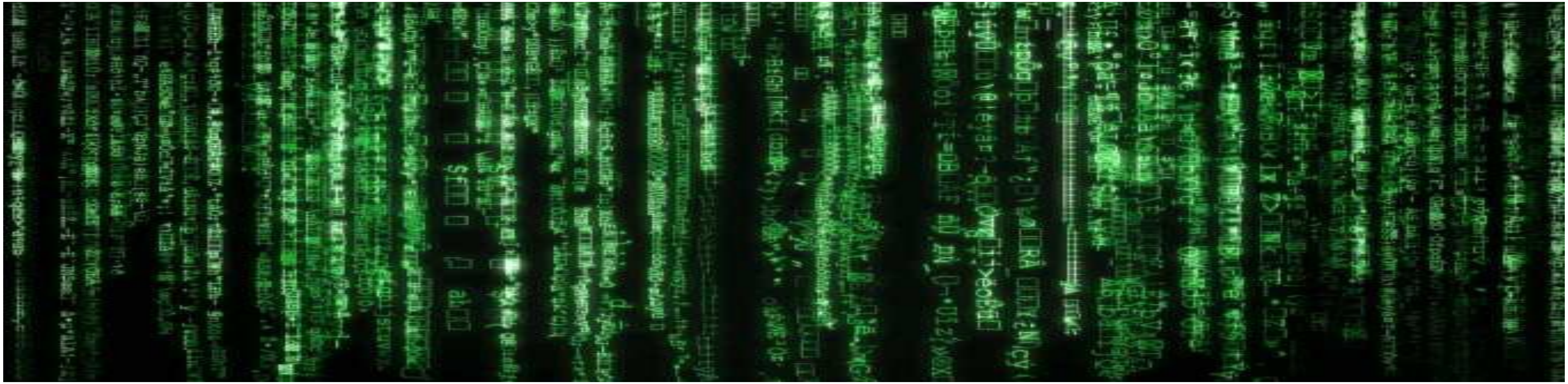


INF 110 Discovering Informatics

Means and Prediction

NORTHERN ARIZONA UNIVERSITY

The Mean

- Also known as “arithmetic average”
- A central value of a finite set of numbers
- The sum of the values divided by the number of values.

```
not_symmetric = make_array(2, 3, 3, 9)
np.average(not_symmetric)
np.mean(not_symmetric)
```

Basic Properties of the Mean

- It doesn’t have to be part of the collection of values.
- It doesn’t have to be an integer even if all the values are integers.
- It must be between the smallest and largest values.
- It doesn’t have to be half way between the minimum and maximum.
- The mean is in the same units (miles, kg, etc) as the values.

Basic Properties of the Mean

- Each value in a collection is weighted by it's ***proportion.***

```
not_symmetric = make_array(2, 3, 3, 9)
np.average(not_symmetric)
np.mean(not_symmetric)
```

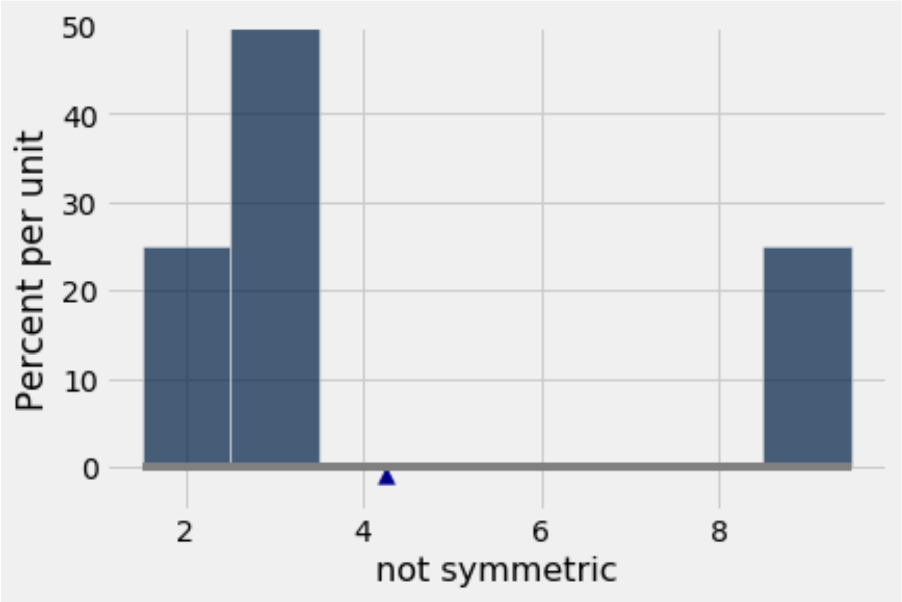
$$
\begin{aligned}
\text{mean} &= 4.25 \\
&= \frac{2+3+3+9}{4} \\
&= 2\cdot\frac{1}{4} \ + \ 3\cdot\frac{1}{4} \ + \ 3\cdot\frac{1}{4} \ + \ 9\cdot\frac{1}{4} \\
&= 2\cdot\frac{1}{4} \ + \ 3\cdot\frac{2}{4} \ + \ 9\cdot\frac{1}{4} \\
&= 2\cdot 0.25 \ + \ 3\cdot 0.5 \ + \ 9\cdot 0.25
\end{aligned}
$$

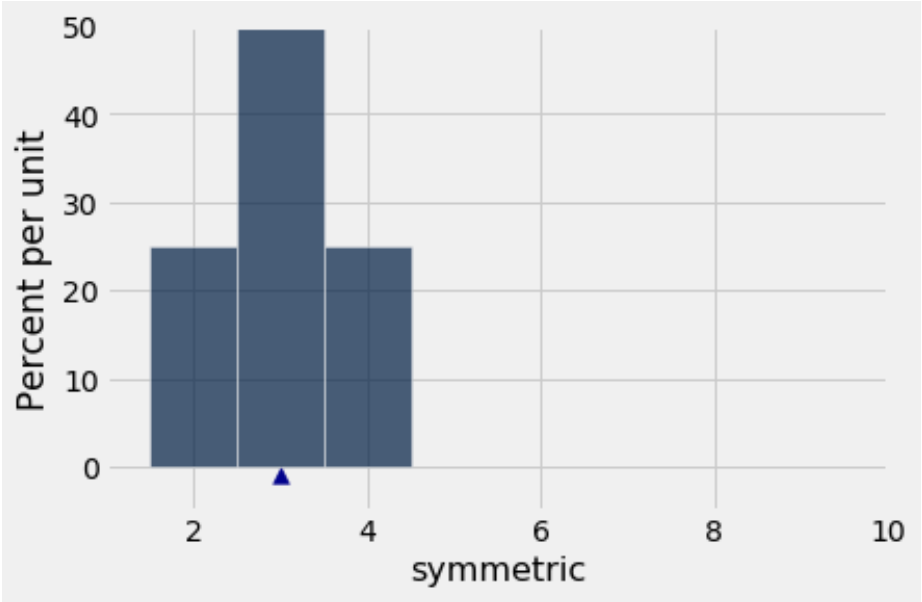
If two collections have the same distribution, they have the same mean.

```
same_distribution = make_array(2, 2, 3, 3, 3, 3, 9, 9)
np.mean(same_distribution)
```

Basic Properties of the Mean

- The mean is the center of gravity or balance point.

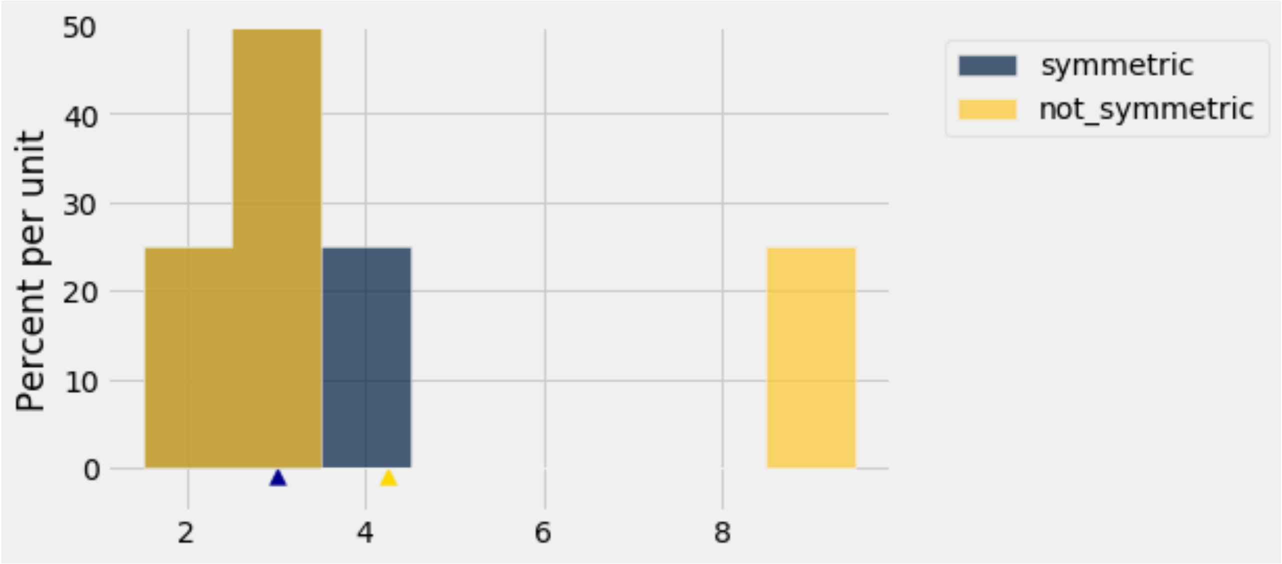




Mean vs. Median

- In a symmetrical distribution, they are the same.
- In an asymmetrical (or ***skewed***) distribution, the mean is pulled away from the median

Here the **blue** median and mean are 3.
The **gold** median is 3, but the mean is 4.25.



Variability

- We saw in the previous histograms that values can spread around the mean.
- But how do we measure how far they are from the mean?

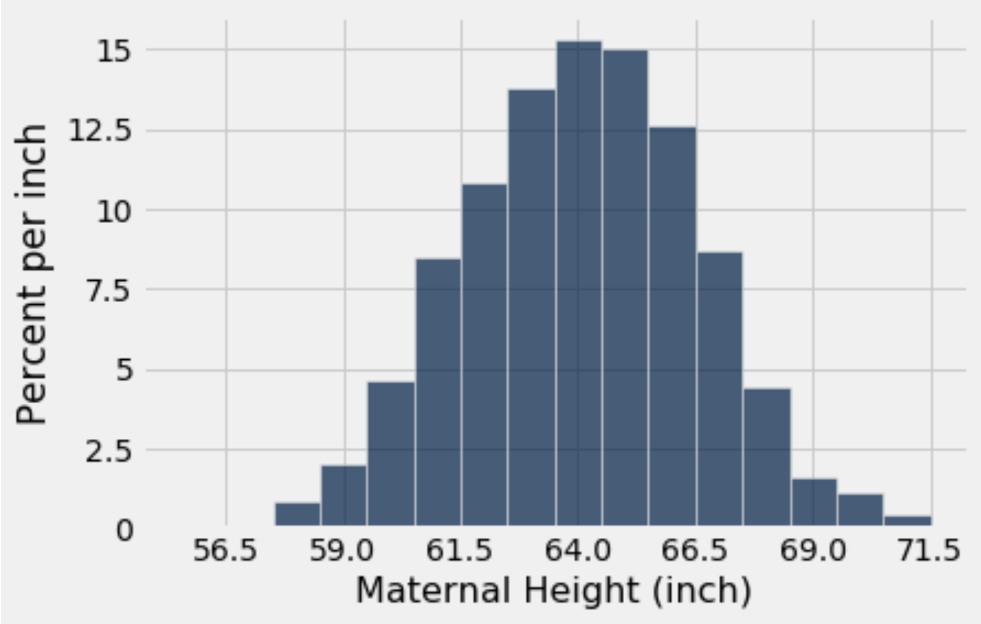
- ***Variance:*** the mean squared deviation from the average
- ***Standard deviation:*** the root mean square of deviations from average.
  - ***Can use np.std()***

Live Code Variability

- Use numpy to:
  - Calculate an average for an array of numbers
  - Measure deviations from the average.
  - Calculate the ***variance***.
  - Calculate the ***standard deviation***.

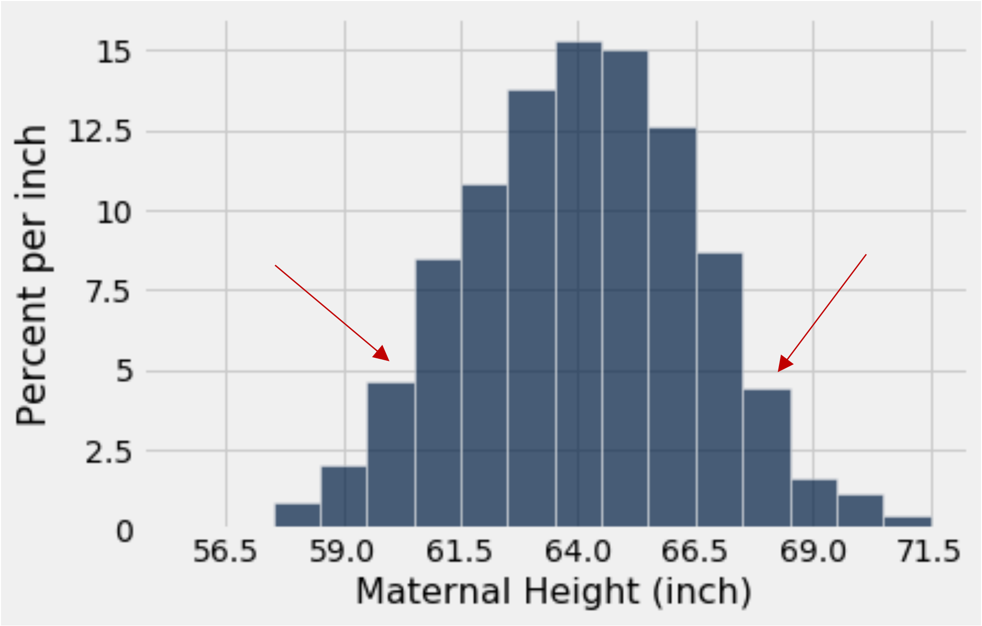
Standard Deviation and the Normal Curve

- SD is not easy to identify in most histograms.
- But it is when the data is in a bell shaped distribution.
  - The SD is the distance between the mean and the points of inflection on either side.



Standard Deviation and the Normal Curve

- SD is not easy to identify in most histograms.
- But it is when the data is in a bell shaped distribution.
  - The SD is the distance between the mean and the points of inflection on either side.



Mean of 64
SD of 2.5

The Standard Normal Curve

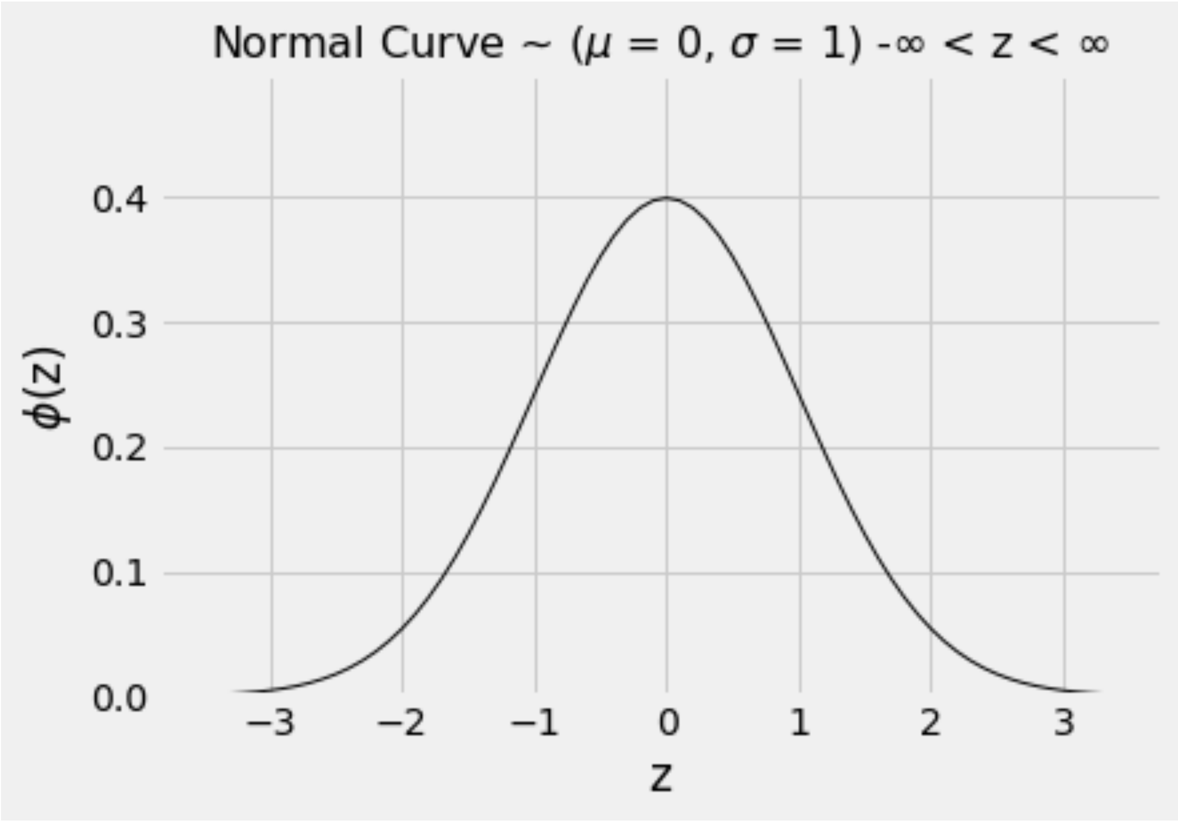
Bell-shaped histograms given in standard units

The standard normal curve has an equation:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$

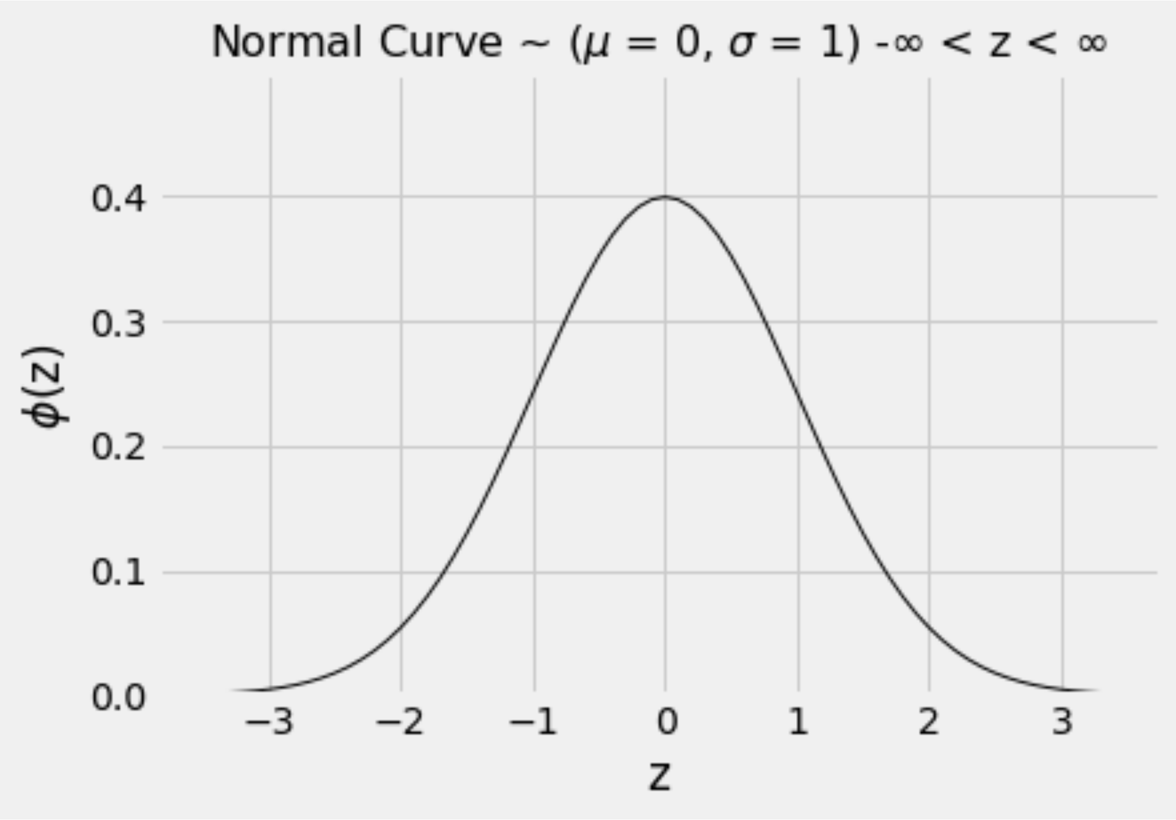
But we will think of it as a smoothed outline of a histogram that:

- Is measured in standard units
- Has a bell shaped distribution.

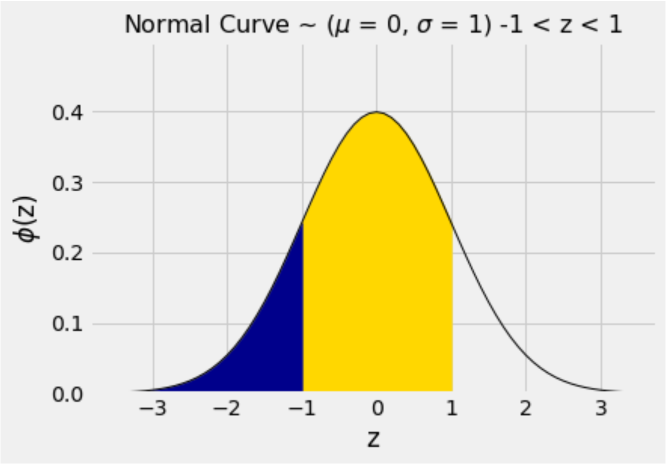


The Standard Normal Curve

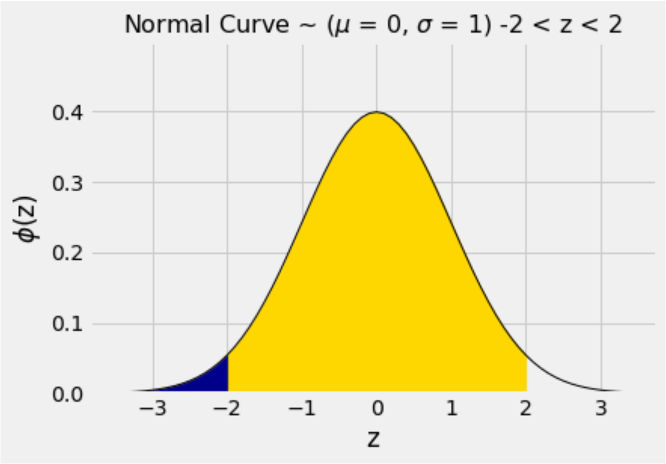
- The total area under the curve is 1.
- The curve is symmetric about 0.
  - Mean and median both = 0.
- The points of inflection are at -1 and +1.
- A normally distributed variable has a SD of 1.



The Standard Normal Curve – Area Under the Curve



Yellow = AUC between *z* = -1 and *z* = 1



Yellow = AUC between *z* = -2 and *z* = 2

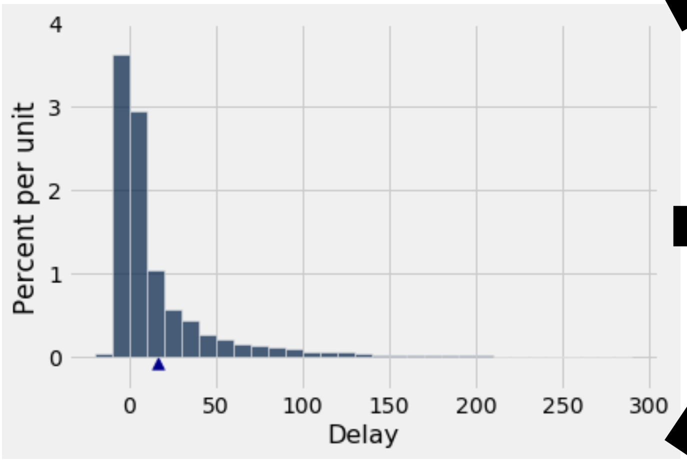
Percent in Range	All distributions: bound	Normal distribution: approximation
Average ± 1 SD	At least 0%	About 68%
Average ± 2 SDs	At least 75%	About 95%
Average ± 3 SDs	At least 88.8888…%	About 99.73%

Live Code The Central Limit Theorem

The probability distribution of the sum or average of a large random sample will be roughly normal, regardless of the distribution of the population from which the sample is drawn.

Variability of the Sample Mean

- The distribution of the mean of a large sample will be roughly normal (CLT)
- However, with larger samples, these distributions will cluster closer to the mean (meaning there is less variability).



[calculate the mean for random samples, then repeat 10,000 times]