

# DISTRIBUTIONS

# DISTRIBUTIONS

- A distribution describes all of the probable outcomes of a **variable**.
- In a discrete distribution, the sum of all the individual probabilities must equal 1
- In a continuous distribution, the area under the probability curve equals 1

## DISCRETE DISTRIBUTIONS

- Discrete probability distributions are also called **probability mass functions**:

Uniform Distribution

Binomial Distribution

Poisson Distribution

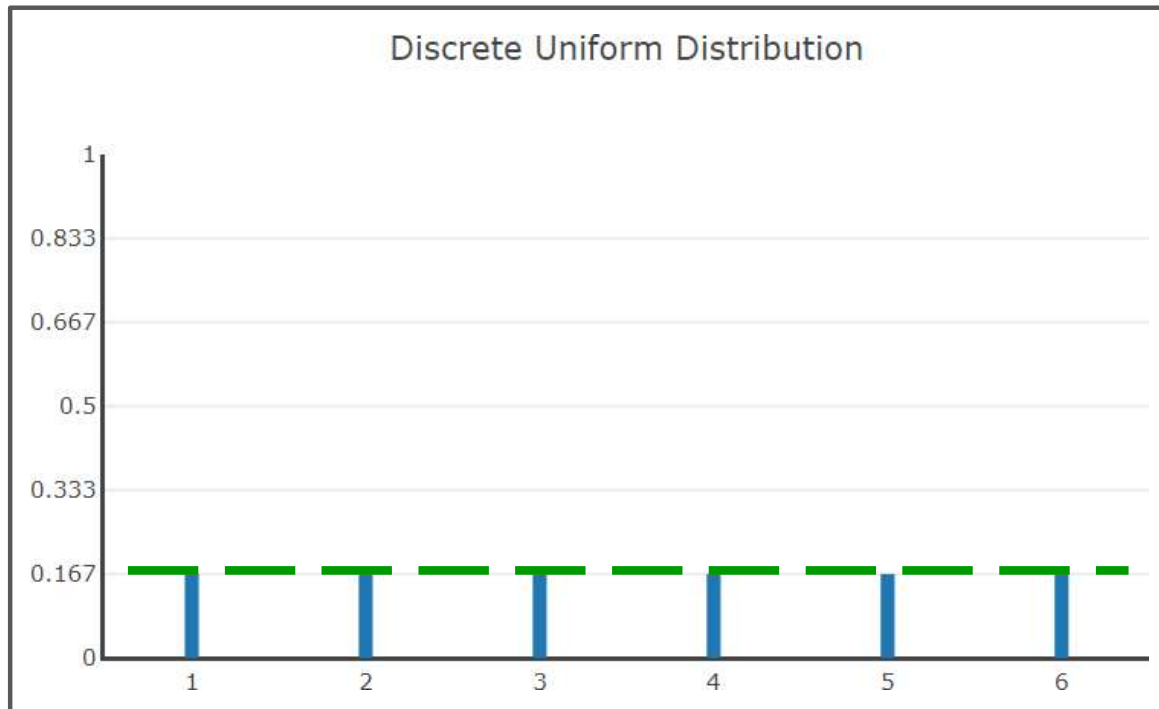
# UNIFORM DISTRIBUTION

- Rolling a fair die has 6 discrete, equally probable outcomes
- You can roll a 1 or a 2, but not a 1.5
- The probabilities of each outcome are evenly distributed across the sample space



# UNIFORM DISTRIBUTION

- Rolling a fair die:



heights are all the same, add up to 1

## BINOMIAL DISTRIBUTION

- “Binomial” means there are two discrete, mutually exclusive outcomes of a trial.

heads or tails

on or off

sick or healthy

success or failure

## BERNOULLI TRIAL

- A **Bernoulli Trial** is a random experiment in which there are only two possible outcomes
  - success or failure
- A series of  $n$  trials will follow a binary distribution as long as
  - a) the probability of success  $p$  is constant
  - b) trials are independent of one another

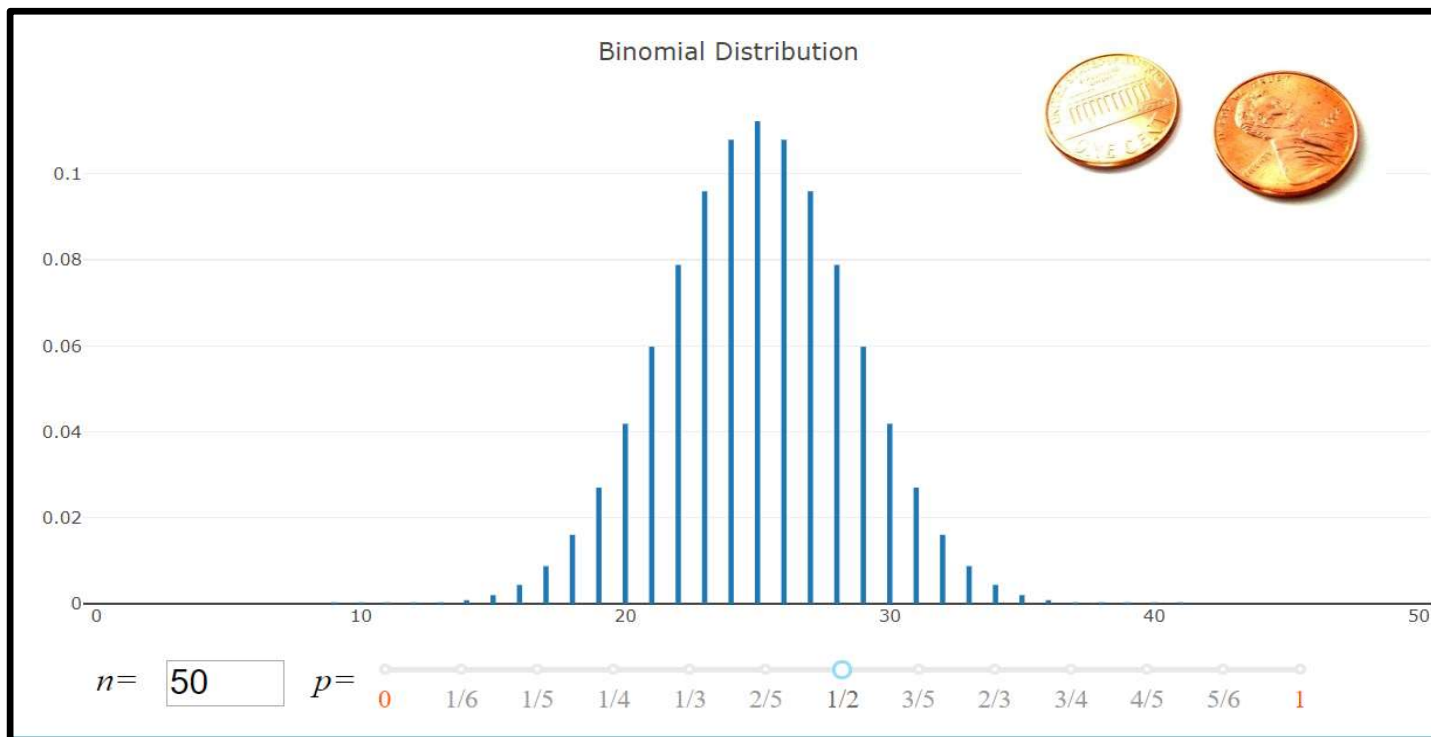
## BINOMIAL PROBABILITY MASS FUNCTION

- Gives the probability of observing  $x$  successes in  $n$  trials
- $p$  is the probability of a successful outcome
- Assumes that  $p$  is fixed for all trials

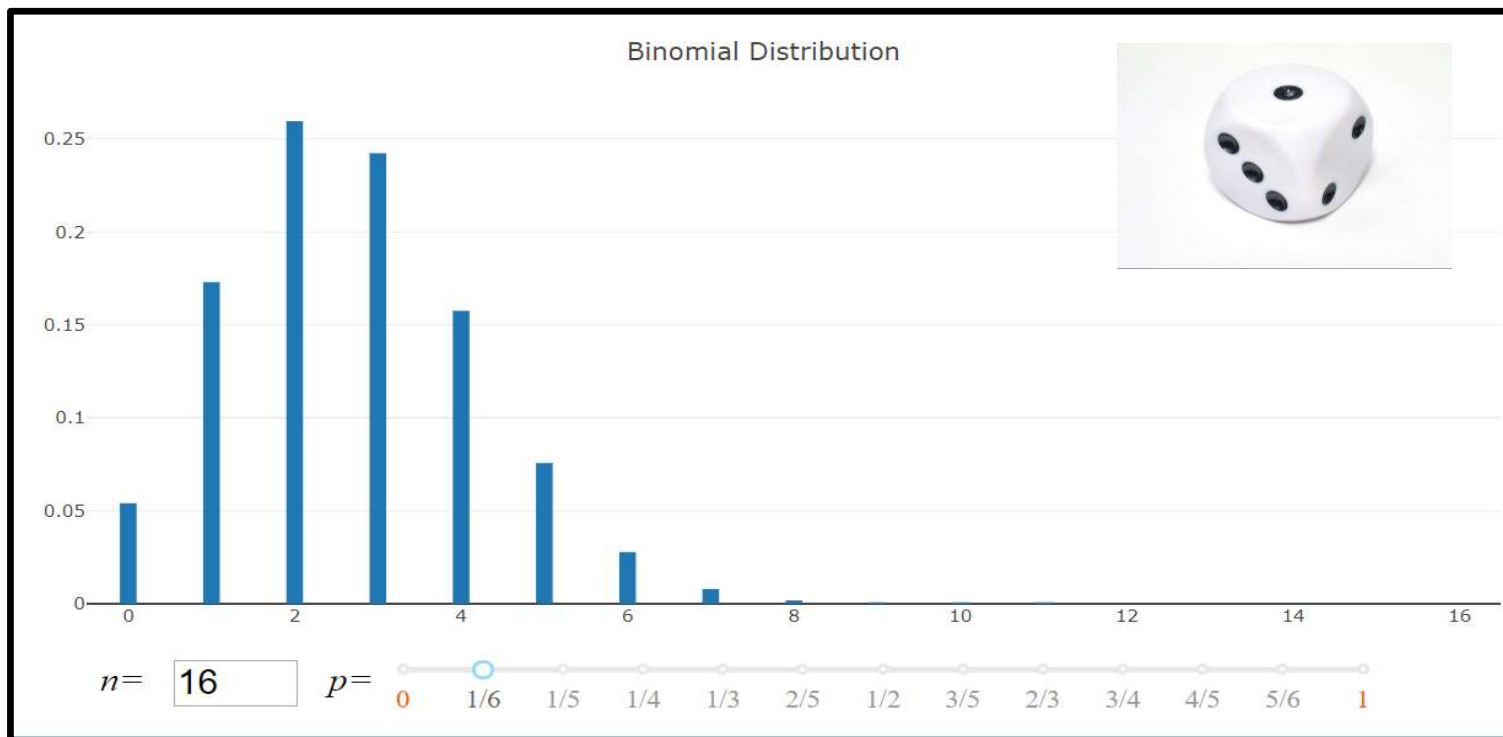
$$f(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$



# BINOMIAL DISTRIBUTION



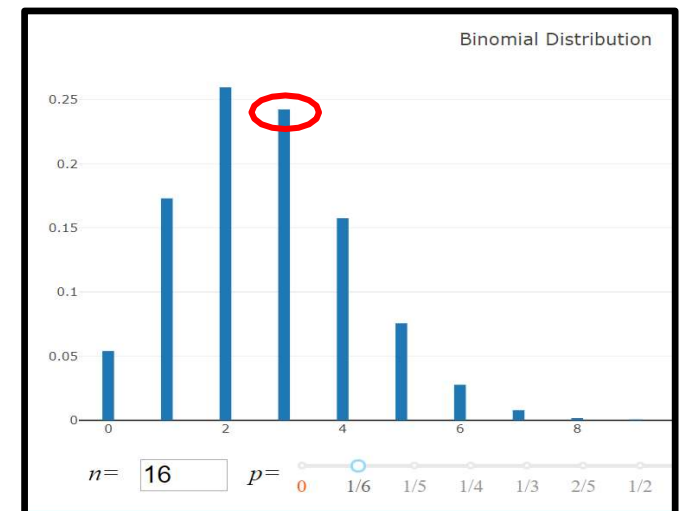
# BINOMIAL DISTRIBUTION



# BINOMIAL DISTRIBUTION

## EXERCISE

- If you roll a die 16 times, what is the probability that a five comes up 3 times?
- Based on the chart, it should be just shy of 0.25
- $x = 3, n = 16, p = 1/6$



## BINOMIAL DISTRIBUTION EXERCISE

$$\begin{aligned}P(x: n, p) &= \binom{n}{x} (p)^x (1 - p)^{(n-x)} \\&= \left( \frac{n!}{x! (n - x)!} \right) (p)^x (1 - p)^{(n-x)} \\&= \left( \frac{16!}{3! (13)!} \right) (1/6)^3 (5/6)^{(13)} \\&= \left( \frac{16 \cdot 15 \cdot 14}{3 \cdot 2} \right) \left( \frac{1^3}{6^3} \right) \left( \frac{5^{13}}{6^{13}} \right) = 0.242\end{aligned}$$

## USING PYTHON

- If you roll a die 16 times, what is the probability that a five comes up 3 times?

```
>>> from scipy.stats import binom
```

```
>>> binom.pmf(3,16,1/6)
```

```
0.24231376033713251
```

## POISSON DISTRIBUTION

- A binomial distribution considers the number of successes out of  $n$  trials
- A **Poisson Distribution** considers the number of successes *per unit of time*\* over the course of many units

\*or any other continuous unit, e.g. *distance*

## POISSON DISTRIBUTION

- Calculation of the Poisson **probability mass function** starts with a mean expected value

$$E(X) = \mu$$

- This is then assigned to "lambda"

$$\lambda = \frac{\# \text{ occurrences}}{\text{interval}} = \mu$$

## POISSON DISTRIBUTION

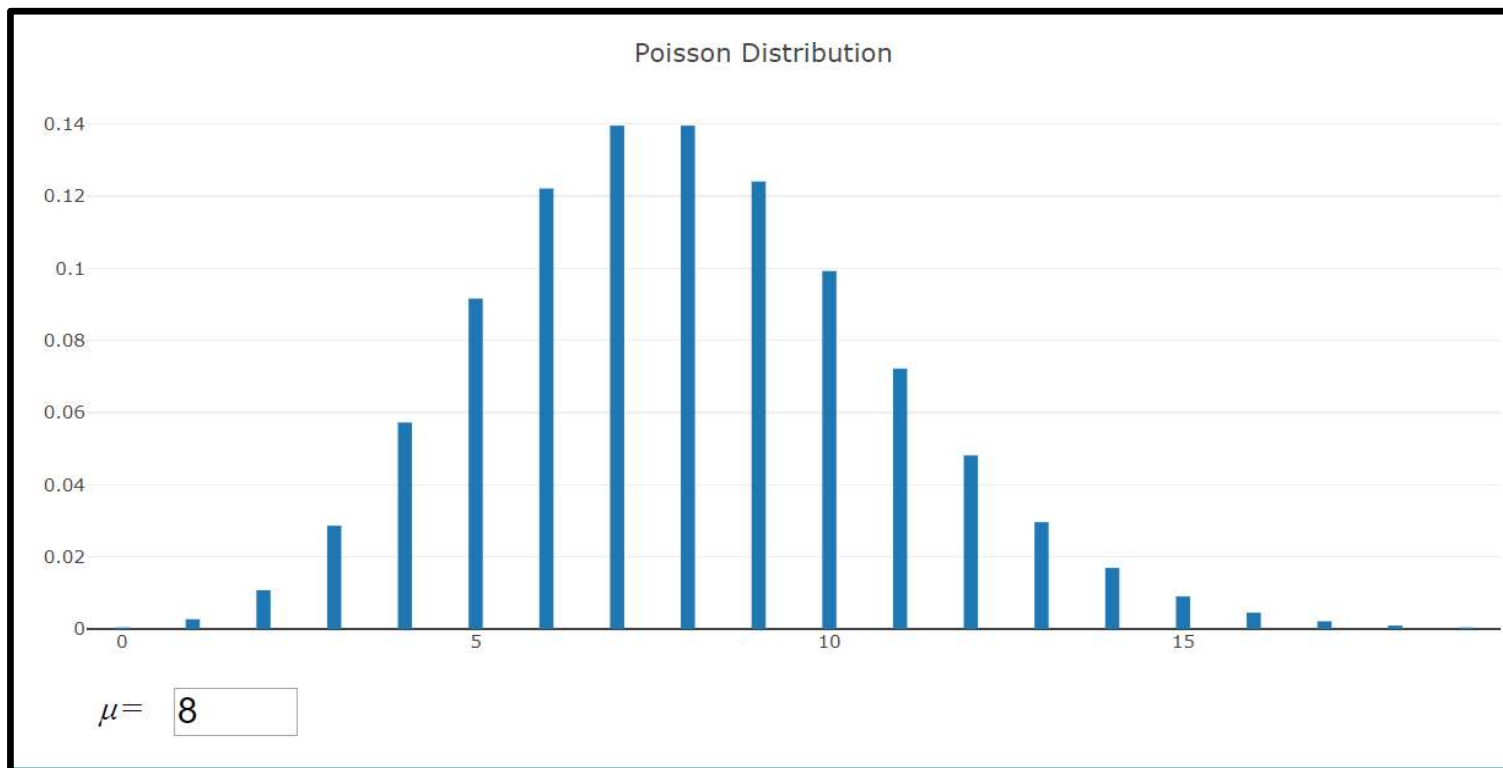
- The equation becomes

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where  $e$  = Euler's number = 2.71828 ...



# POISSON DISTRIBUTION



## POISSON DISTRIBUTION EXERCISE #1

- A warehouse typically receives 8 deliveries between 4 and 5pm on Friday.
- What is the probability that **only 4 deliveries** will arrive between 4 and 5pm this Friday?



## POISSON DISTRIBUTION EXERCISE #1

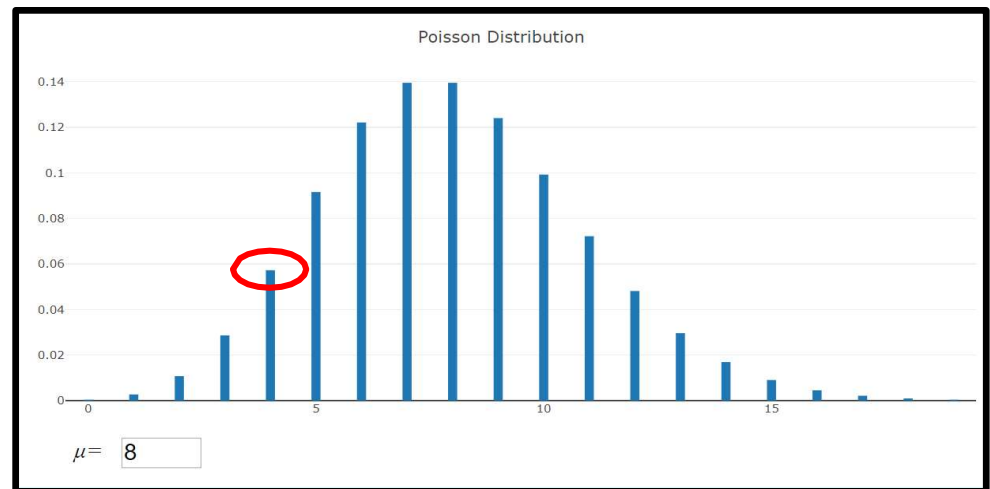
$$x = 4 \quad \lambda = 8$$

$$\begin{aligned} P(x) &= \frac{\lambda^x e^{-\lambda}}{x!} = \frac{8^4 \cdot 2.71828^{-8}}{4!} \\ &= \frac{4096 \cdot \left(\frac{1}{2980.96}\right)}{24} = \mathbf{0.0572} \end{aligned}$$

## POISSON DISTRIBUTION EXERCISE #1

$$= \frac{4096 \cdot \left( \frac{1}{2980.96} \right)}{24} = \mathbf{0.0572}$$

This agrees with  
our chart!



## POISSON DISTRIBUTION

- The **cumulative mass function** is simply the sum of all the discrete probabilities
- The probability of seeing *fewer than* 4 events in a Poisson Distribution is:

$$\begin{aligned} P(X: x < 4) &= \sum_{i=0}^3 \frac{\lambda^i e^{-\lambda}}{i!} \\ &= \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!} \end{aligned}$$

## POISSON DISTRIBUTION

- Remember that the sum of all possibilities equals 1
- The probability of seeing *at least* 1 event is one minus the probability of seeing none:

$$\begin{aligned} P(X: x \geq 1) &= 1 - P(X: x = 0) \\ &= 1 - \frac{\lambda^0 e^{-\lambda}}{0!} = 1 - e^{-\lambda} \end{aligned}$$

## POISSON DISTRIBUTION EXERCISE #2

- A warehouse typically receives 8 deliveries between 4 and 5pm on Friday.
- What is the probability that **fewer than 3** will arrive between 4 and 5pm this Friday?



## POISSON DISTRIBUTION EXERCISE #2

$$\begin{aligned}P(X: x < 3) &= \sum_{i=0}^2 \frac{\lambda^i e^{-\lambda}}{i!} = \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} \\&= \frac{8^0 \cdot 2.71828^{-8}}{0!} + \frac{8^1 \cdot 2.71828^{-8}}{1!} + \frac{8^2 \cdot 2.71828^{-8}}{2!} \\&= \frac{1 \cdot \left(\frac{1}{2980.96}\right)}{1} + \frac{8 \cdot \left(\frac{1}{2980.96}\right)}{1} + \frac{64 \cdot \left(\frac{1}{2980.96}\right)}{2} \\&= \mathbf{0.0137}\end{aligned}$$



## POISSON DISTRIBUTION—PARTIAL INTERVALS

- The Poisson Distribution assumes that the probability of success during a small time interval is proportional to the entire length of the interval.
- If you know the expected value  $\lambda$  over an hour, then the expected value over one minute of that hour

is:

$$\lambda_{minute} = \frac{\lambda_{hour}}{60}$$

## POISSON DISTRIBUTION EXERCISE #3

- A warehouse typically receives 8 deliveries between 4 and 5pm on Friday.
- What is the probability that no deliveries arrive between 4:00 and 4:05 this Friday?



## POISSON DISTRIBUTION EXERCISE #3

$$x = 0 \quad \lambda_{1 \text{ hour}} = 8$$

$$\lambda_{5 \text{ minutes}} = \frac{\lambda_{1 \text{ hour}}}{60/5} = \frac{8}{12} = 0.6667$$

$$\begin{aligned} P(x) &= \frac{\lambda^x e^{-\lambda}}{x!} = \frac{0.67^0 \cdot 2.71828^{-0.6667}}{0!} \\ &= \mathbf{0.5134} \end{aligned}$$

# USING PYTHON

What is the probability that only 4 deliveries will arrive between 4 and 5 pm this Friday?

```
>>> from scipy.stats import poisson  
>>> poisson.pmf(4,8)  
0.057252288495362
```

# USING PYTHON

What is the probability that fewer than 3 will arrive between 4 and 5pm this Friday?

```
>>> from scipy.stats import poisson  
>>> poisson.cdf(2,8)  
0.013753967744002971
```

## USING PYTHON

What is the probability that no deliveries arrive  
between 4:00 and 4:05 this Friday?

```
>>> from scipy.stats import poisson
```

```
>>> poisson.pmf(0,8/12)
```

```
0.51341711903259202
```

# CONTINUOUS DISTRIBUTIONS

- Continuous probability distributions are also called *probability density functions (pdf)*:

Normal Distribution

Exponential Distribution

Beta Distribution

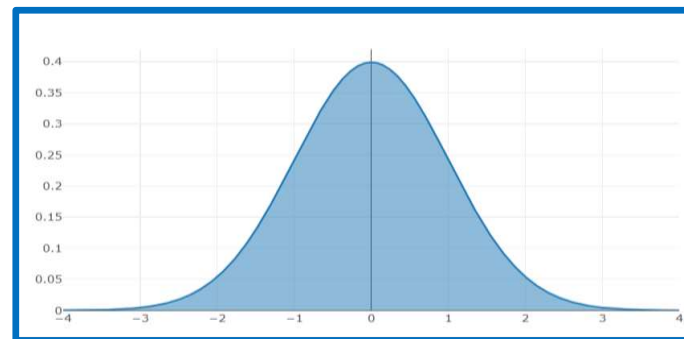
# NORMAL DISTRIBUTION

- Many real life data points follow a normal distribution:
- People's Heights and Weights
- Population Blood Pressure
- Test Scores
- Measurement Errors



# NORMAL DISTRIBUTION

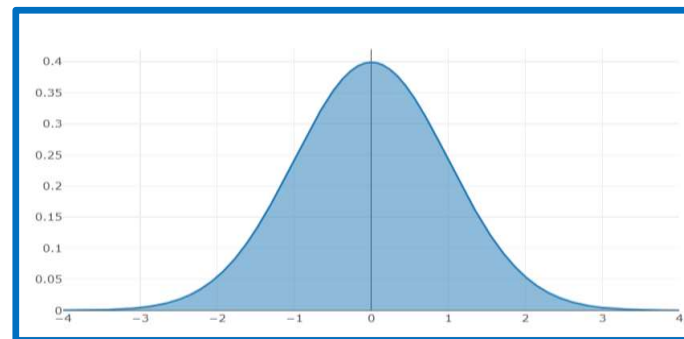
- These data sources tend to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



Normal Distribution

# NORMAL DISTRIBUTION

- We use a continuous distribution to model the behavior of these data sources.
- Notice the continuous line and area in this PDF.

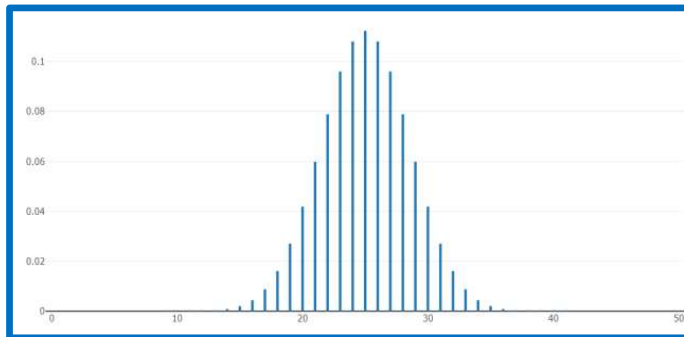


Normal Distribution

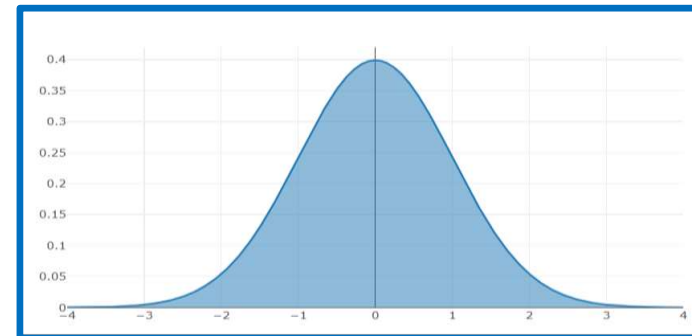
# NORMAL DISTRIBUTION

- Unlike discrete distributions, where the sum of all the bars equals one, in a normal distribution the *area under the curve* equals one

Binomial Distribution



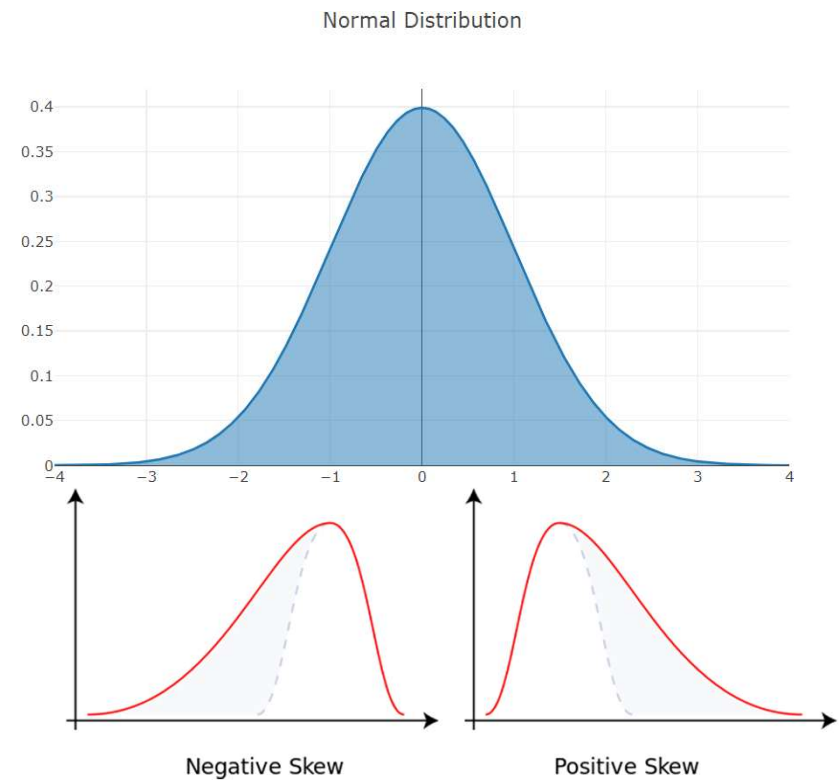
Normal Distribution



# NORMAL DISTRIBUTION

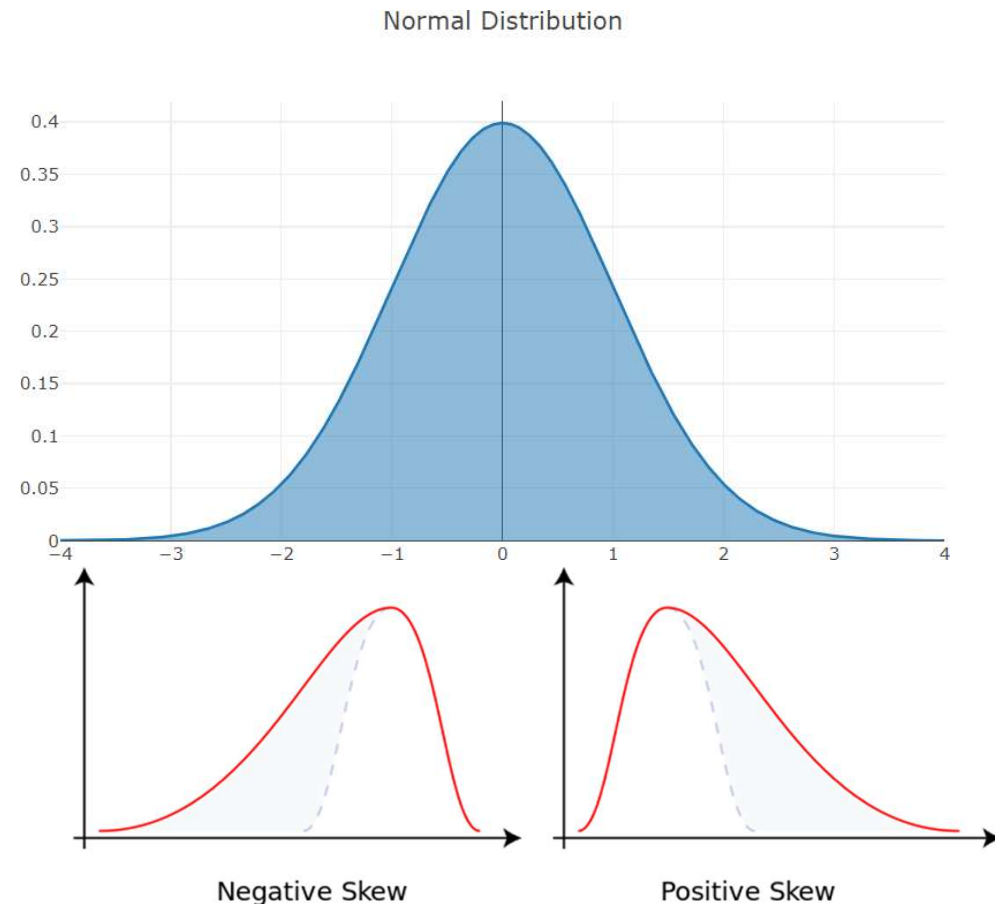
- Also called the **Bell Curve** or **Gaussian Distribution**
- Always symmetrical

Asymmetrical curves display **skew** and are *not* normal

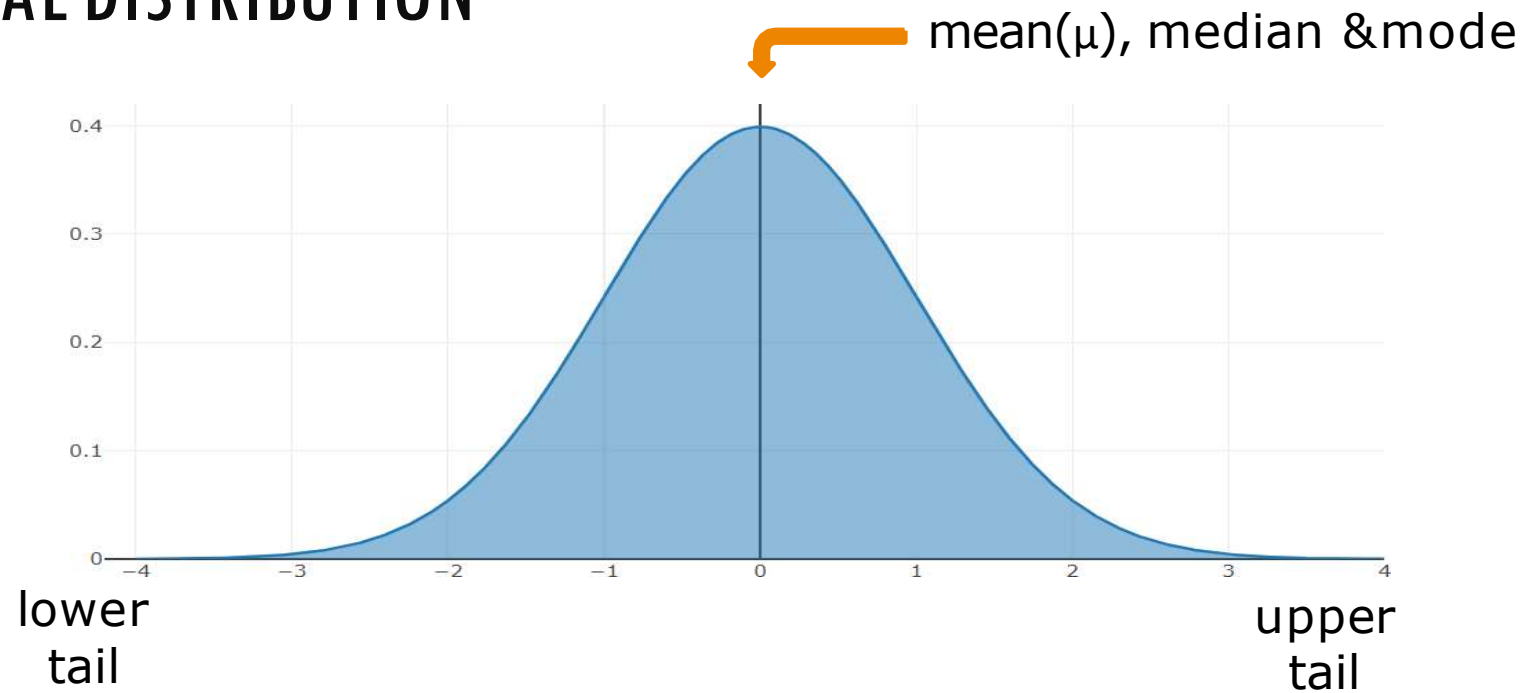


# NORMAL DISTRIBUTION

- The probability of a *specific outcome* is zero
- We can only find probabilities over a *specified interval* or range of outcomes



# NORMAL DISTRIBUTION

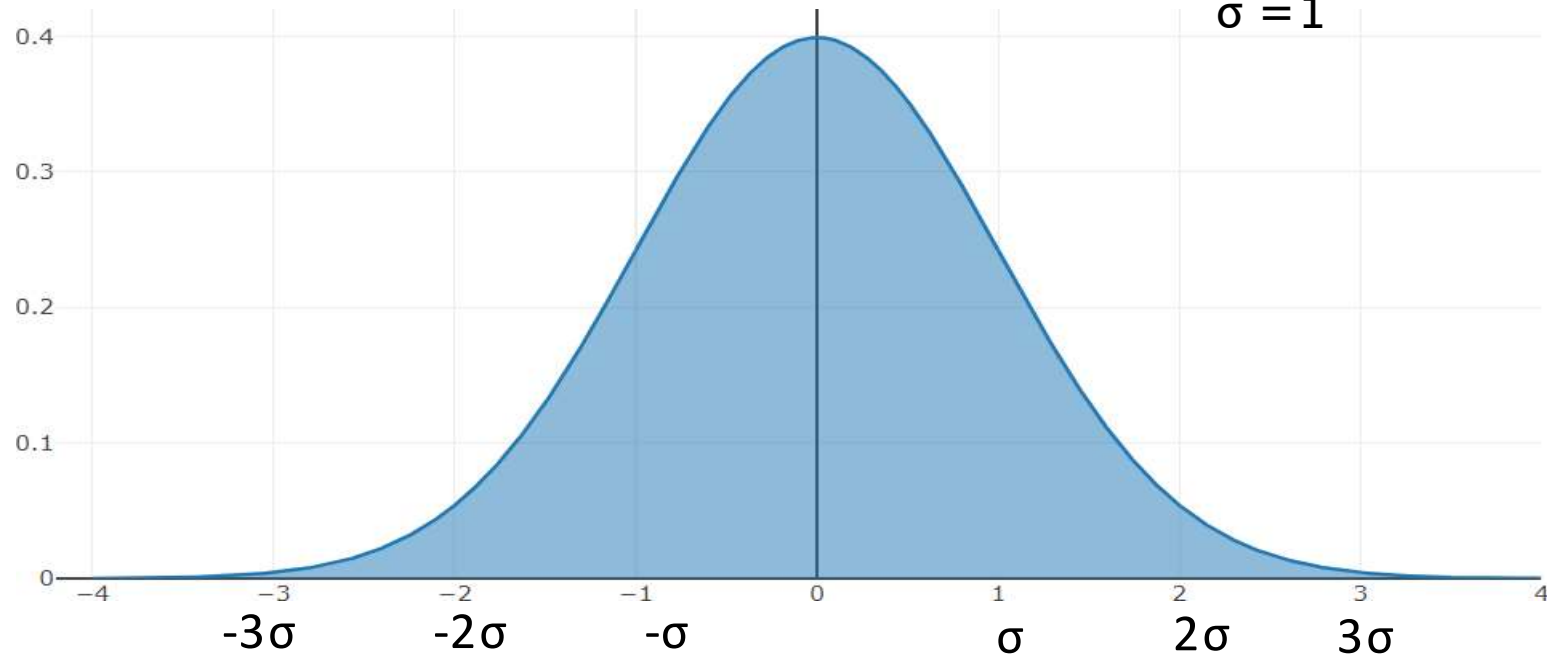


# STANDARD NORMAL DISTRIBUTION

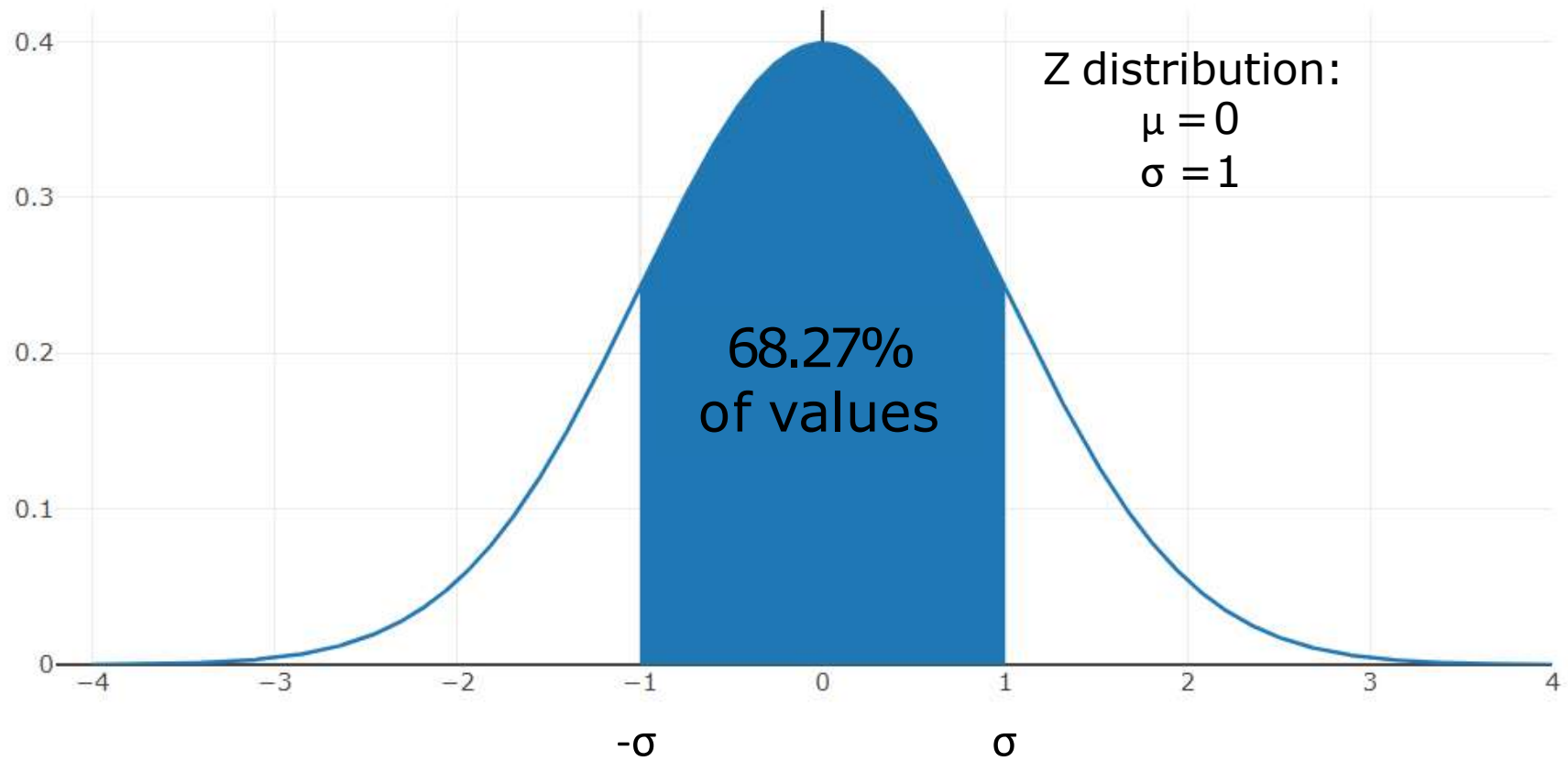
Z distribution:

$$\mu = 0$$

$$\sigma = 1$$



# STANDARD NORMAL DISTRIBUTION



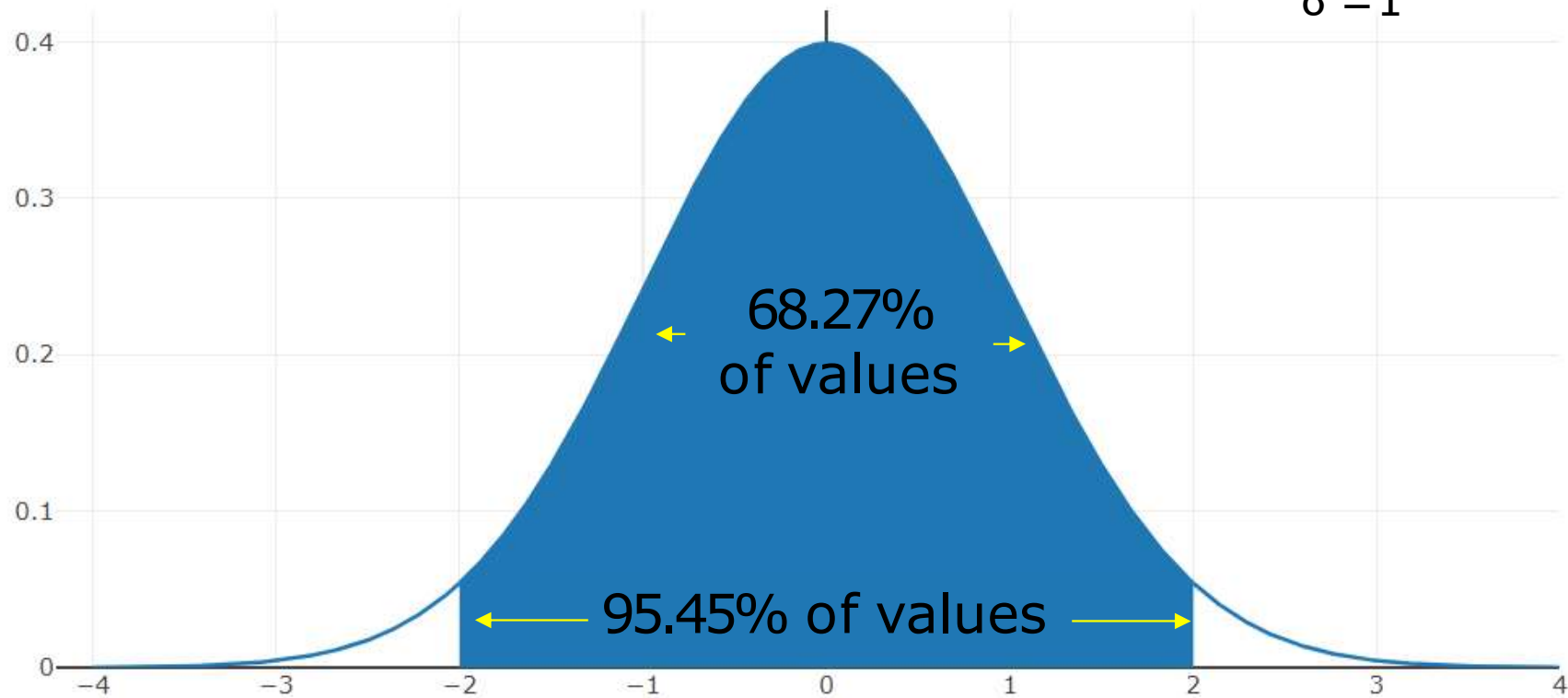


# STANDARD NORMAL DISTRIBUTION

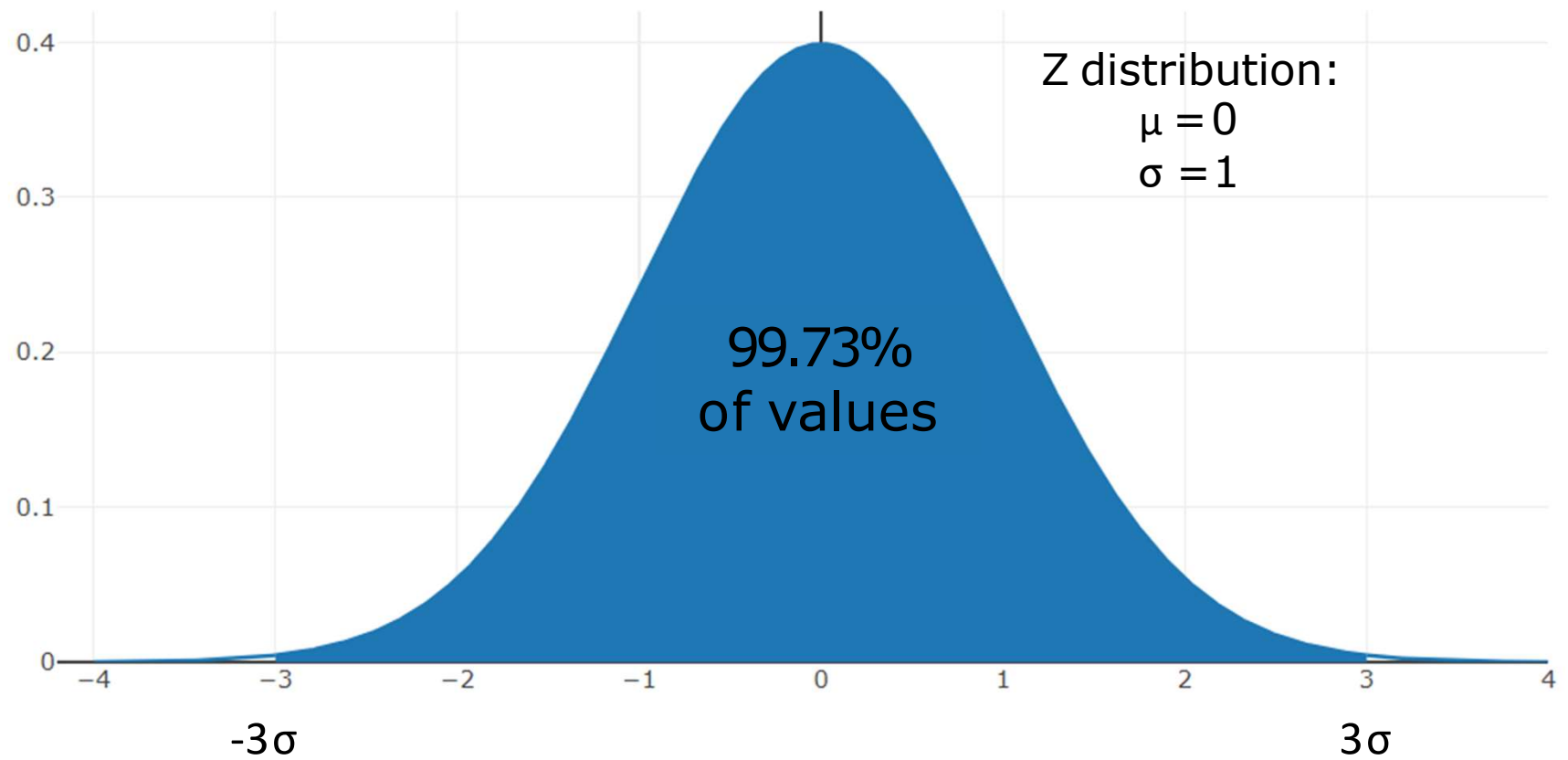
Z distribution:

$$\mu = 0$$

$$\sigma = 1$$



# STANDARD NORMAL DISTRIBUTION

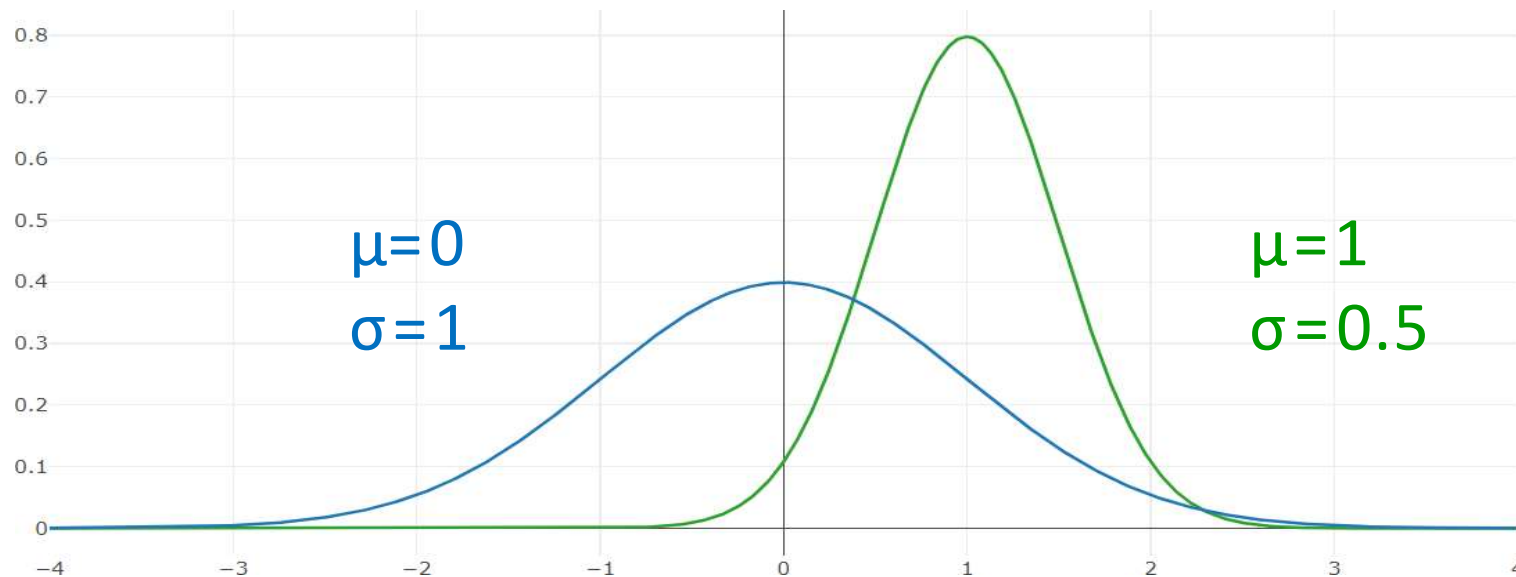


## NORMAL DISTRIBUTION

- All normal curves exhibit the same behavior:
  - symmetry about the mean
  - 99.73% of values fall within three standard deviations
- However, the mean does not have to be zero, and  $\sigma$  does not have to equal one.

# NORMAL DISTRIBUTION FORMULA

Other populations can be normal as well:



## NORMAL DISTRIBUTION

- If we determine that a population approximates a normal distribution, then we can make some powerful inferences about it once we know its mean and standard deviation

## NORMAL DISTRIBUTION

- In the Statistics section of the course, we will be using sampling, standard error, and hypothesis testing to evaluate experiments.
- A large part of this process is understanding how to "standardize" a normal distribution.

# NORMAL DISTRIBUTION

- We can take any normal distribution and standardize it to a standard normal distribution.



# NORMAL DISTRIBUTION

- We'll be able to take any value from a normal distribution and standardize it through a **Z-score**.





# NORMAL DISTRIBUTION

- Using this Z Score, we can then calculate a particular x value's percentile.



## NORMAL DISTRIBUTION

- Recall that a percentile is a way of saying "What percentage falls below this value".
- Meaning a 95 percentile value indicates that 95 percent of all other data points fall below this value.

## NORMAL DISTRIBUTION

- For example if a student scores a 1700 on their SATs and this score is in the 90 percentile, then we know 90% of all other students scored less than 1700.

## NORMAL DISTRIBUTION

- If we can model our data as a normal distribution, we can convert the values in the normal distribution to a standard normal distribution to calculate a percentile.

## NORMAL DISTRIBUTION

- For example, we can have a normal distribution of test point scores with some mean and standard deviation.
- We can then use a Z score to figure out the percentile of any particular test score.

## NORMAL DISTRIBUTION FORMULA

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

Where:

$\mu$  = mean

$e = 2.71828$

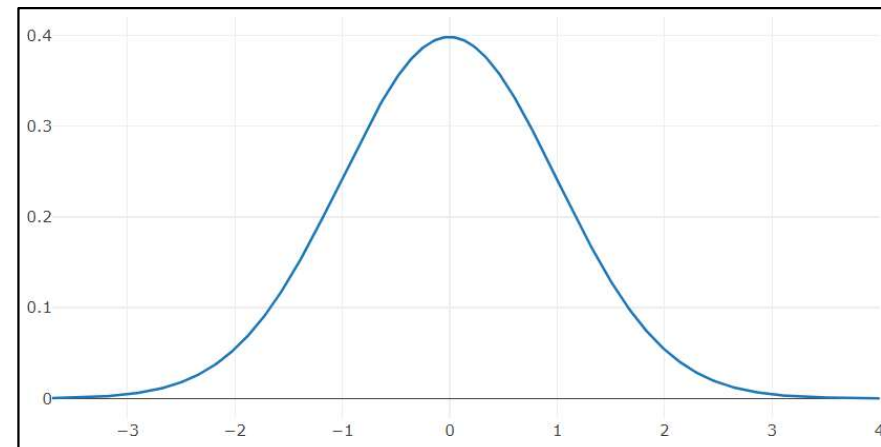
$\sigma$  = standard deviation

$\pi = 3.14159$

## NORMAL DISTRIBUTION FORMULA

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

This produced our plot with a mean of 0 and a standard deviation of 1:



## Z-SCORES AND Z-TABLE

- To gain insight about a specific value  $x$  in other normal populations, we *standardize*  $x$  by calculating a z-score:

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

- We can then determine  $x$ 's *percentile* by looking at a z-table

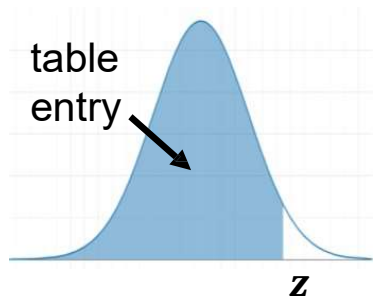


## HOW TO LOOK UP Z-SCORES

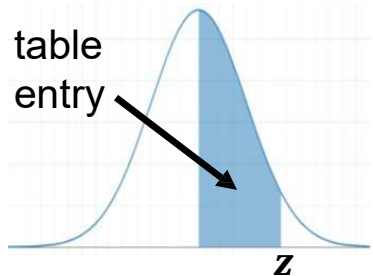
- A z-table of **Standard Normal Probabilities** maps a particular z-score to the area under a normal distribution curve to the left of the score.
- Since the total area under the curve is 1, probabilities are bounded by 0 and 1

# HOW TO LOOK UP Z-SCORES

- Different tables serve different purposes:



$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141



$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141

## Z-SCORES IN PYTHON

```
>>> from scipy import stats
```

```
>>> z = .70
```

```
>>> stats.norm.cdf(z)
```

```
0.75803634777692697
```

```
>>> p = .95
```

```
>>> stats.norm.ppf(p)
```

```
1.6448536269514722
```

## Z-SCORE EXERCISE

- A company is looking to hire a new database administrator.
- They give a standardized test to applicants to measure their technical knowledge.
- Their first applicant, Amy, scores an 87
- Based on her score, is Amy exceptionally qualified?

## Z-SCORE EXERCISE

- To decide how well an applicant scored, we need to understand the population.
- Based on thousands of previous tests, we know that the mean score is 75 out of 100, with a standard deviation of 7 points.

## Z-SCORE EXERCISESOLUTION

- First, convert Amy's score to a standardized z-score using the formula

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{87 - 75}{7} = \mathbf{1.7143}$$

## Z-SCORE EXERCISESOLUTION

- Next, look up 1.7143 on a z-table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

## Z-SCORE EXERCISE SOLUTION

- 0.9564 represents the area to the left of Amy's score
- This means that Amy outscored 95.64% of others who took the same test.

