Revised: 3/1/2021

# Probability Distributions I

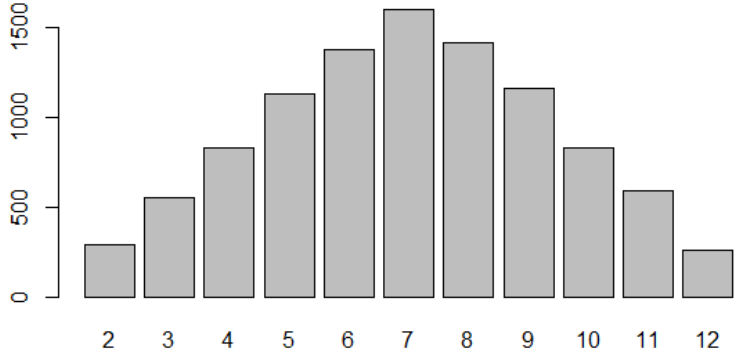
Michael C. Hackett

Assistant Professor, Computer Science

Community College of Philadelphia

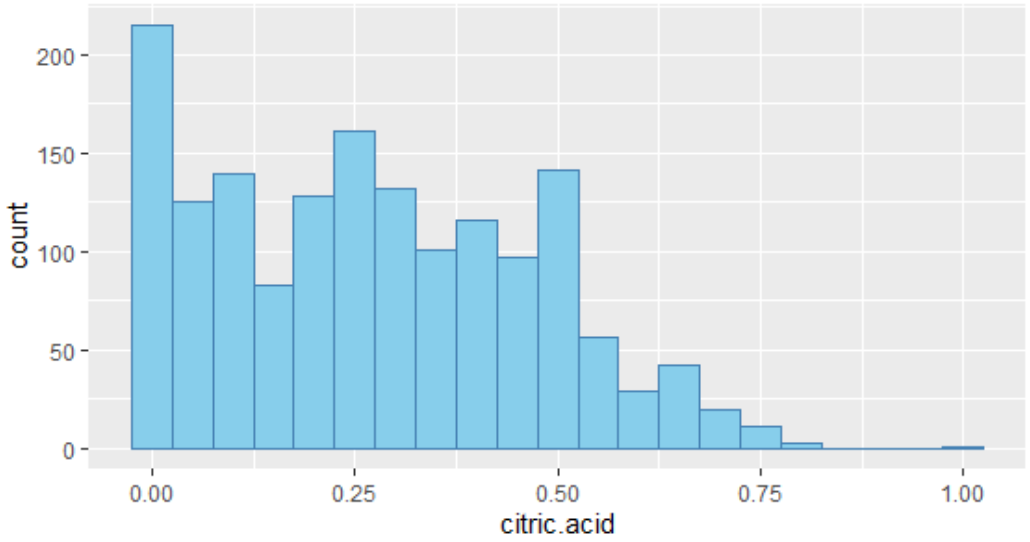
# Continuous Distributions

- In previous lectures, we worked with discrete numerical variables
  - The side of a coin flip or a number on the Roulette wheel, for example.
  - We could not have outcomes like a sum of 6.5 when rolling dice or landing on 3.99999 on a Roulette wheel.



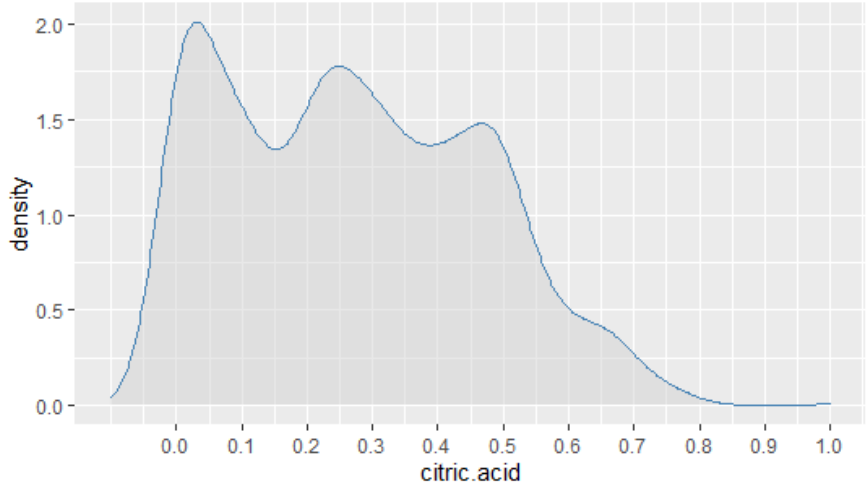
# Continuous Distributions

- Below is a plot from a data set that contains information related to red variants of the Portuguese "Vinho Verde" wine.
  - It plots the citric acid, between 0 (least citric) and 1 (most citric), using a ggplot histogram



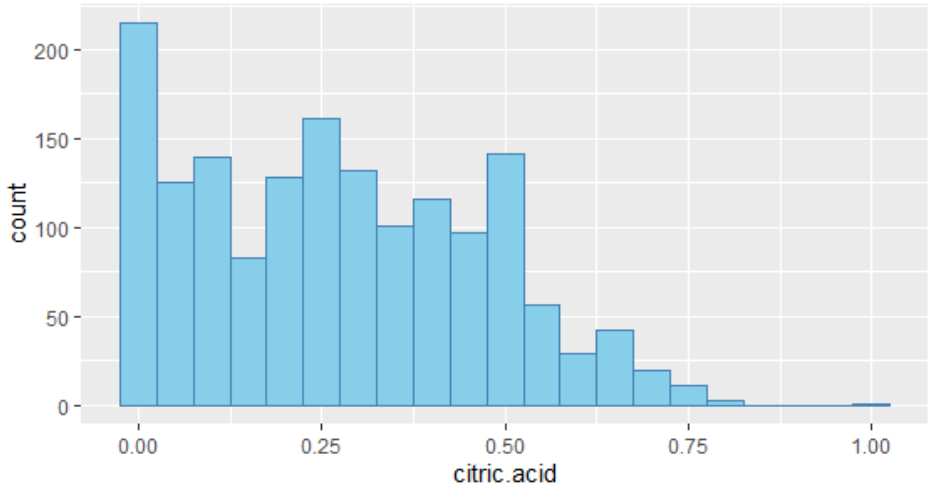
# Continuous Distributions

- Below is a **density plot** (or a **continuous distribution**) of the same citric acid data.
  - See the posted CSCI 118 Data Visualization I slides (particularly the introduction to ggplot2 and the section on Visualizing Distributions) and Module download/sample code

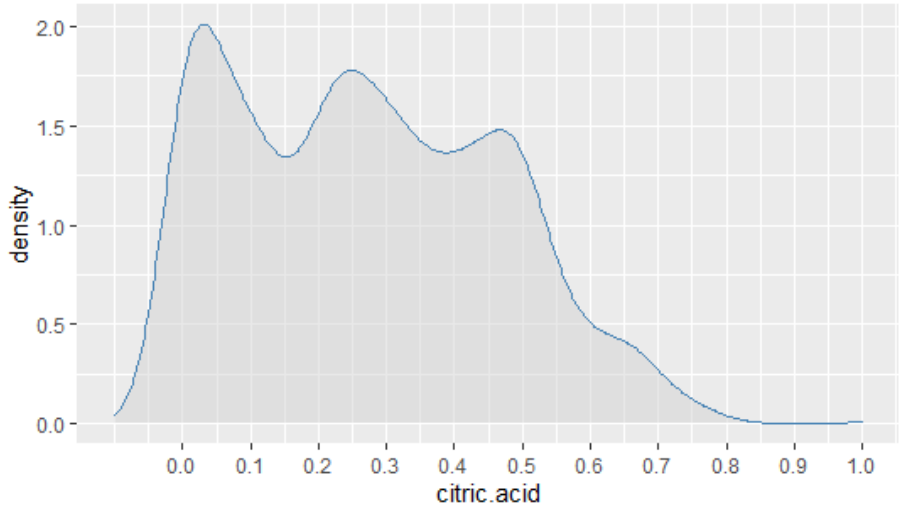


# Continuous Distributions

- Like histograms, density plots also visualize numerical distributions
- Unlike histograms, density plots visualize continuous numerical data



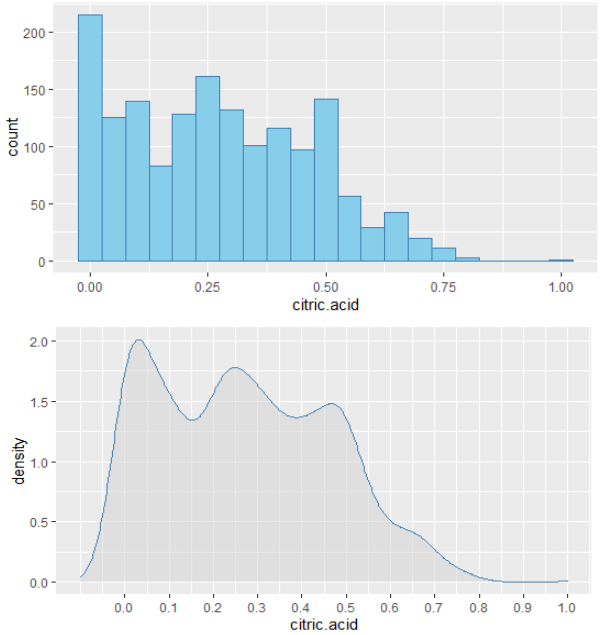

Discrete numerical

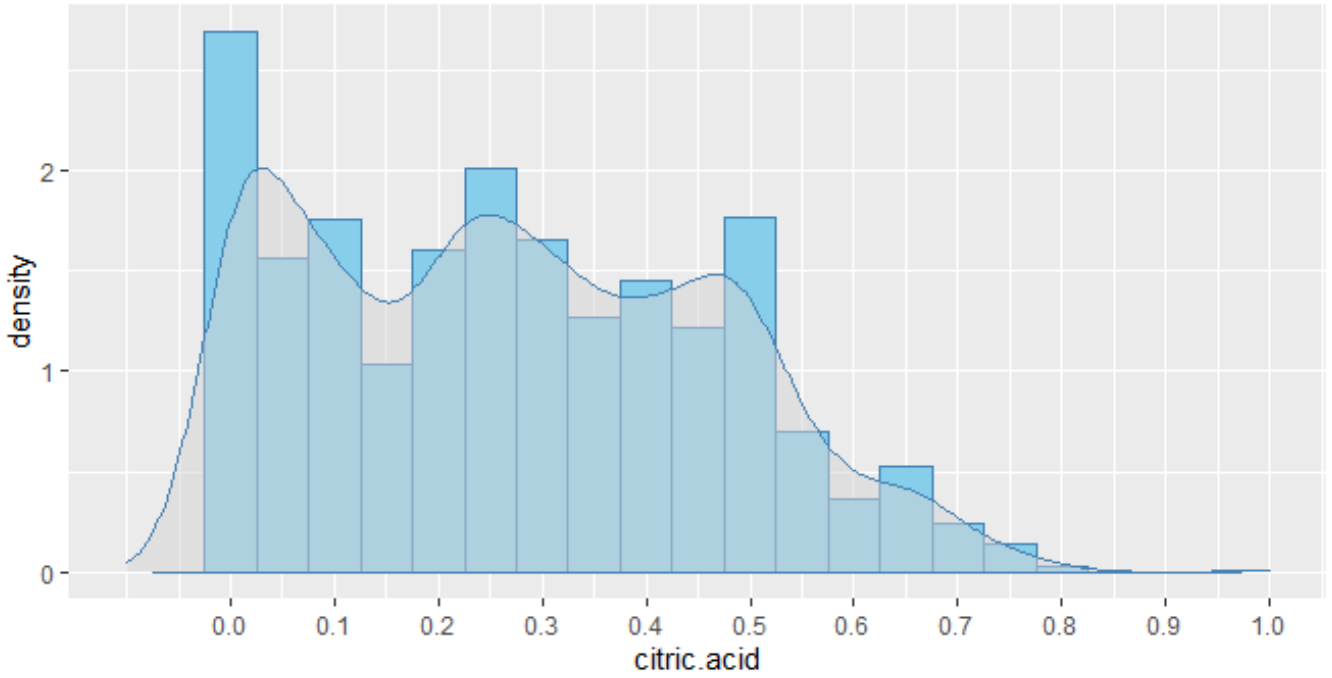



Continuous numerical

# Continuous Distributions

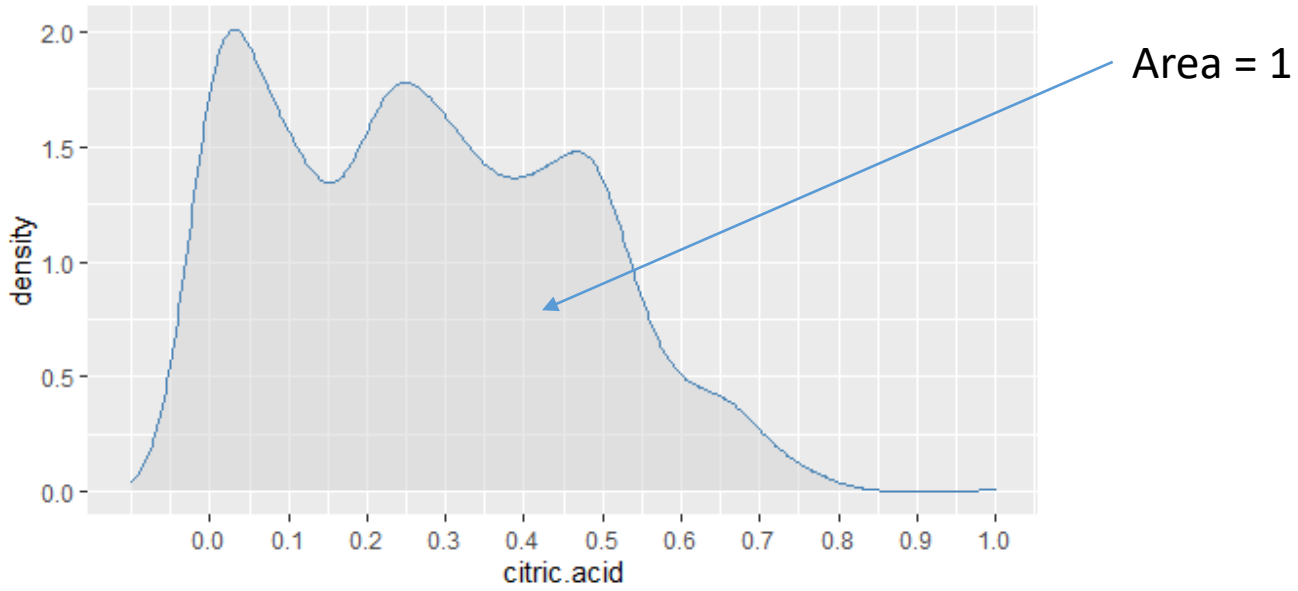
- Density plots show where the data is concentrated
  - The previous histogram with a density plot overlaid



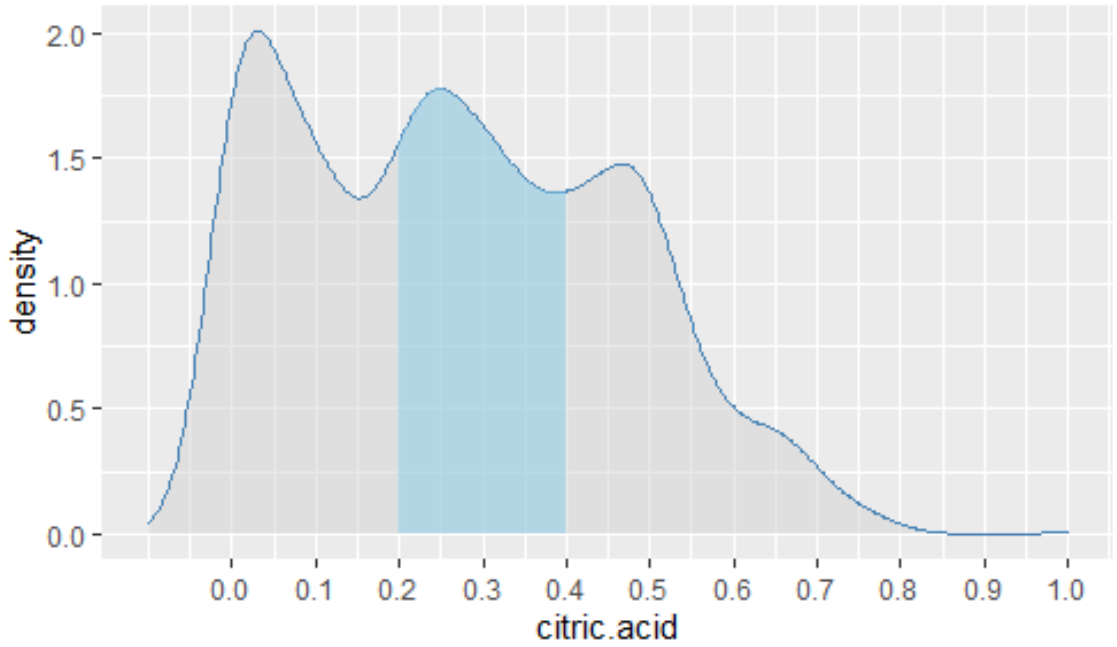
# Continuous Distributions

- The area under the curve of a continuous distribution is equal to 1
  - Typically, finding the area or (sections of the area) under a curve requires integration.



# Continuous Distributions

- For example, the *area* of the shaded section below represents the probability of a sample having a citric acid value between .2 and .4



# Continuous Distributions

- Finding this area normally requires calculus
  - But since we have the dataset, we can easily subset the observations in this range to find the probability:

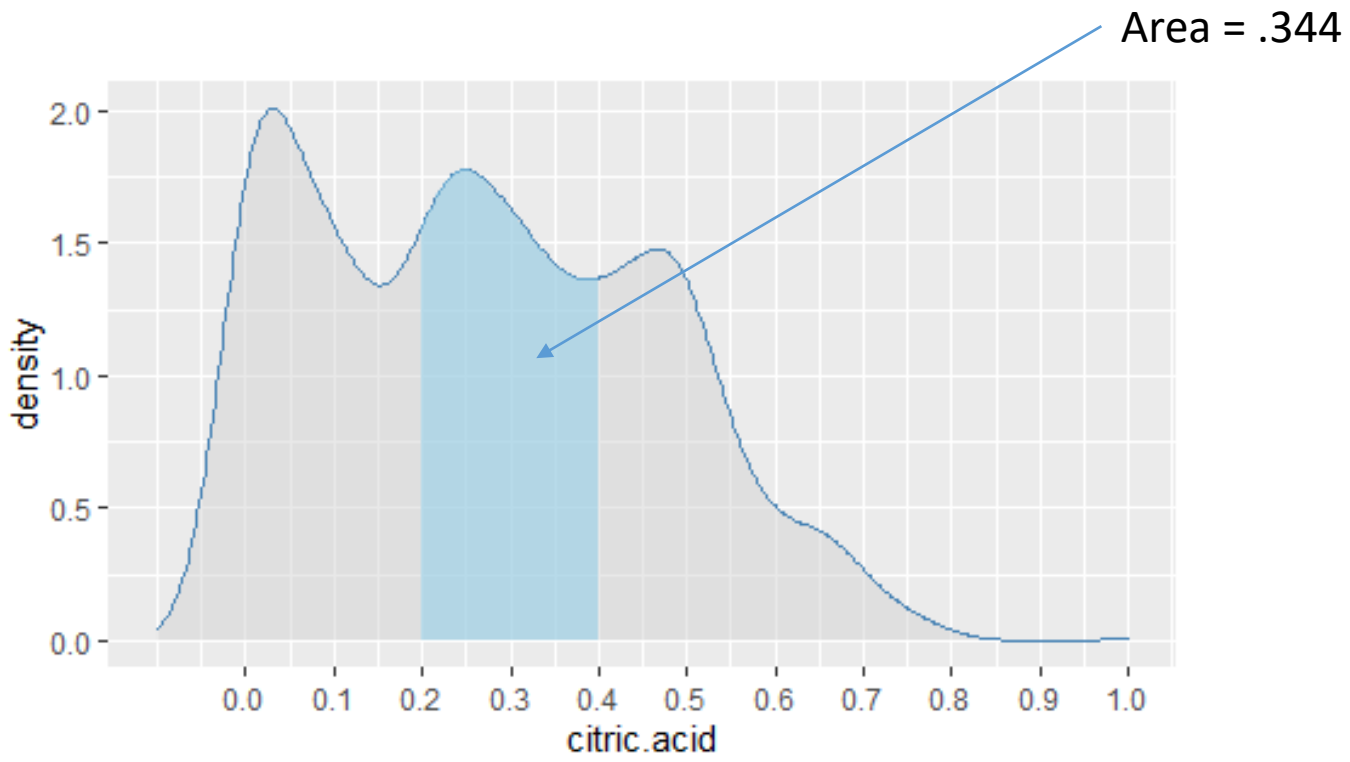
$$P = \frac{Total\ observations\ with\ citric\ acid\ value\ between\ .2\ and\ .4}{Total\ observations}$$

```
count(subset(winedata, winedata$citric.acid >= .2 & winedata$citric.acid <= .4)) / count(winedata)
```

```
> count(subset(winedata, winedata$citric.acid >= .2 & winedata$citric.acid <= .4))/count(winedata)
         n
1 0.343965
```

$$P = .344 = 34.4\%$$

# Continuous Distributions



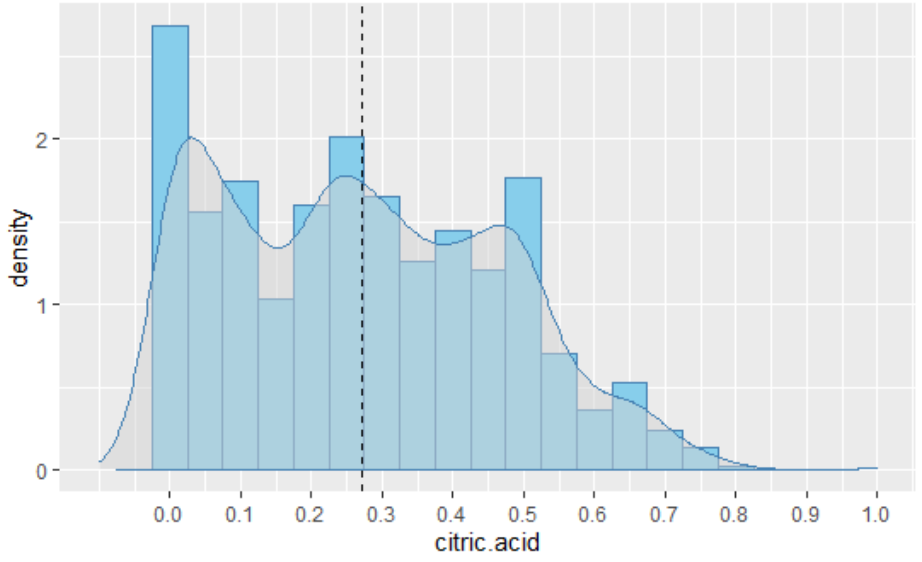
# Continuous Distributions

- The mean of this distribution is 0.271

  **`mean(winedata$citric.acid)`** or

  **`summary(winedata$citric.acid)`**

```
> summary(winedata$citric.acid)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.090   0.260   0.271   0.420   1.000
```

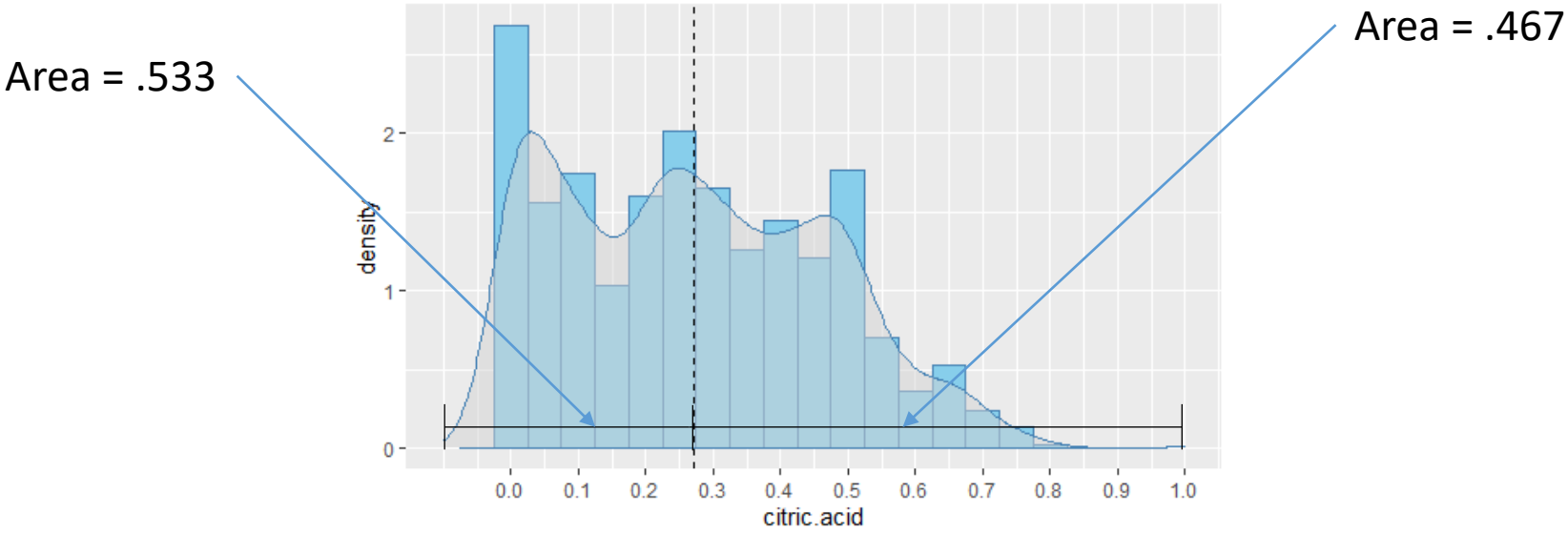


# Continuous Distributions

- However, this will not always indicate the 50% mark

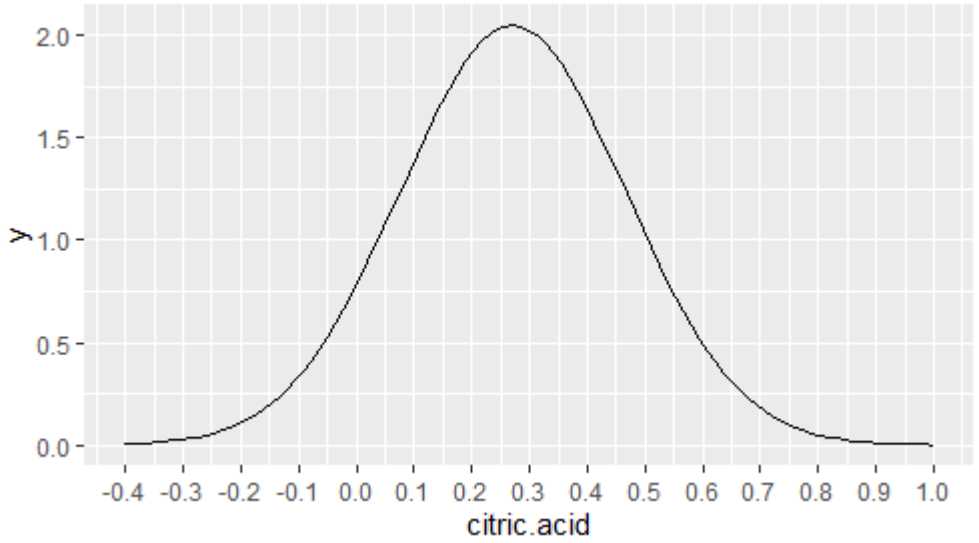
```
count(subset(winedata, winedata$citric.acid < mean(winedata$citric.acid))) / count(winedata)
count(subset(winedata, winedata$citric.acid > mean(winedata$citric.acid))) / count(winedata)
```

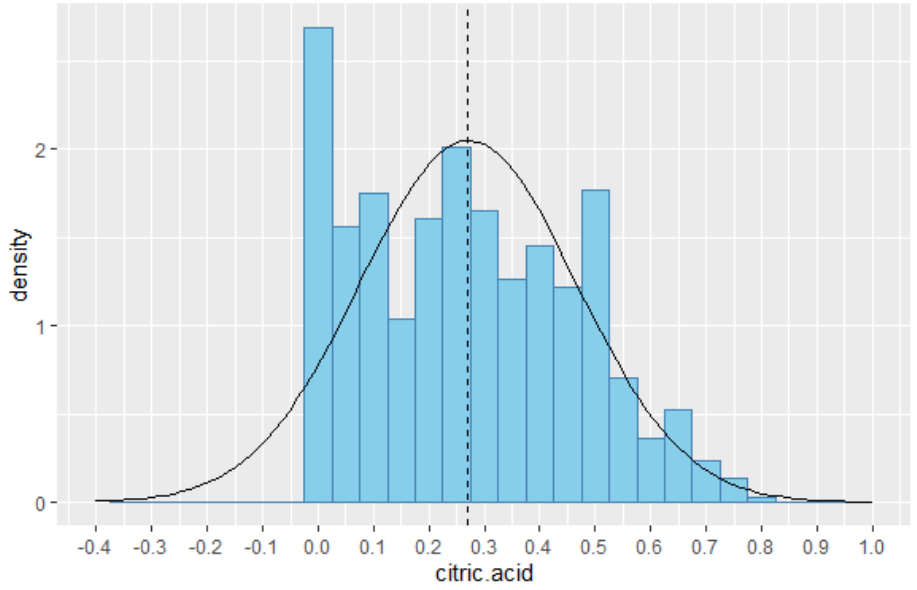
```
> count(subset(winedata, winedata$citric.acid < mean(winedata$citric.acid))) / count(winedata)
         n
1 0.532833
> count(subset(winedata, winedata$citric.acid > mean(winedata$citric.acid))) / count(winedata)
         n
1 0.467167
```



# The Normal Distribution

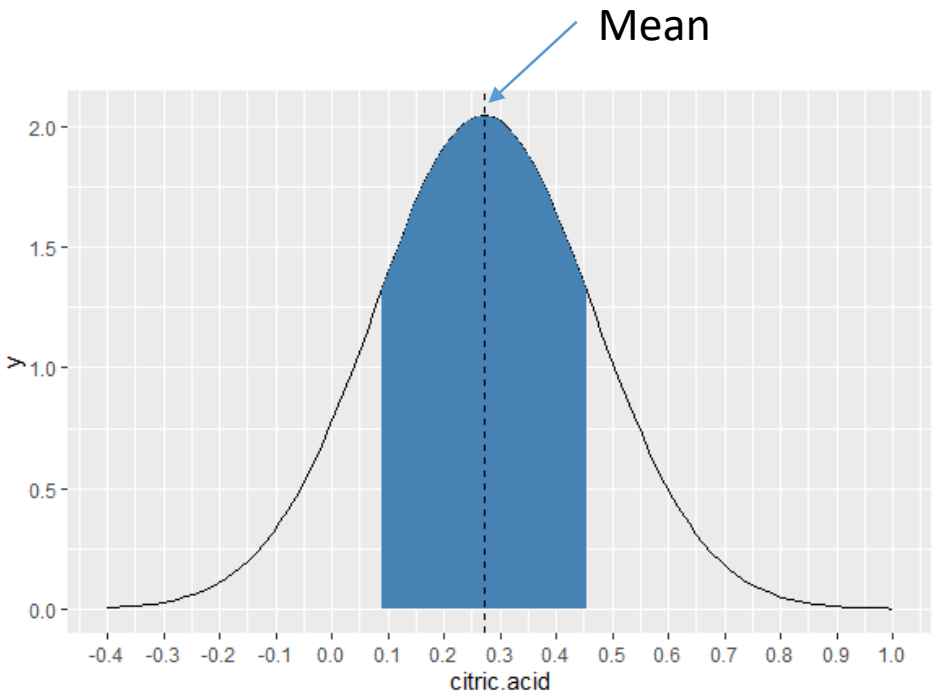
- The **normal distribution** is a continuous probability distribution model that fits a distribution to a symmetric, unimodal, bell-shaped curve



# The Normal Distribution

- One standard deviation from the mean
  - $\sigma = .195$

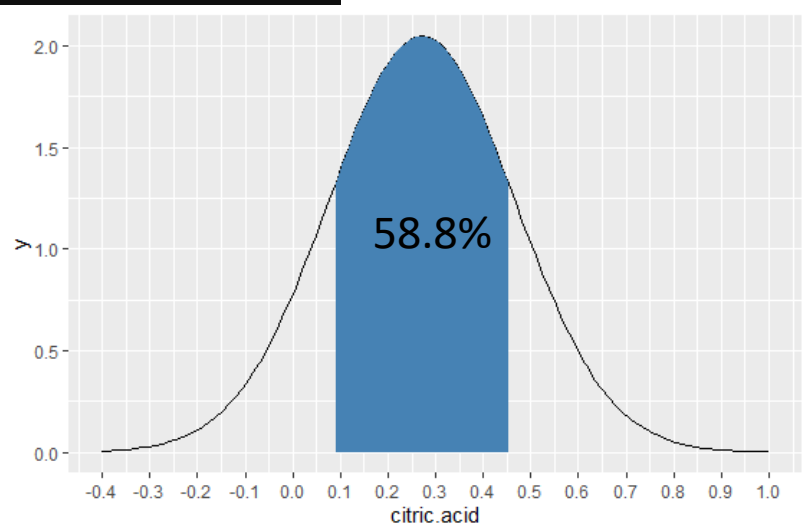


# The Normal Distribution

- Probability an observation is within one standard deviation from the mean:

```
count(subset(winedata, winedata$citric.acid >= m_citric-sd_citric &
                       winedata$citric.acid <= m_citric+sd_citric)) / count(winedata)
```

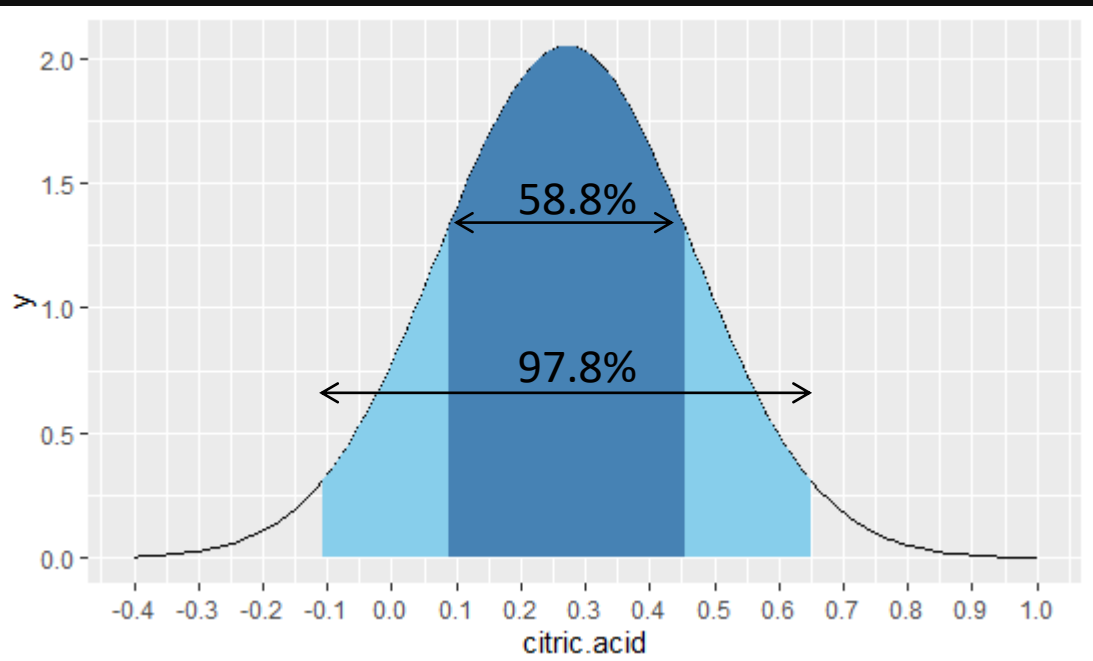
```
> count(subset(winedata, winedata$citric.acid >= m_citric-sd_citric & winedata$citric.acid <= m_citric+s
d_citric)) / count(winedata)
          n
1 0.5878674
```



# The Normal Distribution

- Two standard deviations from the mean

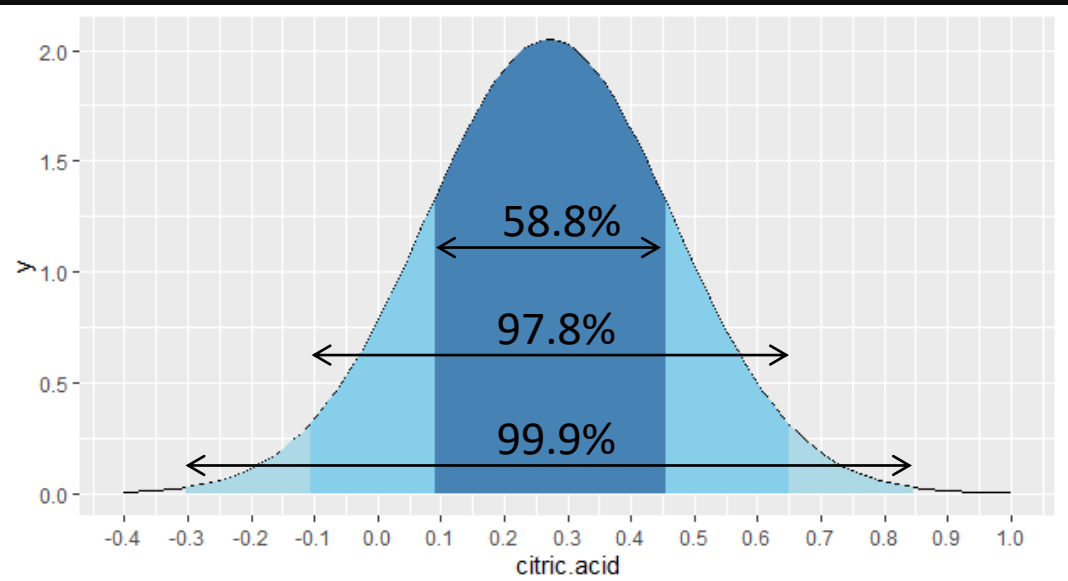
```
> count(subset(winedata, winedata$citric.acid >= m_citric-(2*sd_citric) & winedata$citric.acid <= m_citr
ic+(2*sd_citric))) / count(winedata)
          n
1 0.9781113
```



# The Normal Distribution

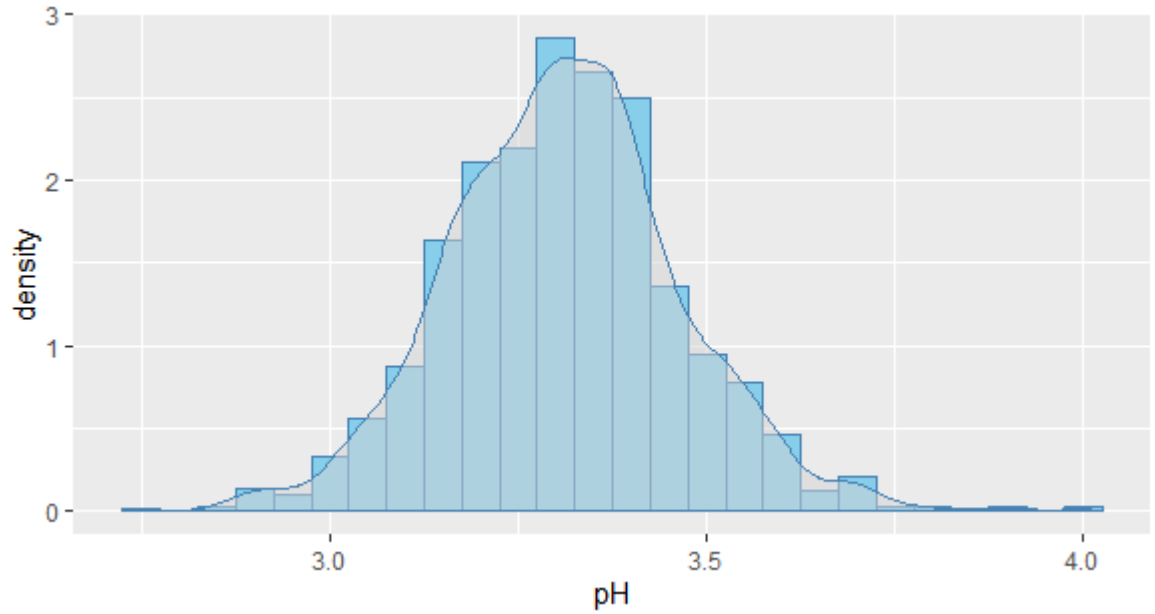
- Three standard deviations from the mean

```
> count(subset(winedata, winedata$citric.acid >= m_citric-(3*sd_citric) & winedata$citric.acid <= m_citr
ic+(3*sd_citric))) / count(winedata)
          n
1 0.9993746
```



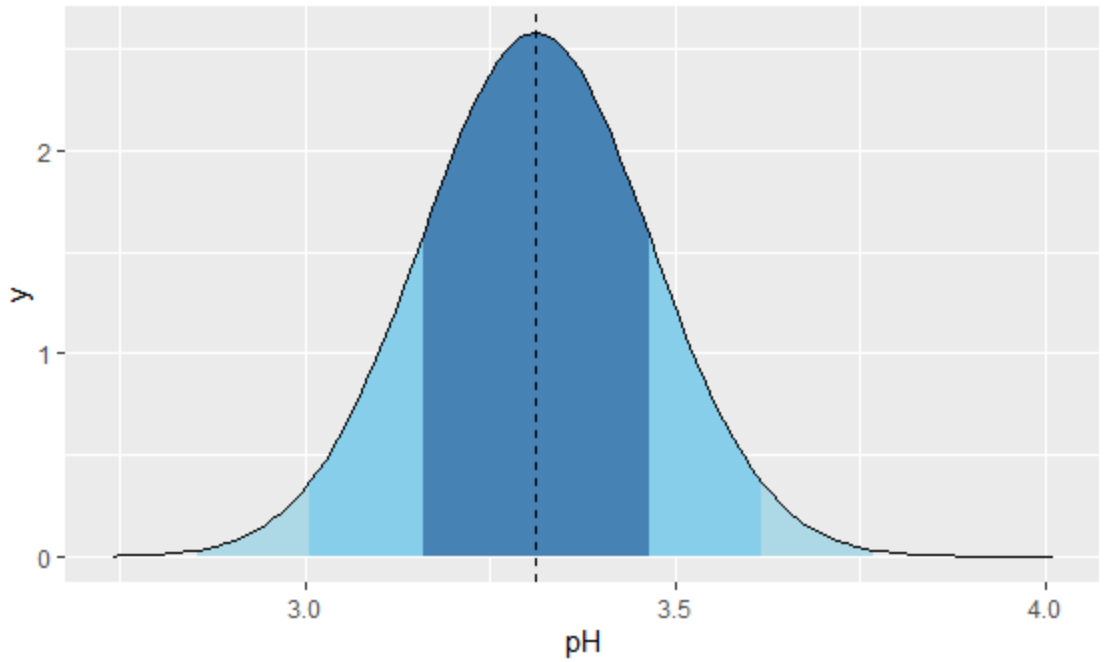
# The Normal Distribution

- As previously mentioned, the citric acid data was positively skewed
  - The pH variable is more symmetrical



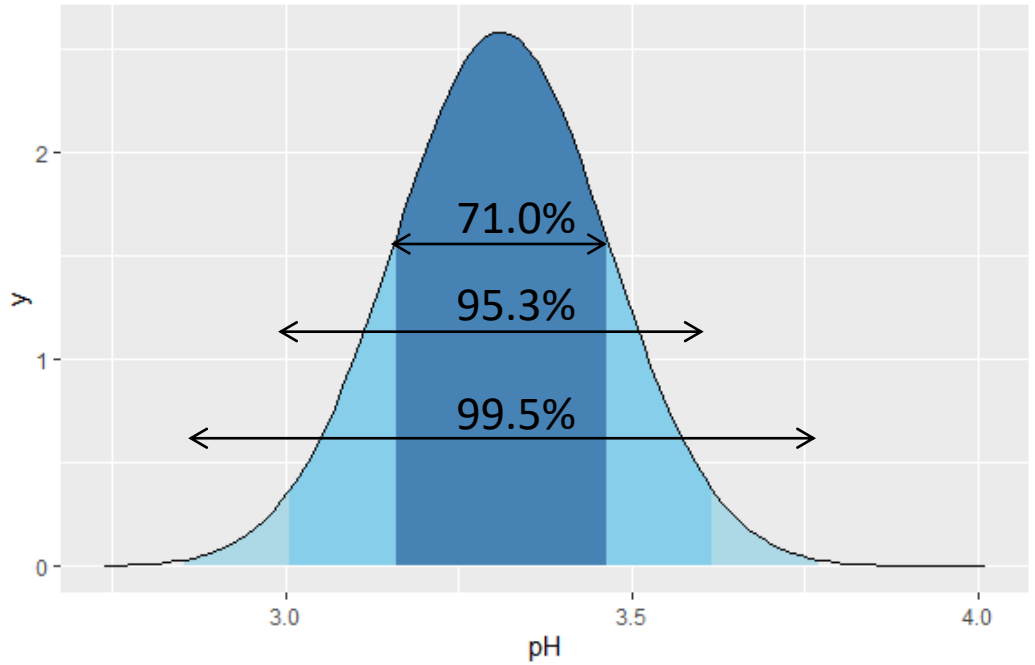
# The Normal Distribution

- $\mu = 3.31$
- $\sigma = 0.154$



# The Normal Distribution

```
> count(subset(winedata, winedata$pH >= mean(winedata$pH)-sd(winedata$pH) & winedata$pH <= mean(winedata
$pH)+sd(winedata$pH))) / count(winedata)
         n
1 0.710444
> count(subset(winedata, winedata$pH >= mean(winedata$pH)-2*sd(winedata$pH) & winedata$pH <= mean(wineda
ta$pH)+2*sd(winedata$pH))) / count(winedata)
         n
1 0.9530957
> count(subset(winedata, winedata$pH >= mean(winedata$pH)-3*sd(winedata$pH) & winedata$pH <= mean(wineda
ta$pH)+3*sd(winedata$pH))) / count(winedata)
         n
1 0.9949969
```

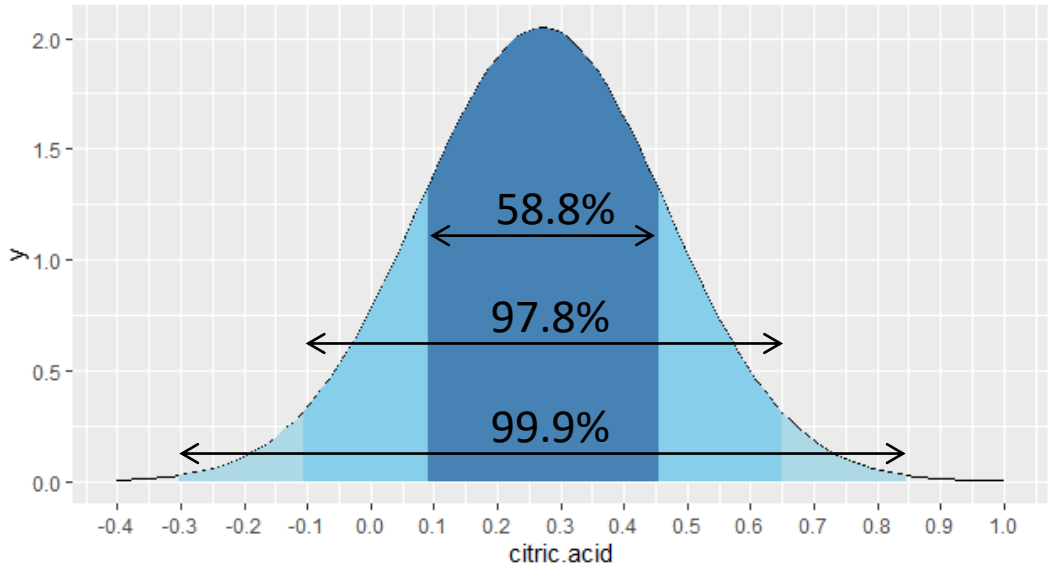


# The Normal Distribution

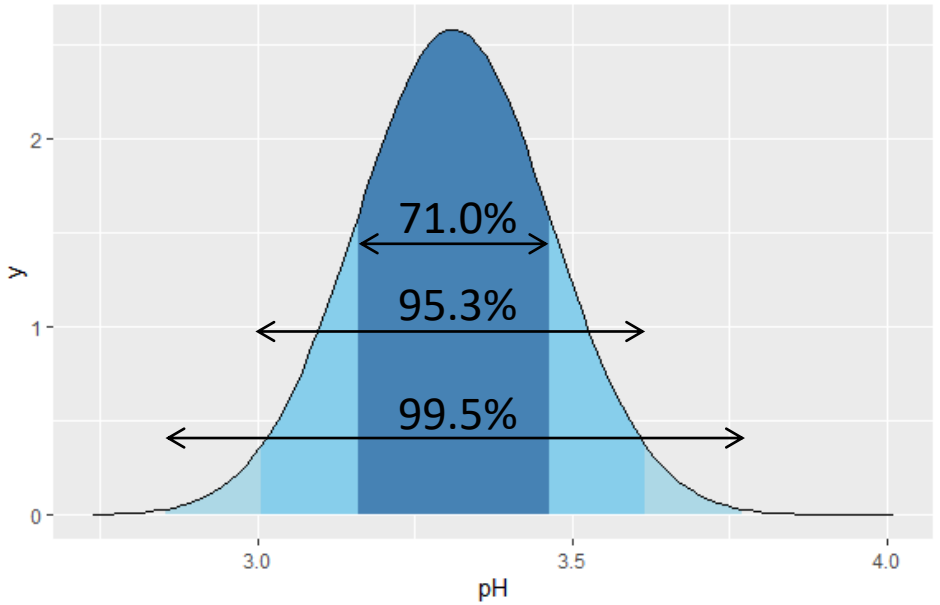
- The **Empirical Rule** (also known as the **Three Sigma Rule** or **68-95-99.7 Rule**) states that almost all data (99.7%) in a distribution falls within three standard deviations from the mean.
  - 68% falls within one standard deviation from the mean
  - 95% falls within two standard deviations from the mean

- The data we've used (citric acid and pH) are not perfectly normal
  - This rule is more of an estimate

# The Normal Distribution




Distribution was skewed, leading the actual proportions to be very different from those estimated by the Empirical Rule




Distribution was much more uniform, leading the actual proportions to be less different from those estimated by the Empirical Rule

# Z-scores

- The **Z-score** of an observation is the number of standard deviations it falls above or below the mean
  - Positive Z-score indicates above the mean
  - Negative Z-score indicates below the mean

- Z-scores are used to put data onto a standardized scale when comparing two different distributions

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

# Z-scores

- We have two datasets: Video game sales (with critic score) and Disneyland guest ratings.
  - Disneyland ratings: 1-5
  - Video game critic scores: 0-100

- We'll say we have a video game with a critic score of 75.2 and a Disneyland rating of 4.6
  - With respect to their data sets, which rating/score is better?
  - In other words, is it better to get a 75.2 critic score or a 4.6 guest score?

# Z-scores

- They are two very different things being compared, but Z-scores allow us to make that comparison.

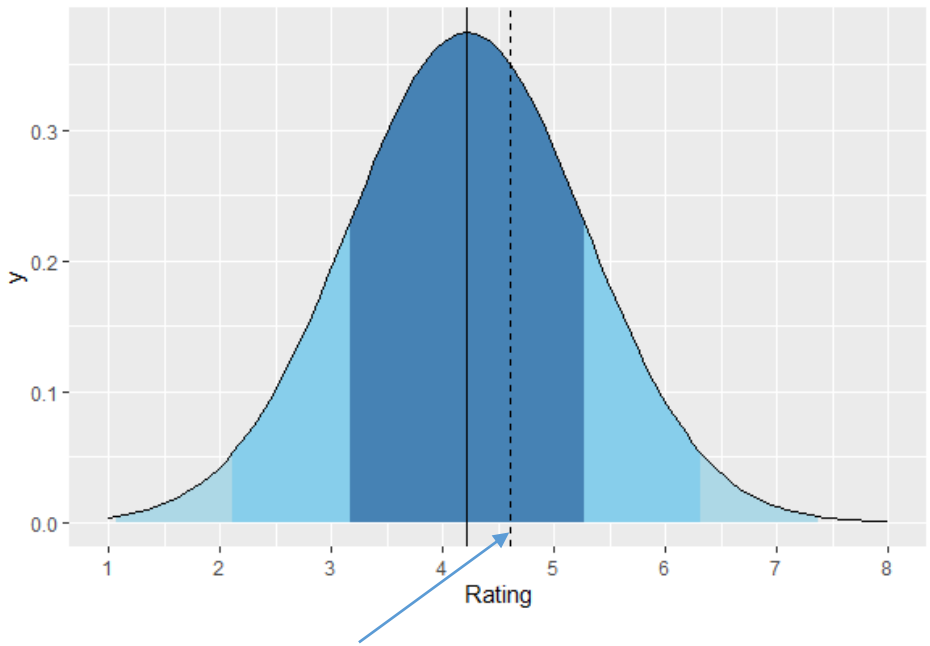
```
> mean(disneydata$Rating)
[1] 4.217695
> sd(disneydata$Rating)
[1] 1.063371
>
> mean(vgdata$Critic_Score, na.rm=TRUE)
[1] 68.96768
> sd(vgdata$Critic_Score, na.rm=TRUE)
[1] 13.93816
```

$$Z = \frac{x - \mu}{\sigma} = \frac{4.6 - 4.2177}{1.0634} = 0.36$$

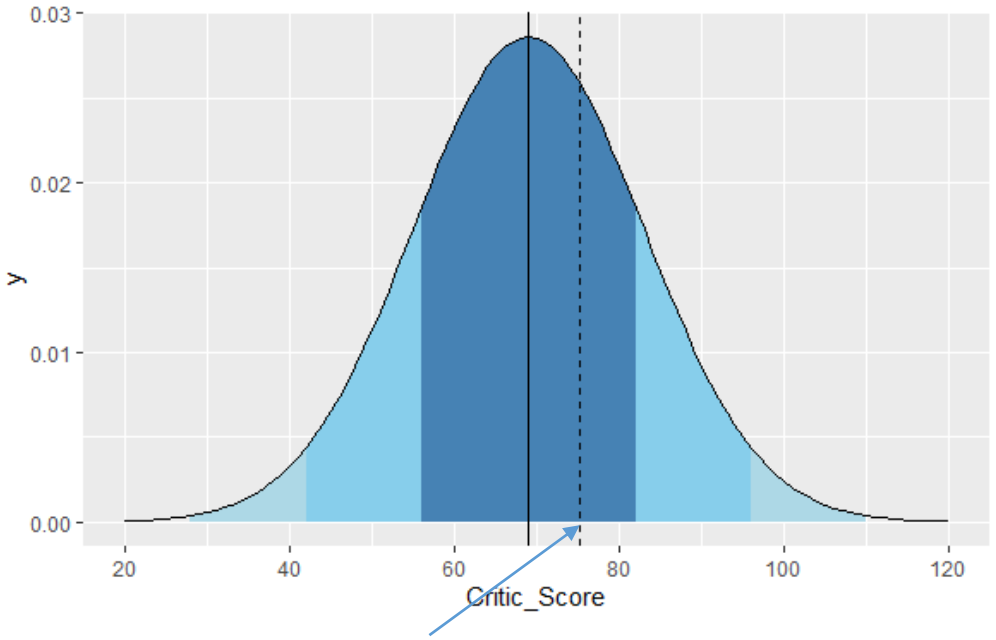
$$Z = \frac{x - \mu}{\sigma} = \frac{75.2 - 68.97}{13.94} = 0.45$$

- The Disney rating is 0.38 standard deviations from the mean
- The video game rating is 0.45 standard deviations from the mean
  - Greater positive distance from the mean

# Z-scores




0.38 standard deviations

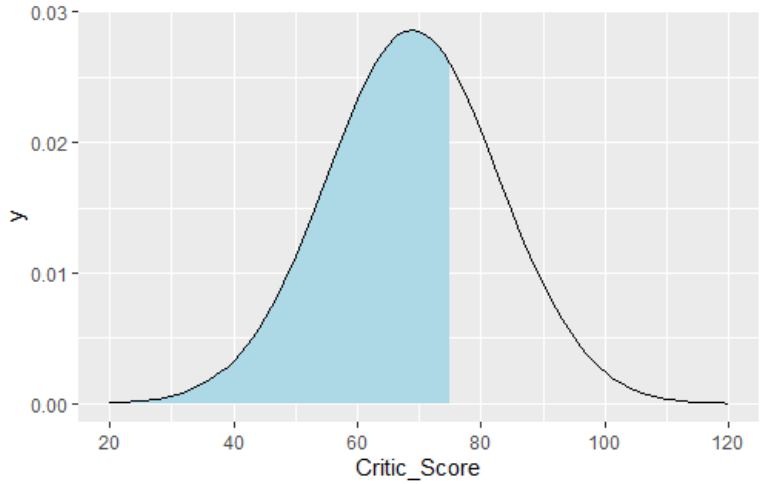



0.45 standard deviations

# Distribution Tails

- The tail of a distribution is also useful.

- For example, all values to the left of the 75.2 critic score are within the same percentile

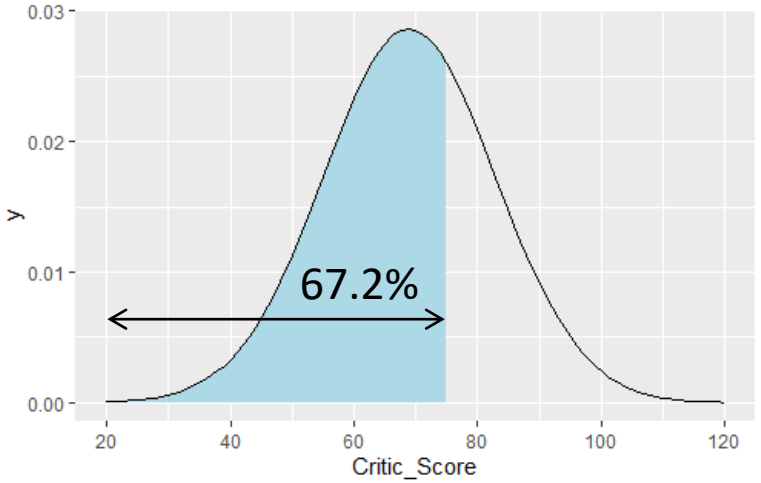


# Distribution Tails

- To find out the area of the tail (such as the shaded portion shown below), R's pnorm function accepts a Z score and will return the left tail proportion

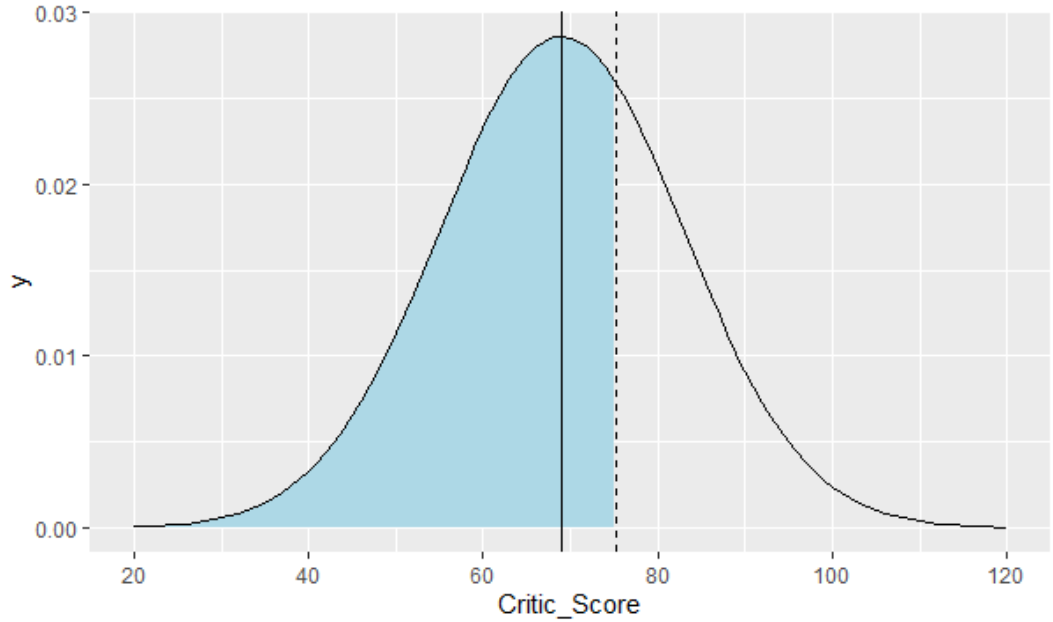
```
> pnorm(zscore_vg)
[1] 0.6726133
```

- 67.3% of ratings are below 75.2
  - Thus, 32.7% are above it



# Distribution Tails

- Since the Z-score is greater than the mean, it is a positive Z-score



# Distribution Tails

- Let's now say we have a Disney review of 2.5
  - Is this a positive or negative Z-score?

```
> mean(disneydata$Rating)
[1] 4.217695
> sd(disneydata$Rating)
[1] 1.063371
```

$$Z = \frac{x - \mu}{\sigma} = \frac{2.5 - 4.2177}{1.0634} = -1.615$$

- Right away, we can already see we have a negative Z-score without plotting the distribution

# Distribution Tails

```
> pnorm(zscore_disney)
[1] 0.05311964
```

- 5.3% of ratings are below 2.5
- Negative Z-score