Chapter 3 – PROBABILITY 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

Random Variables

A random variable is a numerical outcome of a random experiment.

Types:

Discrete Random Variable: Takes specific values (e.g., dice roll: 1, 2, 3...).

Continuous Random Variable: Takes any value in a range (e.g., height: 5.5 ft).

Example: Tossing a coin → Heads = 1, Tails = 0.

Discrete Random Variables:

Random variables that take a countable number of distinct values.

Values are specific and separate (e.g., integers).
Represented using a **Probability Mass Function (PMF)**.
Examples:
Number of heads in 3 coin tosses (0, 1, 2, 3).
Number of customers in a store (0, 1, 2...).

Discrete Random Variables:

Random variables that take a countable number of distinct values.

Values are specific and separate (e.g., integers).
Represented using a **Probability Mass Function (PMF)**.
Examples:
Number of heads in 3 coin tosses (0, 1, 2, 3).
Number of customers in a store (0, 1, 2...).

Continuous Random Variables:

Random variables that can take an infinite number of values within a range.

Values are not countable but can take any value in a continuous interval.

Represented using a **Probability Density Function (PDF).**

Examples:

Height of people (e.g., 5.5, 5.75 feet).

Time taken to complete a task (e.g., 2.35 seconds).

Discrete Probability Distributions

Discrete Distributions

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

Uniform Distribution

Binomial Distribution

Poisson Distribution

Uniform Distribution

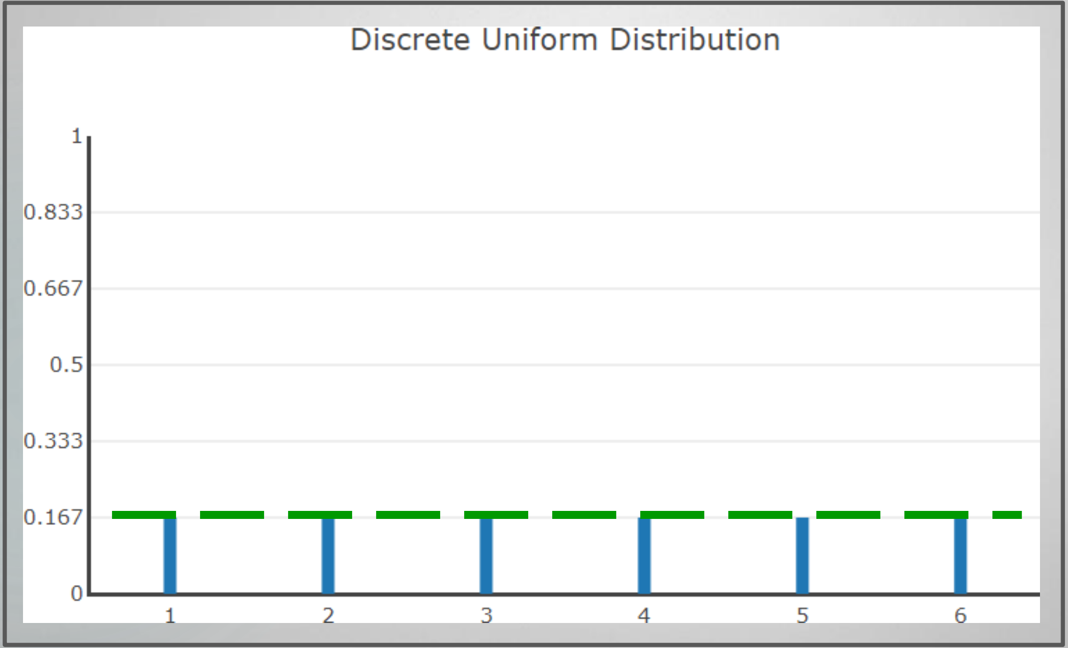
Uniform Distribution

- Rolling a fair die has 6 discrete, equally probable outcomes

- You can roll a 1or 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 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

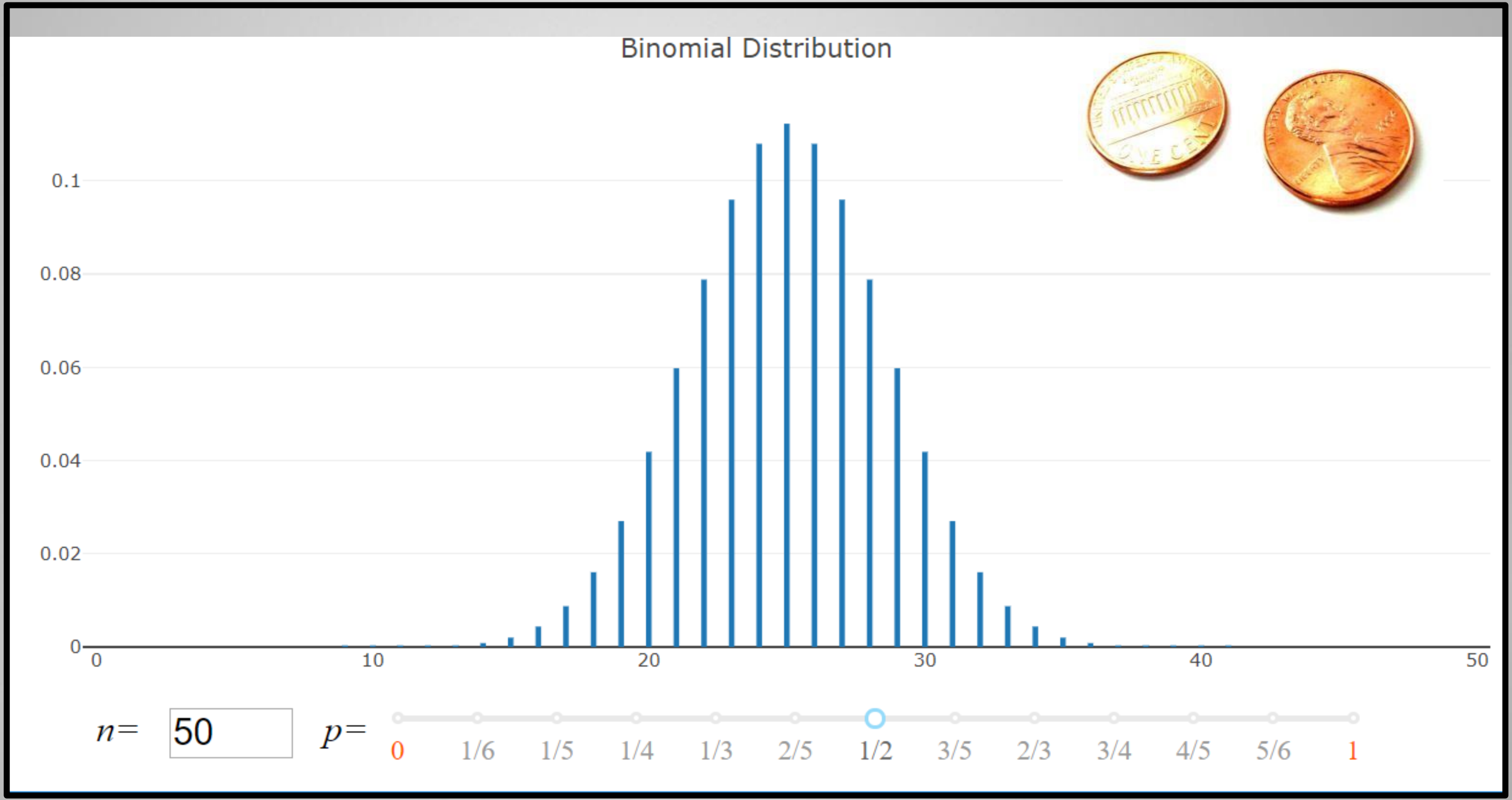
- A Bernoulli Trial is a random experiment in which there are only two possible outcomes - success or failure
- A series of trials *n* will follow a binary distribution so long as
  a) the probability of success *p* is constant
  b) trials are independent of one another

Binomial Probability Mass Function

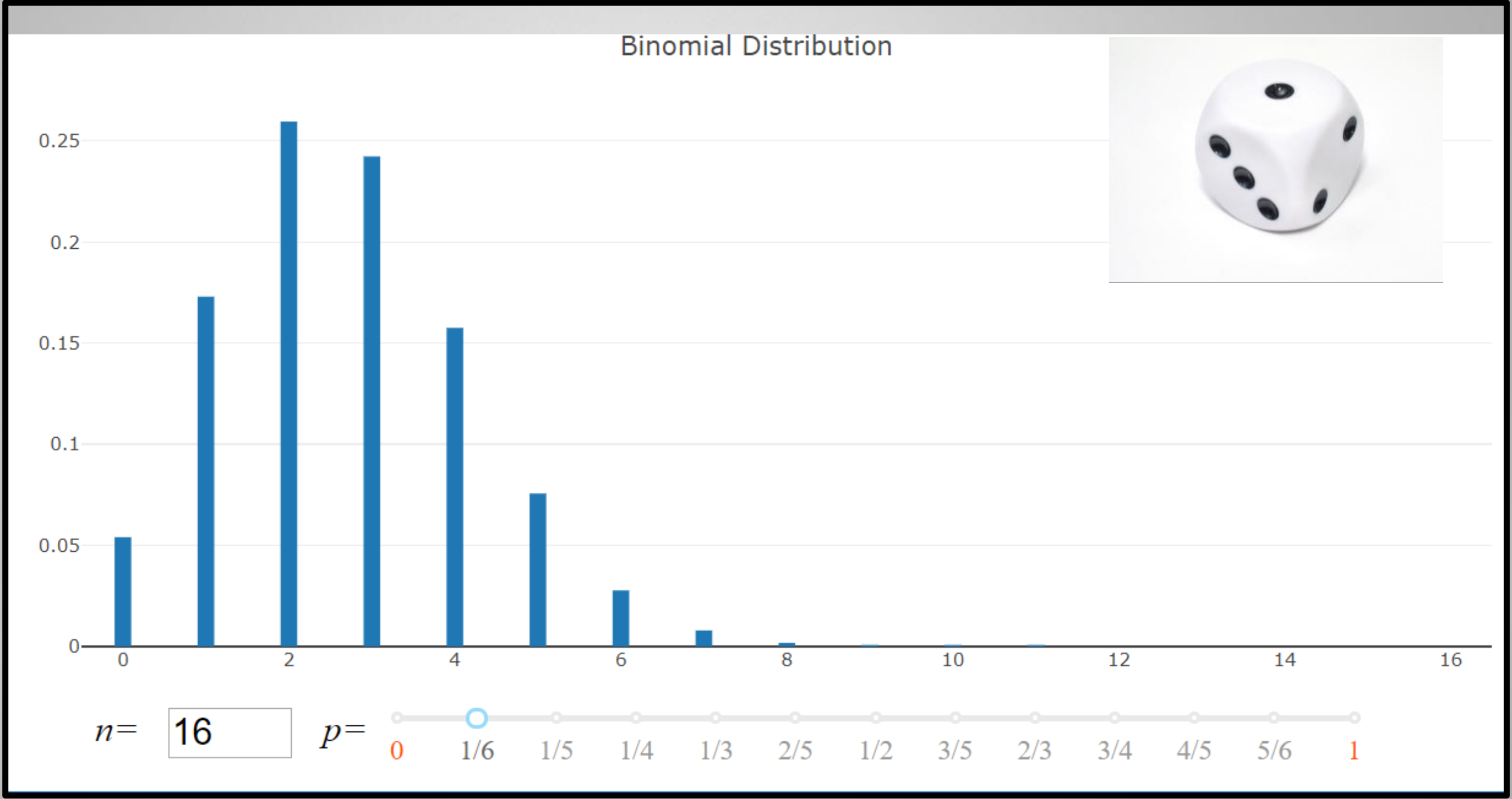
- Gives the probability of observing *x* successes in *n* trials
- The probability of success on a single trial is denoted by *p*
- Assumes that *p* is fixed for all trials

$$P(x\colon n, p) = \binom{n}{x}(p)^x(1-p)^{(n-x)}$$

Binomial Distribution

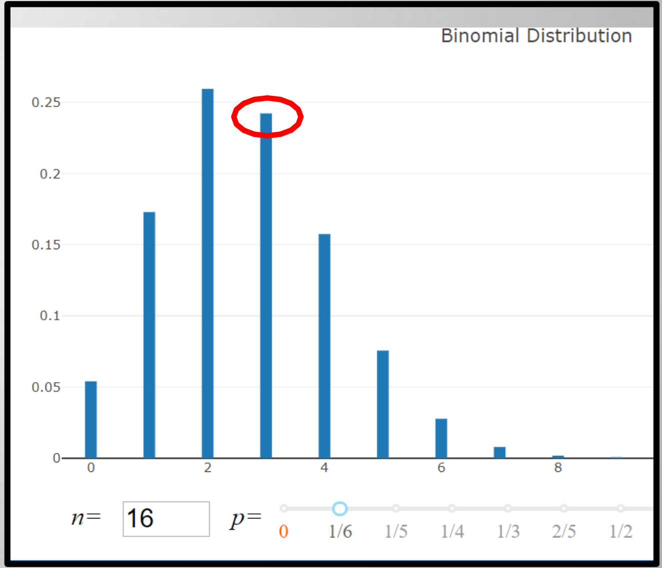


Binomial Distribution



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

$$P(x:n,p) = \binom{n}{x}(p)^x(1-p)^{(n-x)}$$

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

$$= \left(\frac{16!}{3!\,(13)!}\right)(1/6)^3(5/6)^{(13)}$$

$$= \left(\frac{16\cdot 15\cdot 14}{3\cdot 2}\right)\left(\frac{1^3}{6^3}\right)\left(\frac{5^{13}}{6^{13}}\right) = 0.242$$

Poisson Distribution

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"

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

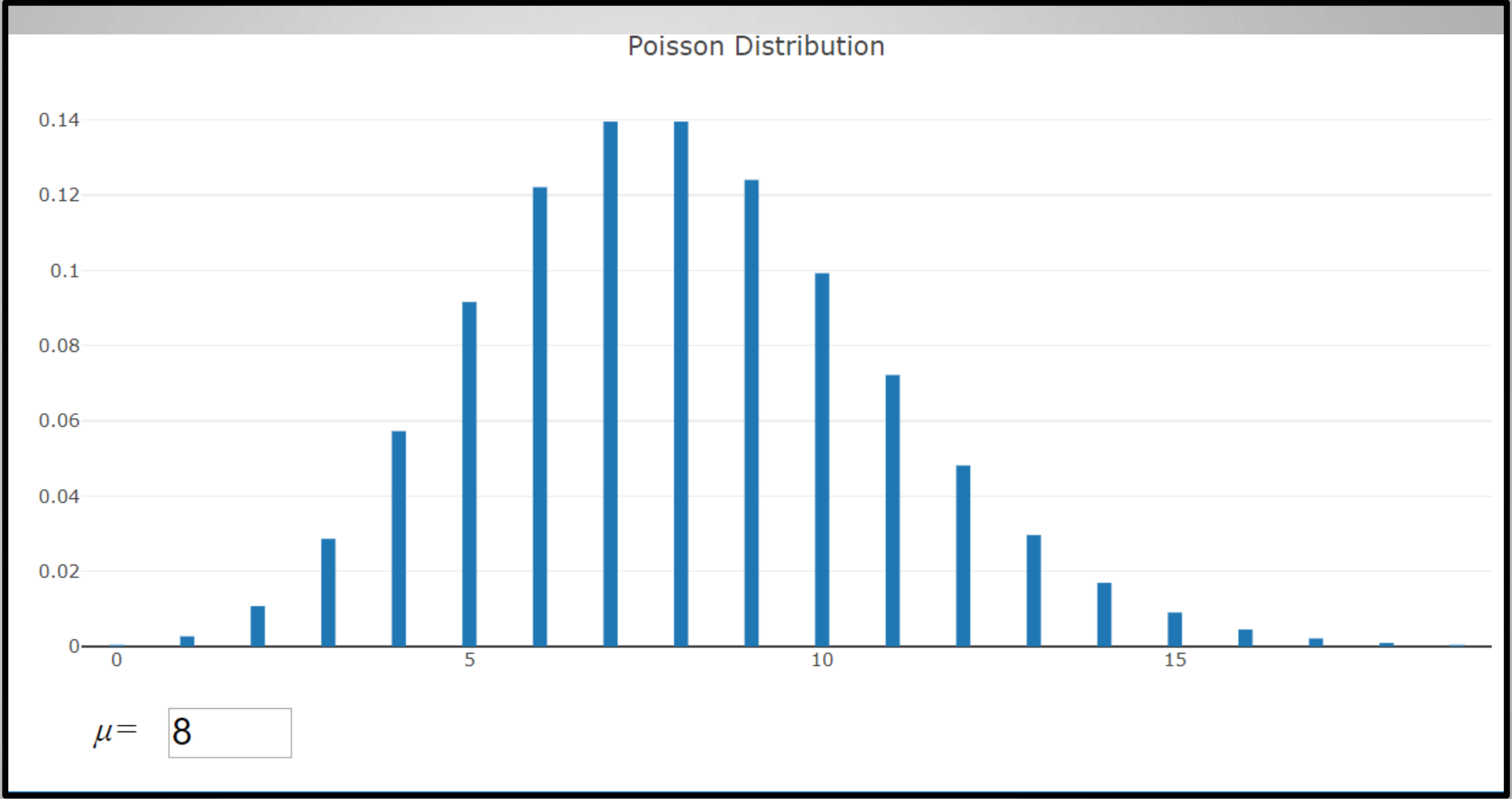
Poisson Distribution

- The equation becomes

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

where $e = Euler's\ number = 2.71828 \ldots$

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

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

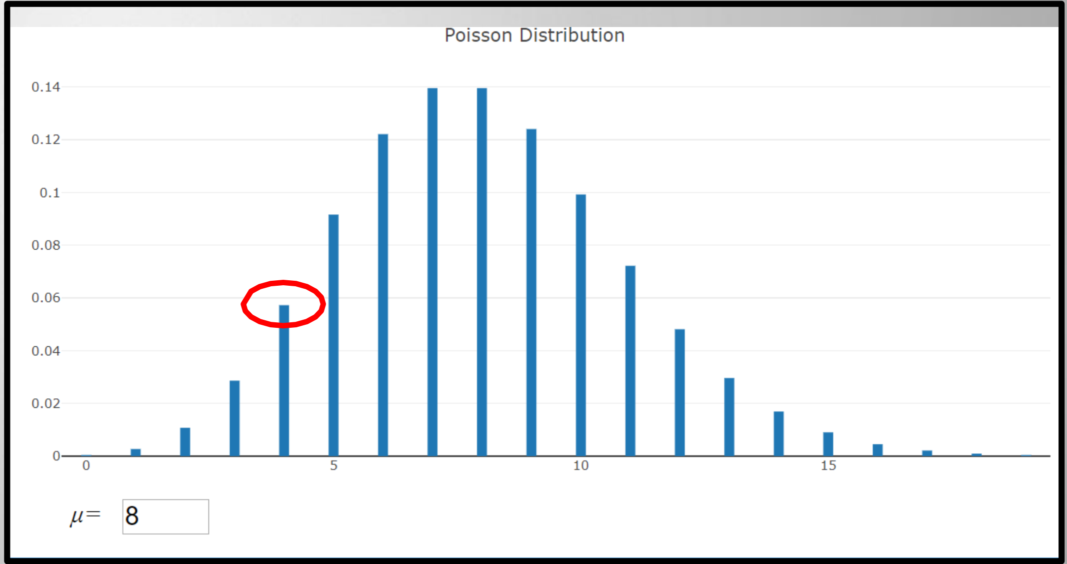
$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{8^4 \cdot 2.71828^{-8}}{4!}$$

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

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:

$$P(X: x < 4) = \sum_{i=0}^{3} \frac{\lambda^i e^{-\lambda}}{i!}$$

$$= \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!}$$

Poisson Distribution

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

$$P(X:x \geq 1) = 1 - P(X:x = 0)$$

$$= 1 - \frac{\lambda^0 e^{-\lambda}}{0!} = 1 - e^{-\lambda}$$

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

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

$$= \frac{8^0 \cdot 2.71828^{-8}}{0!} + \frac{8^1 \cdot 2.71828^{-8}}{1!} + \frac{8^2 \cdot 2.71828^{-8}}{2!}$$

$$= \frac{1 \cdot \left(\frac{1}{2980.96}\right)}{1} + \frac{8 \cdot \left(\frac{1}{2980.96}\right)}{1} + \frac{64 \cdot \left(\frac{1}{2980.96}\right)}{2}$$

$$= \mathbf{0.0137}$$

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 λ 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

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

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

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{0.67^0 \cdot 2.71828^{-0.6667}}{0!}$$

$$= \mathbf{0.5134}$$

Continuous Probability Distributions

Continuous Distributions

- Continuous probability distributions are also called *probability density functions*:

Normal Distribution

Exponential Distribution

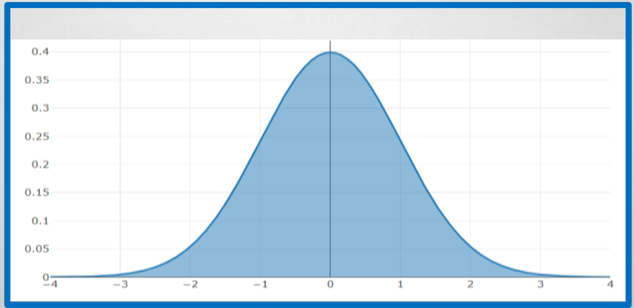
Beta Distribution

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

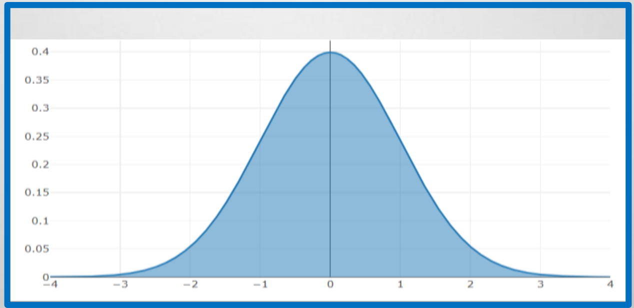
- These data sources tend to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



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.

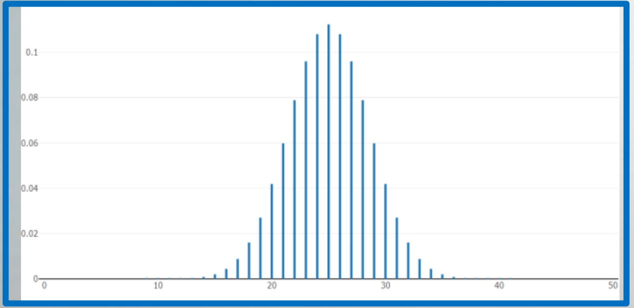


Normal Distribution

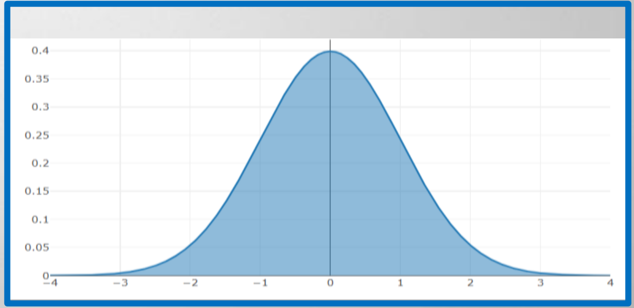
Normal Distribution

- 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

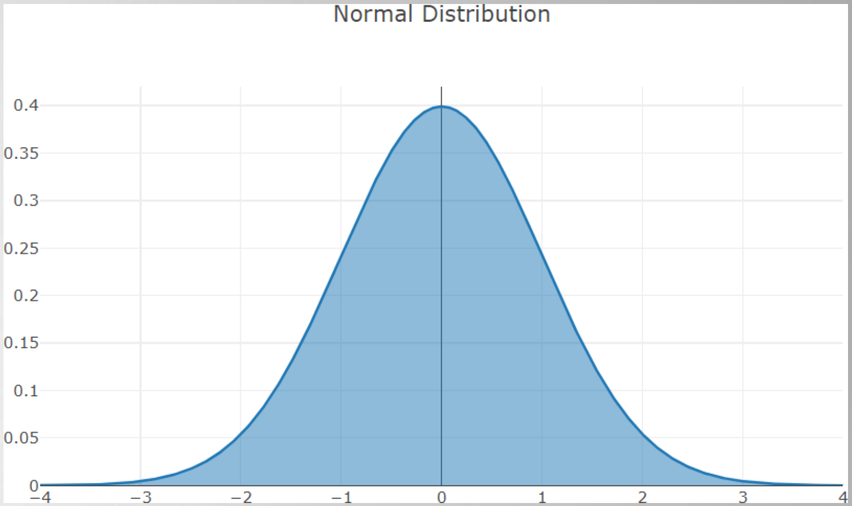


Normal Distribution

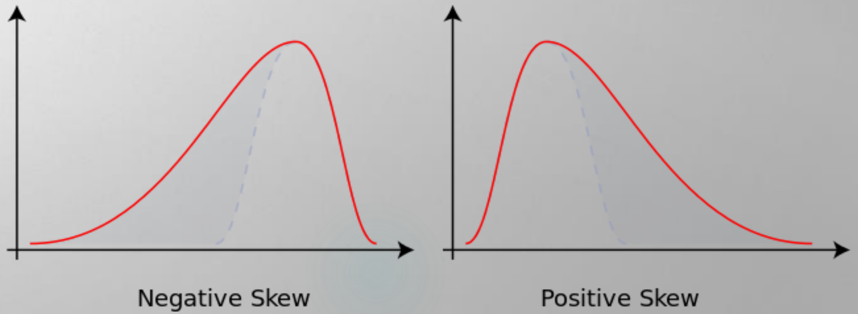


Normal Distribution

- also called the **Bell Curve** or **Gaussian Distribution**
- always symmetrical

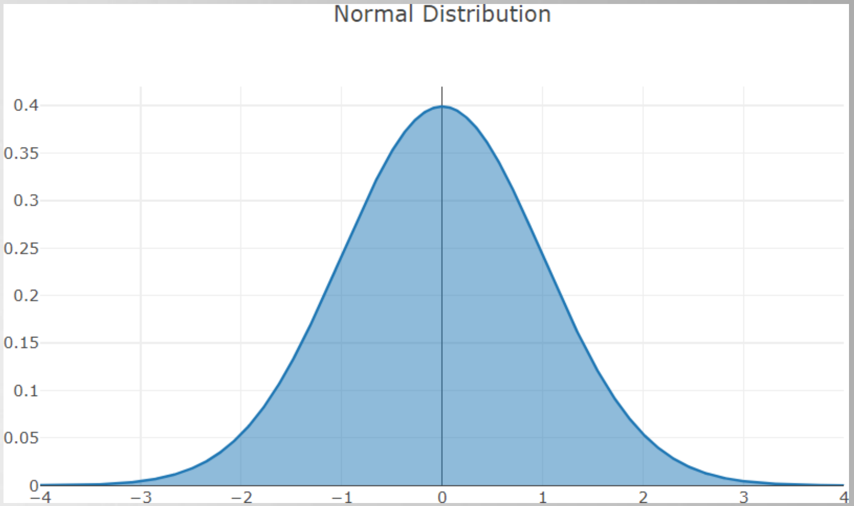


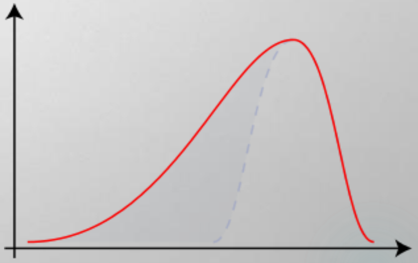
asymmetrical curves display **skew** and are *not* normal



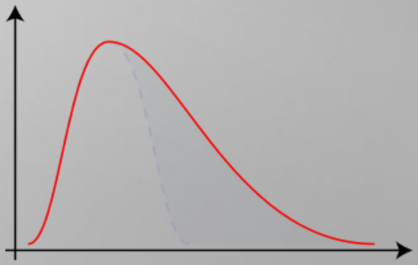
Normal Distribution

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



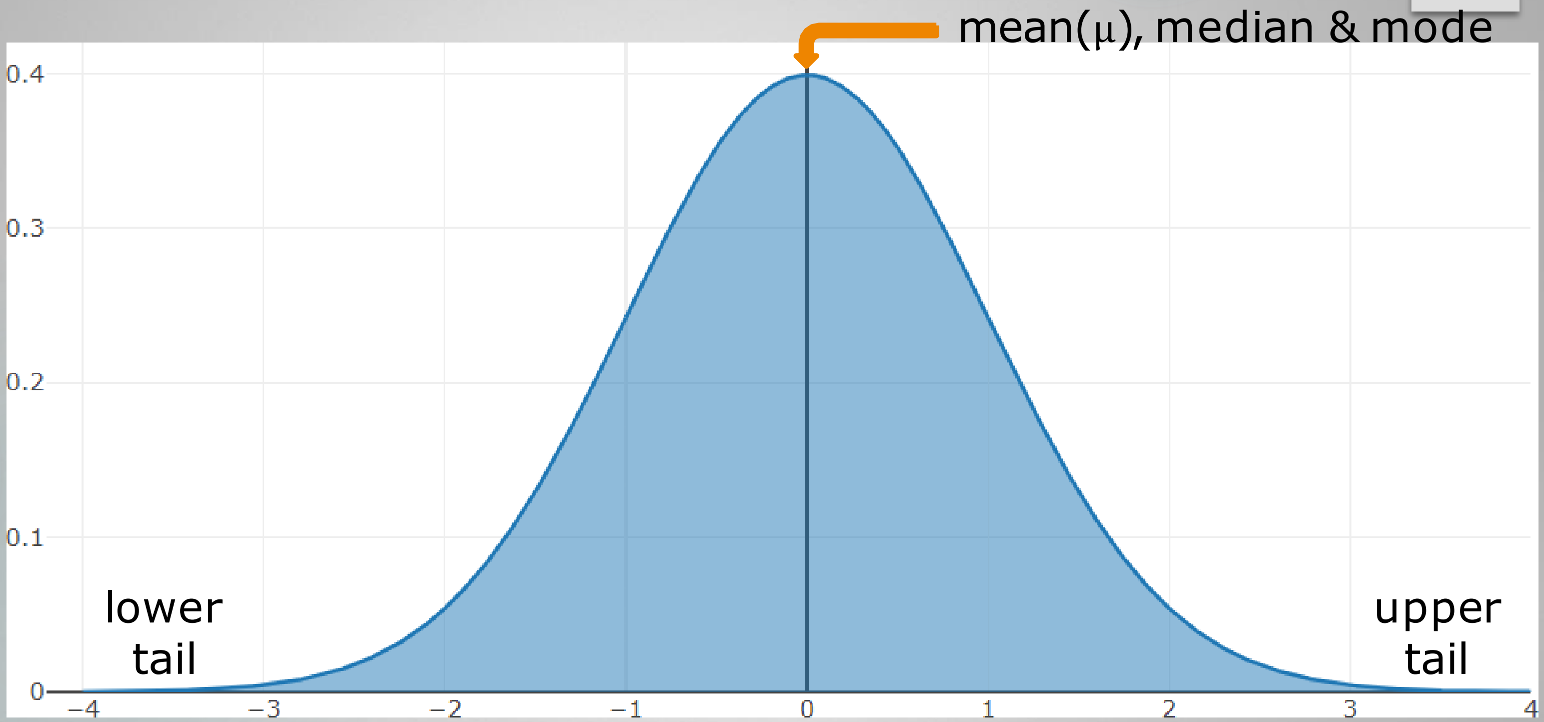


Negative Skew

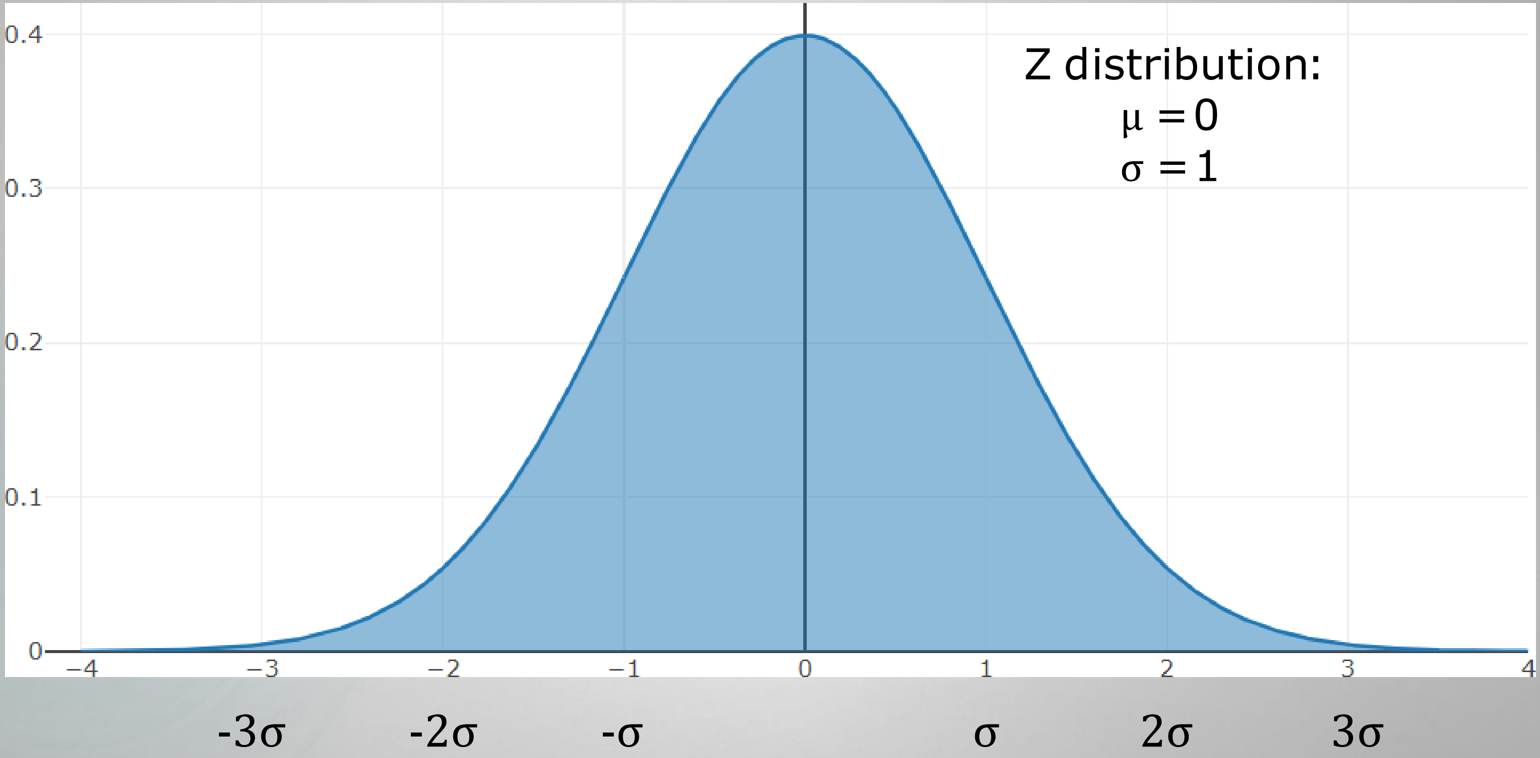


Positive Skew

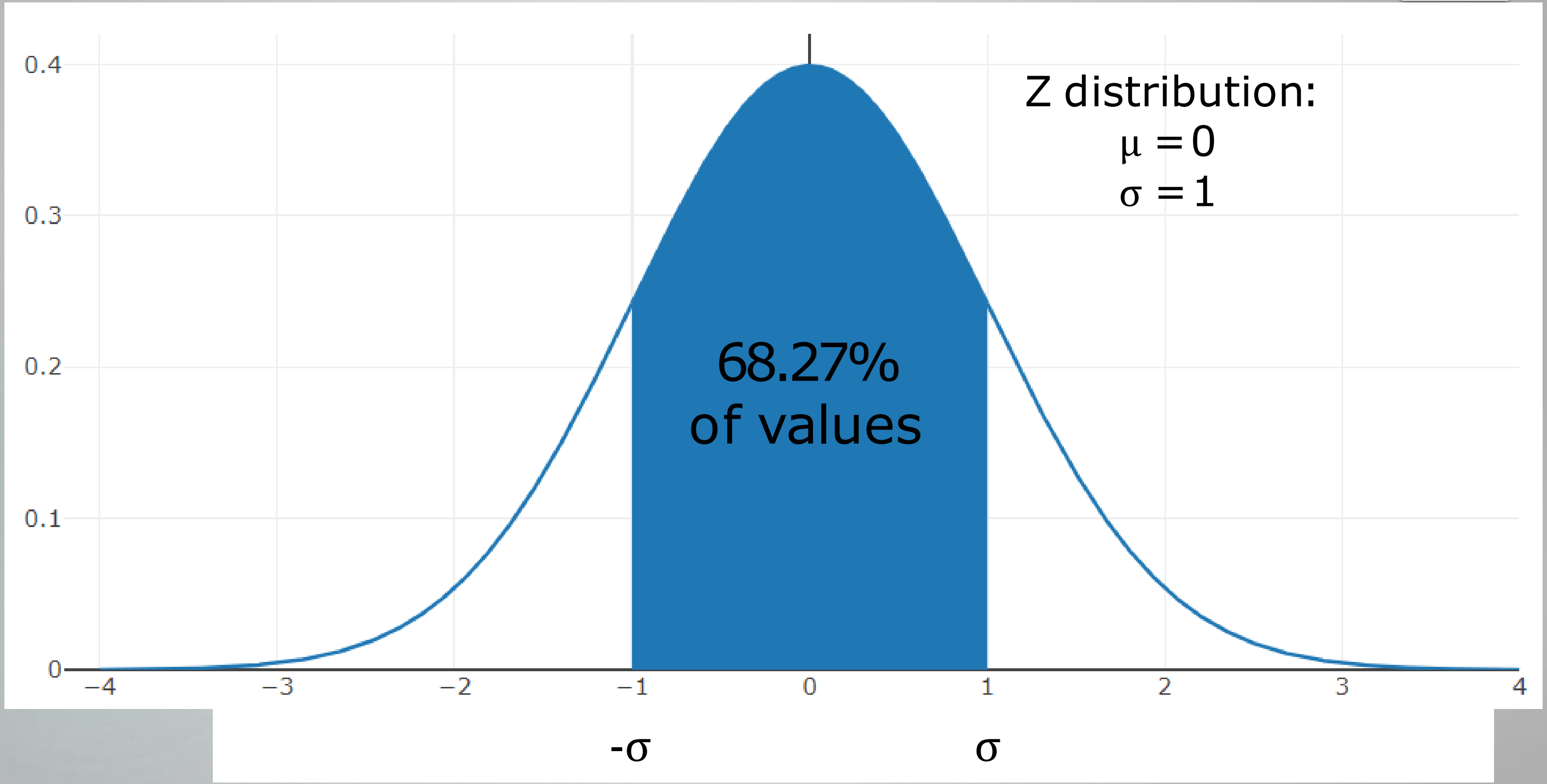
Normal Distribution



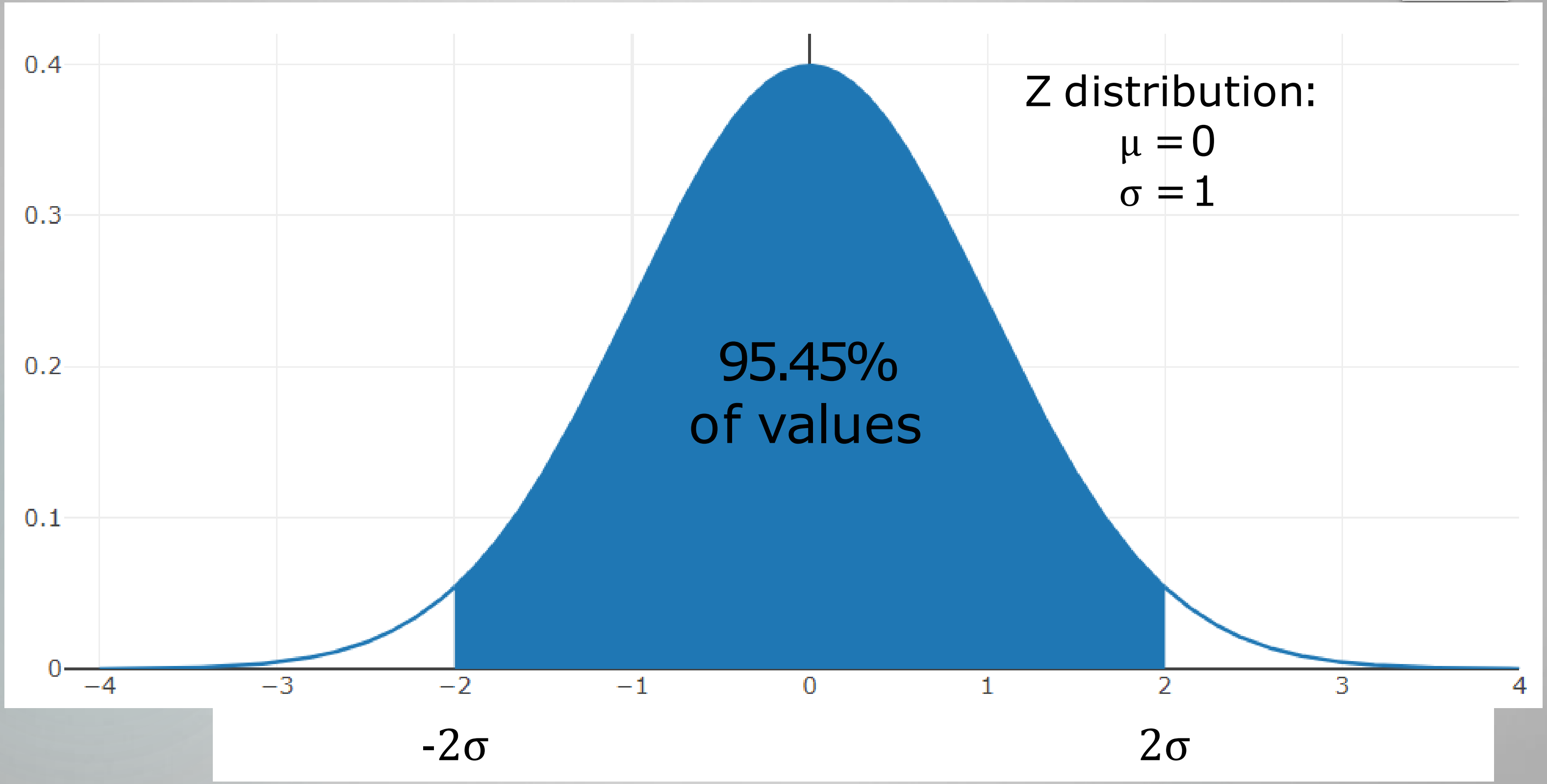
Standard Normal Distribution



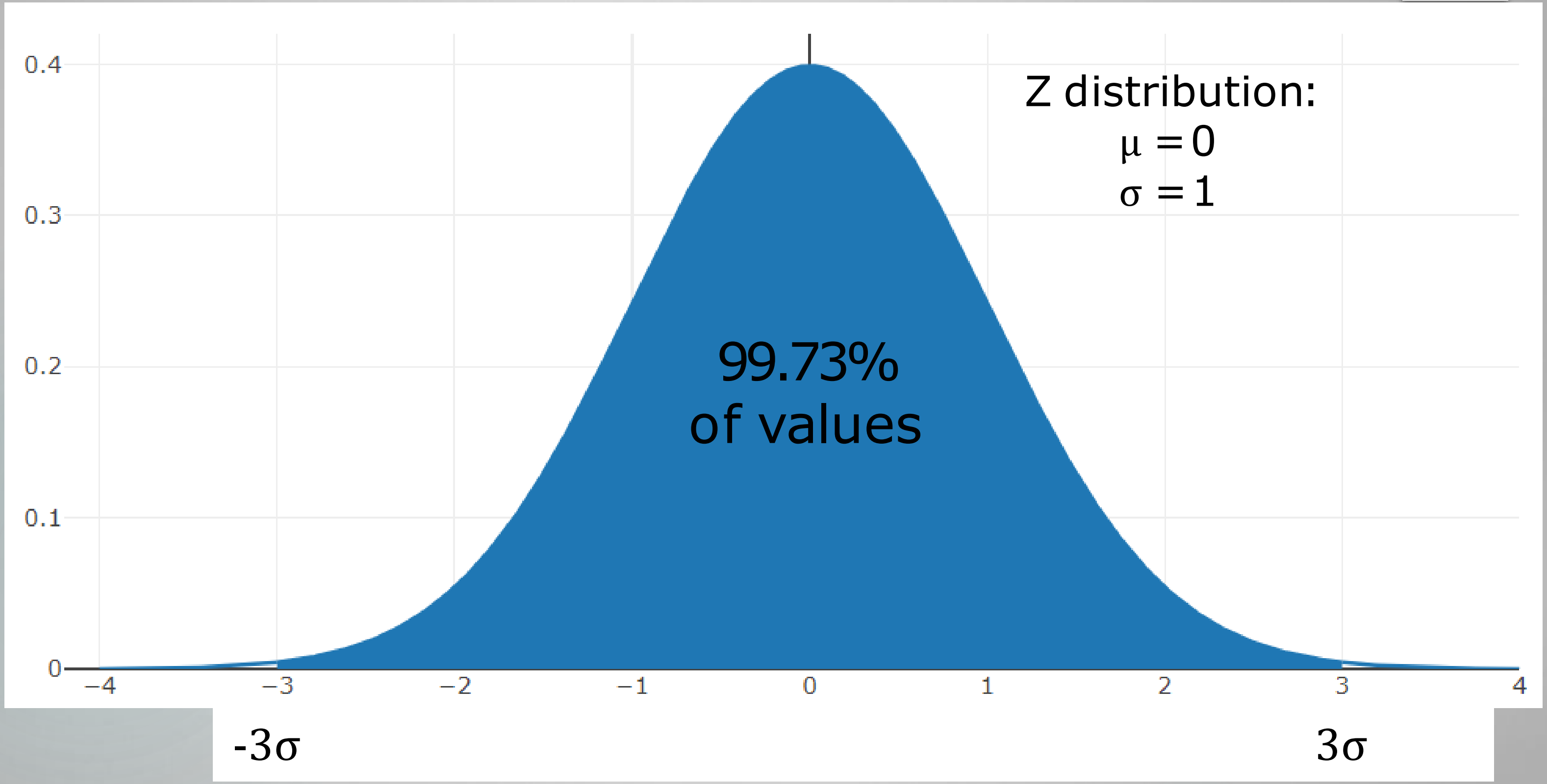
Standard Normal Distribution



Standard Normal Distribution



Standard Normal Distribution

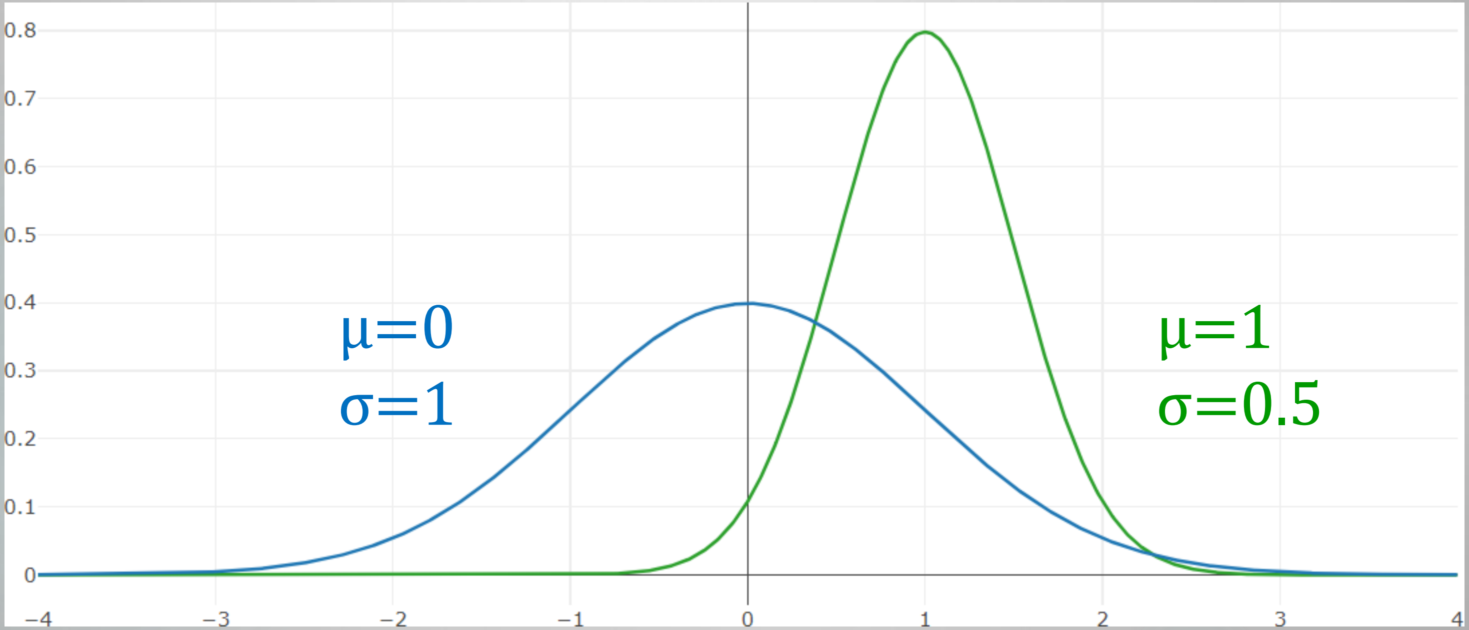


Normal Distribution

- 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 σ 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 Formulas and Z Scores

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, than 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:

μ = mean

σ = standard deviation

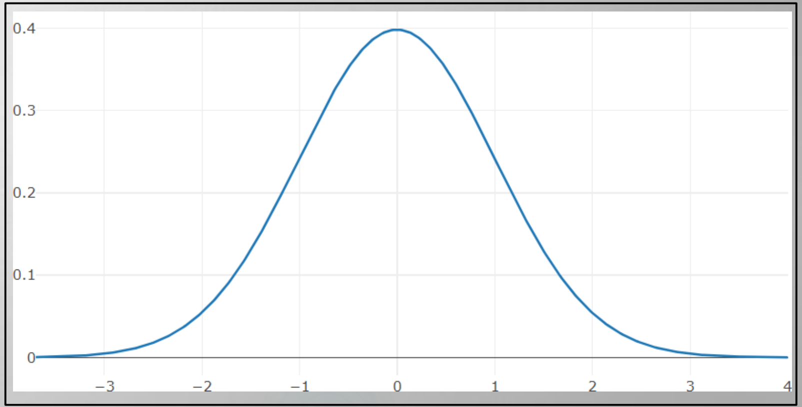
e = 2.71828

π = 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:

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

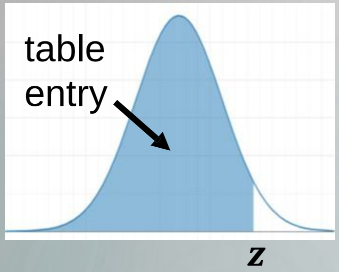
- We can then determine x's *percentile* by looking at a z-table

How to Look Up Z-Scores

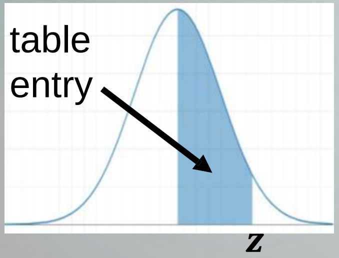
- A z-table of Standard Normal Probabilities maps a particular z-score to the area under a normal distribution curve to the left of the score.
- Since the total area under the curve is 1, probabilities are bounded by 0 and 1

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-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 Exercise Solution

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

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

$$= \frac{87 - 75}{7} = \mathbf{1.7143}$$

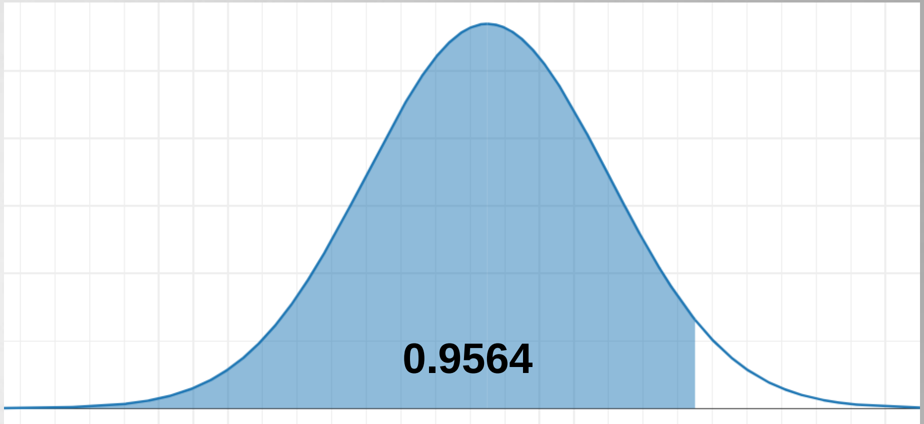
Z-Score Exercise Solution

- 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.

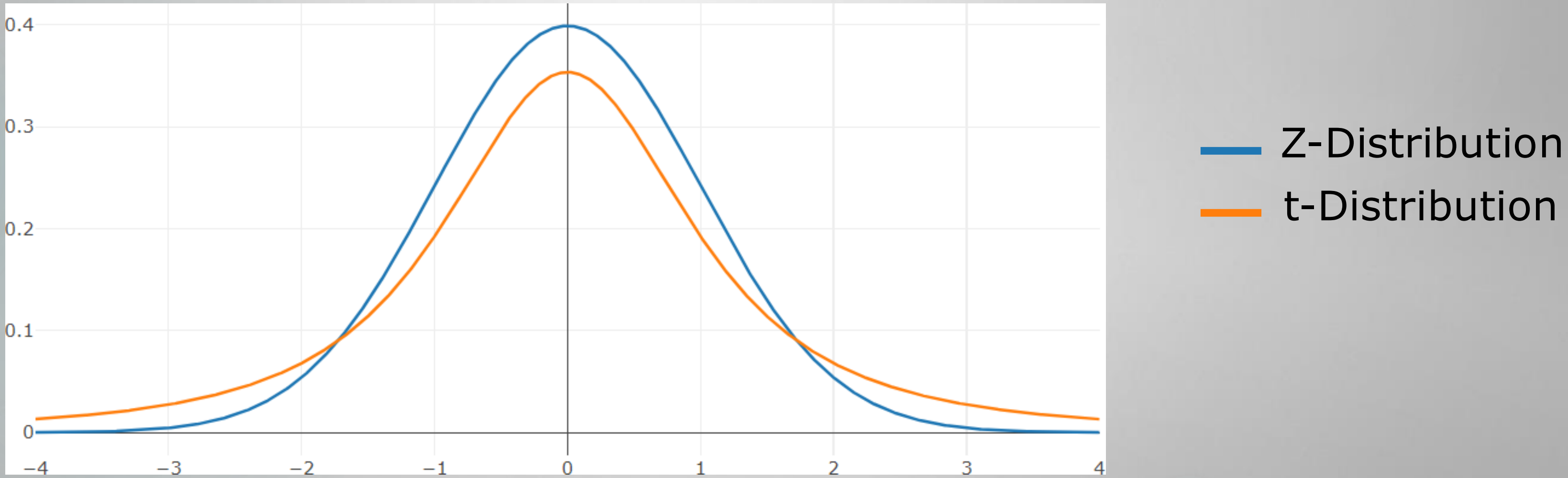
Student's T-Distribution

Student's T-Distribution

- Developed by William Sealy Gossett while he was working at Guinness Brewery
- Published under the pseudonym "Student" as Guinness wouldn't let him use his name.
- Goal was to select the best barley from small samples, when the population standard deviation was unknown!

Student's t-Distribution

- t-Distributions have fatter tails than normal Z-Distributions



Student's t-Distribution Formula

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where;

v is the degree of freedom (sample size – 1)

π is 3.14285

Γ: Gamma function (generalization of factorial).

Student's t-Distribution Characteristics

Symmetrical: Similar to the normal distribution but with heavier tails.

Degrees of Freedom (v):

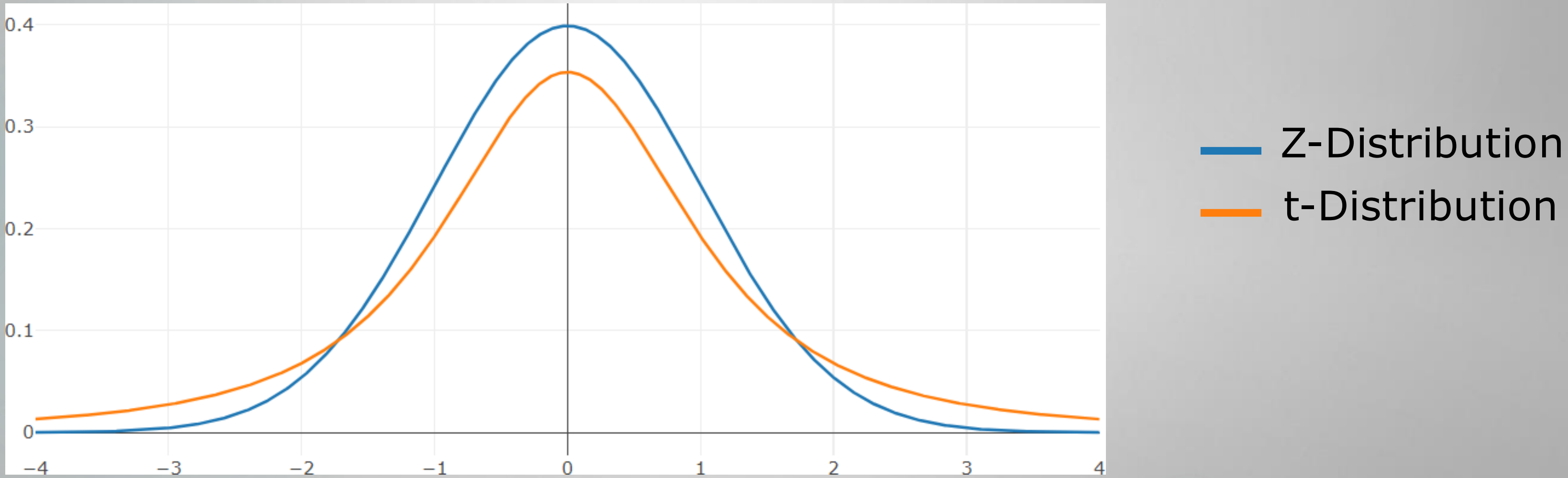
- The shape of the distribution depends on v.
- As $v \rightarrow \infty$, it approaches the standard normal distribution.

Applications:

Used for hypothesis testing when the population standard deviation is unknown, particularly in small samples.

Student's t-Distribution

- They approach a normal distribution as the degrees of freedom increase.



Probability Distribution implementation in R

d Function: Probability Density or Mass Function

Returns the likelihood of a value occurring.

Usage: dnorm(x, mean, sd)
for Normal Distribution.

p Function: Cumulative Probability Function

Returns the probability of a value being less than or equal to a given number.

Example: pnorm(x, mean, sd)

q Function: Quantile Function

Returns the value corresponding to a specific cumulative probability.

Example: qnorm(prob, mean, sd).

r Function: Random Sampling

Generates random numbers from a specified distribution.

Example: rnorm(n, mean, sd).

Binomial Distribution

Usage in R:

- dbinom(x, size = 1, prob) → PMF.
- pbinom(x, size = 1, prob) → CDF.
- qbinom(prob, size = 1, prob) → Quantiles.
- rbinom(n, size = 1, prob) → Random samples.

Poisson Distribution

Usage in R:

- dpois(x,lambda) → PMF.
- ppois(x,lambda) → CDF.
- qpois(prob, lambda) → Quantiles.
- rpois(n, lambda) → Random samples.

Normal Distribution

Usage in R:

- dnorm(x,mean,sd) → PMF.
- pnorm(x,mean,sd) → CDF.
- qnorm(prob,mean,sd) → Quantiles.
- rnorm(n,mean,sd) → Random samples.

t-Distribution

Usage in R:

- dt(x,df) → PMF.
- pt(x,df) → CDF.
- qt(prob,df) → Quantiles.
- rt(n,df) → Random samples.