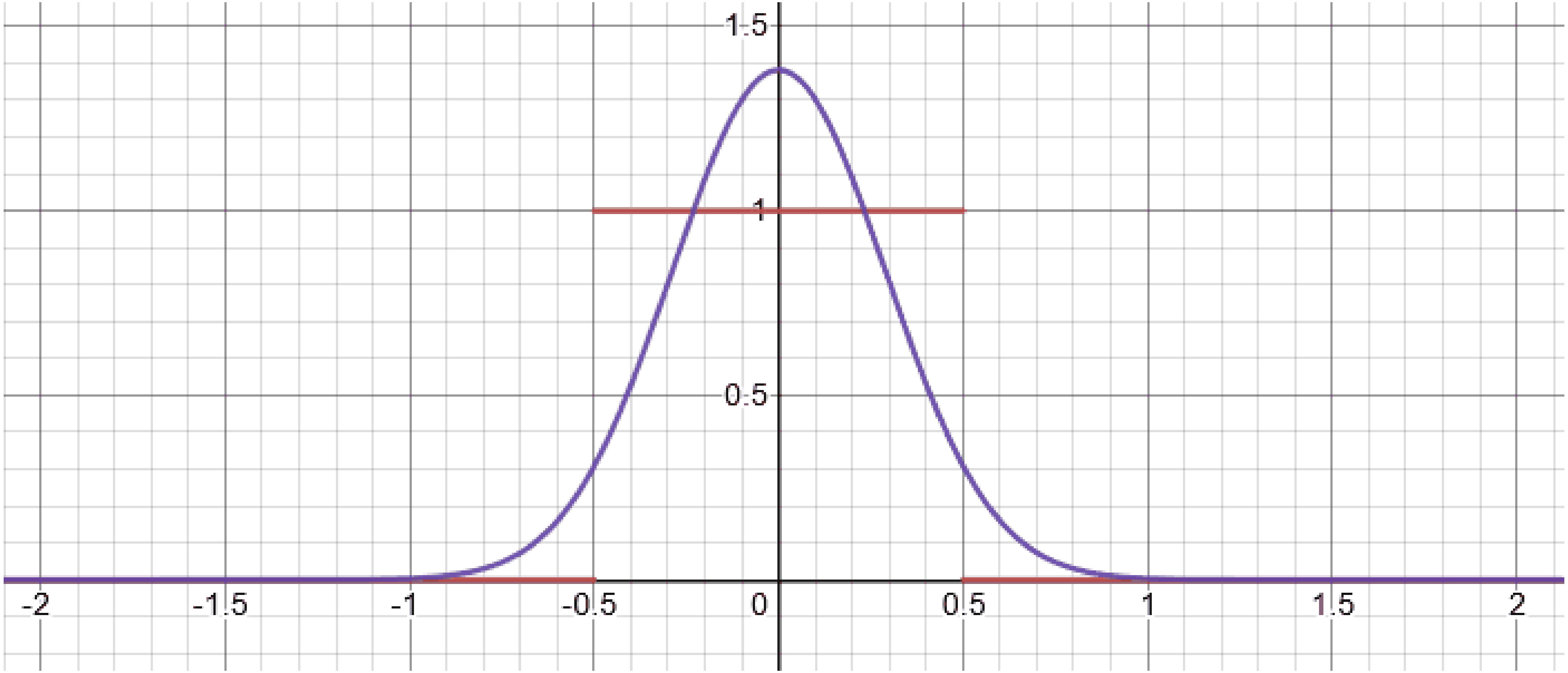
where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent



“Proof by example” -- uniform

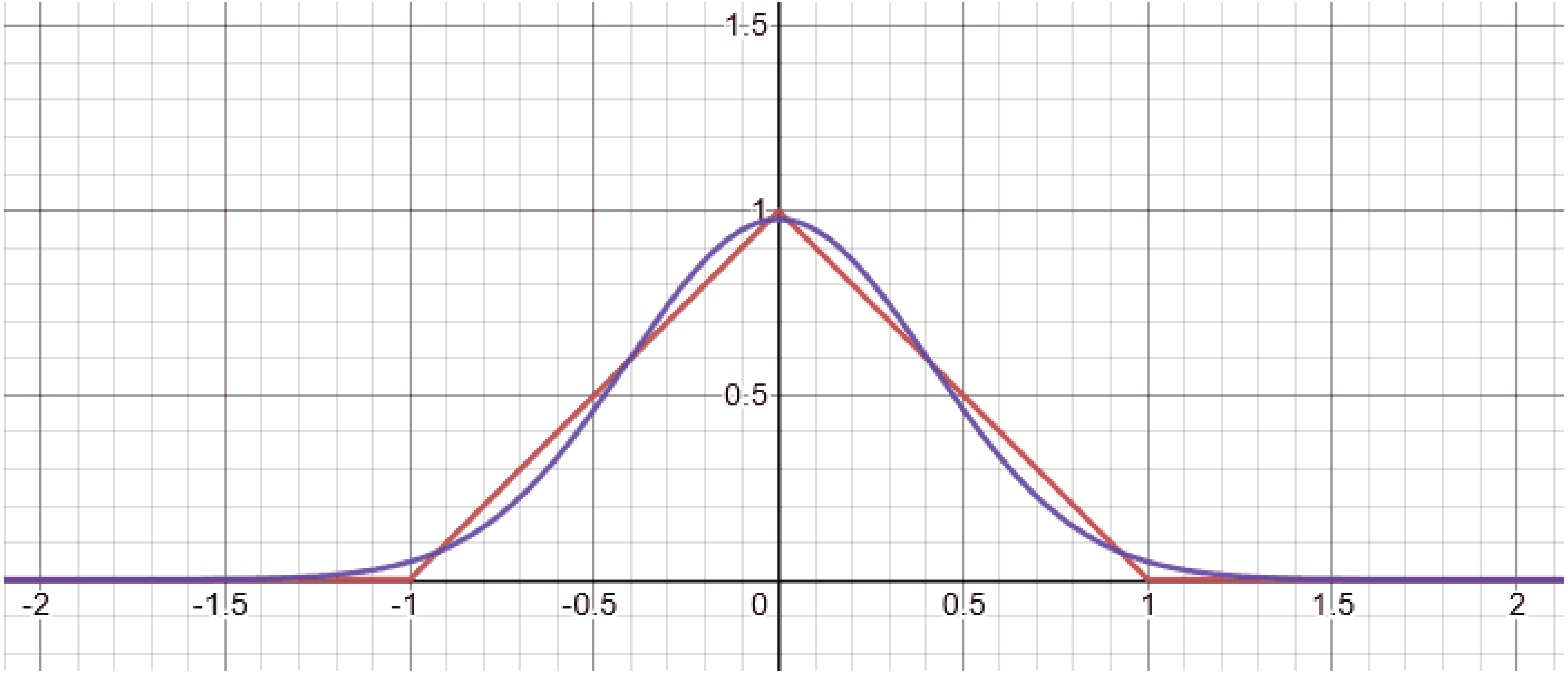
$$\boldsymbol{X = X_1}$$

where each $X_i \sim \text{Unif}(0,1)$ and is independent



“Proof by example” -- uniform

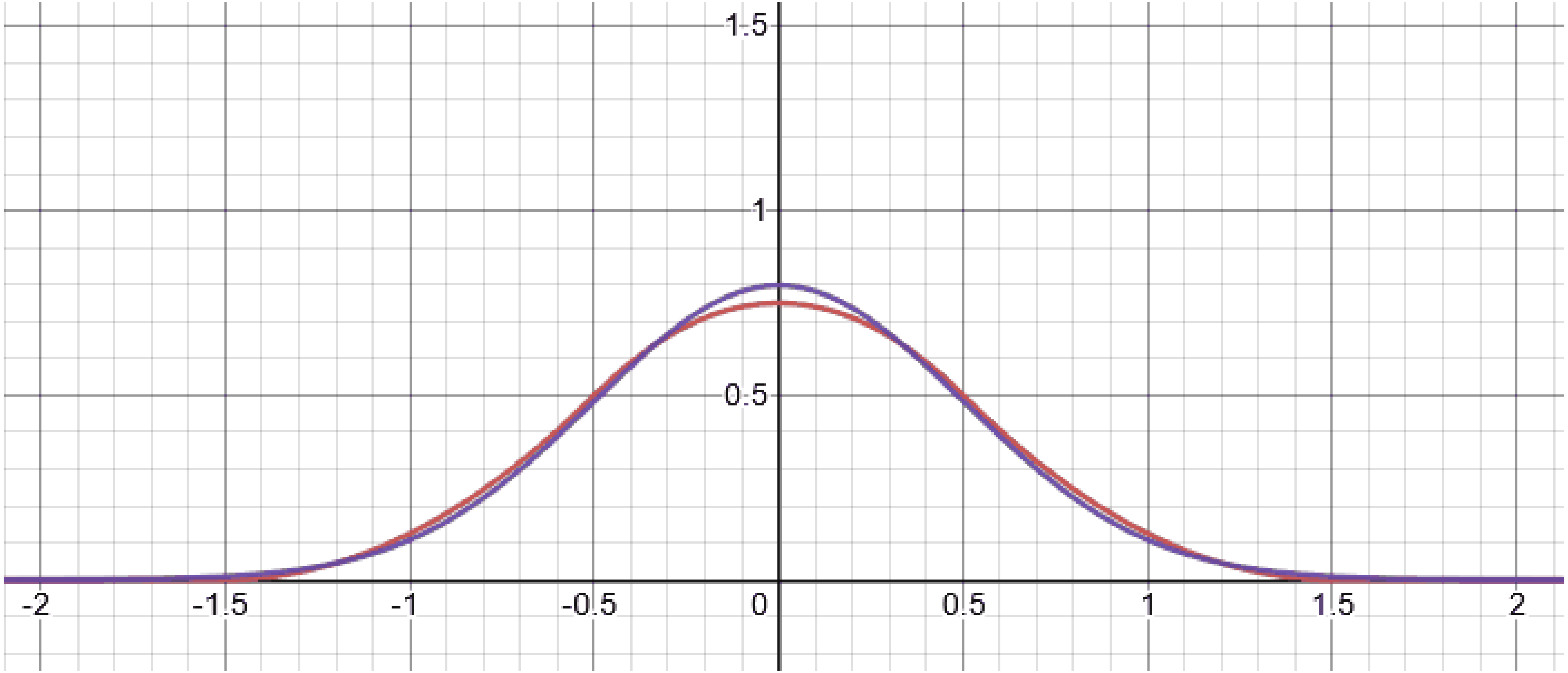
$$X = X_1 + X_2$$
where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent



"Proof by example" -- uniform

$$\boldsymbol{X = X_1 + X_2 + X_3}$$

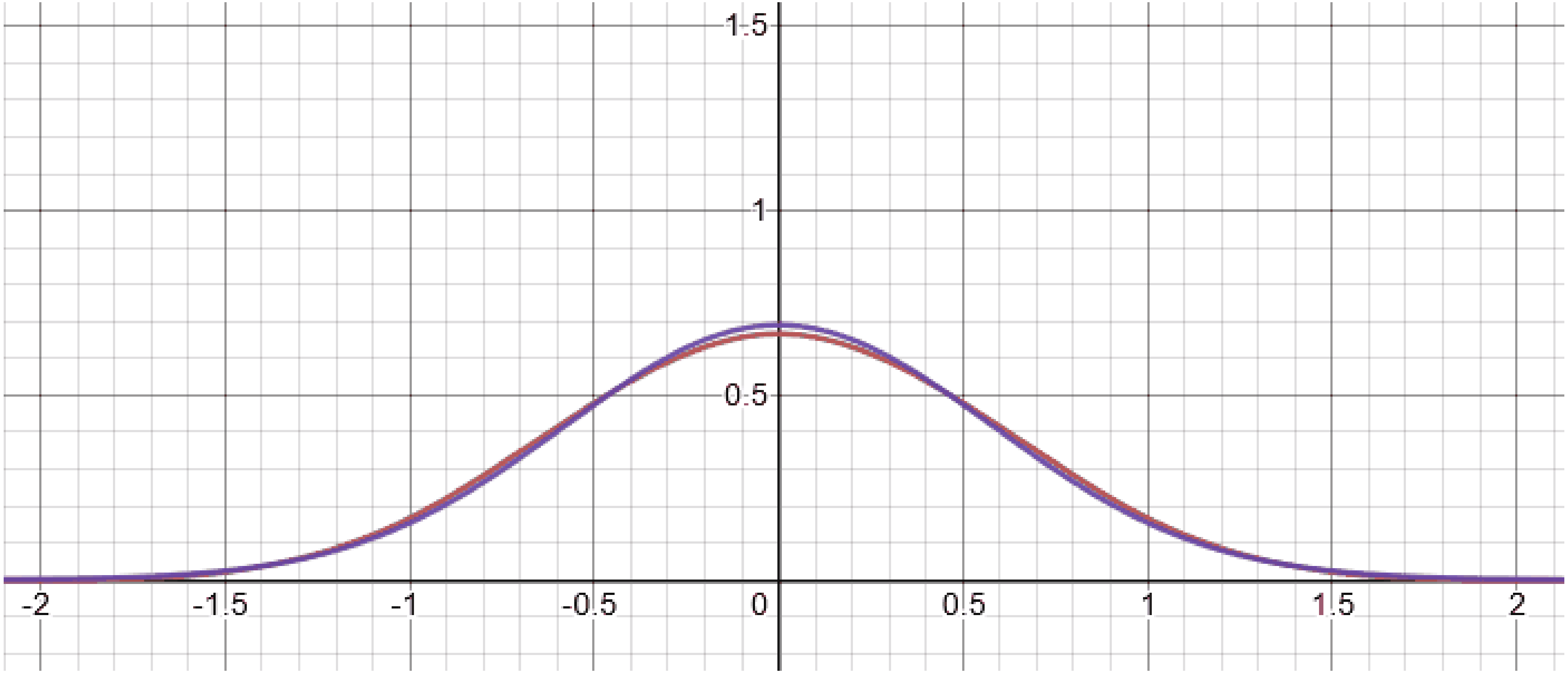
where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent



"Proof by example" -- uniform

$$\boldsymbol{X} = \boldsymbol{X_1} + \boldsymbol{X_2} + \boldsymbol{X_3} + \boldsymbol{X_4}$$

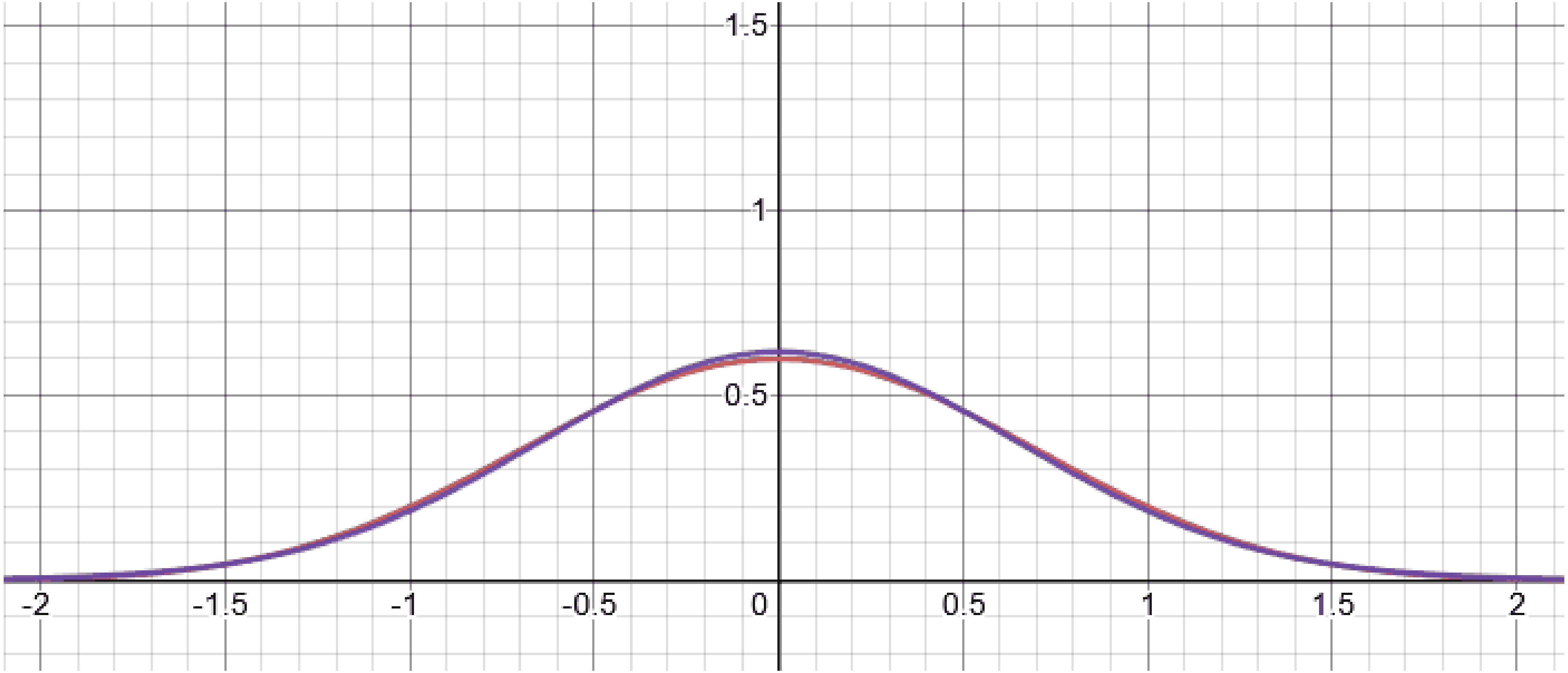
where each $X_i \sim \text{Unif}(0,1)$ and is independent



“Proof by example” -- uniform

$$\boldsymbol{X = X_1 + X_2 + X_3 + X_4 + X_5}$$

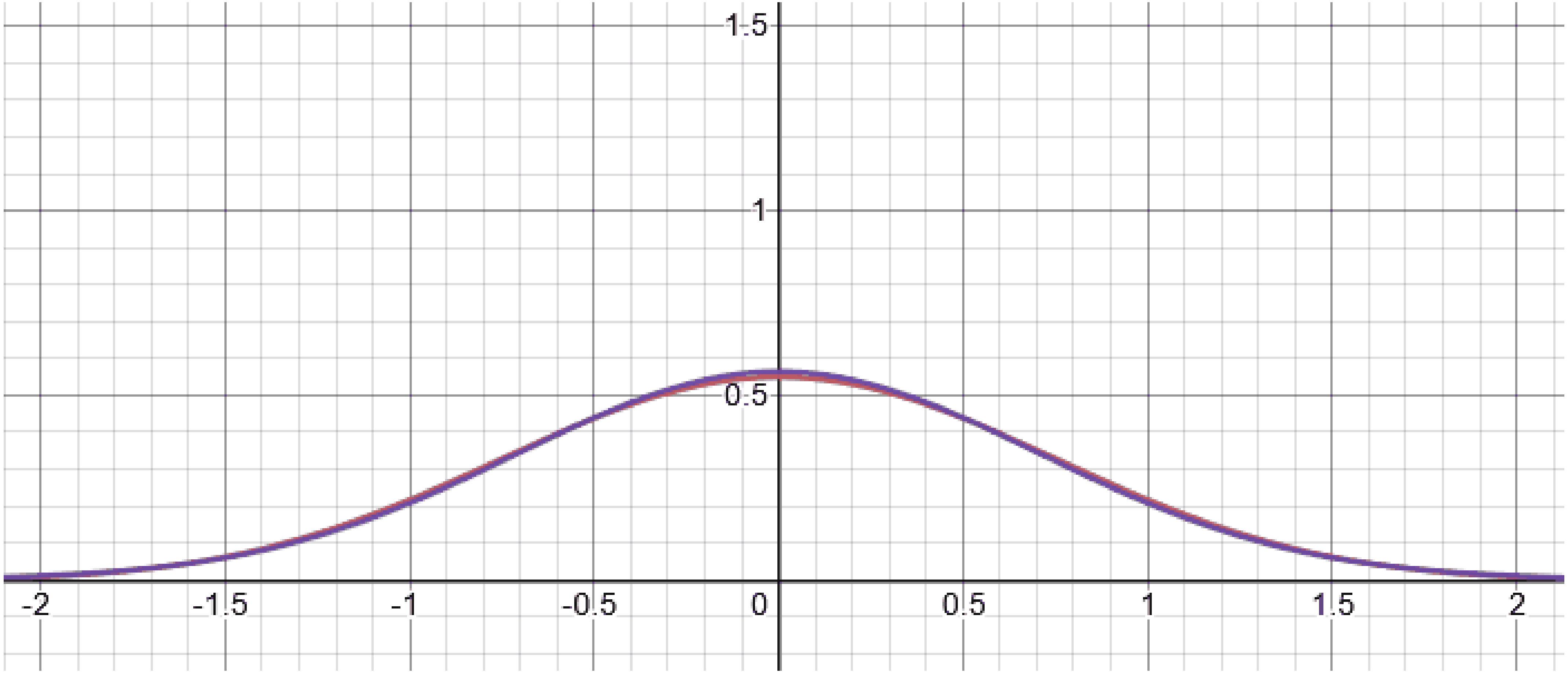
where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent



"Proof by example" -- uniform

$$\boldsymbol{X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6}$$

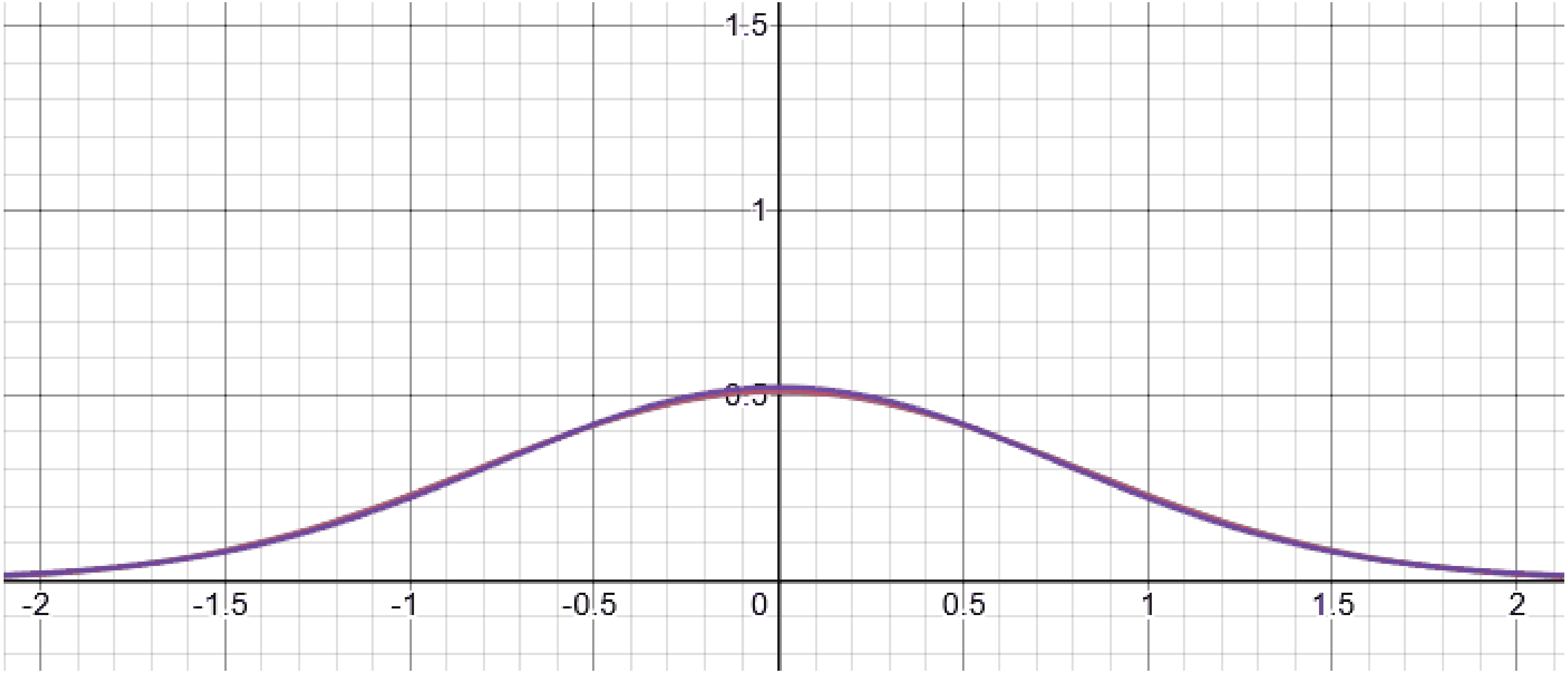
where each $X_i \sim \text{Unif}(0,1)$ and is independent



“Proof by example” -- uniform

$$\boldsymbol{X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}$$

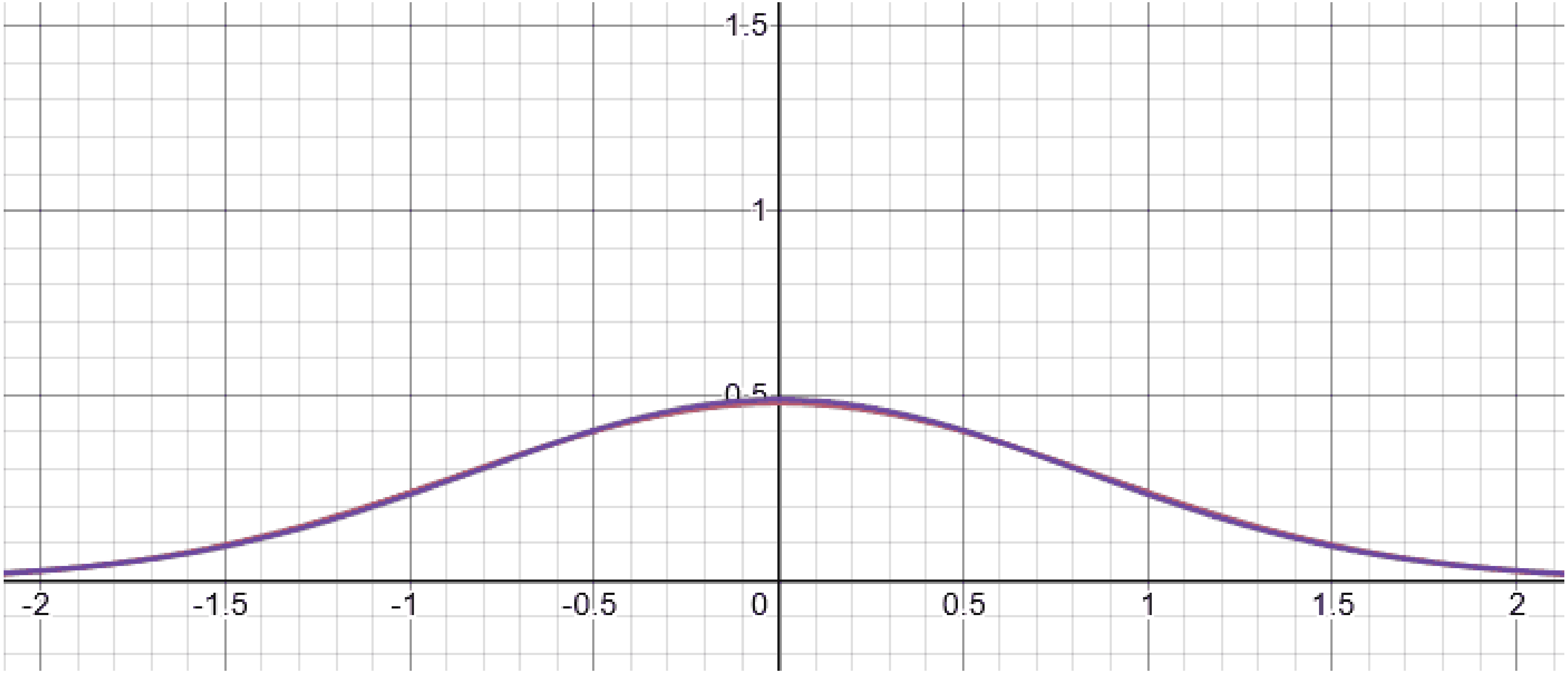
where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent



“Proof by example” -- uniform

$$\boldsymbol{X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8}$$

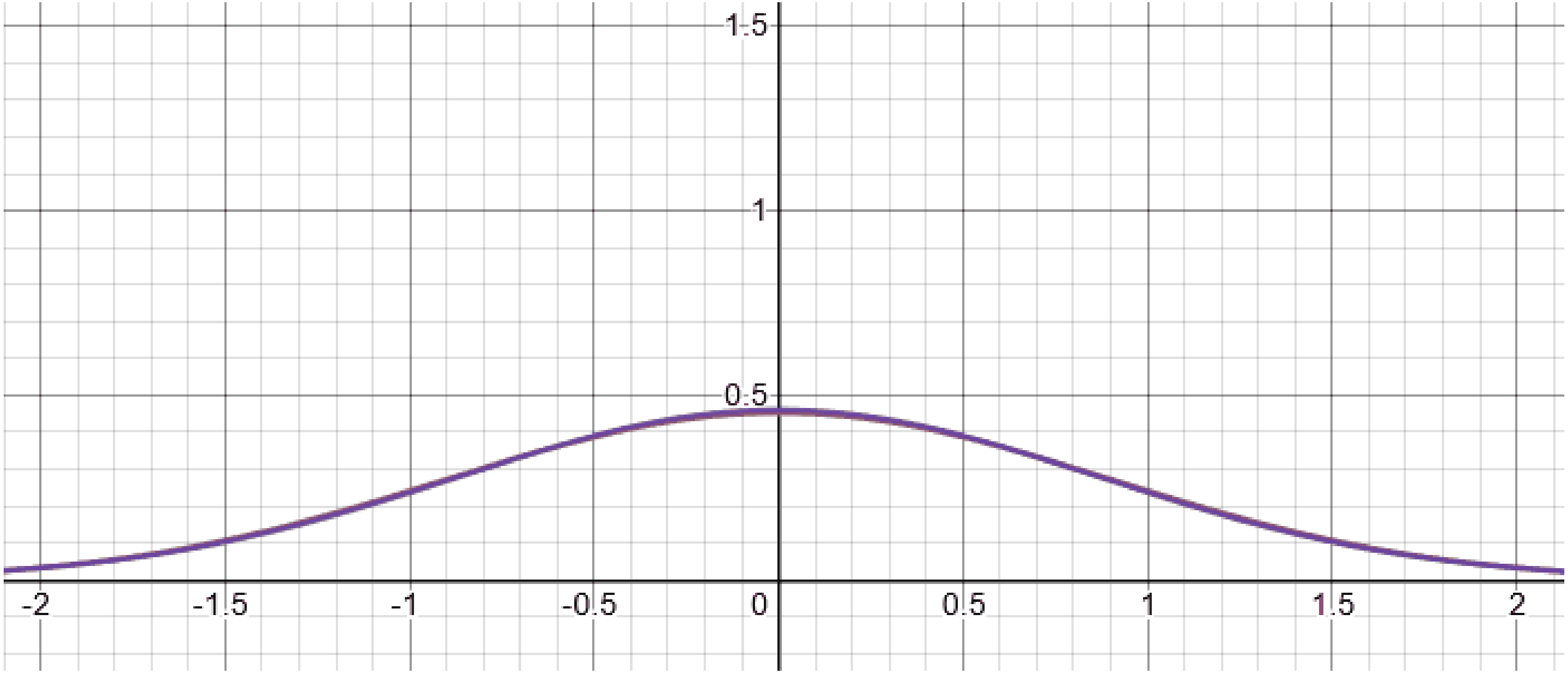
where each $X_i \sim \text{Unif}(0,1)$ and is independent



"Proof by example" -- uniform

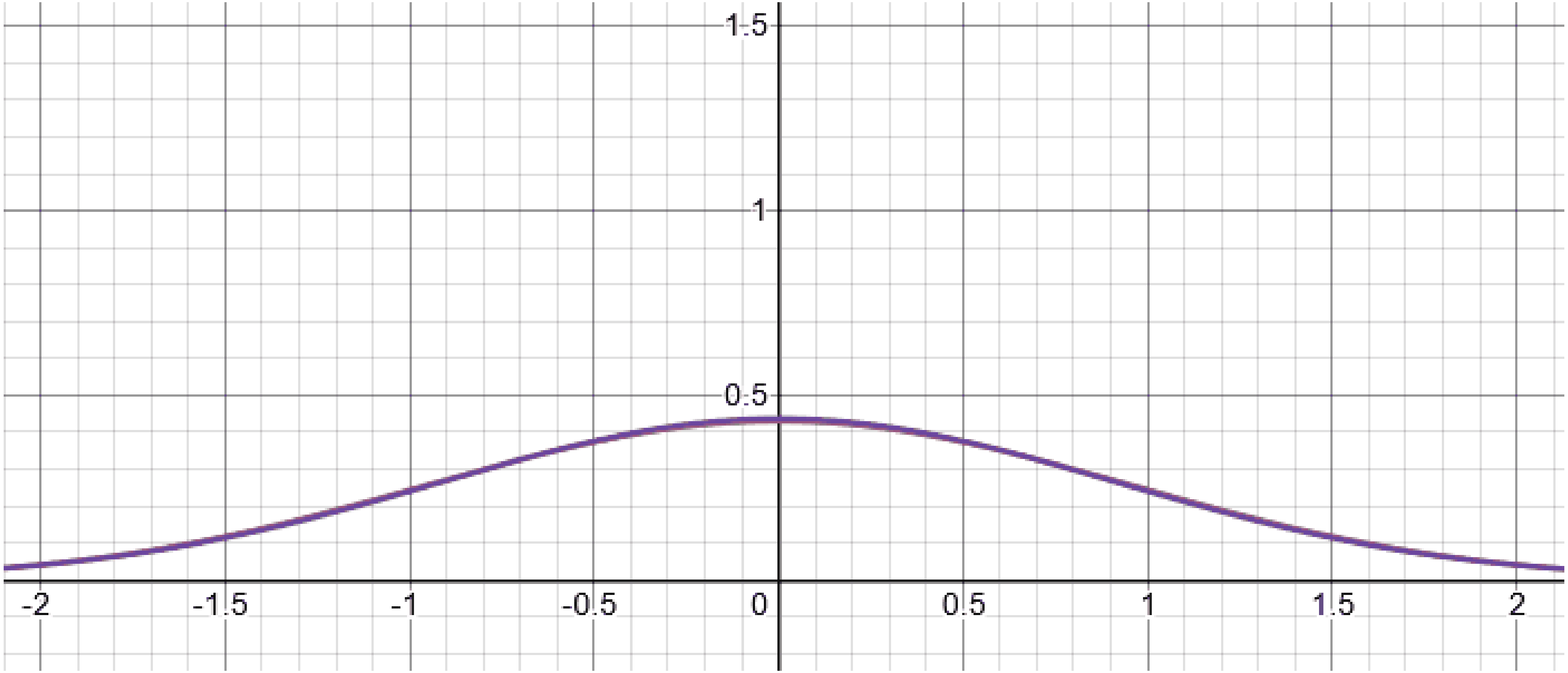
$$\boldsymbol{X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9}$$

where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent

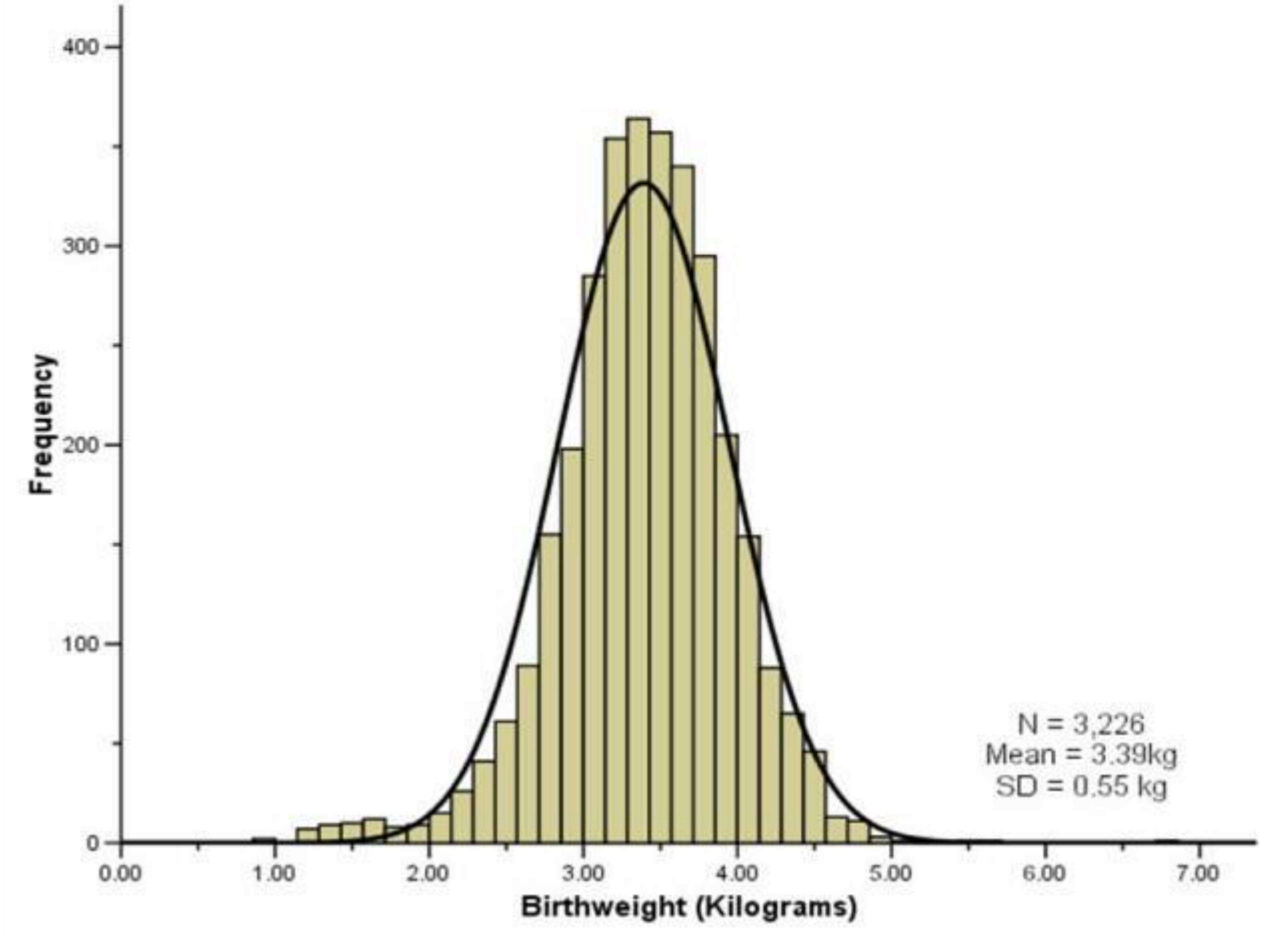


"Proof by example" -- uniform

$$X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10}$$
where each $X_i \sim \mathrm{Unif}(0,1)$ and is independent



"Proof by real-world"



birthweight

A lot of real-world bell-curves can be explained as:

1. The random variable comes from a combination of independent factors.

2. The CLT says the distribution will become like a bell curve.

Theory vs. Practice

> The formal theorem statement is "in the limit"

You might not get exactly a normal distribution for any finite n (e.g. if you sum discrete, the sum is always discrete and will be discontinuous for every finite n.

> In practice, the approximations get very accurate very quickly (at least with a few tricks we'll see soon).

They won't be exact (unless the X_i are normals) but it's close enough to use even with relatively small n.

Using the Central Limit Theorem

Let’s start with the case when we are using CLT to approximate a sum of ***continuous*** i.i.d random variables as normal

Outline of CLT steps

1. Setup the problem (e.g., $X = \sum_{i=1}^{n} X_i$, X_i are i.i.d., and we want $\mathbb{P}(X \leq k)$)
   Write event you are interested in, in terms of sum of random variables.

⚠️ we're going to be adding one more step here when we talk about discrete RVs! ⚠️

2. Apply CLT (e.g., approx X as $Y \sim N(n\mu, n\sigma^2)$ -> $\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k)$
   Approximate sum of RVs as normal with appropriate mean and variance

from here, we're working with a normal distribution, which we've worked with before!

3. Compute probability approximation using Phi table
   > *Standardize* $\left(Z = \frac{N-\mu}{\sigma}\right)$ -> $\mathbb{P}(Y \leq k) = \mathbb{P}\left(\frac{Y-\mu}{\sigma} \leq \frac{k-\mu}{\sigma}\right) = \mathbb{P}\left(Z \leq \frac{k-\mu}{\sigma}\right)$
   > *Write in terms of* $\Phi(z) = \mathbb{P}(Z \leq z)$
   > *Look up in table*

Lightbulbs 💡

You buy lightbulbs that each burn out according to an exponential distribution with parameter of $\lambda = 1.8$ lightbulbs per year.

You buy a 10 pack of (independent) light bulbs. What is the probability that your 10-pack lasts at least 5 years?

Let X_i be the time it takes for lightbulb i to burn out.

Let X be the total time. Estimate $\mathbb{P}(X \geq 5)$.

Lightbulbs 💡

You buy lightbulbs that each burn out according to distribution $\mathrm{Exp}(1.8)$ lightbulbs per year. Estimate the probability your 10-pack (independent) lasts at least 5 years?

1. Setup the Problem:

2. Apply CLT.

3. Compute Probability.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

Lightbulbs 💡

True value (uses a distribution not in our zoo) is ≈ 0.58741

You buy lightbulbs that each burn out according to distribution $\text{Exp}(1.8)$ lightbulbs per year. Estimate the probability your 10-pack (independent) lasts at least 5 years?

1. Setup the Problem: Let X_i be the time it takes for lightbulb i to burn out. $X_i \sim \text{Exp}(1.8)$ and $\mu = \mathbb{E}[X_i] = \frac{1}{1.8}$ and $\sigma^2 = Var(X_i) = \frac{1}{1.8^2}$. Let X be the total time and $X = \sum_{i=1}^{10} X_i$. We are interested in $\mathbb{P}(X \geq 5)$.

2. Apply CLT. Because the X_i's are i.d.d, we can apply CLT and X can be approximated by $Y \sim \mathcal{N}(10 \cdot \frac{1}{1.8}, 10 \cdot \frac{1}{1.8^2})$. $\mathbb{P}(X \geq 5) \approx \mathbb{P}(Y \geq 5)$

3. Compute Probability.

$$\mathbb{P}(Y \geq 5) = \mathbb{P}\left(Z \geq \frac{5 - 10/1.8}{\sqrt{10}/1.8}\right) \quad \textit{standardize}$$

$$\approx \mathbb{P}(Z \geq -0.32) = \mathbb{P}(Z \leq 0.32) \text{(symmetry)}$$

$$= \Phi(0.32) \approx .62552 \quad \textit{plug into z-table}$$

Using the Central Limit Theorem

Now, let's try the case when we are using CLT to approximate a sum of ***discrete*** i.i.d random variables as normal

Factory Widgets

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing at most 940 non-defective widgets?

Factory Widgets 🕰️ - *Exact Answer*

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing <u>at most</u> 940 non-defective widgets?

X is the number of non-defective widgets. Let $X \sim \text{Bin}(1000, .95)$

Our goal: $\mathbb{P}(X \le 940)$?

That's a big summation: $\sum_{k=0}^{940} \binom{1000}{k} (.95)^k \cdot (.05)^{1000-k} \approx .08673$

Factory Widgets 🕰️ - *Exact Answer*

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing *<u>at most</u>* *940 non-defective widgets?*

X is the number of non-defective widgets. Let $X \sim \text{Bin}(1000, .95)$

Our goal: $\mathbb{P}(X \leq 940)$?

That's a big summation: $\sum_{k=0}^{940} \binom{1000}{k} (.95)^k \cdot (.05)^{1000-k} \approx .08673$

What does the CLT give? Binomial is sum of i.i.d bernoullis -> can use CLT!

Factory Widgets 🕰️ - *CLT*

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing <u>at most</u> 940 non-defective widgets?

1. Setup the Problem:

2. Apply CLT.

3. Compute Probability.

Factory Widgets 🕰️ - *CLT*

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing <u>*at most*</u> *940 non-defective widgets?*

1. Setup the Problem: X is the number of non-defective widgets.
$X = \sum_{i=1}^{1000} X_i$ where X_i is 1 if the i'th widget is non-defective. ***Goal***: $\mathbb{P}(X \leq 940)$

2. Apply CLT. X is sum of i.i.d RVs each with $\mu = \mathbb{E}[X_i] = p = .95$ and $\text{Var}(X_i) = p(1-p) = .0475$, we can approximate X with $Y \sim \mathcal{N}(1000 \cdot 0.95, 1000 \cdot 0.0475)$.
So, $\mathbb{P}(X \leq 940) \approx \mathbb{P}(Y \leq 940)$

3. Compute Probability.

$$\mathbb{P}(Y \leq 940) = \mathbb{P}\left(Z \leq \frac{940 - 1000 \cdot 0.95}{\sqrt{1000 \cdot 0.0475}}\right) \qquad \textit{standardize}$$

$$\approx \Phi(-1.45) = 1 - \Phi(1.45) \qquad \textit{write in terms of } \boldsymbol{\Phi}$$

$$\approx 1 - .92647 = .07353. \qquad \textit{plug into z-table}$$

Factory Widgets 🕰️ - *CLT*

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing <u>*at most*</u> *940 non-defective widgets?*

1. Setup the Problem: X is the number of non-defective widgets.
$X = \sum_{i=1}^{1000} X_i$ where X_i is 1 if the i'th widget is non-defective. ***Goal***: $\mathbb{P}(X \leq 940)$

2. Apply CLT. X is sum of i.i.d RVs each with $\mu = \mathbb{E}[X_i] = p = .95$ and $\text{Var}(X_i) = p(1-p) = .0475$, we can approximate X with $Y \sim \mathcal{N}(1000 \cdot 0.95, 1000 \cdot 0.0475)$.
So, $\mathbb{P}(X \leq 940) \approx \mathbb{P}(Y \leq 940)$

3. Compute Probability.

$$\mathbb{P}(Y \leq 940) = \mathbb{P}\left(Z \leq \frac{940 - 1000 \cdot 0.95}{\sqrt{1000 \cdot 0.0475}}\right)$$ *standardize*

$$\approx \Phi(-1.45) = 1 - \Phi(1.45)$$ *write in terms of* $\boldsymbol{\Phi}$

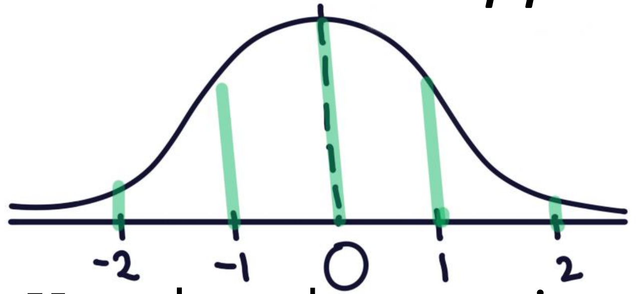
$$\approx 1 - .92647 = .07353.$$ *plug into z-table*

The exact probability is .08673. We're off by ~1.3%!

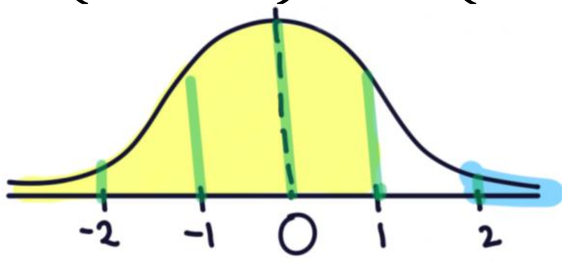
There's are some problems ☹

When approximating a discrete distribution like binomial with a continuous normal distribution, there are some problems that arise!

> $\mathbb{P}(X = 2) > 1$ (we can use the binomial PMF).
> But, ***when we approximate to the normal, continuous***, Y, $\mathbb{P}(Y = 2) = 0$



> X only takes on integers, so $\mathbb{P}(X \leq 1) + \mathbb{P}(X \geq 2) = 1$.
> But, ***when we approximate to the normal, continuous***, Y,
> $\mathbb{P}(Y \leq 1) + \mathbb{P}(Y \geq 2) < 1$



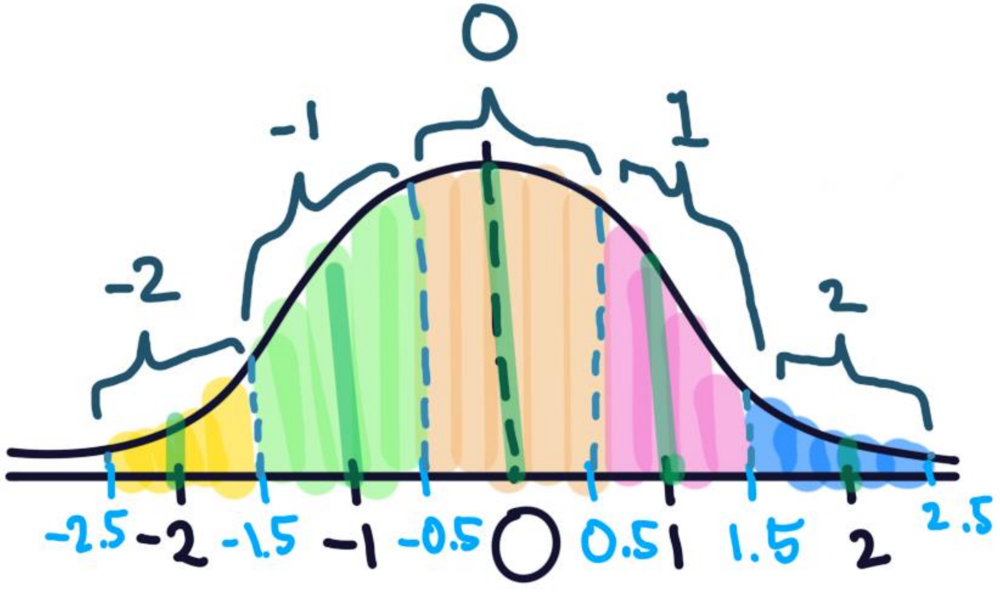
Continuity Correction

The binomial distribution is **discrete**, but the normal is **continuous**.

Let's correct for that (called a "*continuity correction*")

⭐ Assign each value in the discrete range to a continuous interval

Here the support of X *is* $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$



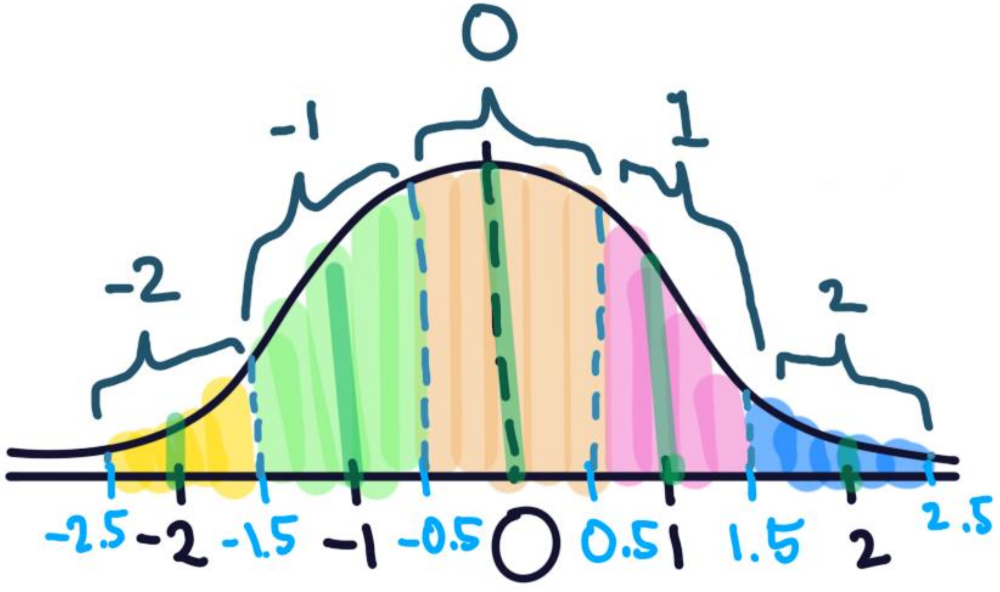
Continuity Correction

The binomial distribution is **discrete**, but the normal is **continuous**.

Let's correct for that (called a "*continuity correction*")

⭐ Assign each value in the discrete range to a continuous interval

Here the support of X is $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$



e.g.,

$\mathbb{P}(X = 2)$ ->

$\mathbb{P}(X \geq 2)$ ->

$\mathbb{P}(X > 1)$ ->

$\mathbb{P}(X \leq 1)$ ->

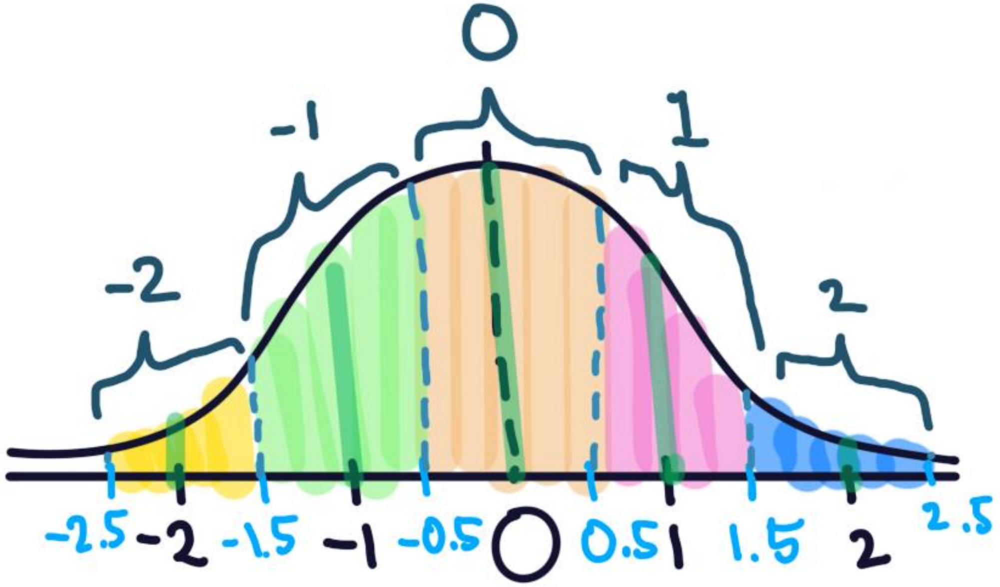
Continuity Correction

The binomial distribution is **discrete**, but the normal is **continuous**.

Let's correct for that (called a "*continuity correction*")

⭐ Assign each value in the discrete range to a continuous interval

Here the support of X is $\{\dots, -2, -1, 0, 1, 2, \dots\}$



e.g.,

$\mathbb{P}(X = 2)$ -> $\mathbb{P}(1.5 \le X \le 2.5)$

$\mathbb{P}(X \ge 2)$ -> $\mathbb{P}(X \ge 1.5)$

$\mathbb{P}(X > 1)$ -> $\mathbb{P}(X \ge 1.5)$

$\mathbb{P}(X \le 1)$ -> $\mathbb{P}(X \le 1.5)$

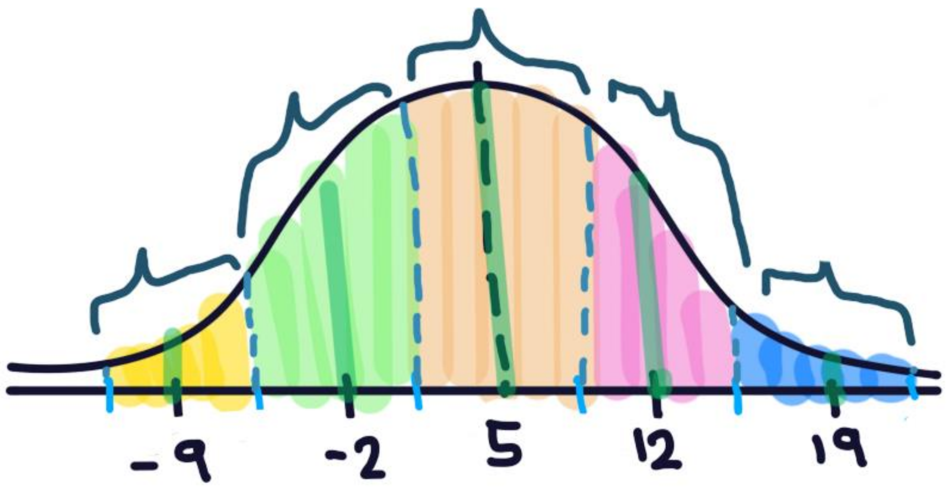
Continuity Correction

The binomial distribution is **discrete**, but the normal is **continuous**.

Let's correct for that (called a "*continuity correction*")

⭐ Assign each value in the discrete range to a continuous interval

Here the support of X is $\{\dots, -2, 5, 12, 19 \dots\}$



e.g.,

$\mathbb{P}(X = -2)$ ->

$\mathbb{P}(X \geq 5)$ ->

$\mathbb{P}(X < 12)$ ->

$\mathbb{P}(X \geq 0)$ ->

Fill out the poll everywhere:
pollev.com/cse312

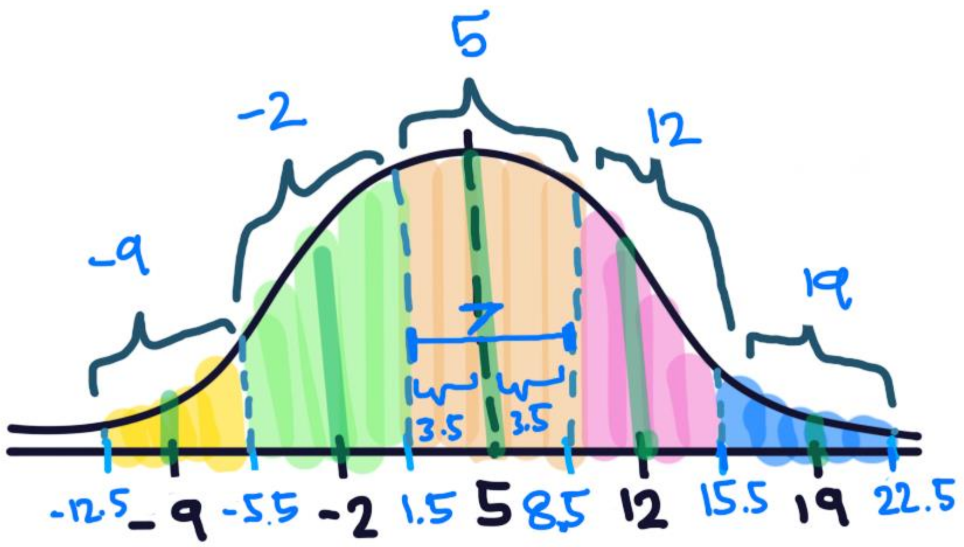
Continuity Correction

The binomial distribution is **discrete**, but the normal is **continuous**.

Let's correct for that (called a "*continuity correction*")

⭐ Assign each value in the discrete range to a continuous interval

Here the support of X is $\{\ldots, -2, 5, 12, 19 \ldots\}$



e.g.,

$\mathbb{P}(X = -2)$ ->

$\mathbb{P}(X \geq 5)$ ->

$\mathbb{P}(X < 12)$ ->

$\mathbb{P}(X \geq 0)$ ->

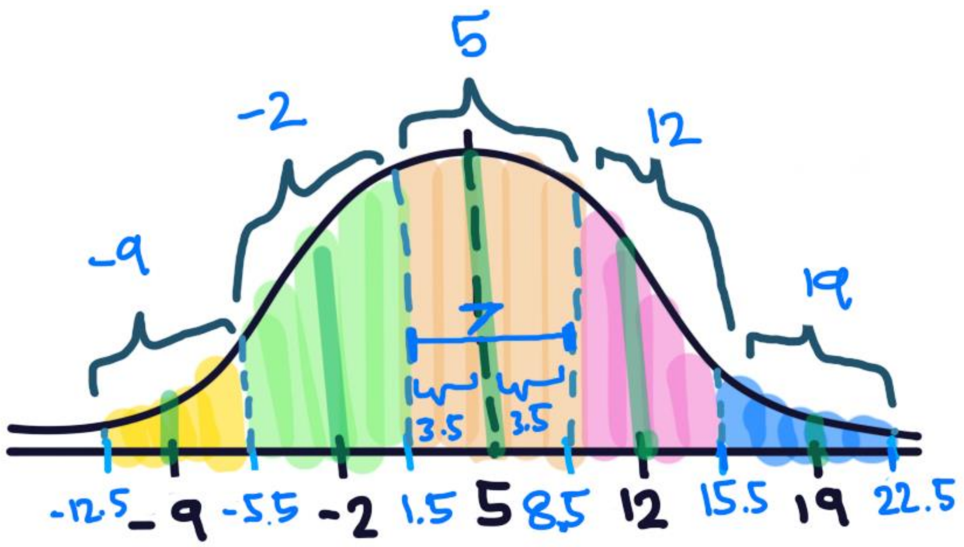
Continuity Correction

The binomial distribution is **discrete**, but the normal is **continuous**.

Let's correct for that (called a "*continuity correction*")

⭐ Assign each value in the discrete range to a continuous interval

Here the support of X is $\{\ldots, -2, 5, 12, 19 \ldots\}$



e.g.,

$\mathbb{P}(X = -2)$ -> $\mathbb{P}(-5.5 \leq X \leq 1.5)$

$\mathbb{P}(X \geq 5)$ -> $\mathbb{P}(X \geq 1.5)$

$\mathbb{P}(X < 12)$ -> $\mathbb{P}(X \leq 8.5)$

$\mathbb{P}(X \geq 0)$ -> $\mathbb{P}(X \geq 1.5)$

Outline of CLT steps

1. Setup the problem (e.g., $X = \sum_{i=1}^{n} X_i$, X_i are i.i.d., and we want $\mathbb{P}(X \leq k)$)
   Write event you are interested in, in terms of sum of random variables.

   ⭐ Apply *continuity correction here* if RVs are discrete.

2. Apply CLT (e.g., approx X as $Y \sim N(n\mu, n\sigma^2)$ -> $\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k)$
   Approximate sum of RVs as normal with appropriate mean and variance

from here, we're working with a normal distribution, which we've worked with before!

3. Compute probability approximation using Phi table
   > *Standardize* $(Z = \frac{N-\mu}{\sigma})$ -> $\mathbb{P}(Y \leq k) = \mathbb{P}\left(\frac{Y-\mu}{\sigma} \leq \frac{k-\mu}{\sigma}\right) = \mathbb{P}\left(Z \leq \frac{k-\mu}{\sigma}\right)$

   > *Write in terms of* $\Phi(z) = \mathbb{P}(Z \leq z)$

   > *Look up in table*

Factory Widgets - *CLT* **with continuity correction**

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%. *Your factory will produce 1000 (possibly defective) widgets. What is the probability of producing* *__at most__* *940 non-defective widgets?*

1. Setup the Problem: X is the number of non-defective widgets.
$X = \sum_{i=1}^{1000} X_i$ where X_i is 1 if the i'th widget is non-defective. We want to find $\mathbb{P}(X \le 940)$.

Because $\boldsymbol{X}$ is discrete, we use continuity correction: $\mathbb{P}(\boldsymbol{X} \le \mathbf{940}) = \mathbb{P}(\boldsymbol{X} \le \mathbf{940.5})$

2. Apply CLT. X is sum of i.i.d RVs each with $\mu = \mathbb{E}[X_i] = p = .95$ and $\mathrm{Var}(X_i) = p(1-p) = .0475$, we can approximate X with $Y \sim \mathcal{N}(1000 \cdot 0.95, 1000 \cdot 0.0475)$.
So, $\mathbb{P}(X \le 940.5) \approx \mathbb{P}(Y \le 940.5)$

3. Compute Probability.

$$\mathbb{P}(Y \le 940.5) = \mathbb{P}\left(Z \le \frac{940.5 - 1000 \cdot 0.95}{\sqrt{1000 \cdot 0.0475}}\right)$$ *standardize*

$$\approx \Phi(-1.38) = 1 - \Phi(1.38)$$ *write in terms of* $\boldsymbol{\Phi}$

$$\approx 1 - .91621 = .08379.$$ *plug into z-table*

The exact probability is .08673. Still an approximation, but very close now! :D

Sometimes, we are solving for something that is not the probability

For example, we might be looking for the value of the expectation, variance, or some other parameter that makes the probability be a certain value.

In this case, we will still follow the exact same approach! But will end up doing a reverse z-table lookup at the end.

Lying about the time…….

Dr. Evelyn, is studying the amount of time her students spend on a specific assignment. The time the i'th student spends on the assignment is a random variable X_i with a mean of μ=4 hours and a standard deviation of $\sigma = 1.5$ hours. Dr. Evelyn wants to find out how many students n she needs to survey so that the probability that the average time spent on the assignment by the students is within 30 minutes (0.5 hours) of the mean is at least 95%.

Lying about the time…….

Dr. Evelyn, is studying the amount of time her students spend on a specific assignment. The time the i’th student spends on the assignment is a random variable X_i with a mean of μ=4 hours and a standard deviation of $\sigma = 1.5$ hours. Dr. Evelyn wants to find out **how many students $\boldsymbol{n}$ she needs to survey so that the probability that the average time spent on the assignment by the students is within 30 minutes (0.5 hours) of the mean is at least 95%.**

Still follow the same approach: