Revised: 3/20/2021

Probability IV

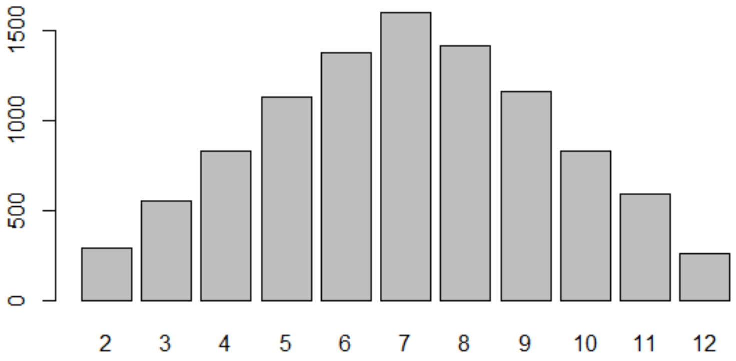
Michael C. Hackett

Assistant Professor, Computer Science

Community
College
of Philadelphia

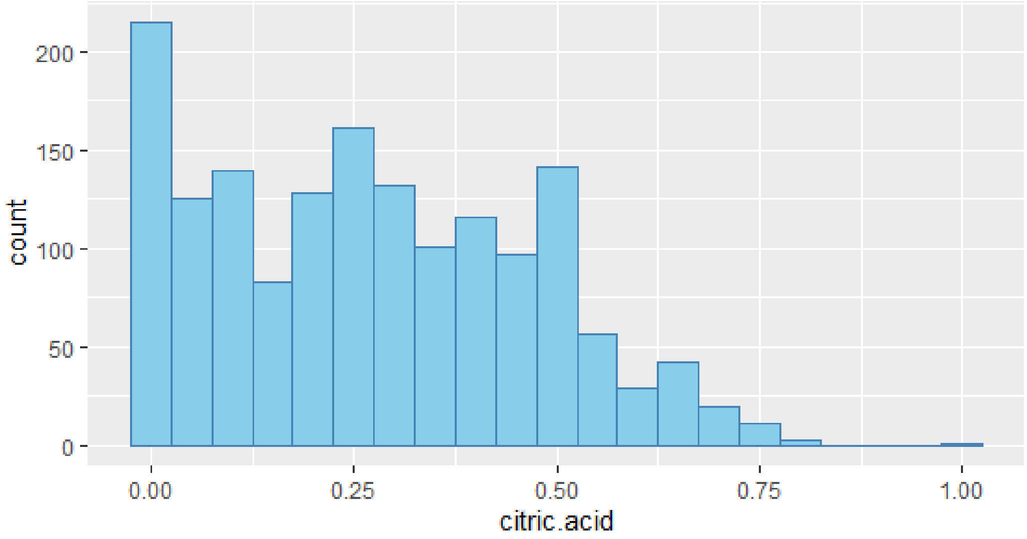
Continuous Distributions

- 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.



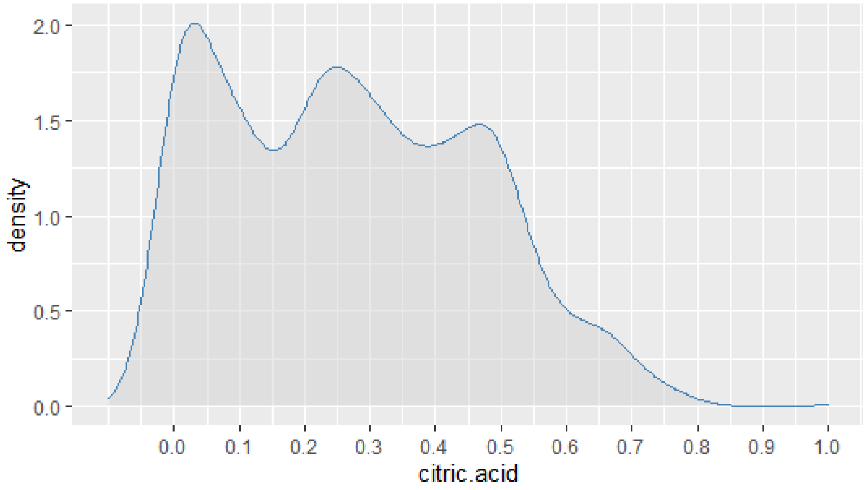
Continuous Distributions

- 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



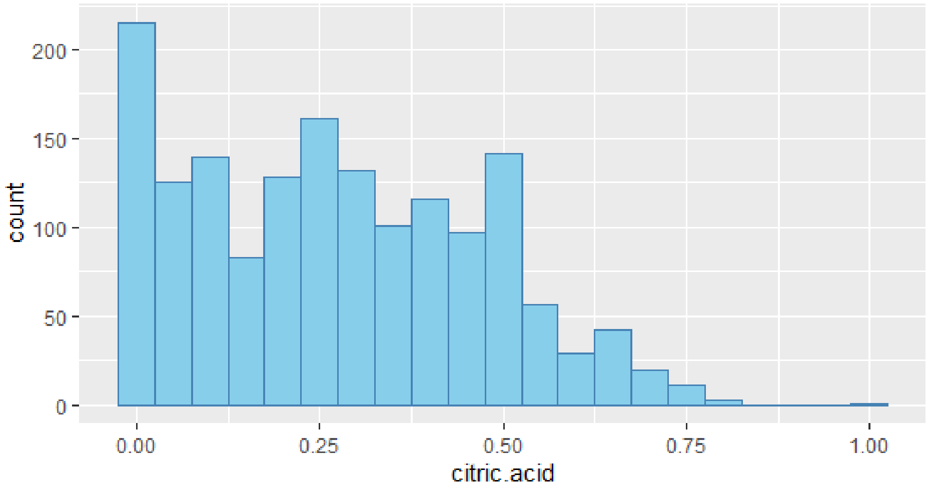
Continuous Distributions

- 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

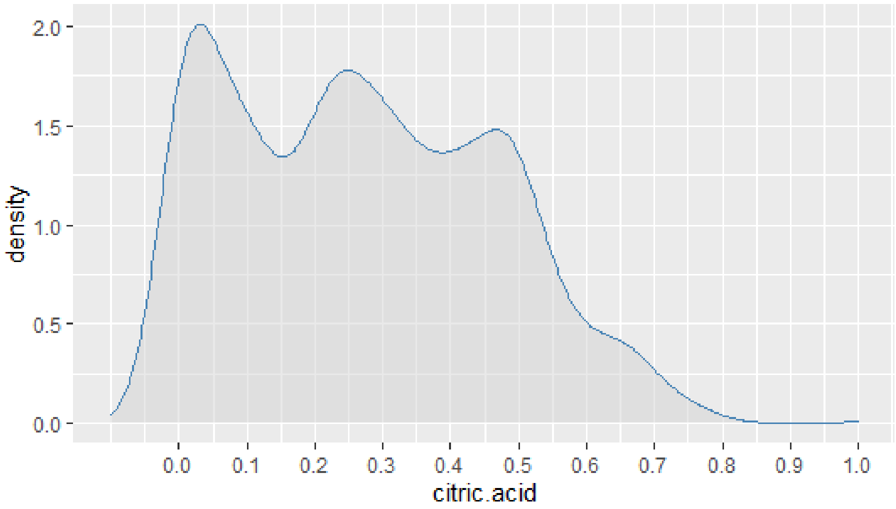


Continuous Distributions

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



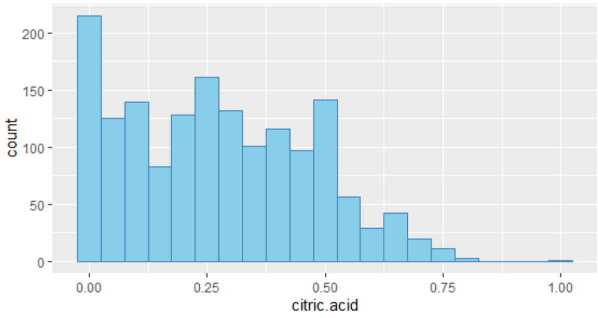
Discrete numerical

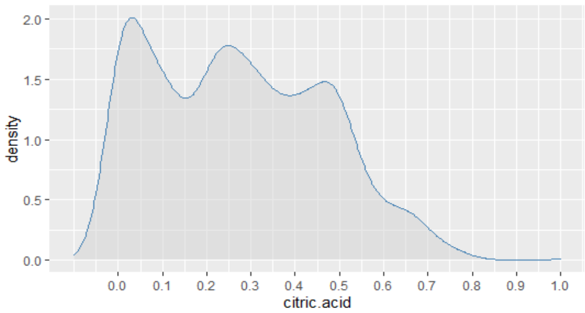


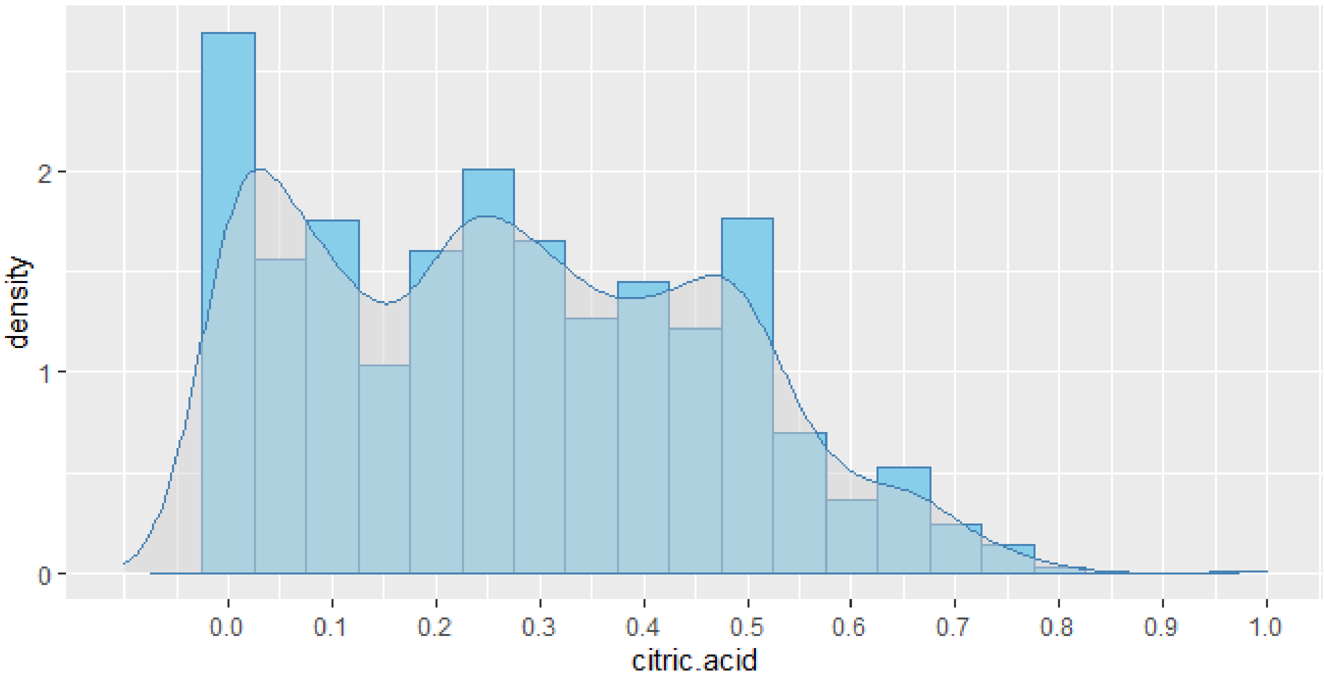
Continuous numerical

Continuous Distributions

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

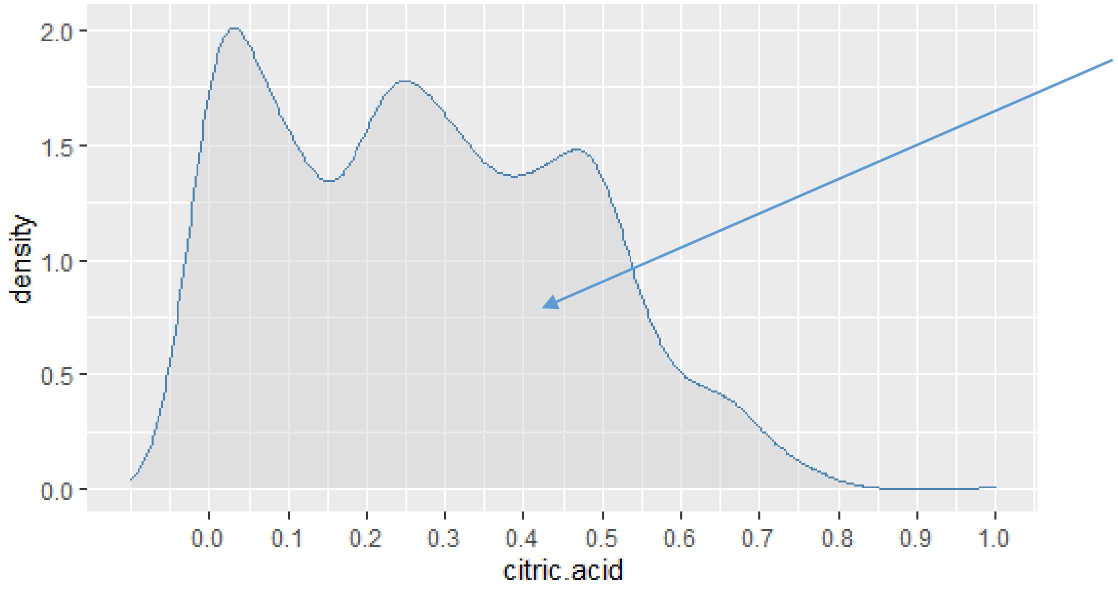






Continuous Distributions

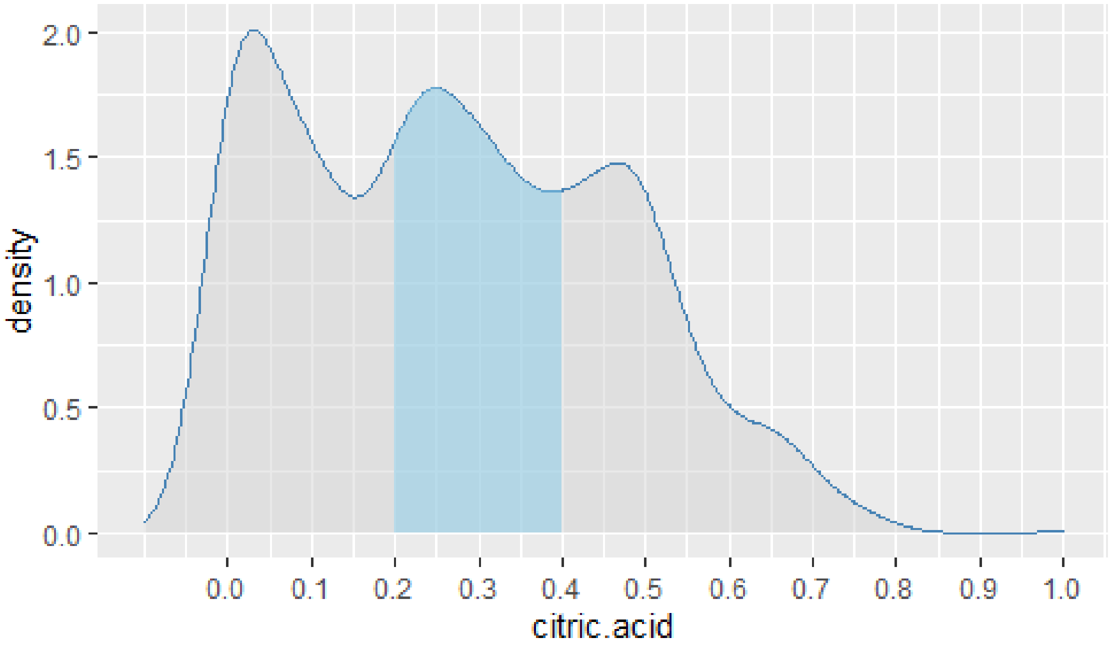
- The area under the curve of a continuous distribution is equal to 1
  - Typically, finding the area or (sections of the area) under a curve requires integration.



Area = 1

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

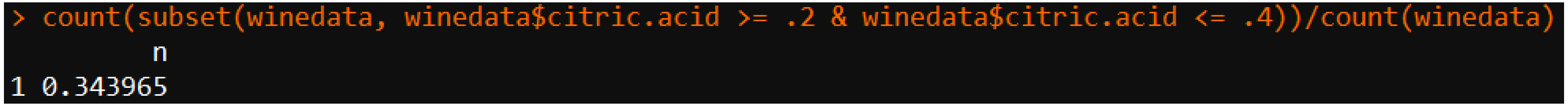


Continuous Distributions

- 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:

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

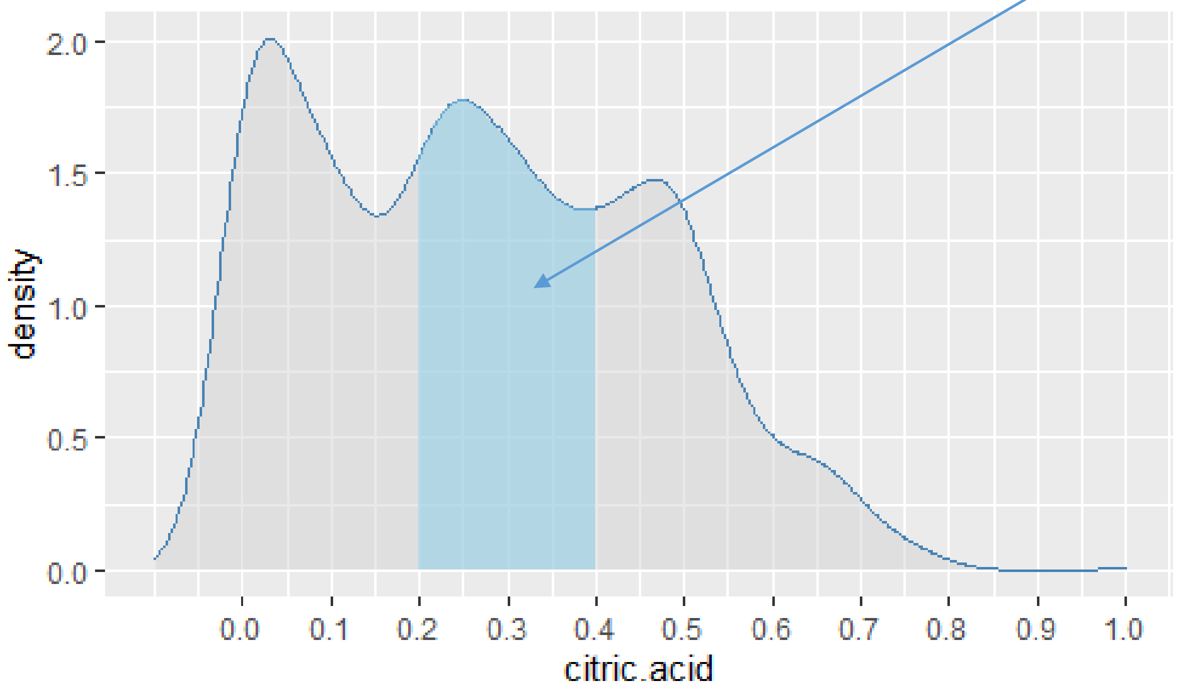
```
count(subset(winedata, winedata$citric.acid >= .2 & winedata$citric.acid <= .4)) / count(winedata)
```



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

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

Continuous Distributions



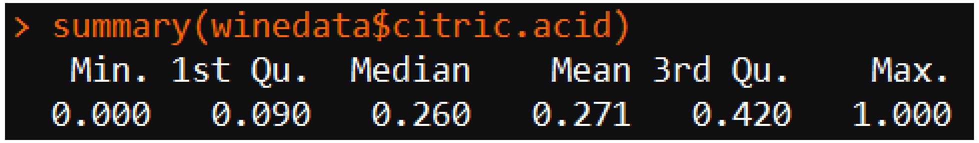
Area = .344

Continuous Distributions

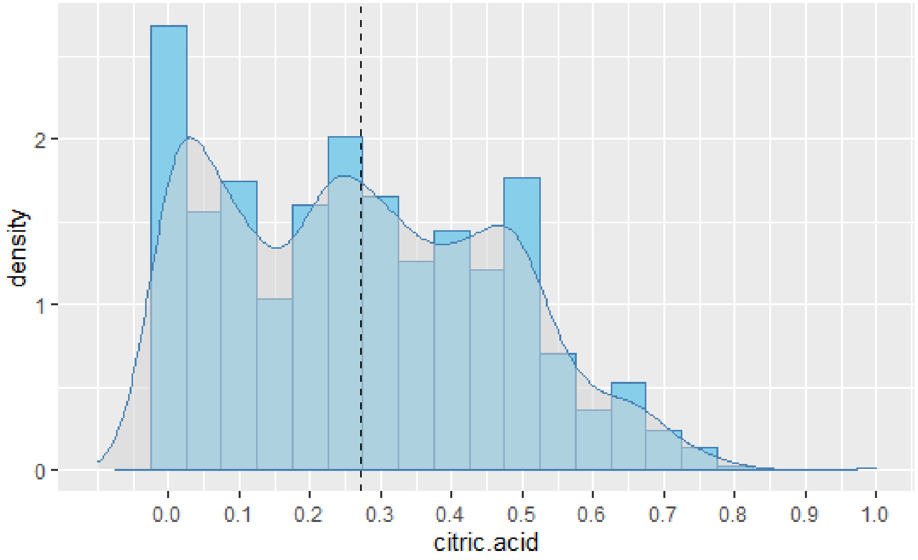
- The mean of this distribution is 0.271

  `mean(winedata$citric.acid)` or

  `summary(winedata$citric.acid)`



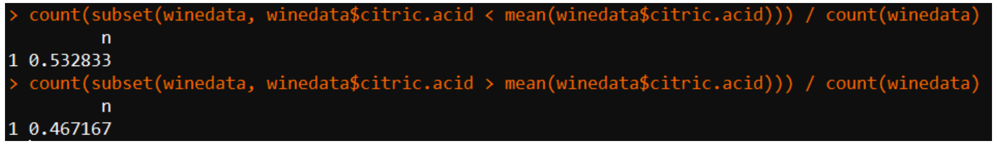
```
> summary(winedata$citric.acid)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.090   0.260   0.271   0.420   1.000 
```



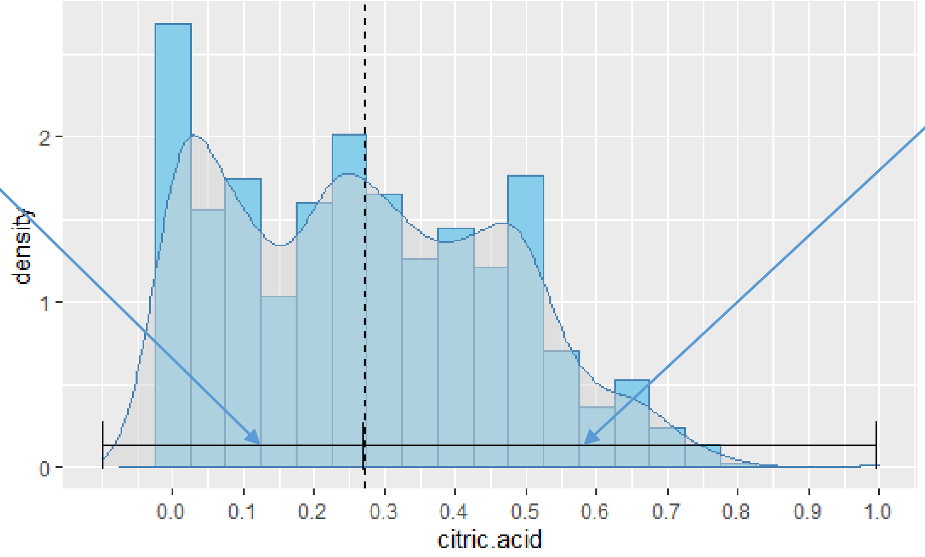
Continuous Distributions

- 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
```

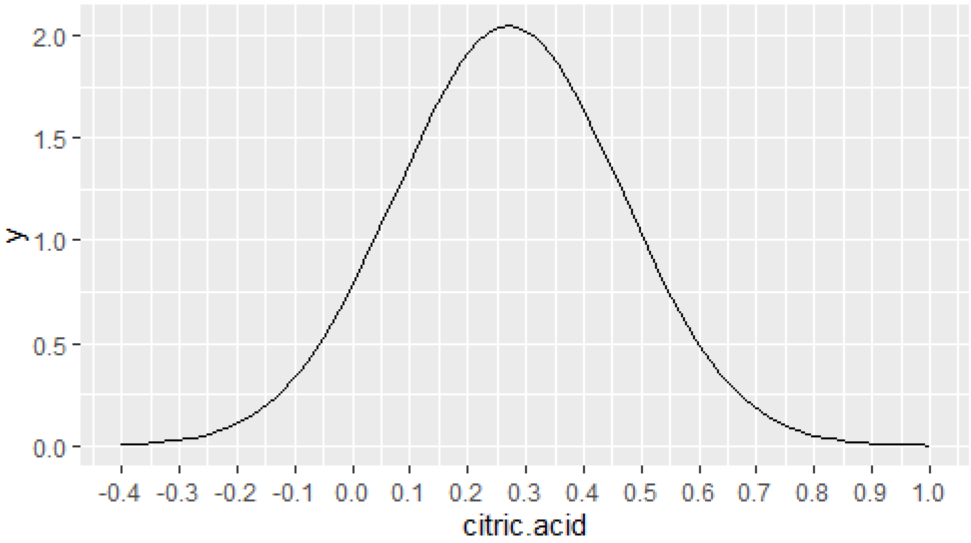


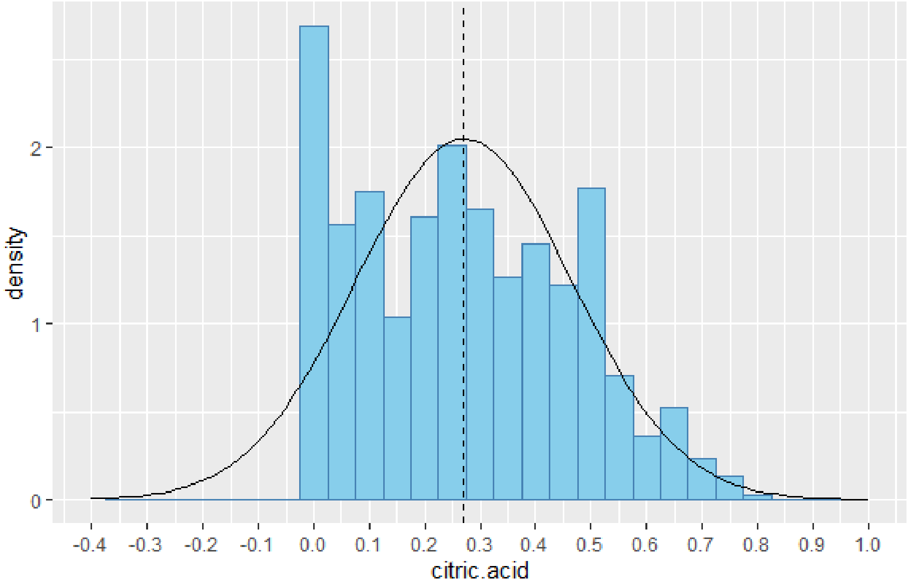
Area = .533

Area = .467

The Normal Distribution

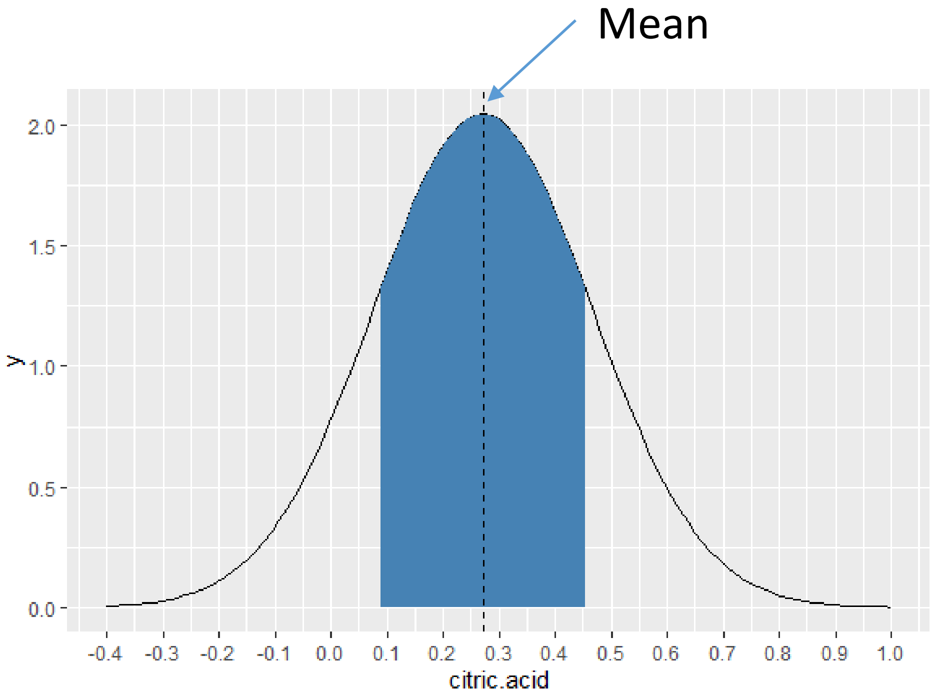
- The **normal distribution** is a continuous probability distribution model that fits a distribution to a symmetric, unimodal, bell-shaped curve





The Normal Distribution

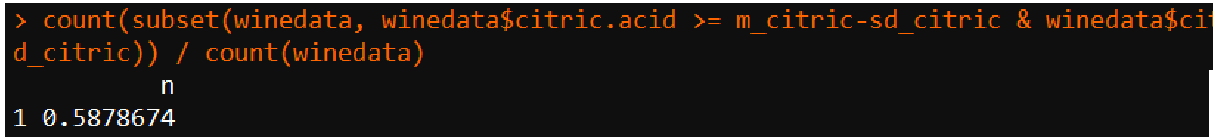
- One standard deviation from the mean
  - $\sigma = .195$



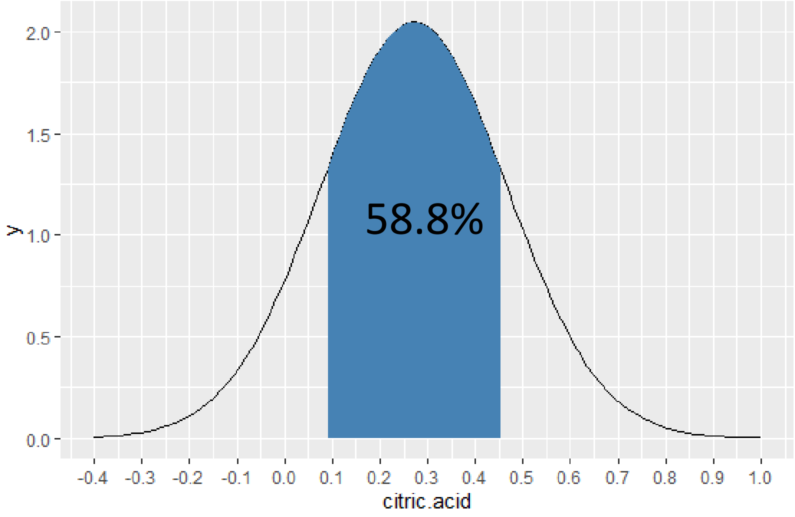
The Normal Distribution

- 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)
```

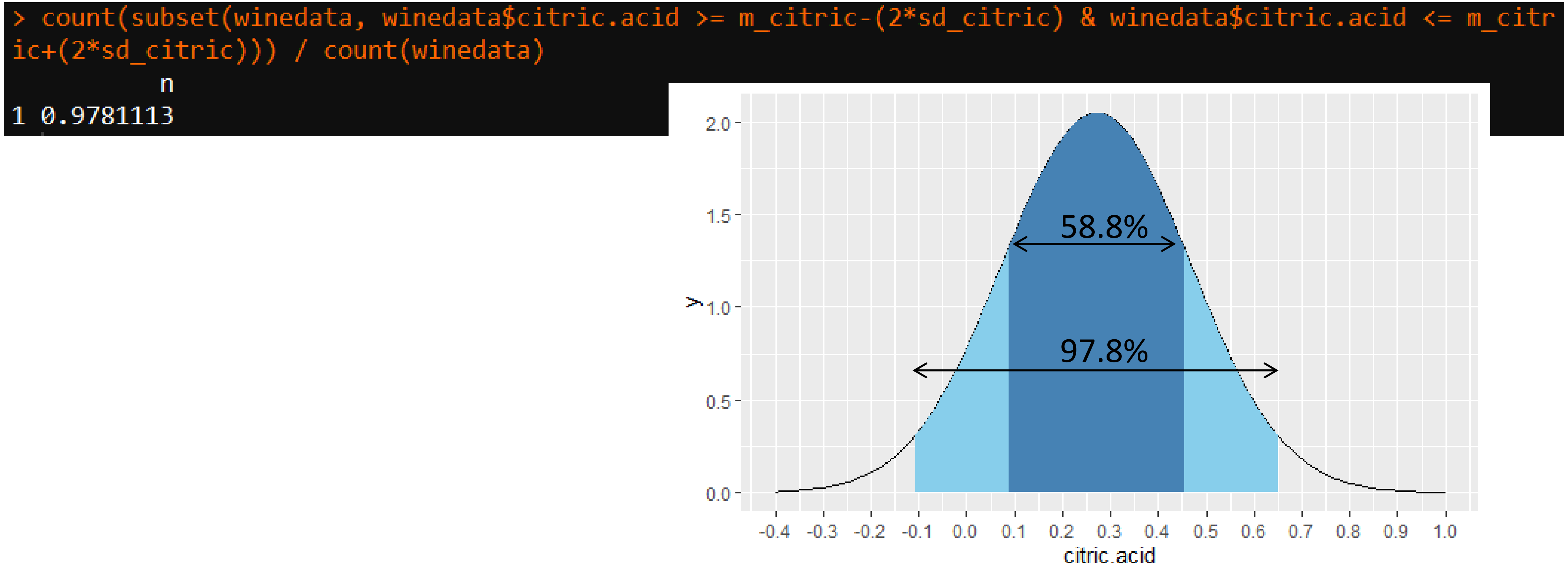


```
> count(subset(winedata, winedata$citric.acid >= m_citric-sd_citric & winedata$citric.acid <= m_citric+s
d_citric)) / count(winedata)
          n
1 0.5878674
```



The Normal Distribution

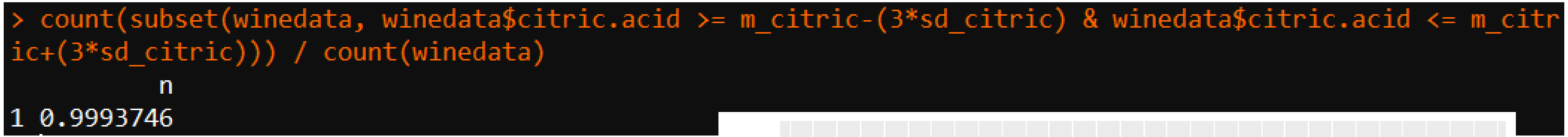
- Two standard deviations from the mean



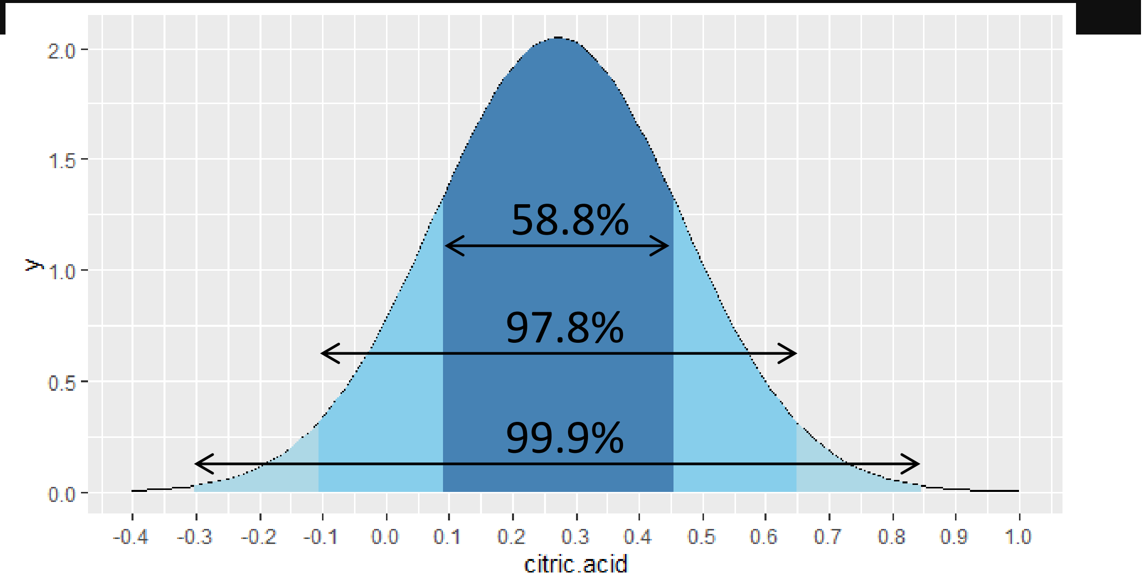
```
> count(subset(winedata, winedata$citric.acid >= m_citric-(2*sd_citric) & winedata$citric.acid <= m_citr
ic+(2*sd_citric))) / count(winedata)
          n
1 0.9781113
```

The Normal Distribution

- Three standard deviations from the mean

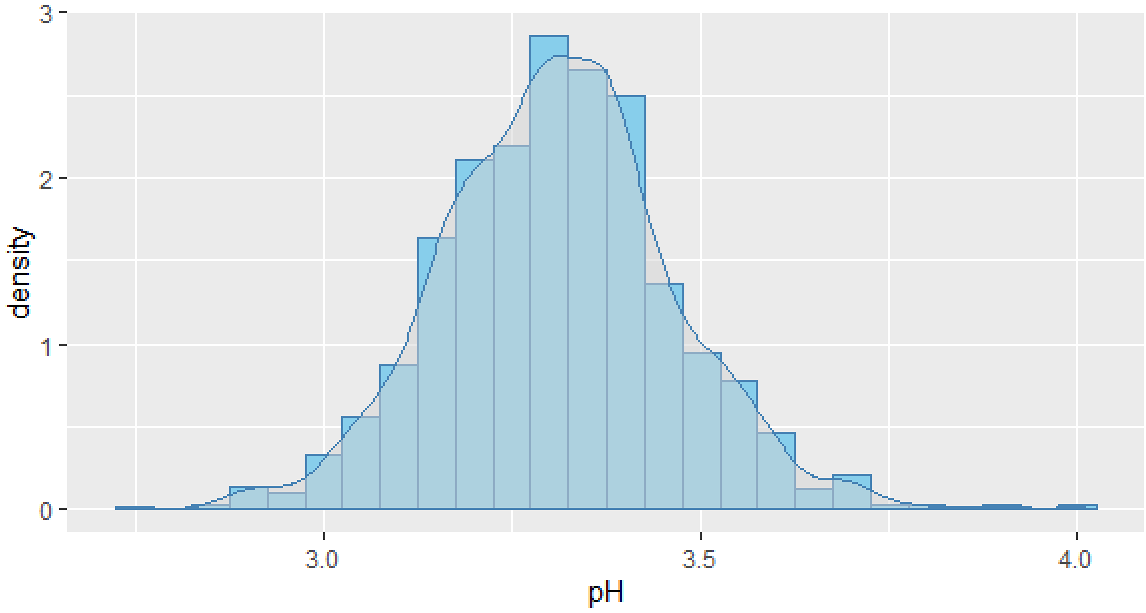


```
> count(subset(winedata, winedata$citric.acid >= m_citric-(3*sd_citric) & winedata$citric.acid <= m_citr
ic+(3*sd_citric))) / count(winedata)
          n
1 0.9993746
```



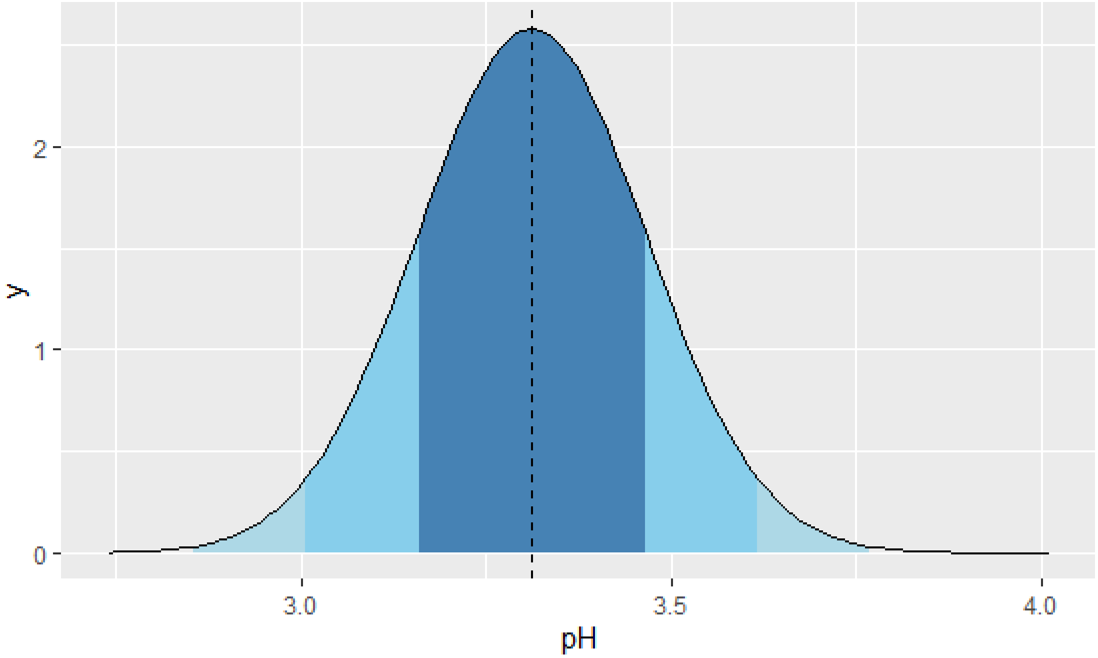
The Normal Distribution

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

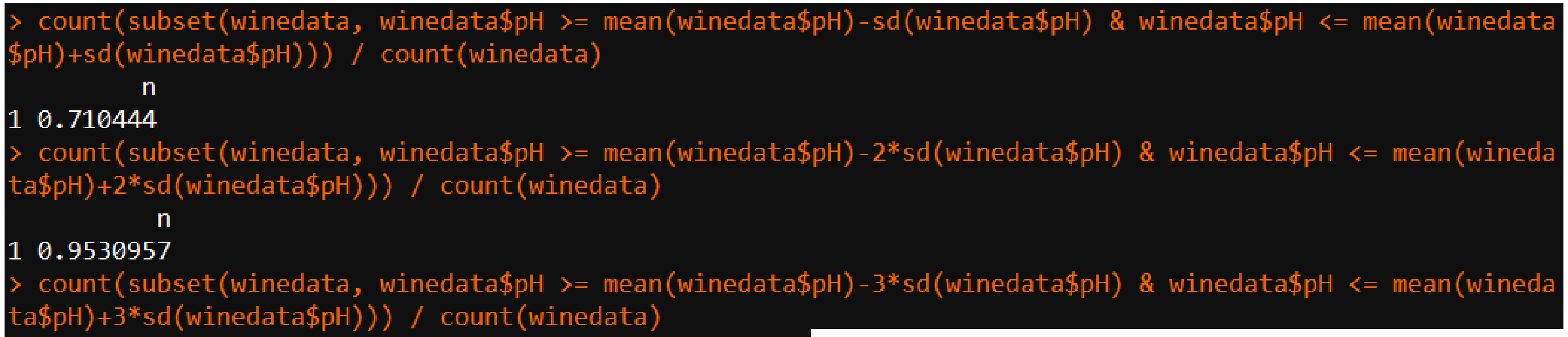


The Normal Distribution

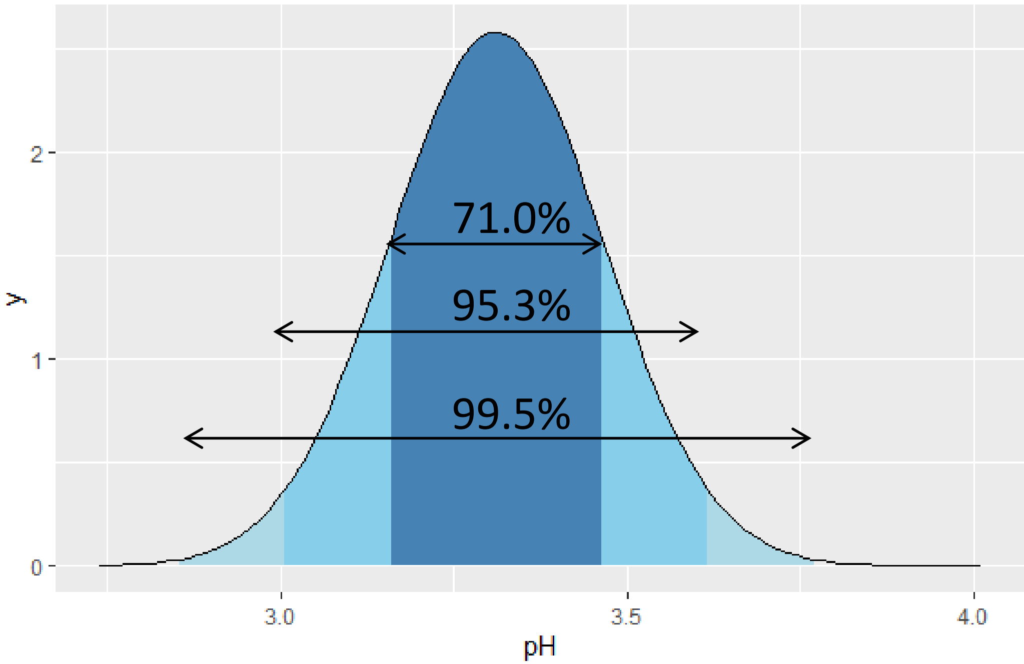
- $\mu = 3.31$
- $\sigma = 0.154$



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
```

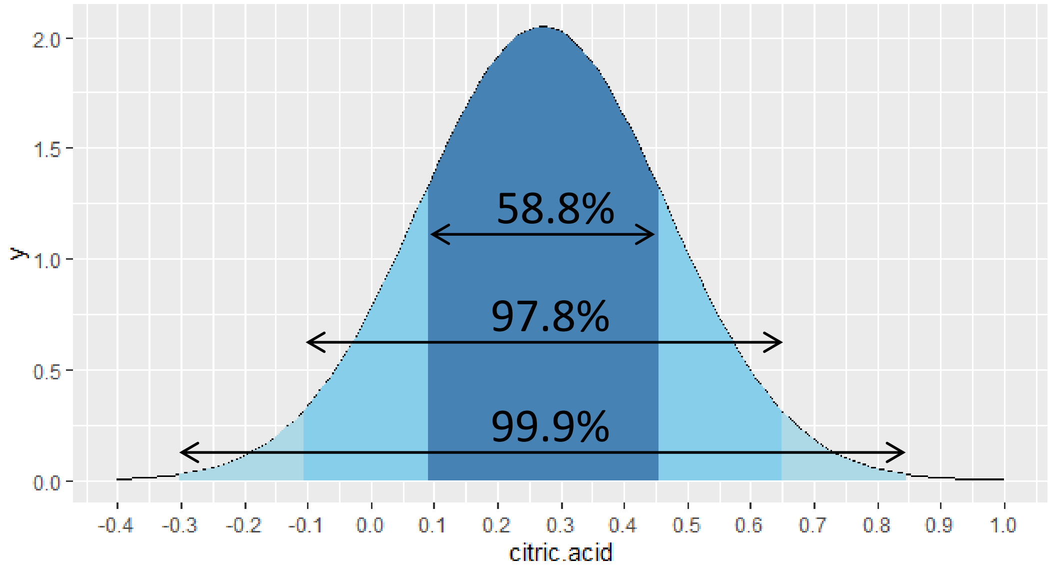


The Normal Distribution

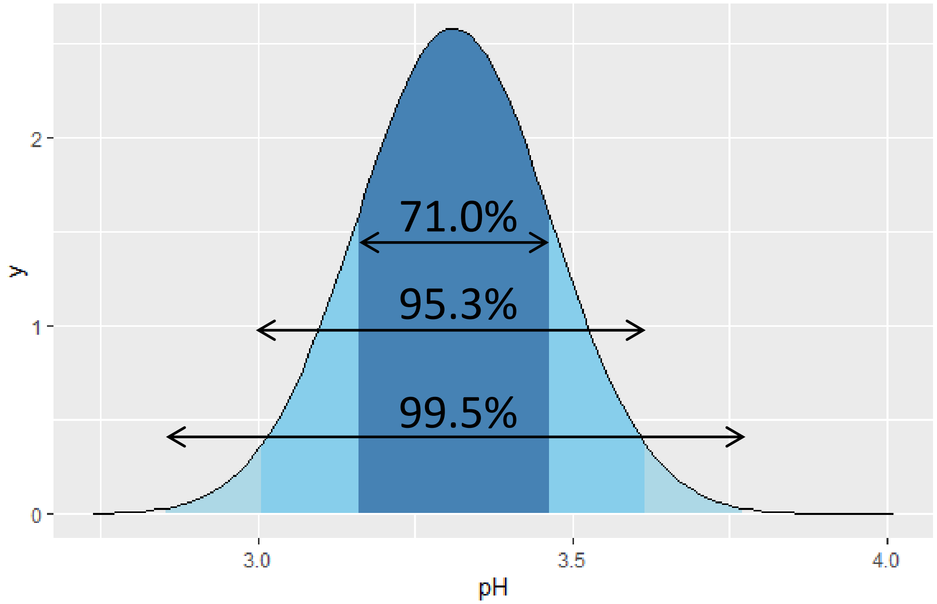
- The **Empirical Rule** (also known as the **Three Sigma Rule** or **68-95-99.7 Rule**) states that almost all data (99.7%) in a distribution falls within three standard deviations from the mean.
  - 68% falls within one standard deviation from the mean
  - 95% falls within two standard deviations from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

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

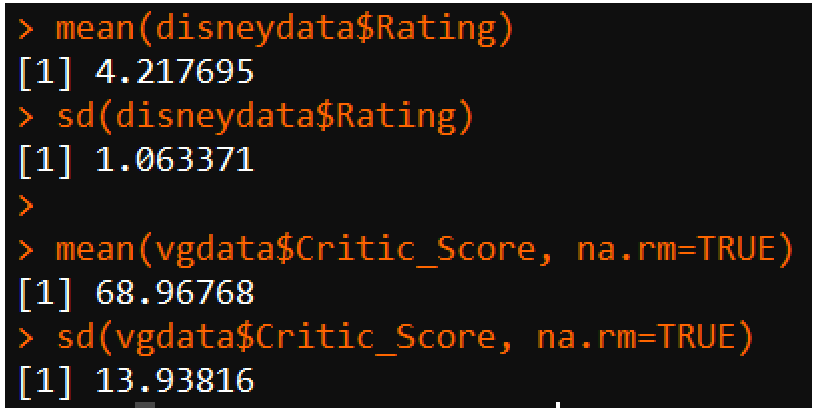
Z-scores

- 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.



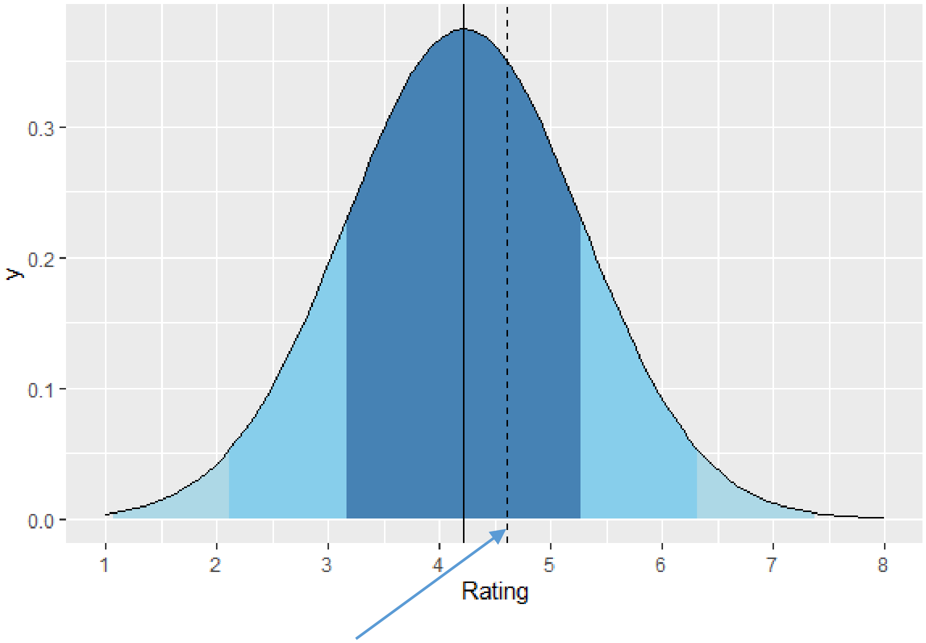
```
> mean(disneydata$Rating)
[1] 4.217695
> sd(disneydata$Rating)
[1] 1.063371
>
> mean(vgdata$Critic_Score, na.rm=TRUE)
[1] 68.96768
> sd(vgdata$Critic_Score, na.rm=TRUE)
[1] 13.93816
```

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

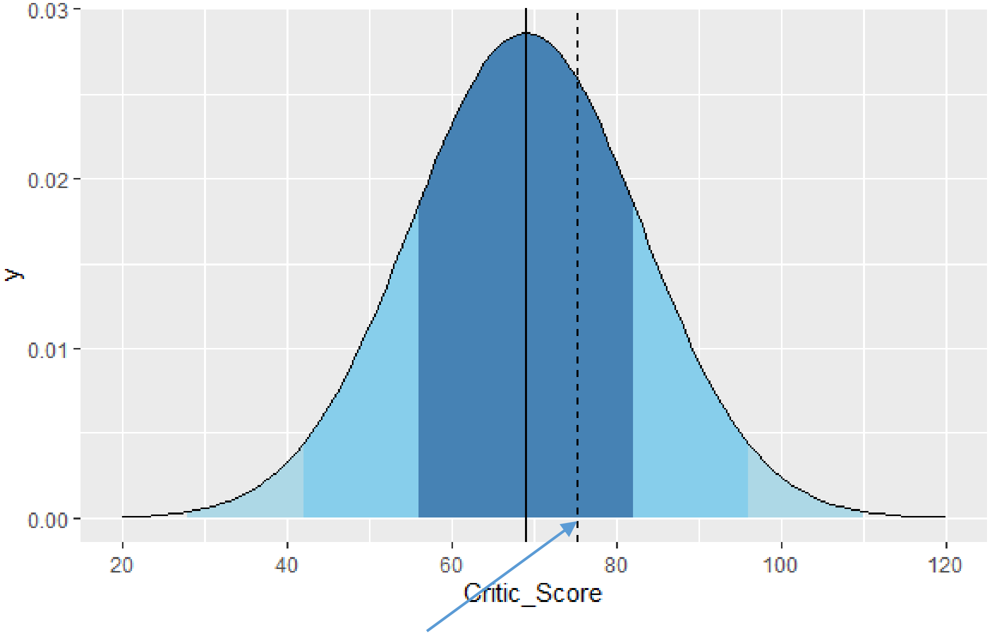
$$Z = \frac{x - \mu}{\sigma} = \frac{75.2 - 68.97}{13.94} = 0.45$$

- 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

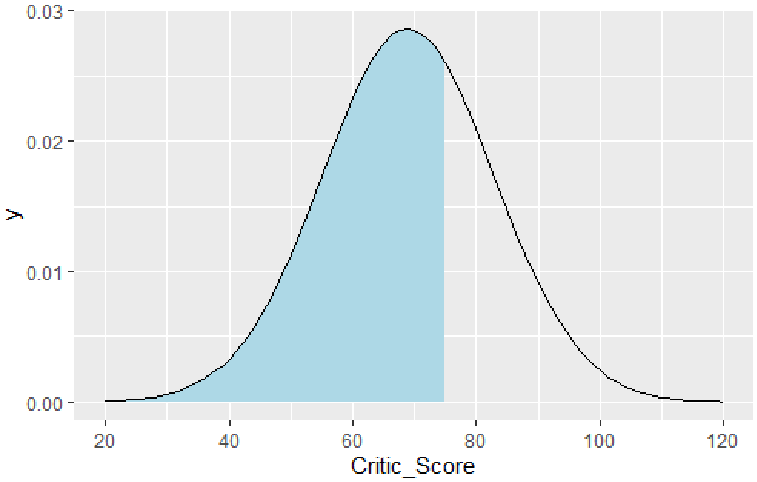


0.45 standard deviations

Distribution Tails

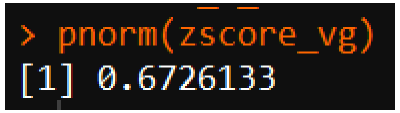
- The tail of a distribution is also useful.

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



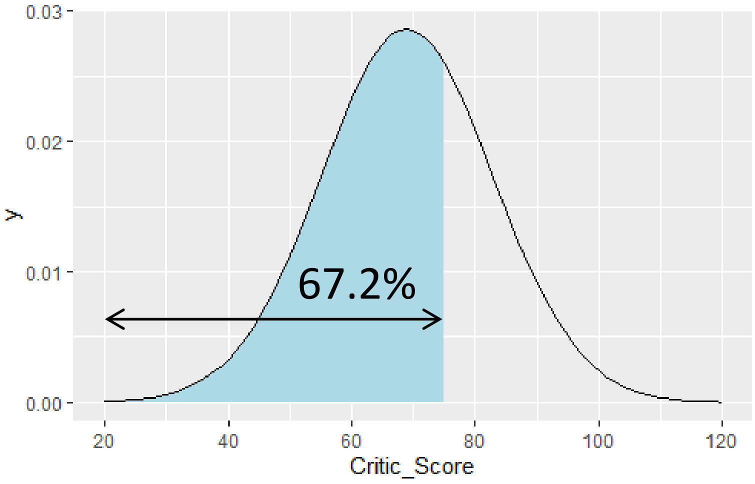
Distribution Tails

- 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



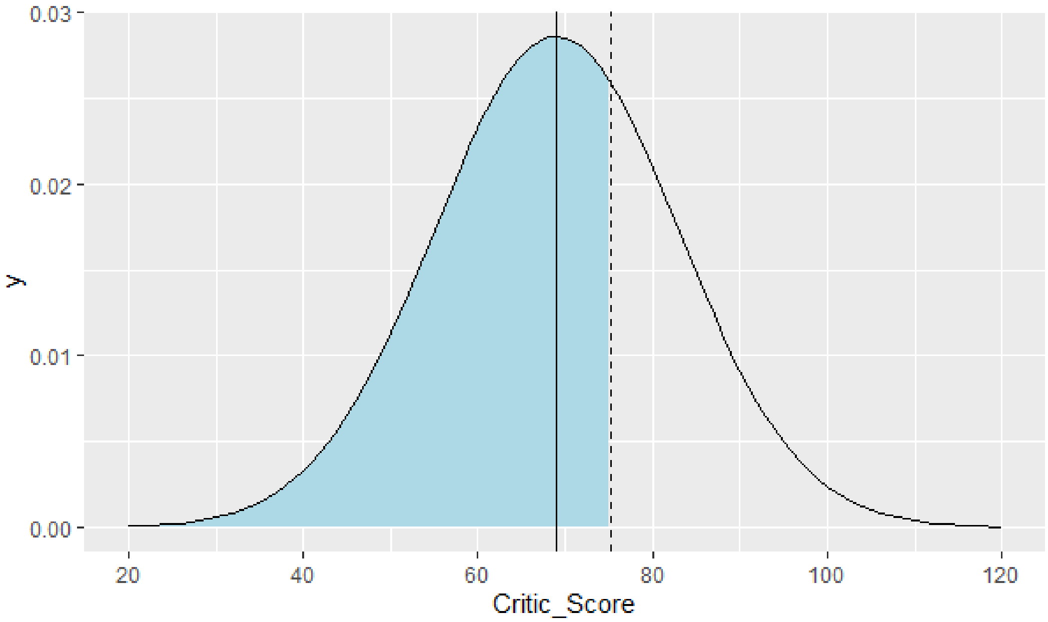
```
> pnorm(zscore_vg)
[1] 0.6726133
```

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



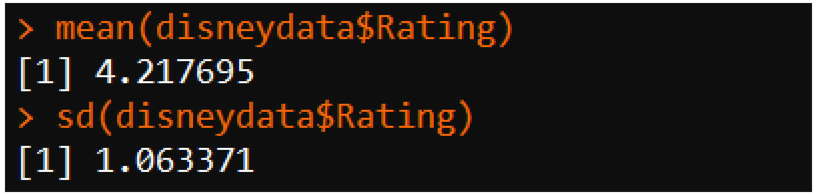
Distribution Tails

- 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?

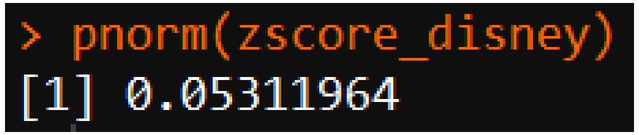


```
> mean(disneydata$Rating)
[1] 4.217695
> sd(disneydata$Rating)
[1] 1.063371
```

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

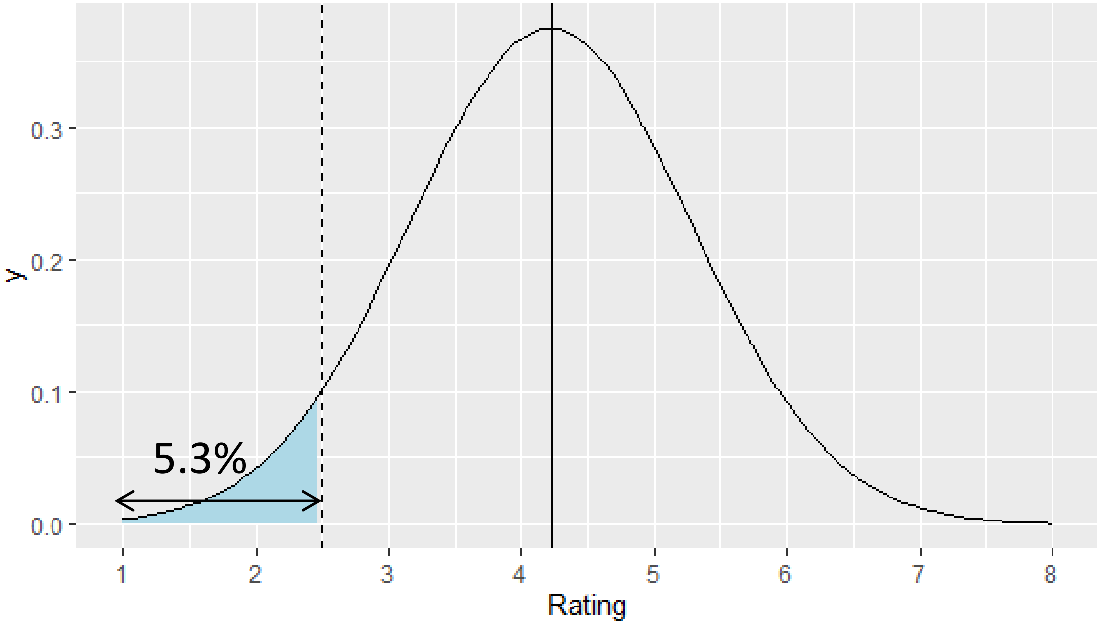
- Right away, we can already see we have a negative Z-score without plotting the distribution

Distribution Tails



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

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



Bernoulli Distribution

- Distribution of a random variable with only two possible outcomes: 1 or 0
  - Success or failure, heads or tails, yes or no, etc.
  - “Bernoulli random variable”

- As an example, let’s say a software testing team found that 80% of the unit tests ran in the past year “passed”.
  - Probability of passing: $\boldsymbol{p} = 80\% = .80$
  - Probability of failing: $\boldsymbol{q} = \mathbf{1} - \boldsymbol{p} = 1 - .80 = .20 = 20\%$

Bernoulli Distribution

- A sample of the unit test results:
  0 1 1 0 1 1 1 1 1 0 1 0 1 1 1
    - 1 indicates the test passed
    - 0 indicates the test failed

- The proportion of successes in a Bernoulli distribution *is* the sample mean

$$\hat{p} = \frac{Number\ of\ successes}{Number\ of\ tests} = \frac{11}{15} = 0.733 = 73.3\%$$

$$\mu = \frac{Sum\ of\ tests}{Number\ of\ tests} = \frac{11}{15} = 0.733 = 73.3\%$$

Bernoulli Distribution

- Since 0 and 1 are numeric, Bernoulli random variables have a mean (shown on the last slide), variance, and standard deviation
  - $\mu = p$
  - $\sigma^2 = p(1 - p)$
  - $\sigma = \sqrt{p(1 - p)}$

Geometric Distribution

- Distribution that describes how many trials it takes to observe a success for independent and identically distributed (**iid**) Bernoulli random variables.

- We'll use the same example of the software testing team finding that 80% of the unit tests ran in the past year "passed".
  - Which means 20% of the tests "failed"

Geometric Distribution

- When randomly selecting tests:
  - The probability that the first test selected is “pass” is 80% or 0.80
    - P(first test selected is “pass”) = 0.8 or 80%
  - The probability that the second test selected is “pass”
    - P(first test selected is “fail” and second test selected is “pass”)

    0.2 × 0.8 = 0.16 or 16%
  - The probability that the third test selected is “pass”
    - P(first test selected is “fail” and second test selected is “fail” and third test selected is “pass”)

    0.2 × 0.2 × 0.8 = 0.032 or 3.2%

Geometric Distribution

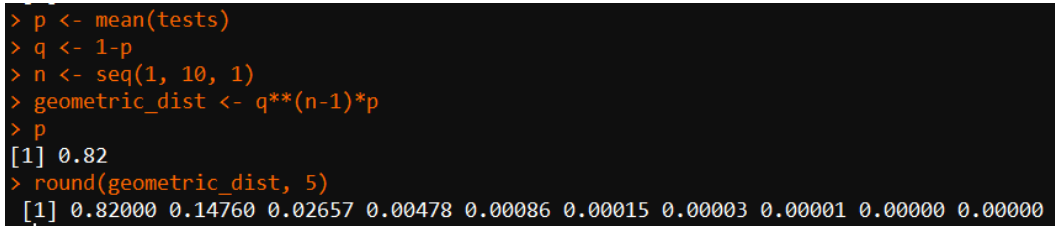
- The previous slide could be generalized to
  - $P(success\ on\ n^{th}\ test)\ = (0.2)^{n-1}\times 0.7$
  - $P(success\ on\ n^{th}\ test)\ = (q)^{n-1}\times p$
    - $q = 1 - p$

  - $P(success\ on\ n^{th}\ test)\ = \boldsymbol{(1-p)^{n-1}\times p}$

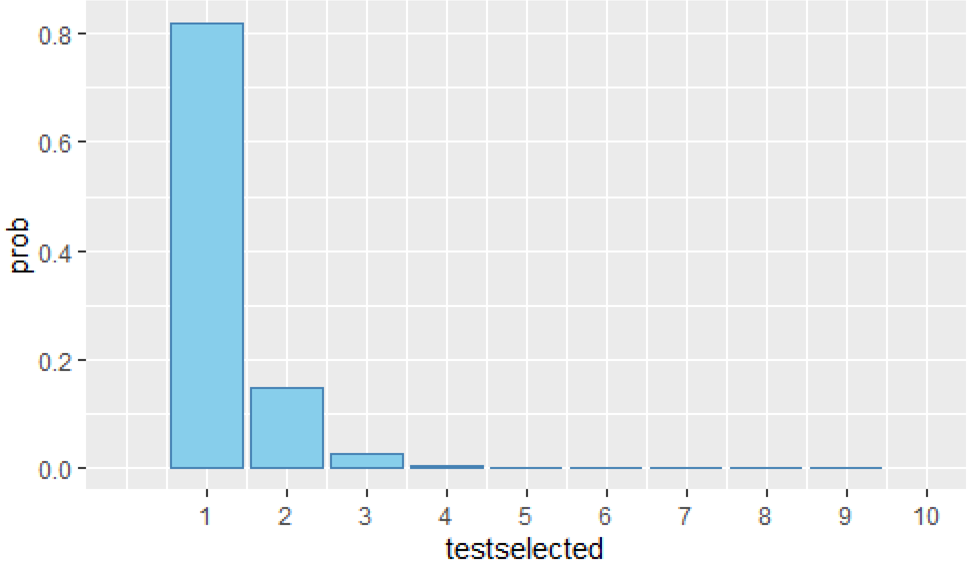
- The probability of success on the fourth test
  - $P\left(success\ on\ 4^{th}\ test\right) = (1-p)^{n-1} \times p = (1-0.8)^{4-1} \times 0.8 = 0.0064$
  - To be clear, this is the probability of needing to sample 4 tests until finding a successful/”passed” test is 0.64%

Geometric Distribution

- The geometric probability distribution decreases exponentially

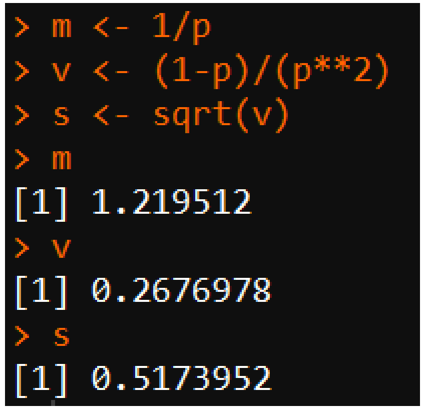


```
> p <- mean(tests)
> q <- 1-p
> n <- seq(1, 10, 1)
> geometric_dist <- q**(n-1)*p
> p
[1] 0.82
> round(geometric_dist, 5)
 [1] 0.82000 0.14760 0.02657 0.00478 0.00086 0.00015 0.00003 0.00001 0.00000 0.00000
```



Geometric Distribution

- Mean, variance, and standard distribution of a geometric distribution:
  - $\mu = \frac{1}{p}$
  - $\sigma^2 = \frac{1-p}{p^2}$
  - $\sigma = \sqrt{\frac{1-p}{p^2}}$



```
> m <- 1/p
> v <- (1-p)/(p**2)
> s <- sqrt(v)
> m
[1] 1.219512
> v
[1] 0.2676978
> s
[1] 0.5173952
```

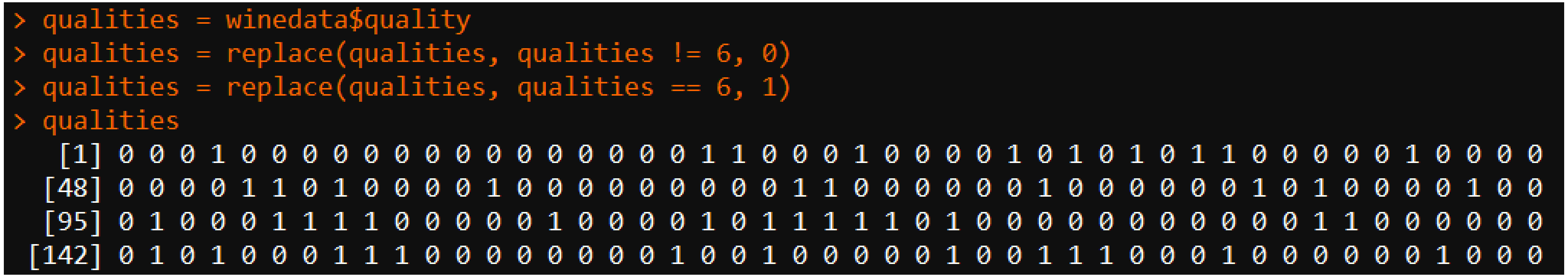
Geometric Distribution

- The wine quality dataset includes a “quality” variable

- Selecting observations randomly, we are interested in the probability of selecting a “6” wine on the first try, the second try, and the third try
  - A wine rated with a 6 will be a “success” or 1, and all others will be a “failure” or 0

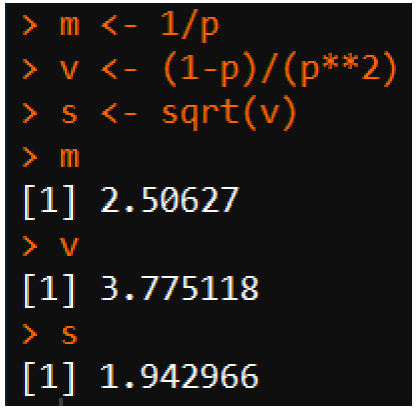
Geometric Distribution

- Preparing the data set:



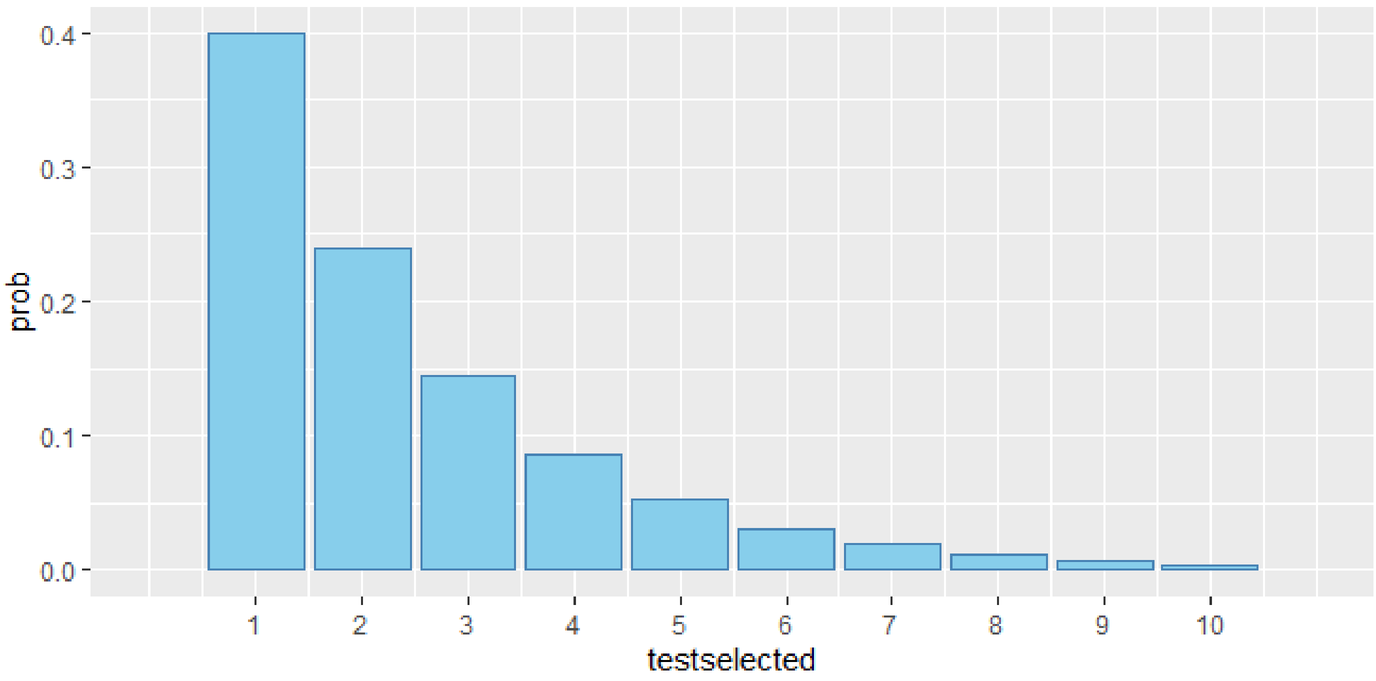
```
> qualities = winedata$quality
> qualities = replace(qualities, qualities != 6, 0)
> qualities = replace(qualities, qualities == 6, 1)
> qualities
   [1] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 0 0 1 0 0 0 0
  [48] 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0
  [95] 0 1 0 0 0 1 1 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0
 [142] 0 1 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0
```

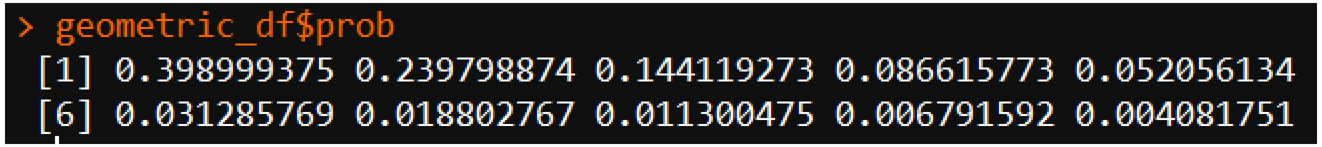
- Mean, variance, standard deviation:



```
> m <- 1/p
> v <- (1-p)/(p**2)
> s <- sqrt(v)
> m
[1] 2.50627
> v
[1] 3.775118
> s
[1] 1.942966
```

Geometric Distribution

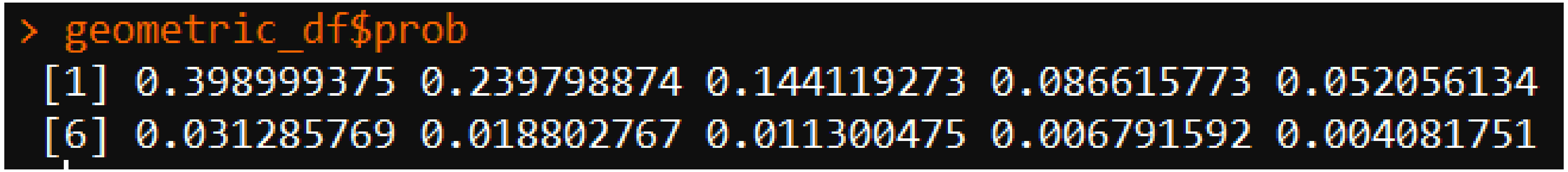




```
> geometric_df$prob
 [1] 0.398999375 0.239798874 0.144119273 0.086615773 0.052056134
 [6] 0.031285769 0.018802767 0.011300475 0.006791592 0.004081751
```

Geometric Distribution

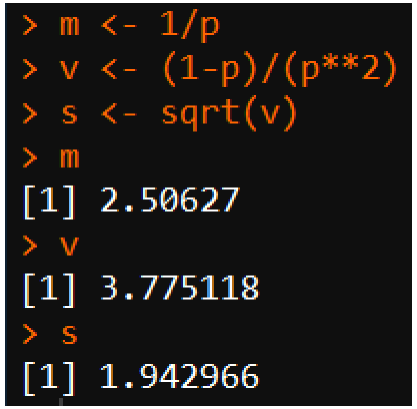
- If we were to randomly sample from the wine quality data set:
  - The probability that…
    - the first observation selected has a quality of 6 (“success”) is 0.3989 or ~40%
  - The probability that…
    - the first observation selected does not have a quality of 6 (“fail”) and the second observation is a success is .2397 or ~24%
  - The probability that…
    - the first and second observations selected are failures and the third observation selected is a success is .1441 or ~14%



```
> geometric_df$prob
 [1] 0.398999375 0.239798874 0.144119273 0.086615773 0.052056134
 [6] 0.031285769 0.018802767 0.011300475 0.006791592 0.004081751
```

Geometric Distribution

- Mean, variance, standard deviation:



```
> m <- 1/p
> v <- (1-p)/(p**2)
> s <- sqrt(v)
> m
[1] 2.50627
> v
[1] 3.775118
> s
[1] 1.942966
```

- On average, we will need to sample 2.5 (around 2 or 3) observations to get a wine with a “6” quality.

Binomial Distribution

- Distribution that describes the number of successes in a fixed number of trials.
- The probability of observing exactly k successes in n independent trials is given by

$$P = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

Binomial Distribution

- Mean, variance, and standard distribution of a binomial distribution:
  - $\mu = np$
  - $\sigma^2 = np(1-p)$
  - $\sigma = \sqrt{np(1-p)}$

Binomial Distribution

- Four conditions to check:
  - The trials are independent
  - The number of trials, n, is fixed
  - Each trial outcome can be classified as a success or failure
  - The probability of a success, p, is the same for each trial

Binomial Distribution

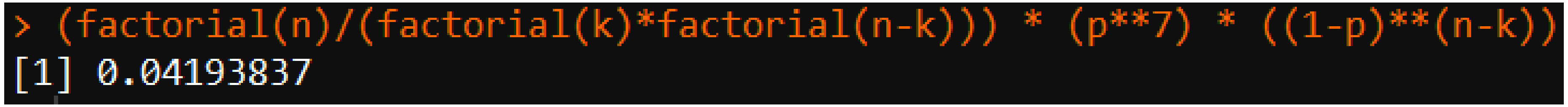
- Selecting 10 observations randomly, what is the probability that 7 will be rated a 6 (a success)?

k = 7

n = 10

p = .399

$$P = \frac{10!}{7!\,(10-7)!}0.399^7(1-0.399)^{10-7} = 0.042 = 4.2\%$$



```
> (factorial(n)/(factorial(k)*factorial(n-k))) * (p**7) * ((1-p)**(n-k))
[1] 0.04193837
```

Binomial Distribution

- Selecting 10 observations randomly, what is the probability that 5 will be rated a 6 (a success)?

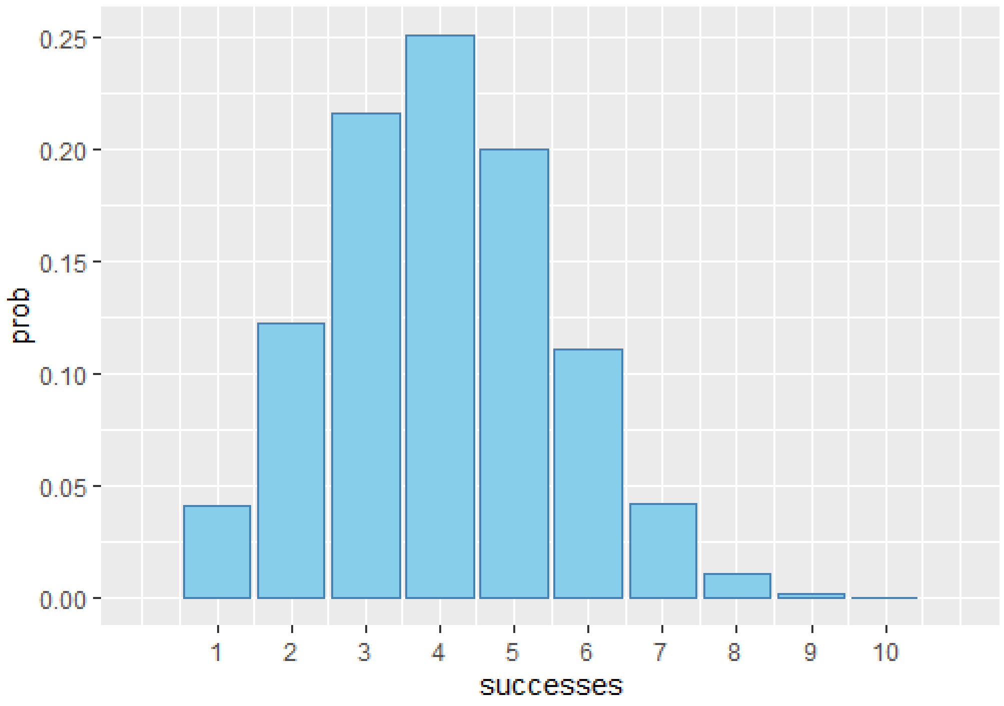
  k = 5

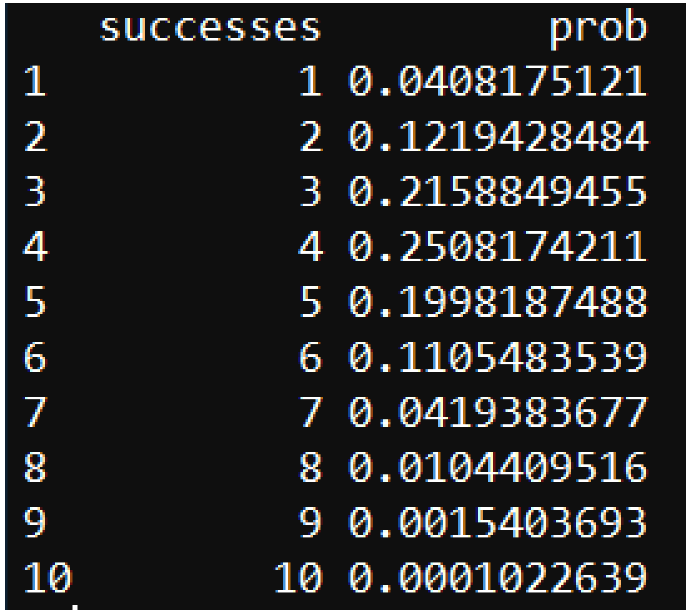
  n = 10

  p = .399

$$P = \frac{10!}{5!\,(10-5)!}\,0.399^{5}(1-0.399)^{10-5} = 0.199 = 19.9\%$$

Binomial Distribution





```
    successes         prob
1           1 0.0408175121
2           2 0.1219428484
3           3 0.2158849455
4           4 0.2508174211
5           5 0.1998187488
6           6 0.1105483539
7           7 0.0419383677
8           8 0.0104409516
9           9 0.0015403693
10         10 0.0001022639
```

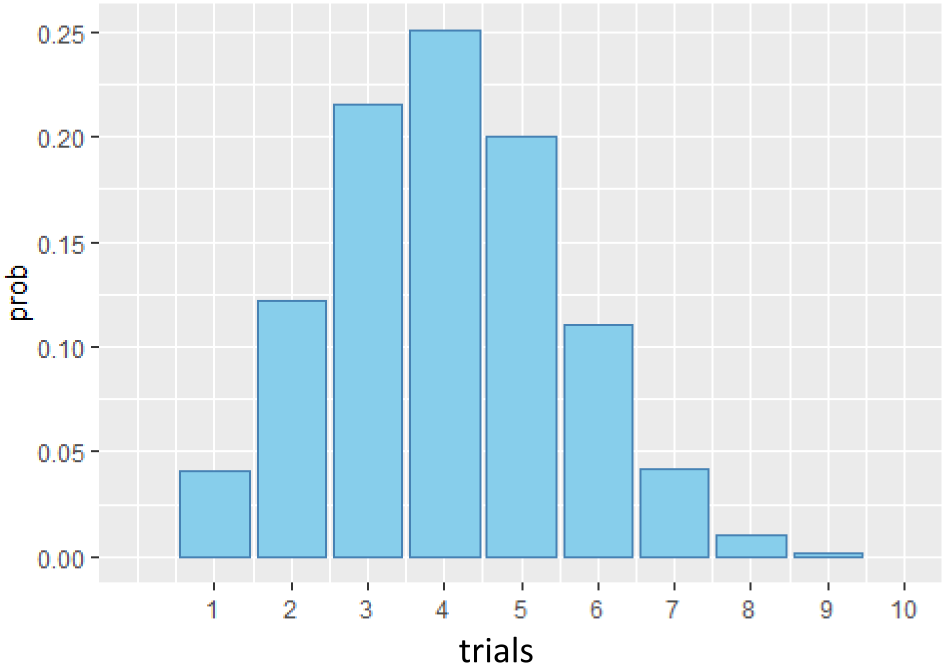
Binomial Distribution

- Normal Approximation of the Binomial Distribution:
  - The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 10.

  - $np = 10 \ * 0.399 = 3.99$
  - $n(1-p) = 10(1-0.399) = 6.01$

  - Cannot approximate to normal with only 10 trials

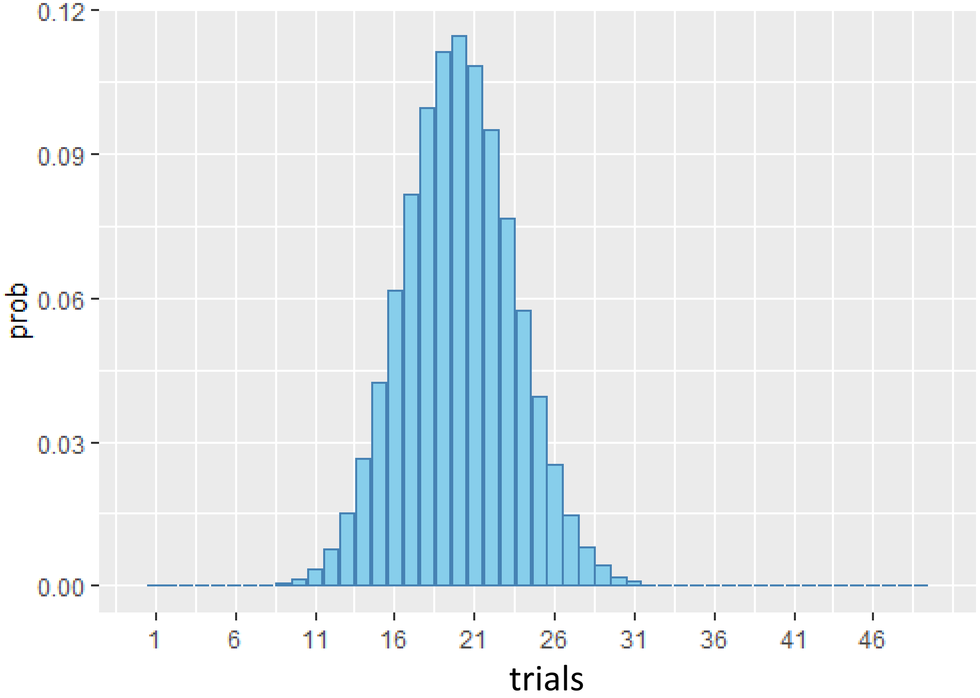
Binomial Distribution



n=10; can't approximate to normal

$$np = 10 \ * 0.399 = 3.99$$

$$n(1 - p) = 10(1 - 0.399) = 6.01$$



n=50; can approximate to normal

$$np = 50 \ * 0.399 = 19.95$$

$$n(1 - p) = 50(1 - 0.399) = 30.05$$

Negative Binomial Distribution

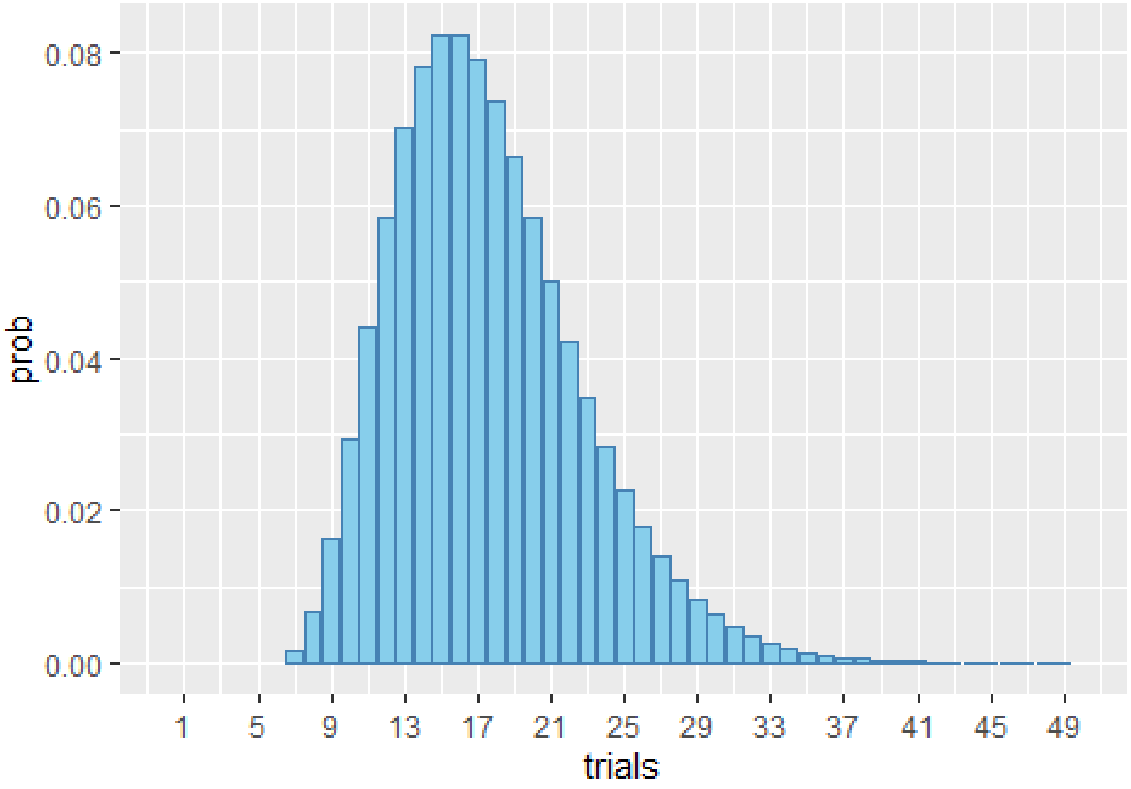
- Distribution that describes the probability of observing the k^{th} success on the n^{th} trial.

- The probability of observing the k^{th} success on the n^{th} trial, where all trials are independent:

$$P = \frac{(n-1)!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k}$$

Negative Binomial Distribution

- Probability of observing the 7^{th} success on the n^{th} trial.
  - n between 1 and 50



Poisson Distribution

- Distribution that estimates the number of events in a large population over some unit time.
  - Days, weeks, months, etc.

- How many events we expect to observe is the rate, λ

- $P(observe\ k\ events) = \frac{\lambda^k e^{-\lambda}}{k!}$

Poisson Distribution

- *A coffee shop serves an average of 75 customers per hour during the morning rush.*
  - λ = 75
  - √λ = 8.66

- *Would it be unusually low if only 60 customers showed up in one hour during this time of day?*
  - *Z* = $\frac{x-\mu}{\sigma} = \frac{60\ -75}{8.66} = -1.73$
  - 2 standard deviations = $75 \pm 2 \times 8.66 = (57.68, 92.32)$
  - No. 60 customers are within 2 standard deviations from the mean

- *Calculate the probability that this coffee shop serves 70 customers in one hour during this time of day.*
  - $P(70) = \frac{75^{70}e^{-75}}{70!}$ = 0.040 = 4%