Disclaimer: Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

Content

- Gaussian Distribution ("Normal" Distribution)
- 68/85/99 rule (Empirical Rule)
- Z-Score
- Percent Point Function (PPF)
- Standard Normal Distribution
  - Standardization

Gaussian Distribution

Imagine you are a Data Scientist and are collecting data about the heights of a college's students.

In fact, let's represent these values with a random variable X.

> **Q1. What kind of random variable would X be? Discrete or Continuous?**

Continuous.

> **Q2. Do you expect any kind of pattern in the height values?**

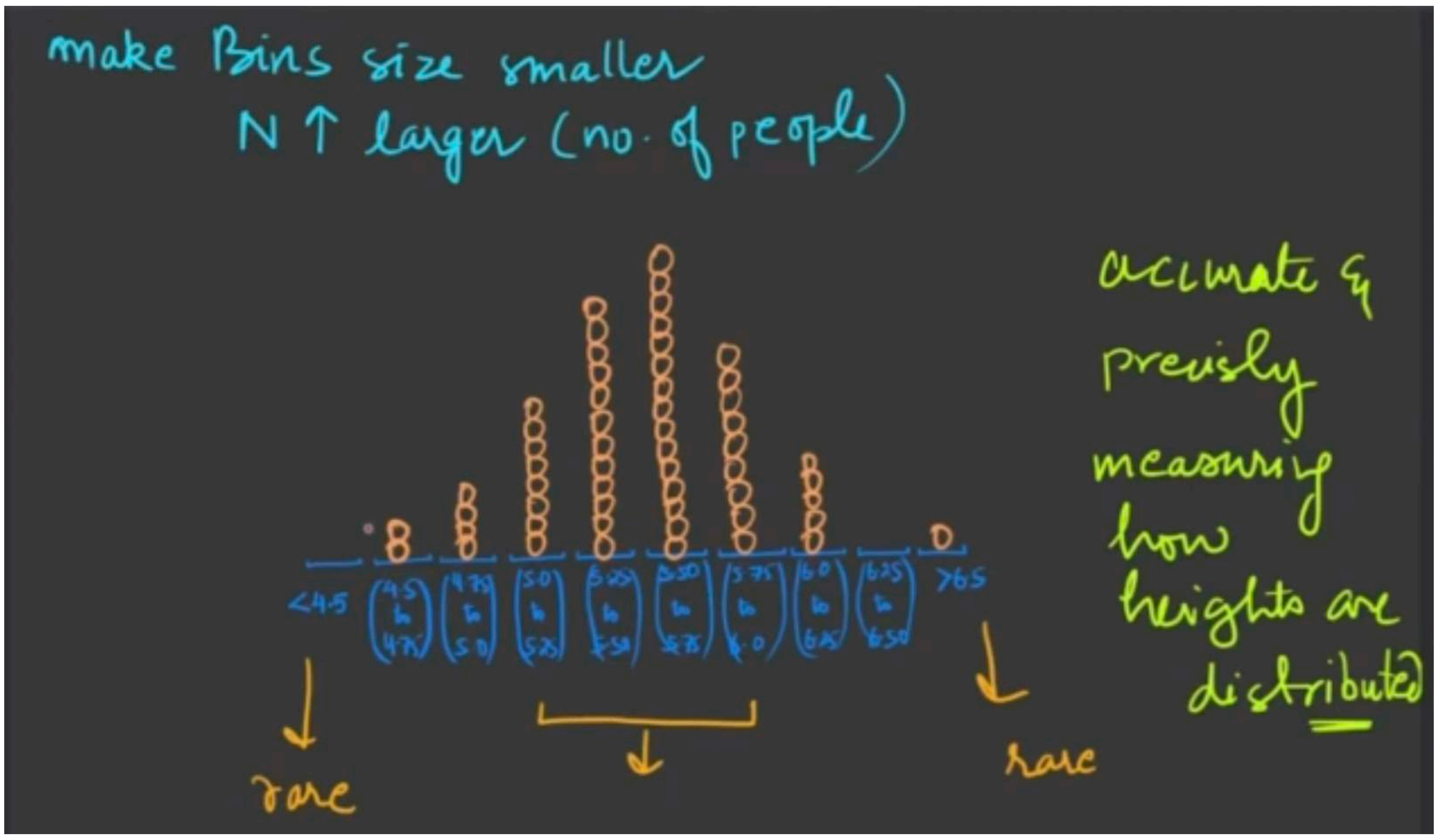
Assuming that this college is in India, we can assert

- On average, the height values should be something like 5.5 feet to 5.75 feet (These values are not important)
- Though, there would also be people that are shorter or taller than this range, we know that they will be comparatively less in number.

So, if we want to plot a histogram to visualize the frequency/count of these height values,

Would you agree that we get a plot that looks something like as shown below?

- Since X is continuous, we will have to count the frequency for values belonging to a certain range (bin the data values)
- Suppose we have 2 groups of people having height 5 feet and 5.1 feet
  - This does not mean there can't be an individual with height 5.005 feet
  - Hence in order to include everyone, we bin the data while plotting.

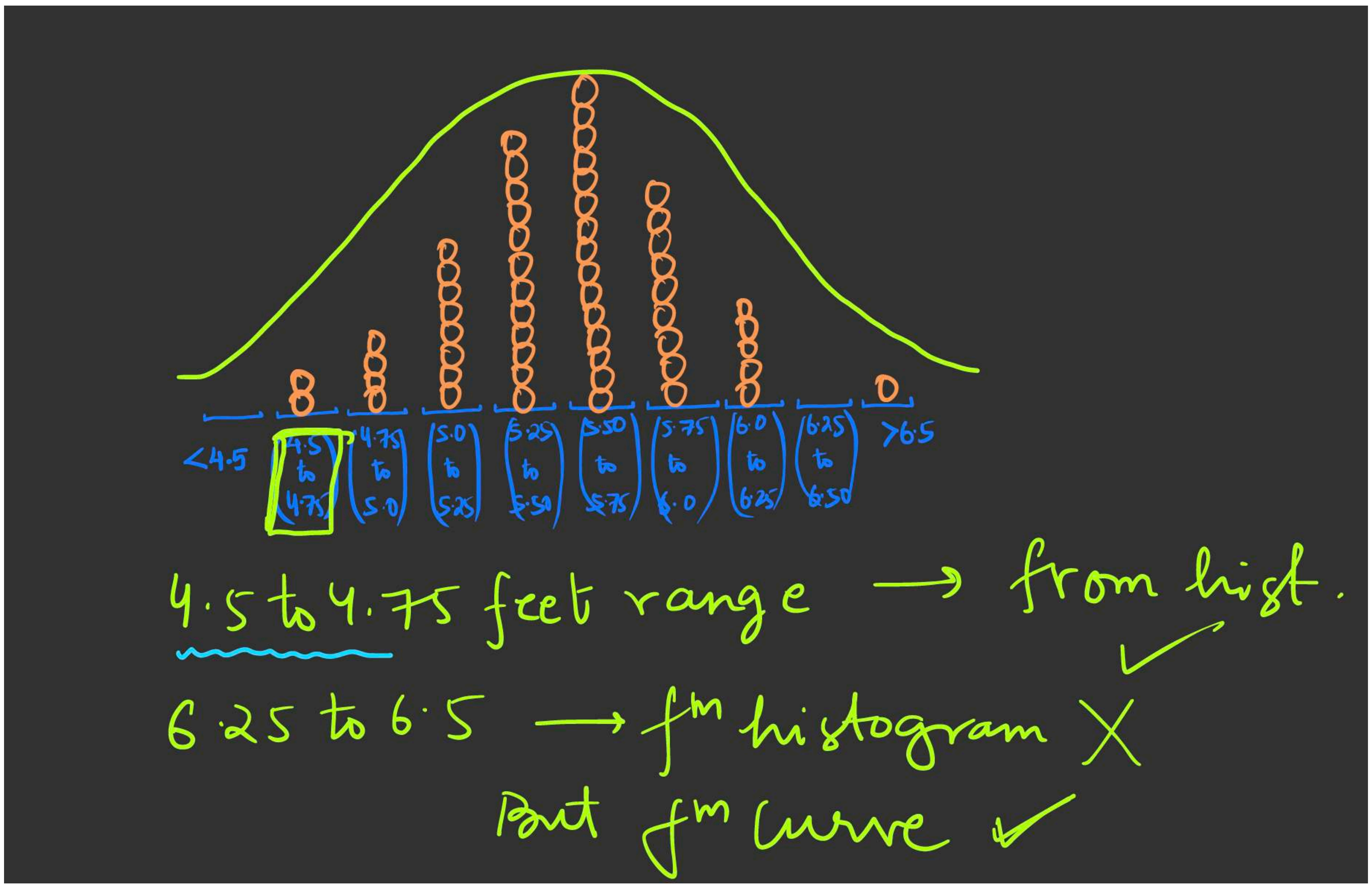


> **3. What Can we estimate this histogram with a curve as shown in the image**

- It covers the same information about the frequency of people belonging to each bin/range.

There are also some additional advantages to estimating using such a curve:

1. **Drawing insights about unseen data**
   - In the histogram, notice that we did not encounter anyone whose height belonged to the bin 6.25 to 6.5 feet. This does not mean that you will not find anyone belonging to that range.
   - It is a chance factor, that no one of that height was present in this college.
   - From the histogram, you won't be able to calculate the probability of a person having that height. But using this curve, you can calculate.

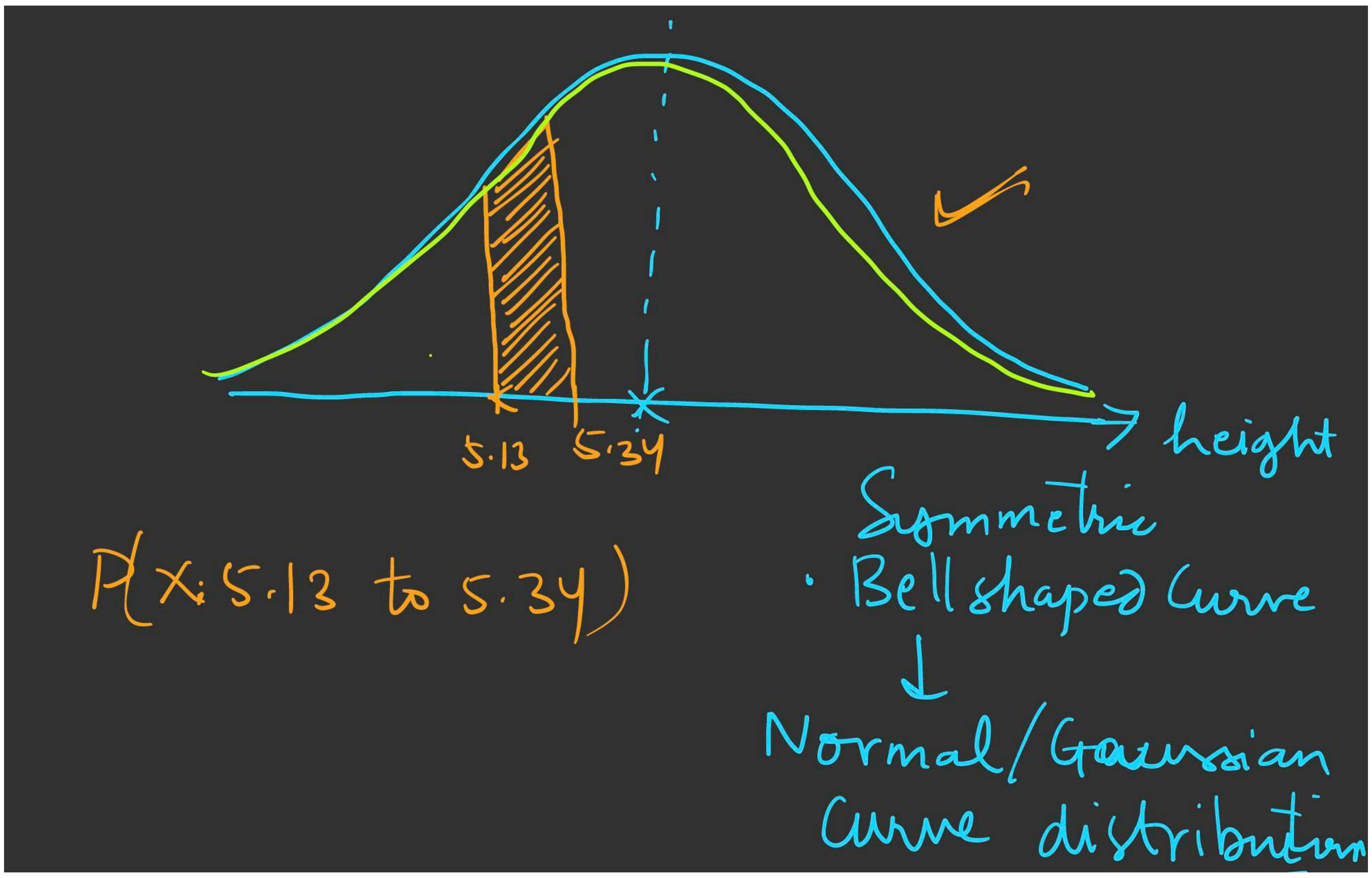


2. **The Plot is not limited by bin size**
    - Using the histogram, we could only calculate the probability of people having height within the given bin size, i.e. P(X: 5.25 to 5.5), P(X: 5.75 to 6.0), etc.
    - However, using the curve, we can find the probability of people having a height in a bin of any size: P(X: 5.13 to 5.34), etc

3. **Note: that the curve is symmetric, while the histogram was not**

This type of symmetric bell-shaped probability distribution curve is known as **Gaussian / Normal / Bell shaped Curve Distribution**

In fact, this is a very special distribution that has its own set of properties, that can be leveraged to draw insights and observations.



Let's look at one of those properties.

68/95/99 rule (Empirical rule)

Once you collected the data, you observed the following:

- Mean / Average = 65 inches
    - (Note: 1 feet = 12 inches)

- Standard Deviation = 2.5 inches

We know that the point corresponding to the peak here would be the mean, i.e. 65

> **Q1. What did the standard deviation represent?**

We saw that SD is a measure of the spread of the data

- A higher SD means a higher spread

Since SD measures the spread.

Let's mark the points that are away from the mean by a SD

- One SD to the right would be: $65 + 2.5 = 67.5$
- One SD to the left would be: $65 - 2.5 = 62.5$

Notice, on the region between 1 standard deviation of the mean, i.e. between 62.5 & 67.5 (shaded yellow)

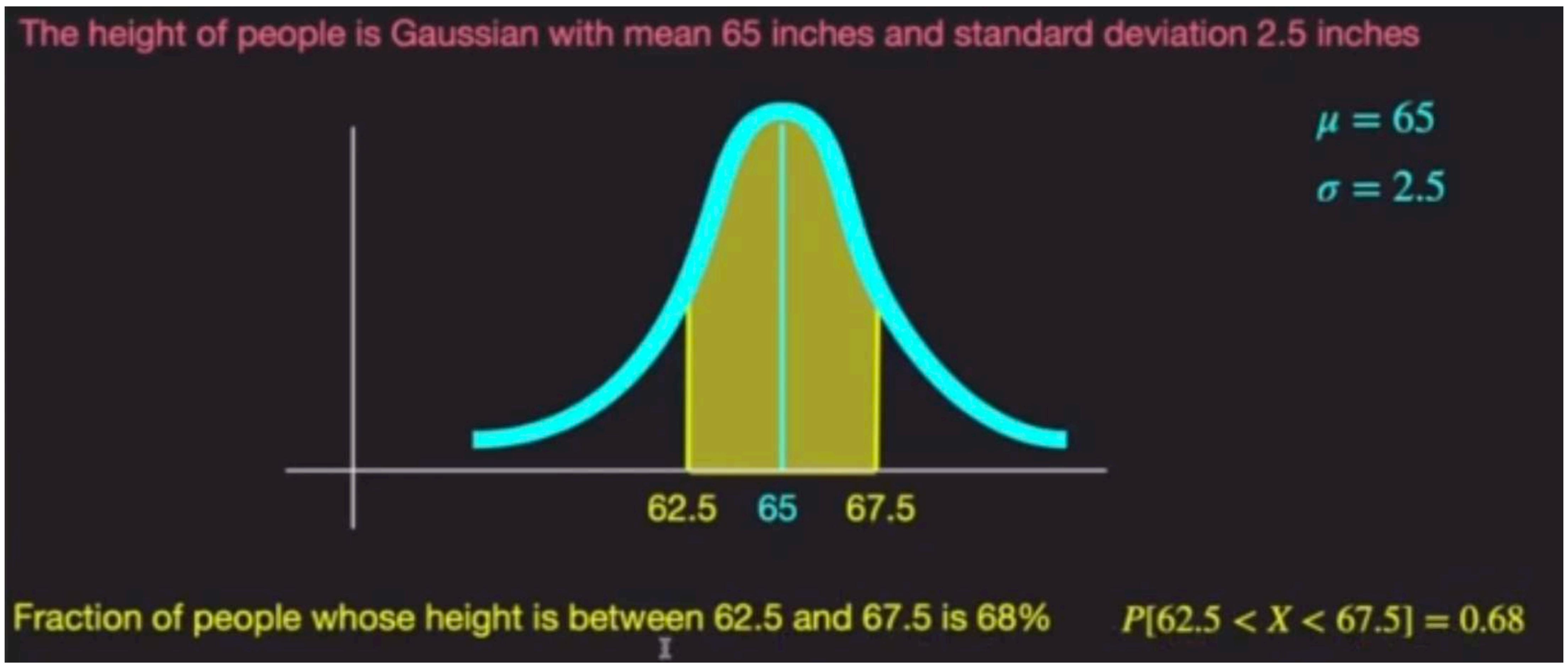
> **Q2. How much of the population do you think would be within this region?**

Since this region is around the mean, frequency of people is high.

So the fraction should also be high, but how high?

Gaussian distributions have a property that the fraction of population in this region is **exactly 68%**

- This means that the fraction of people whose height is between [62.5, 67.5] is **68%**, i.e. $P(62.5 < X < 67.5) = 0.68$
- We can also write this as: $P(\mu - \sigma < X < \mu + \sigma) = 0.68$



The height of people is Gaussian with mean 65 inches and standard deviation 2.5 inches

$\mu = 65$

$\sigma = 2.5$

Fraction of people whose height is between 62.5 and 67.5 is 68%

$P[62.5 < X < 67.5] = 0.68$

> **Two standard deviations away:** Let's move farther away from the mean, by another standard deviation.

Therefore the points become:

- $65 - 2(2.5) = 60$ on the left
- $65 + 2(2.5) = 70$ on the right

The region enclosed between $[\mu - 2\sigma, \mu + 2\sigma]$ (shaded green) comprises **95%** of the entire population, i.e. $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$

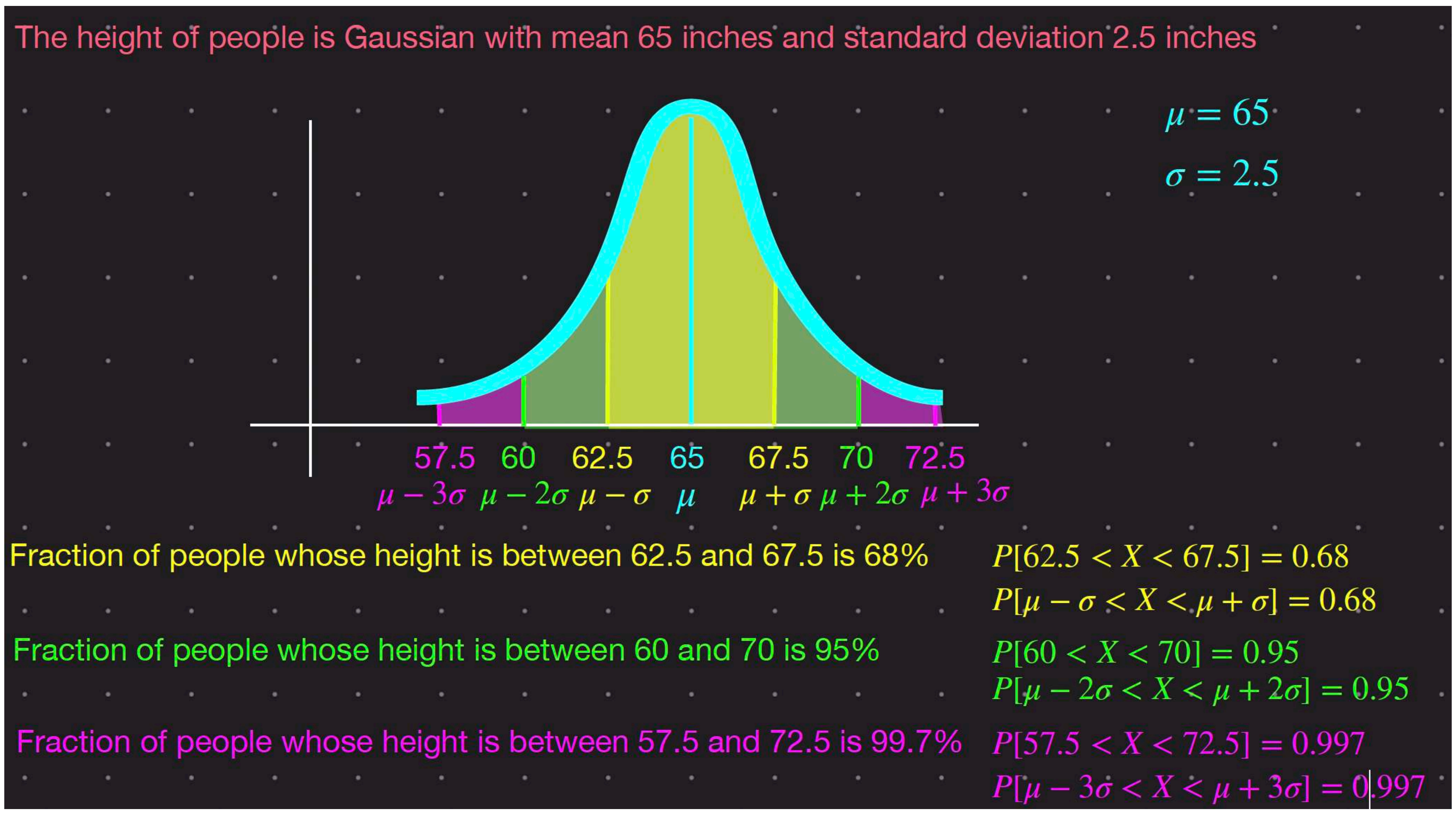
Similarly,

> **Three standard deviations away:** Let's move farther away from the mean, by another standard deviation.

Therefore the points become:

- $65 - 3(2.5) = 57.5$ on the left
- $65 + 3(2.5) = 72.5$ on the right

The region enclosed between $[\mu - 3\sigma, \mu + 3\sigma]$ (shaded purple) comprises **99.7%** of the entire population, i.e. $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$



The height of people is Gaussian with mean 65 inches and standard deviation 2.5 inches

$\mu = 65$

$\sigma = 2.5$

Fraction of people whose height is between 62.5 and 67.5 is 68%

$P[62.5 < X < 67.5] = 0.68$

$P[\mu - \sigma < X < \mu + \sigma] = 0.68$

Fraction of people whose height is between 60 and 70 is 95%

$P[60 < X < 70] = 0.95$

$P[\mu - 2\sigma < X < \mu + 2\sigma] = 0.95$

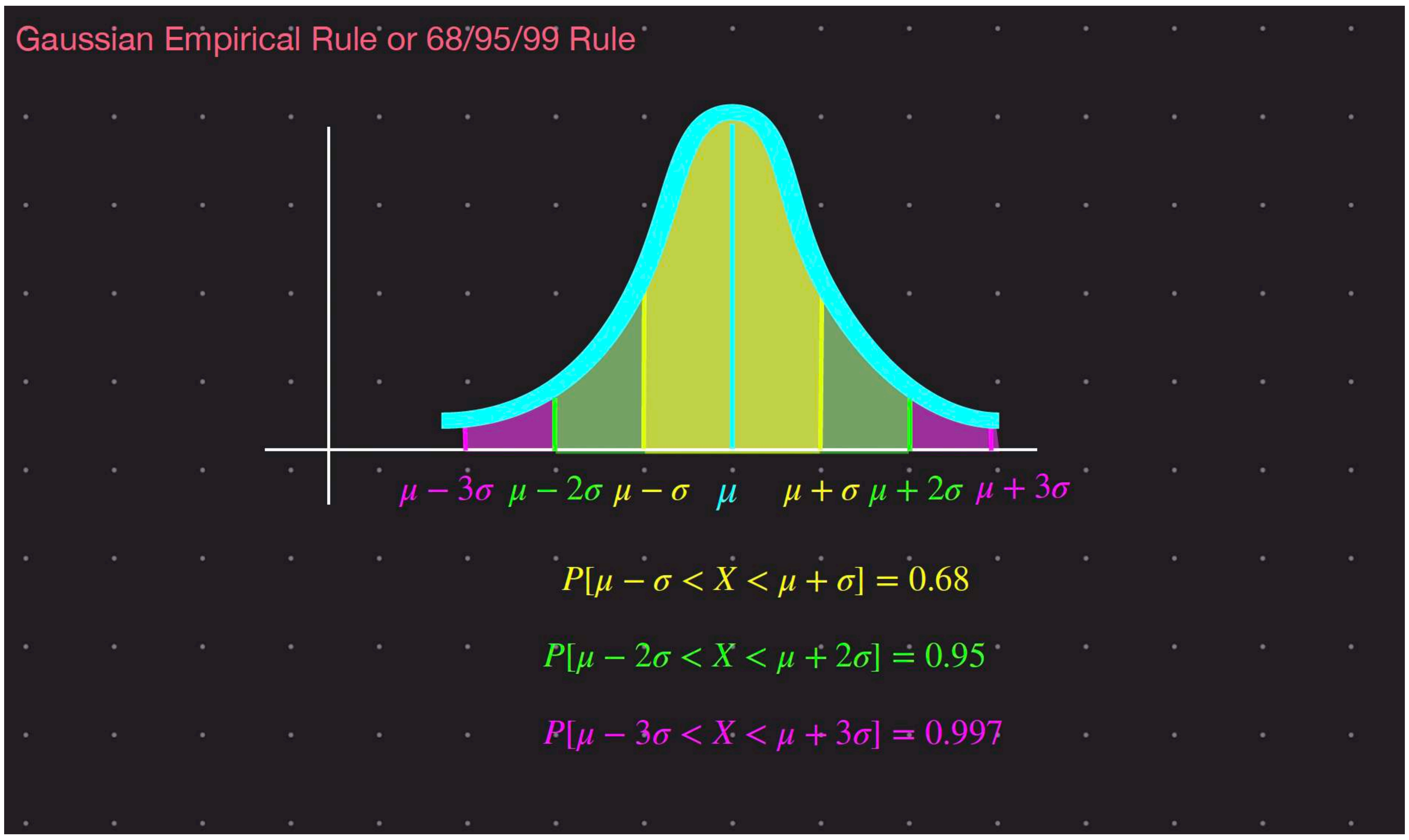
Fraction of people whose height is between 57.5 and 72.5 is 99.7%

$P[57.5 < X < 72.5] = 0.997$

$P[\mu - 3\sigma < X < \mu + 3\sigma] = 0.997$

There is a fancy name to this property.

It is known as **Gaussian Empirical Rule**, more popularly known as the **68/95/99 Rule**



Gaussian Empirical Rule or 68/95/99 Rule

$P[\mu - \sigma < X < \mu + \sigma] = 0.68$

$P[\mu - 2\sigma < X < \mu + 2\sigma] = 0.95$

$P[\mu - 3\sigma < X < \mu + 3\sigma] = 0.997$

Z-score

> **Q1. Is this empirical rule enough for us to find the fraction of population between any random value?**

- Through the Empirical rule, we are only restricted to these numbers (68/95/99.7)
- This means that we won't be able to utilize this rule if we want to know the fraction of people shorter than 66, 69 or some other random value.
- This rule can only give us information if number is within the interval enclosed by points 1, 2 or 3 SD away.

> **Consider the following question.**

```
Suppose The height of people is Gaussian with a mean of 65 inches and a standard deviation of 2.5 inches.
What fraction of people are shorter than 69.1 inches?
```

Solution Approach:

First things first, let's try to figure out some relation between the point 69.1 and the mean & std dev.

We know that

- 70 inches is 2 SD far away from the mean and
    - 70 = 65 + 2(SD)

- 67.5 is 1 SD far away from the mean
  - 67.5 = 65 + 1(SD)

Therefore, 69.1 will be more than 1 SD and less than 2 SD.

Let's say 69.1 is "z" standard deviation away from the mean

Hence, we can write: $69.1 = 65 + z(2.5)$

$z = \frac{69.1-65}{2.5} = 1.64$

Conclusion:

```
69.1 inches is 1.64 Standard Deviation away from the mean
```

This 1.64 is known as the **Z-score** of the point 69.1

It represents the distance between a data point and the mean using standard deviations.

Formula

We can write in this way

$z = \frac{x-\mu}{\sigma}$ This is the formula of Z-score

where

- x = number for which we need to find z-score
- μ = mean value
- σ = standard deviation.

> **Q1. What can be the possible values of this z-score value?**

It can be both positive and negative.

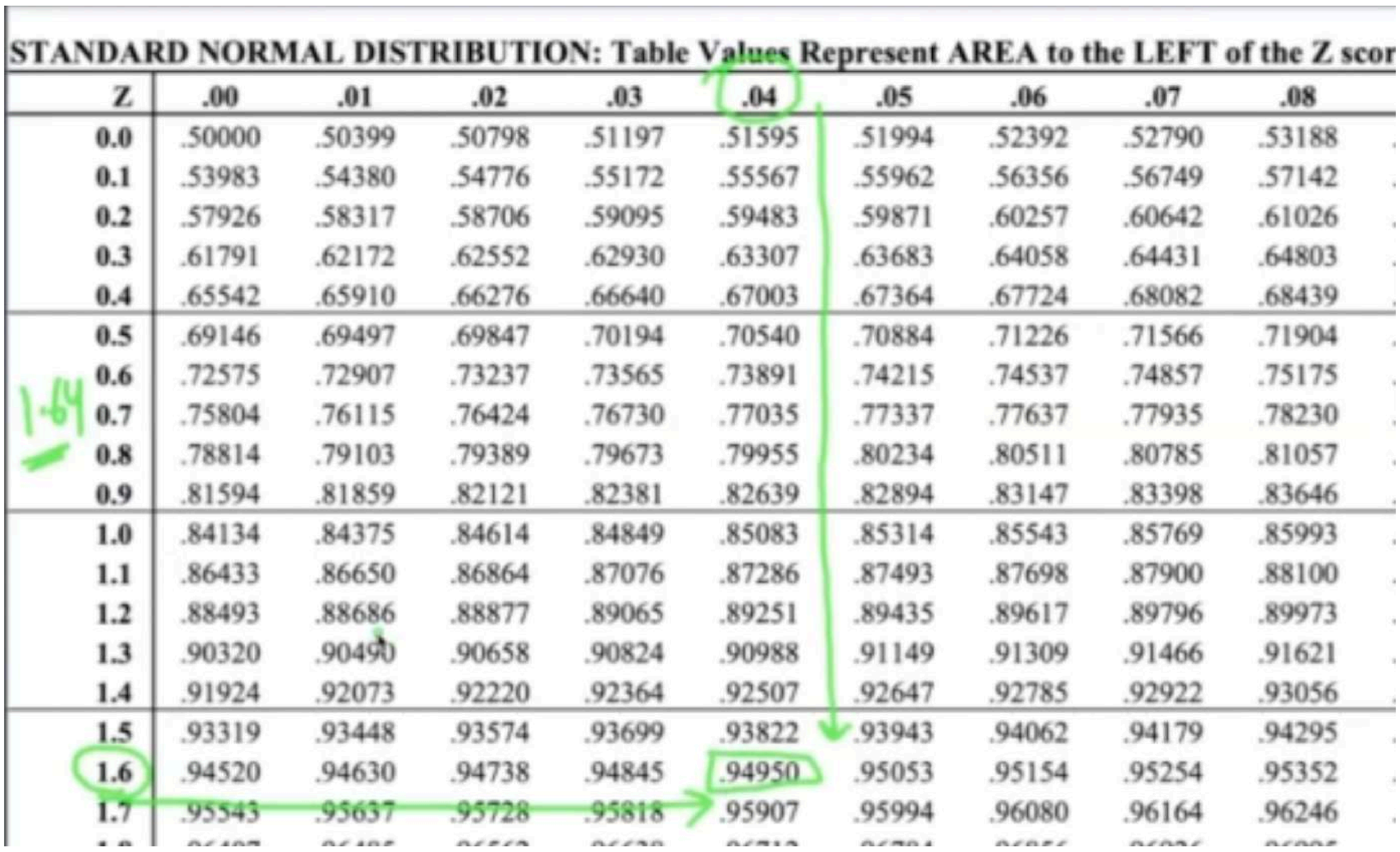
- A z-score = 0, represents that the data point is the **mean** itself
- A positive z-score represents that the data point is to the **right of mean**.
- A negative z-score represents the data point is to the **left of mean**.

Coming back to the question at hand.

> **Q2. How can we find the fraction of people shorter than z-score = 1.64?**

For that, we will have to refer to a **Z-table**: https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf

From the table, we find that 94.95% people have a shorter height than 69.1



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

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246

Q3. Alternately, can we solve this problem using code?

First, let's calculate the z-score

```
# importing libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import norm
```

```
z = (69.1 - 65) / 2.5
z
```

```
    1.6399999999999977
```

First, let's visualize what we are trying to calculate.

In order to find the fraction/probability, we want to calculate the area under the normal distribution curve, up to the point of $x = 69.1$ or $z - score = 1.64$

> **Q4. What exactly is this value? How can we calculate this? PDF/PMF/CDF?**

This can be found using the CDF value.

Since this is in relation to Normal distribution, we will use the `scipy.stats.norm.cdf(Z-score)` to find the percentage of people.

```
# calculate fraction who are shorter than 69.1
norm.cdf(z)
```

```
    0.949497416525896
```

Hence, we can conclude that **94.95%** of people are shorter than 69.1 inches

This is exactly what we found using the Z-table.

Verification of Empirical Rule

Now that we know how to calculate CDF using `norm.cdf(z)`, let's also verify the Empirical rule we learned.

> **Verify 68 Rule**

For this, we will need to calculate the area enclosed between $\mu - \sigma$ and $\mu + \sigma$

What will be the z-score for these points?

- $\mu - \sigma$: -1
- $\mu + \sigma$: +1

```
norm.cdf(1) - norm.cdf(-1)
```

```
    0.6826894921370859
```

Similarly, we can check for 95 and 99.7 Rule.

```
norm.cdf(2) - norm.cdf(-2)
```

```
    0.9544997361036416
```

```
norm.cdf(3) - norm.cdf(-3)
```

```
    0.9973002039367398
```

Now let's look at some other problems

Example on z-score

```
Balls produced by manufacturer have a mean diameter of 50 mm and a std dev of 2 mm.

What would be the diameter corresponding to the z-score of 1.5?
```

Solution:

Here we have given the z-score of a specific number with the mean and SD values too.

So we can easily get the number from the z-score formula

Given:

- Z-Score (z) = 1.5
- Mean (μ) = 50 mm
- std dev (σ) = 2 mm

We know:

$z = \frac{x-\mu}{\sigma}$

x = (1.5 * 2) + 50

x = 53mm

Conclusion:

```
The Diameter corresponding to the z-score of 1.5 would be 53mm
```

PPF (Percent Point Function)

Let's revisit the height example

Consider the following question.

```
The height of people is Gaussian with mean 65 inches and standard deviation 2.5 inches.

One person says:
96% people are shorter than me. What is my height?
```

Solution Approach:

We are given that this person is taller than 96% of the population

- Hence, we can find the z-score corresponding to his height from the **Z-table**
- So, we see that his height should be assigned a z-score value between 1.75 and 1.76
- Now consider the z-score formula: $z = \frac{x-\mu}{\sigma}$
- We can calculate the value of x, by plugging in the other values.

If you think about it, the approach we followed was the **inverse of `norm.cdf()`**, where

- We first calculated the z-score
- Then the fraction of population up to this score was looked up on Z-table (or computed using code)

This inverse method of finding the z-score from given percent/fraction is termed as the **Percent Point Function (PPF)**.

We can calculate this directly using **norm.ppf()** function

`norm.ppf(percentile)` will give the Z-score corresponding to the percentile.

- From that, we can get the height.
  - Here 96% people are shorter than me basically indicating 96 percentile.

```
# what is the height such that 96% people are shorter?
z = norm.ppf(0.96)
z
```

```
1.7506860712521692
```

```
# we know z = (x-mu)/sigma so from this we can get x
x = (z*2.5) + 65
x
```

```
69.37671517813042
```

Standard Normal Distribution

The special case, when a normal distribution has

- $\mu = 0$, and
- $\sigma = 1$

It is known as The **Standard Normal Distribution**, also called the **z-distribution**

- And, it is denoted by $Z(0, 1)$

We can take any Normal Distribution and convert it to The Standard Normal Distribution, using **Standardization**
