Social Sciences Intro to Statistics

Week 3.2 Distributions

Week 3: Learning goal - Articulate the descriptors of normal distribution and skewness.

Introduction

Load packages:

```
library(tidyverse)
library(labelled)
library(patchwork)
library(ggplot2)

# Load ipeds dataset from course website
load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/ipeds/outpu
```

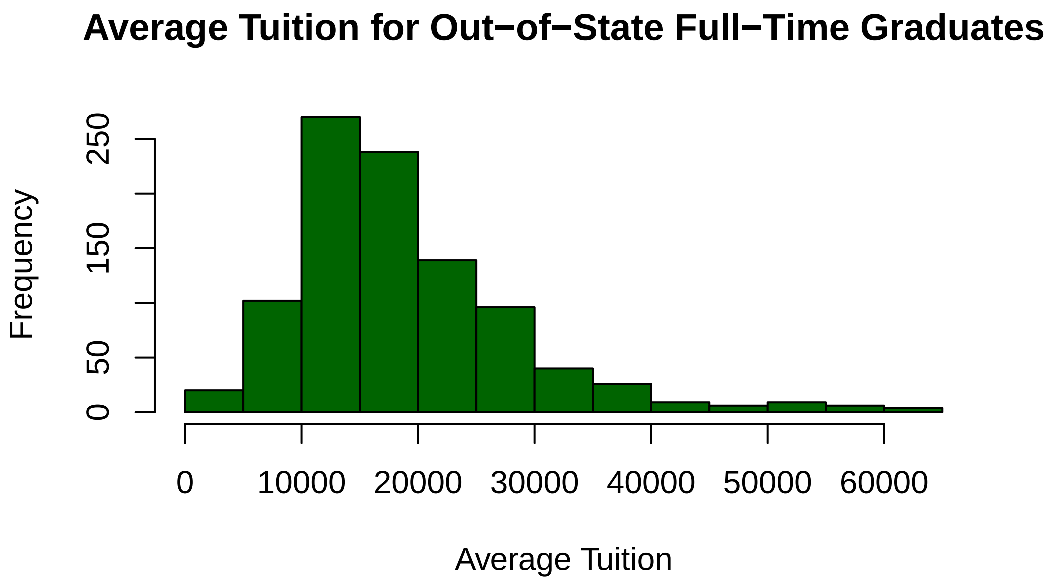
```
#> Rows: 965
#> Columns: 38
#> $ instnm          <chr> "Alabama A & M University", "University of Alabama a~
#> $ unitid          <dbl> 100654, 100663, 100706, 100724, 100751, 100830, 1008~
#> $ opeid6          <chr> "001002", "001052", "001055", "001005", "001051", "0~
#> $ opeid           <chr> "00100200", "00105200", "00105500", "00100500", "001~
#> $ control         <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, ~
#> $ c15basic        <dbl+lbl> 18, 15, 16, 19, 16, 18, 16, 20, 18, 18, 19, 18, ~
#> $ stabbr          <chr+lbl> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
#> $ city            <chr> "Normal", "Birmingham", "Huntsville", "Montgomery", ~
#> $ zip             <chr> "35762", "35294-0110", "35899", "36104-0271", "35487~
#> $ locale          <dbl+lbl> 12, 12, 12, 12, 13, 12, 13, 12, 23, 43, 21, 13, ~
#> $ region          <dbl+lbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
#> $ tuit_grad_res   <dbl> 10128, 8424, 10632, 7416, 11100, 7812, 10386, 15325,~
#> $ fee_grad_res    <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
```

```
#> $ tuit_grad_nres    <dbl> 20160, 19962, 24430, 14832, 31460, 17550, 31158, 153~
#> $ fee_grad_nres     <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_md_res       <dbl> NA, 31198, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_res        <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_md_nres      <dbl> NA, 62714, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_nres       <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_law_res      <dbl> NA, NA, NA, NA, 24080, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_res       <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ tuit_law_nres     <dbl> NA, NA, NA, NA, 44470, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_nres      <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ books_supplies    <dbl> 1600, 1200, 2416, 1600, 800, 1200, 1200, 1800, 998, ~
#> $ roomboard_off     <dbl> 9520, 14330, 11122, 7320, 14426, 10485, 14998, 8020,~
#> $ oth_expense_off   <dbl> 3090, 6007, 4462, 5130, 4858, 4030, 6028, 4600, 3318~
#> $ tuitfee_grad_res  <dbl> 11542, 8424, 11686, 10156, 11790, 8578, 12170, 16225~
#> $ tuitfee_grad_nres <dbl> 21574, 19962, 25484, 17572, 32150, 18316, 32942, 162~
#> $ tuitfee_md_res    <dbl> NA, 34662, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_md_nres   <dbl> NA, 66178, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_law_res   <dbl> NA, NA, NA, NA, 24380, NA, NA, 39325, NA, NA, NA, NA~
#> $ tuitfee_law_nres  <dbl> NA, NA, NA, NA, 44770, NA, NA, 39325, NA, NA, NA, NA~
#> $ coa_grad_res      <dbl> 25752, 29961, 29686, 24206, 31874, 24293, 34396, 306~
#> $ coa_grad_nres     <dbl> 35784, 41499, 43484, 31622, 52234, 34031, 55168, 306~
#> $ coa_md_res        <dbl> NA, 56199, NA, NA, 51282, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_md_nres       <dbl> NA, 87715, NA, NA, 82798, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_law_res       <dbl> NA, NA, NA, NA, 44464, NA, NA, 53745, NA, NA, NA, NA~
#> $ coa_law_nres      <dbl> NA, NA, NA, NA, 64854, NA, NA, 53745, NA, NA, NA, NA~
#> Rows: 200
#> Columns: 4
#> $ norm_dist    <dbl> 42.70513, 50.24400, 61.29008, 45.47494, 44.74406, 47.9912~
#> $ rskew_dist   <dbl> 0.34451771, 0.31359906, 0.09375337, 0.05581678, 0.0744584~
#> $ lskew_dist   <dbl> 0.6554823, 0.6864009, 0.9062466, 0.9441832, 0.9255415, 0.~
#> $ stdnorm_dist <dbl> -1.45897348, 0.04880097, 2.25801577, -0.90501164, -1.0511~
#> [1] 32528.35
#> [1] 31620.8
```

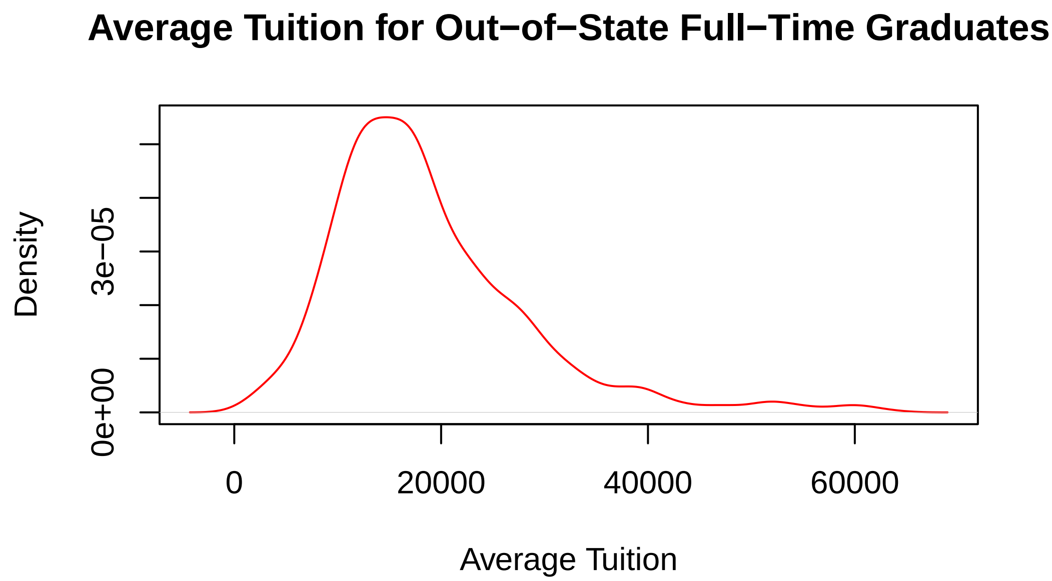
Distributions

Distributions help us further understand our data as it provides a snapshot of the data. Distribution shows us how often each value appears in our dataset (frequency). Distributions tell us where the average value is (central tendency), the spread of the dataset (what the variability is), if the values are evenly spread out (normal) or if there is more values on one side (skewness).

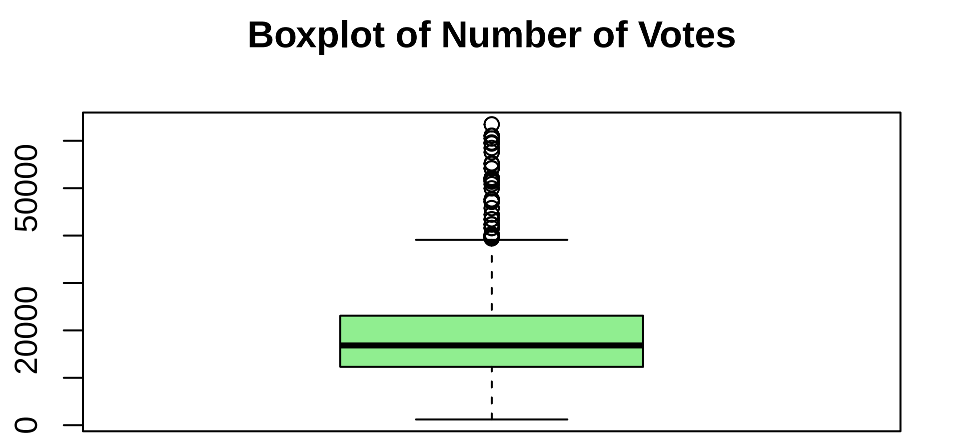
```
# Distribution with a histogram for out-of-state average tuition for full-time graduates
hist(df_ipeds_pop$tuit_grad_nres, breaks = 20, col = "darkgreen", main = "Average Tuition fo
```




```
# Distribution with a density plot
plot(density(df_ipeds_pop$tuit_grad_nres), main = "Average Tuition for Out-of-State Full-Time
```




```r
# Distribution with a box plot
boxplot(df_ipeds_pop$tuit_grad_nres, main = "Boxplot of Number of Votes", col = "lightgreen")
```

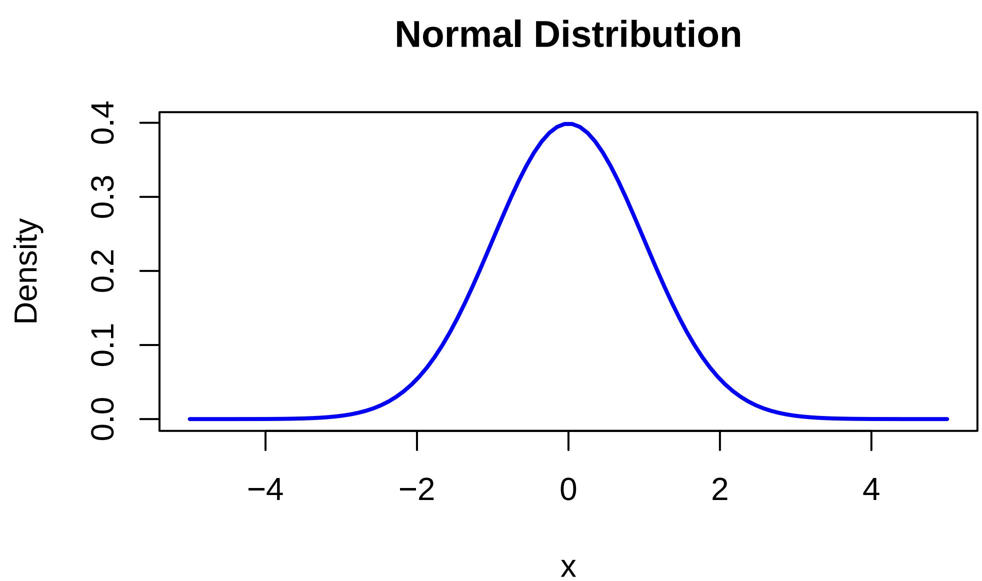


Normal distribution

Normal distributions are continuous probability distributions that are symmetric around the mean. Normal distributions have a bell-shaped curve, where the mean, median, and mode of the distribution are all equal and located at the center of the distribution. The standard deviation of our normal distribution tells us the spread of the distribution. The larger the standard deviation, the wider the normal distribution. The smaller the standard deviation, the narrower the normal distribution. For datasets that have a normal distribution, about 68% of the data will fall within one standard deviation of the mean, and 95% of the data will fall within two standard deviations, and 99.7% of the data falls within three standard deviations.

When the mean, median, and mode are all the same, we are looking at a normal distribution. If the mean and median are equal, we know that the distribution is symmetric or has a “bell” shape.

```r
# Example of normal distribution
x <- seq(-5, 5, length.out = 100)  # Range of x values
y <- dnorm(x, mean = 0, sd = 1)      # PDF values for the normal distribution

# Plot the normal distribution
plot(x, y, type = "l", lwd = 2, col = "blue",
     xlab = "x", ylab = "Density",
     main = "Normal Distribution")
```



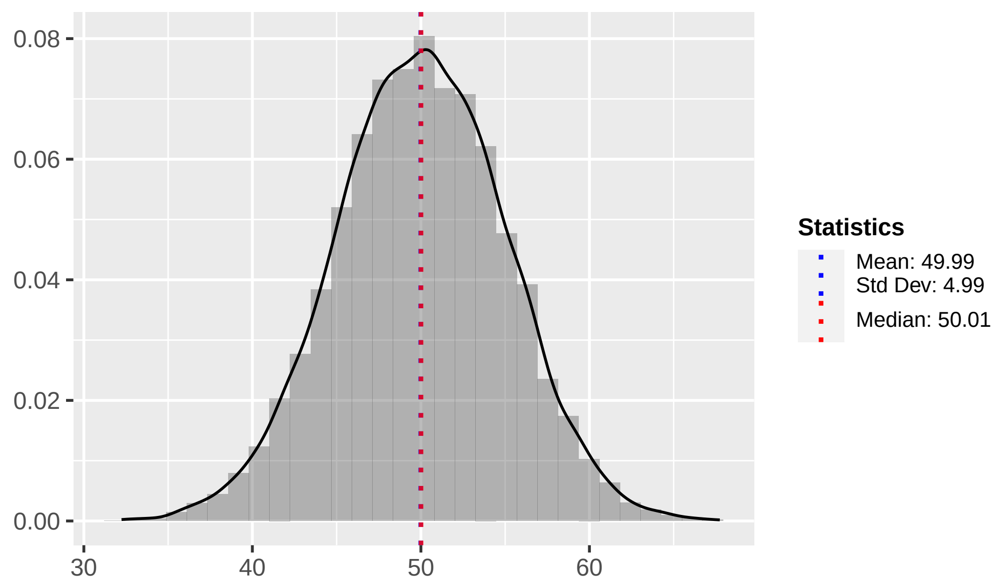
We generated a variable `df_generated_pop$norm_dist` that has a normal distribution and then plot the variable to visualize what a normal distribution looks

- Descriptive statistics about the variable `df_generated_pop$norm_dist`
  - It has a mean of 49.99
  - It has a standard deviation of 4.99
    - Standard deviation is a measure of how far away from the mean observations tend to be
    - we can interpet this standard deviation as follows: on average, observations are 4.99 away from the mean of 49.99

We can also visualize the variable `df_generated_pop$norm_dist`, as shown below. Note the following:

- Symmetric, "bell" shape
- The mean is (nearly) identical to the median

```
plot_distribution(df_generated_pop$norm_dist)
```



Skewness (normal, left-skewed, right-skewed)

Skewness measures the asymmetry of the distribution around its mean. In a normal distribution, the skewness is zero, which means that the distribution is symmetric. When the mean and median are not the same, we know that there is skewdness. There are some unusually extreme values on either side of the distribution. When the distribution leans towards the left side, it is left-skewed or negatively skewed. When the distribution leans towards the right side, it is right-skewed or positively skewed.

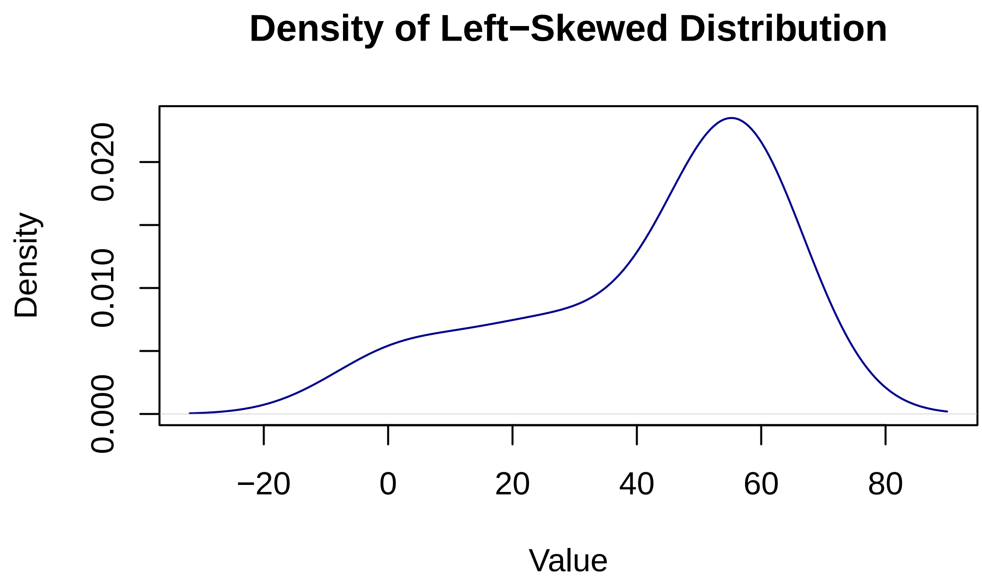
Left-skewed distribution has its mean less than its median, and its median less than its mode. The tail of the distribution extends to the left side. Visually we will see that most of the data points are on the right side of the distribution. And there's value(s) that are unusually small in our dataset. Since this is negatively skewed, the skewness will be less than zero.

- The left tail is longer than right tail, usually due to the presence of more negative outliers than would be expected in a bell shaped variable
  - Negative outliers are defined as observations with very low values (e.g., extreme negative values) compared to most observations
- These negative outliers decrease the value of the mean, such that the value of the mean is lower than the value of the median
- In social science research left-skewed variables are less common than right-skewed variables

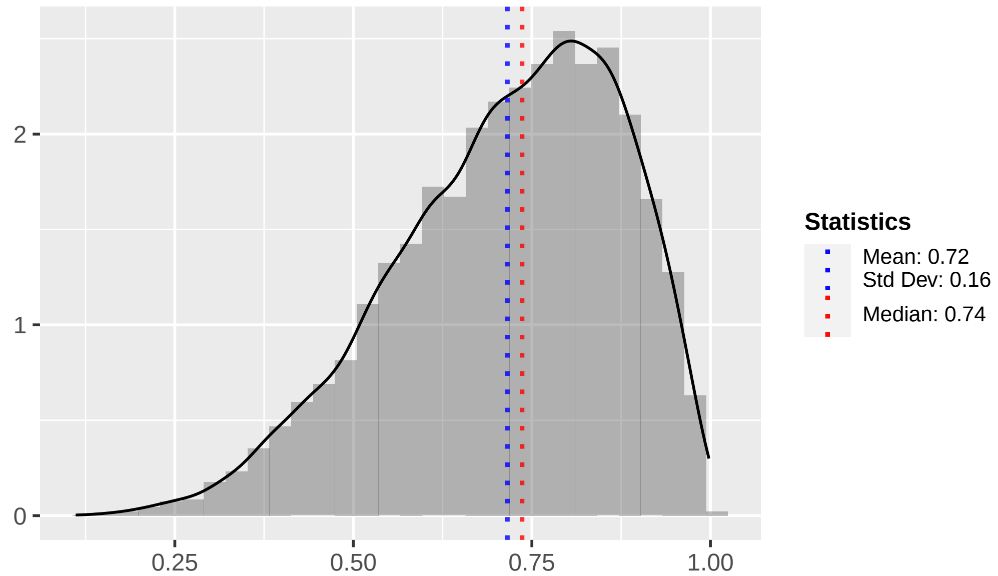
Right-skewed distribution has its mean pulled towards the unusual values, so the mean is greater than its median, and its median greater than its mode. The tail of the distribution extends to the right side. Visually we will see that most of the data points are on the left side of the distribution. And there's value(s) that are unusually large in our dataset. Since this is positively skewed, the skewness will be greater than zero.

- The right "tail" is longer than the left due to the presence of positive outliers, defined as observations with very high values compared to most observations
- There are more positive outliers than you would expect in a bell (normal) shaped variable
- These positive outliers increase the value of the mean, such that the value of the mean is higher than the value of the median
  - Mean > Median
- Real-world variables that tend to be right-skewed
  - such as income; enrollment size, city population

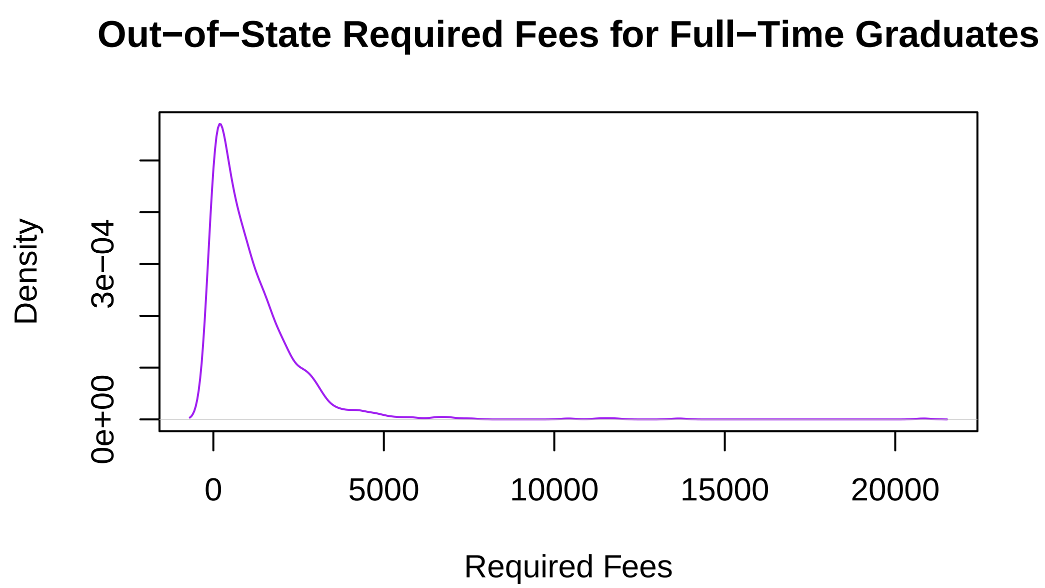
```r
# Example of left-skewed distribution
# We are creating left-skewed dataset
data_left_skewed <- c(1, 20, 35, 55, 56, 56, 56, 57)

# Density plot to show left-skewed distribution
plot(density(data_left_skewed), main = "Density of Left-Skewed Distribution",
     xlab = "Value", col = "darkblue")
```



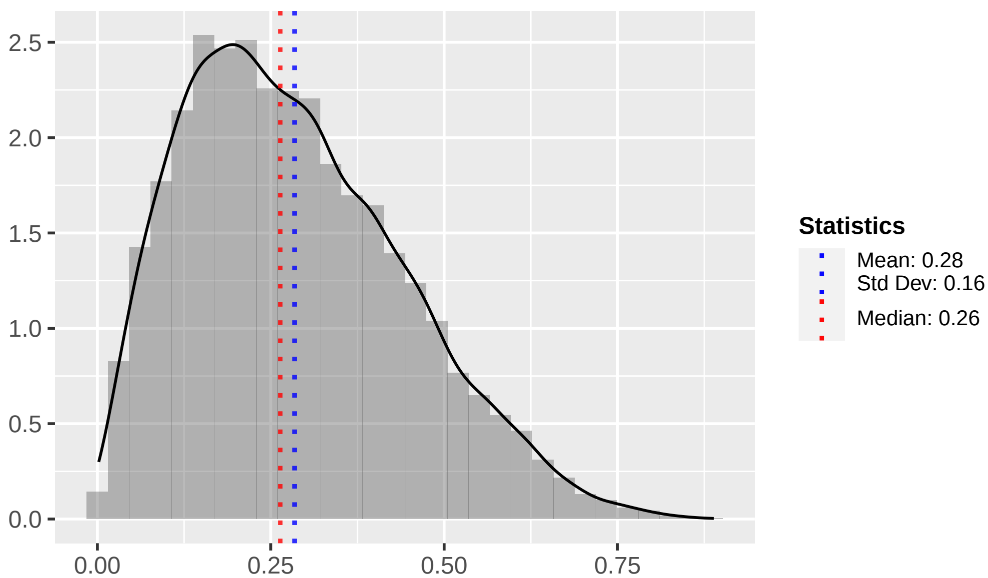

```r
# Another example with the left-skewed distribution that we generated
plot_distribution(df_generated_pop$lskew_dist)
```




```
# Example of right-skewed distribution
# Density plot of out-of-state average tuition for full-time graduates
plot(density(df_ipeds_pop$fee_grad_nres), main = "Out-of-State Required Fees for Full-Time G
```



```r
# Another example with the right-skewed distribution that we generated
plot_distribution(df_generated_pop$rskew_dist)
```

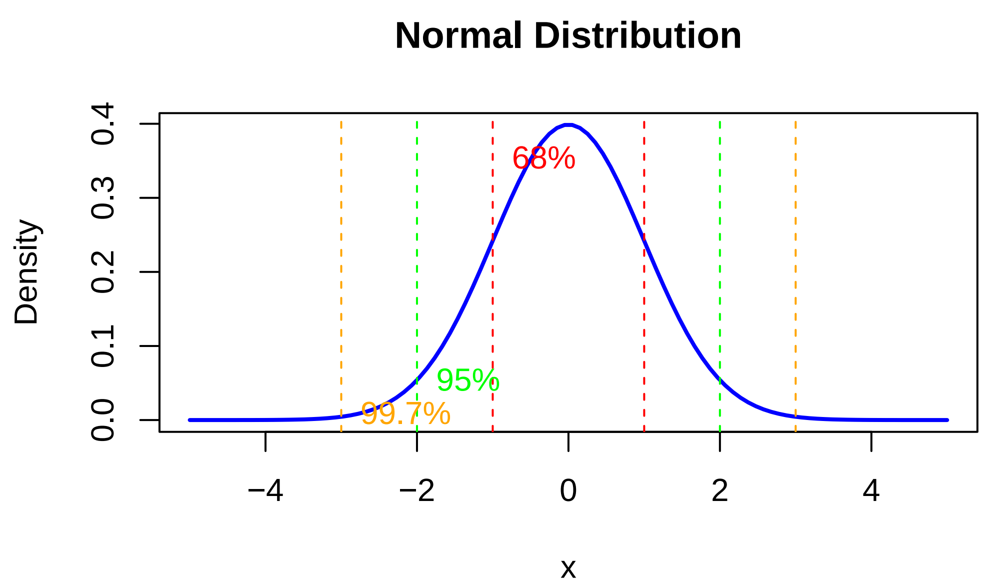


Normal Distributions and the Empirical Rule

The empirical rule states that when you have a normal distribution or approximately a normal, then all of the observed data points fall within 3 standard deviations of the mean.

- About 68% of obs fall within one std. dev of mean
    - i.e., between $x - \hat{\sigma x}$ and $x + \hat{\sigma x}$
- About 95% of obs fall within two std. dev of mean
    - i.e., between $x - 2\hat{\sigma}_x$ and $x + 2\hat{\sigma}_x$
- About 99% of obs fall within three std. dev of mean
    - i.e., between $x - 3\hat{\sigma}_x$ and $x + 3\hat{\sigma}_x$

```r
# Example of normal distribution
x <- seq(-5, 5, length.out = 100)  # Range of x values
y <- dnorm(x, mean = 0, sd = 1)      # PDF values for the normal distribution

# Plot the normal distribution
plot(x, y, type = "l", lwd = 2, col = "blue",
     xlab = "x", ylab = "Density",
     main = "Normal Distribution")

# Add vertical lines for one, two, and three standard deviations
abline(v = c(-1, 1), col = "red", lty = 2)  # One SD
abline(v = c(-2, 2), col = "green", lty = 2)  # Two SD
abline(v = c(-3, 3), col = "orange", lty = 2)  # Three SD

# Add text annotations for the percentages
text(-1, 0.35, "68%", col = "red", pos = 4)
text(-2, 0.05, "95%", col = "green", pos = 4)
text(-3, 0.005, "99.7%", col = "orange", pos = 4)
```



Why is the empirical rule so important for inferential statistics?

- If a variable has an approximately normal distribution, then we know how likely it would be to observe a variable that is a certain number of standard deviations away from the mean
- For example:
  - only about 2.5% of observations have a value higher than two standard deviations or more from the mean;
  - the variable `norm_dist` has a mean of about `50` and a standard deviation of about `5`, so the value of `40` would be about two standard deviations below the mean. the empirical rule tells us that only about 2.5% of observations would have a value less than `40`
- you might say, but most real-life variables are unlikely to have a normal distribution
  - True! But the "sampling distribution" – discussed below – which is the basis for all inferential statistics/hypothesis testing, **always** has a normal distribution so long as our sample size is large enough

Z-scores

The "z-score" of an observation is the number of standard deviations away from the mean.

The z-score formula

- where x is some variable of interest; subscript i refers to observations
- $z_i = (x_i - \bar{x})/(\hat{\sigma}_x)$
- in words:
  - z score for observation i equals the difference between the observation x_i and the mean $\bar{x}$ divided by the standard deviation $\hat{\sigma}_x$
- Intuition behind z-score
  - It is just the difference between an observation value and the mean value, scaled in terms of standard deviations
  - That's why we say that the z-score represents the number of standard deviations away from the mean

Calculating z-score for the variable `norm_dist` from data frame `df_generated_pop`

```r
# components of z-score
mean(df_generated_pop$norm_dist, na.rm = TRUE)
#> [1] 49.98631
sd(df_generated_pop$norm_dist, na.rm = TRUE)
#> [1] 4.991961

#create new variable z_norm_dist
df_generated_pop <- df_generated_pop %>% mutate(
  z_norm_dist = (norm_dist - mean(norm_dist, na.rm = TRUE))/sd(norm_dist, na.rm = TRUE)
)

#list a few observations
df_generated_pop %>% select(norm_dist,z_norm_dist)
#>       norm_dist   z_norm_dist
#> 1      43.07465 -1.384559e+00
#> 2      50.19162  4.112727e-02
#> 3      46.18485 -7.615166e-01
#> 4      51.06153  2.153904e-01
#> 5      57.12769  1.430576e+00
#> 6      53.72240  7.484211e-01
#> 7      53.50115  7.040994e-01
#> 8      48.85323 -2.269816e-01
#> 9      50.98547  2.001536e-01
#> 10     56.03577  1.211840e+00
#> 11     51.59168  3.215918e-01
#> 12     42.88101 -1.423349e+00
#> 13     47.97455 -4.030008e-01
#> 14     54.97693  9.997319e-01
#> 15     54.79409  9.631042e-01
#> 16     54.59044  9.223088e-01
#> 17     49.24515 -1.484703e-01
#> 18     43.88466 -1.222296e+00
#> 19     45.65588 -8.674811e-01
#> 20     44.78757 -1.041422e+00
#> 21     44.48181 -1.102673e+00
#> 22     52.22093  4.476428e-01
#> 23     48.97525 -2.025383e-01
#> 24     58.37816  1.681073e+00
#> 25     49.34339 -1.287914e-01
#> 26     49.00059 -1.974625e-01
#> 27     50.27456  5.774322e-02
#> 28     46.58917 -6.805217e-01
```

```
#> 29     46.36148 -7.261337e-01
#> 30     45.69048 -8.605499e-01
#> 31     49.81238 -3.484117e-02
#> 32     41.84288 -1.631308e+00
#> 33     50.88583  1.801943e-01
#> 34     49.93750 -9.778555e-03
#> 35     48.02841 -3.922098e-01
#> 36     51.75781  3.548715e-01
#> 37     54.39384  8.829251e-01
#> 38     51.02327  2.077260e-01
#> 39     45.56310 -8.860674e-01
#> 40     47.61392 -4.752422e-01
#> 41     48.66130 -2.654297e-01
#> 42     57.92930  1.591155e+00
#> 43     50.23450  4.971850e-02
#> 44     51.78248  3.598133e-01
#> 45     49.39310 -1.188331e-01
#> 46     49.81954 -3.340759e-02
#> 47     45.09426 -9.799851e-01
#> 48     47.82870 -4.322168e-01
#> 49     49.66258 -6.485066e-02
#> 50     54.90947  9.862182e-01
#> 51     47.68974 -4.600541e-01
#> 52     48.87450 -2.227194e-01
#> 53     45.76776 -8.450685e-01
#> 54     50.36523  7.590633e-02
#> 55     48.62482 -2.727370e-01
#> 56     48.06787 -3.843063e-01
#> 57     49.76898 -4.353517e-02
#> 58     45.87053 -8.244814e-01
#> 59     45.72983 -8.526672e-01
#> 60     50.59368  1.216704e-01
#> 61     51.41798  2.867960e-01
#> 62     59.65043  1.935937e+00
#> 63     44.29736 -1.139622e+00
#> 64     43.38941 -1.321505e+00
#> 65     56.14416  1.233553e+00
#> 66     47.25772 -5.465969e-01
#> 67     49.36996 -1.234680e-01
#> 68     53.43859  6.915686e-01
#> 69     53.52600  7.090784e-01
#> 70     54.00739  8.055115e-01
```

```
#> 71      47.68558 -4.608862e-01
#> 72      45.57723 -8.832371e-01
#> 73      41.84536 -1.630812e+00
#> 74      52.81115  5.658784e-01
#> 75      48.37054 -3.236743e-01
#> 76      51.20688  2.445067e-01
#> 77      55.42935  1.090360e+00
#> 78      59.53531  1.912875e+00
#> 79      54.06464  8.169790e-01
#> 80      52.47444  4.984281e-01
#> 81      49.52426 -9.255861e-02
#> 82      55.69390  1.143357e+00
#> 83      52.51157  5.058659e-01
#> 84      47.42293 -5.135017e-01
#> 85      37.65805 -2.469623e+00
#> 86      45.63724 -8.712155e-01
#> 87      54.82044  9.683830e-01
#> 88      54.55398  9.150054e-01
#> 89      59.62904  1.931653e+00
#> 90      48.48547 -3.006524e-01
#> 91      44.72646 -1.053663e+00
#> 92      52.09058  4.215318e-01
#> 93      53.50636  7.051445e-01
#> 94      51.23379  2.498980e-01
#> 95      52.32148  4.677852e-01
#> 96      48.02266 -3.933627e-01
#> 97      53.56535  7.169610e-01
#> 98      55.92506  1.189663e+00
#> 99      40.44425 -1.911485e+00
#> 100     55.57465  1.119468e+00
#> 101     47.65350 -4.673136e-01
#> 102     44.25999 -1.147109e+00
#> 103     55.51999  1.108518e+00
#> 104     48.55375 -2.869733e-01
#> 105     48.61336 -2.750329e-01
#> 106     45.62109 -8.744506e-01
#> 107     50.01640  6.026808e-03
#> 108     51.32796  2.687617e-01
#> 109     49.37988 -1.214815e-01
#> 110     49.35537 -1.263912e-01
#> 111     51.49556  3.023364e-01
#> 112     50.09150  2.107106e-02
```

```
#> 113    44.24969 -1.149171e+00
#> 114    47.85302 -4.273455e-01
#> 115    54.25001  8.541142e-01
#> 116    48.92585 -2.124331e-01
#> 117    46.91293 -6.156659e-01
#> 118    50.54714  1.123460e-01
#> 119    46.46744 -7.049065e-01
#> 120    57.71488  1.548203e+00
#> 121    51.75401  3.541090e-01
#> 122    55.86310  1.177251e+00
#> 123    49.83313 -3.068490e-02
#> 124    51.80722  3.647684e-01
#> 125    55.22788  1.050002e+00
#> 126    50.51606  1.061210e-01
#> 127    51.65898  3.350723e-01
#> 128    42.15721 -1.568342e+00
#> 129    52.20135  4.437213e-01
#> 130    59.40936  1.887644e+00
#> 131    56.74553  1.354021e+00
#> 132    41.82459 -1.634974e+00
#> 133    46.42234 -7.139426e-01
#> 134    45.41753 -9.152279e-01
#> 135    38.19487 -2.362087e+00
#> 136    57.64225  1.533654e+00
#> 137    46.85106 -6.280591e-01
#> 138    44.65954 -1.067070e+00
#> 139    56.22430  1.249607e+00
#> 140    52.29938  4.633591e-01
#> 141    50.83853  1.707183e-01
#> 142    48.62498 -2.727035e-01
#> 143    52.46161  4.958578e-01
#> 144    56.69877  1.344654e+00
#> 145    47.61628 -4.747696e-01
#> 146    42.51602 -1.496464e+00
#> 147    43.99288 -1.200616e+00
#> 148    47.96269 -4.053748e-01
#> 149    45.26126 -9.465320e-01
#> 150    54.25833  8.557803e-01
#> 151    47.05784 -5.866362e-01
#> 152    43.18763 -1.361925e+00
#> 153    47.75181 -4.476197e-01
#> 154    46.78337 -6.416187e-01
```

```
#> 155    47.44627 -5.088257e-01
#> 156    51.28601  2.603586e-01
#> 157    47.90222 -4.174887e-01
#> 158    45.23886 -9.510187e-01
#> 159    52.56197  5.159609e-01
#> 160    41.27866 -1.744334e+00
#> 161    47.51244 -4.955712e-01
#> 162    53.82793  7.695611e-01
#> 163    52.50697  5.049429e-01
#> 164    51.65956  3.351881e-01
#> 165    49.54462 -8.848003e-02
#> 166    46.71659 -6.549972e-01
#> 167    44.19576 -1.159975e+00
#> 168    58.27966  1.661341e+00
#> 169    46.25395 -7.476747e-01
#> 170    53.10960  6.256645e-01
#> 171    42.64097 -1.471433e+00
#> 172    53.99098  8.022232e-01
#> 173    51.39731  2.826549e-01
#> 174    53.28937  6.616761e-01
#> 175    52.96863  5.974248e-01
#> 176    47.11037 -5.761148e-01
#> 177    52.38031  4.795718e-01
#> 178    44.06183 -1.186804e+00
#> 179    53.58844  7.215862e-01
#> 180    50.67349  1.376565e-01
#> 181    57.11040  1.427113e+00
#> 182    50.20365  4.353888e-02
#> 183    48.30795 -3.362123e-01
#> 184    47.20942 -5.562718e-01
#> 185    55.13907  1.032211e+00
#> 186    48.87265 -2.230912e-01
#> 187    43.84185 -1.230871e+00
#> 188    46.20878 -7.567234e-01
#> 189    55.42759  1.090009e+00
#> 190    44.75451 -1.048046e+00
#> 191    51.45200  2.936096e-01
#> 192    47.76213 -4.455518e-01
#> 193    48.90570 -2.164708e-01
#> 194    41.97278 -1.605287e+00
#> 195    50.37420  7.770258e-02
#> 196    47.39041 -5.200168e-01
```

```
#> 197    49.67549 -6.226451e-02
#> 198    55.25685  1.055805e+00
#> 199    41.28921 -1.742221e+00
#> 200    58.66274  1.738080e+00
#> 201    58.16765  1.638902e+00
#> 202    48.11675 -3.745139e-01
#> 203    49.16881 -1.637636e-01
#> 204    49.83924 -2.946168e-02
#> 205    52.56577  5.167222e-01
#> 206    56.34114  1.273012e+00
#> 207    44.69859 -1.059248e+00
#> 208    53.09831  6.234029e-01
#> 209    45.90186 -8.182058e-01
#> 210    47.56129 -4.857857e-01
#> 211    49.54457 -8.848975e-02
#> 212    51.53964  3.111656e-01
#> 213    48.12558 -3.727453e-01
#> 214    46.06072 -7.863831e-01
#> 215    44.91830 -1.015234e+00
#> 216    36.87491 -2.626504e+00
#> 217    48.85084 -2.274599e-01
#> 218    48.67041 -2.636032e-01
#> 219    51.81908  3.671450e-01
#> 220    50.57843  1.186149e-01
#> 221    53.78653  7.612686e-01
#> 222    57.57656  1.520495e+00
#> 223    36.70108 -2.661325e+00
#> 224    46.55910 -6.865459e-01
#> 225    47.86711 -4.245232e-01
#> 226    47.86897 -4.241494e-01
#> 227    50.74152  1.512845e-01
#> 228    62.63905  2.534623e+00
#> 229    49.28300 -1.408886e-01
#> 230    55.16534  1.037474e+00
#> 231    48.21068 -3.556988e-01
#> 232    53.78861  7.616844e-01
#> 233    61.49371  2.305187e+00
#> 234    49.62215 -7.294942e-02
#> 235    64.30624  2.868599e+00
#> 236    42.95605 -1.408316e+00
#> 237    50.38219  7.930323e-02
#> 238    41.13395 -1.773322e+00
```

```
#> 239    48.94736 -2.081247e-01
#> 240    48.48106 -3.015345e-01
#> 241    46.94089 -6.100648e-01
#> 242    49.29790 -1.379047e-01
#> 243    50.60679  1.242952e-01
#> 244    53.66853  7.376301e-01
#> 245    61.13049  2.232426e+00
#> 246    52.88396  5.804632e-01
#> 247    50.82255  1.675168e-01
#> 248    53.97446  7.989144e-01
#> 249    55.18798  1.042010e+00
#> 250    42.73662 -1.452274e+00
#> 251    52.47081  4.977009e-01
#> 252    45.84590 -8.294154e-01
#> 253    54.03994  8.120309e-01
#> 254    51.04705  2.124897e-01
#> 255    48.61216 -2.752728e-01
#> 256    42.65874 -1.467875e+00
#> 257    52.48662  5.008663e-01
#> 258    50.73219  1.494158e-01
#> 259    53.20796  6.453673e-01
#> 260    50.34488  7.182961e-02
#> 261    50.09265  2.130299e-02
#> 262    48.64789 -2.681159e-01
#> 263    48.83822 -2.299875e-01
#> 264    45.26051 -9.466825e-01
#> 265    51.36870  2.769229e-01
#> 266    49.65151 -6.706690e-02
#> 267    47.77669 -4.426366e-01
#> 268    50.19385  4.157435e-02
#> 269    48.42718 -3.123282e-01
#> 270    59.36426  1.878610e+00
#> 271    43.05332 -1.388831e+00
#> 272    53.48605  7.010742e-01
#> 273    44.19035 -1.161058e+00
#> 274    45.36129 -9.264929e-01
#> 275    51.46921  2.970567e-01
#> 276    51.02736  2.085461e-01
#> 277    51.98211  3.998027e-01
#> 278    49.12527 -1.724864e-01
#> 279    45.66979 -8.646950e-01
#> 280    46.24326 -7.498160e-01
```

```
#> 281    41.70097 -1.659737e+00
#> 282    37.68444 -2.464336e+00
#> 283    54.98680  1.001708e+00
#> 284    56.00621  1.205919e+00
#> 285    39.55513 -2.089596e+00
#> 286    55.01385  1.007126e+00
#> 287    48.99485 -1.986115e-01
#> 288    52.01828  4.070490e-01
#> 289    51.44183  2.915718e-01
#> 290    50.00815  4.375430e-03
#> 291    46.63843 -6.706547e-01
#> 292    54.40426  8.850129e-01
#> 293    46.94176 -6.098898e-01
#> 294    44.85565 -1.027785e+00
#> 295    59.83144  1.972197e+00
#> 296    43.64185 -1.270935e+00
#> 297    50.46526  9.594346e-02
#> 298    51.53531  3.102984e-01
#> 299    50.04766  1.229034e-02
#> 300    46.52315 -6.937467e-01
#> 301    49.75506 -4.632365e-02
#> 302    54.45261  8.946985e-01
#> 303    64.05216  2.817700e+00
#> 304    51.63795  3.308598e-01
#> 305    44.77946 -1.043046e+00
#> 306    46.05020 -7.884888e-01
#> 307    50.99080  2.012222e-01
#> 308    49.22594 -1.523182e-01
#> 309    45.42020 -9.146933e-01
#> 310    45.11280 -9.762719e-01
#> 311    57.78380  1.562009e+00
#> 312    49.69629 -5.809699e-02
#> 313    49.44142 -1.091540e-01
#> 314    44.25667 -1.147774e+00
#> 315    57.53105  1.511377e+00
#> 316    55.13040  1.030474e+00
#> 317    44.95571 -1.007741e+00
#> 318    52.54006  5.115729e-01
#> 319    49.09442 -1.786644e-01
#> 320    43.96592 -1.206017e+00
#> 321    53.06544  6.168182e-01
#> 322    58.23491  1.652376e+00
```

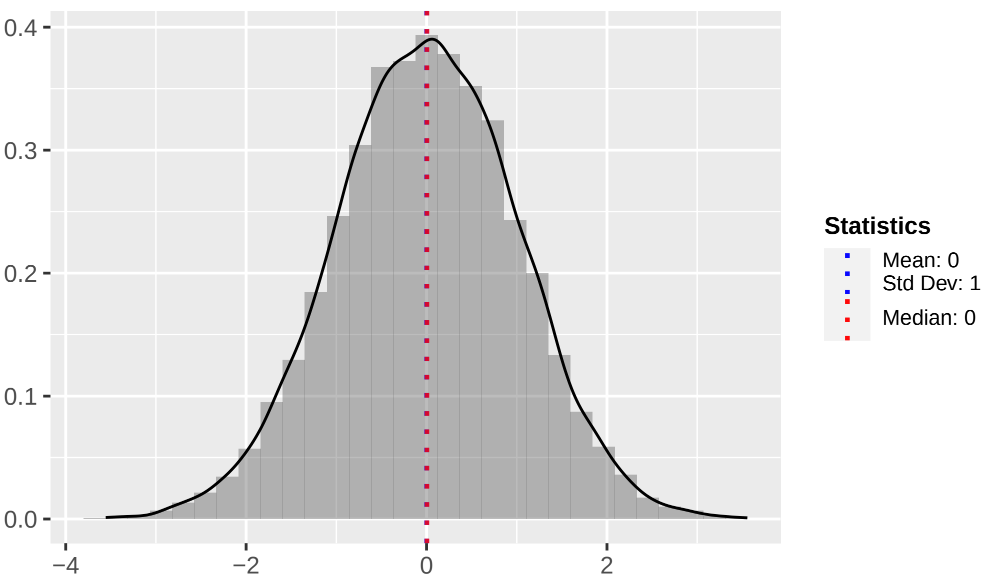
```
#> 323    52.08062  4.195362e-01
#> 324    50.60057  1.230507e-01
#> 325    48.26564 -3.446877e-01
#> 326    41.71115 -1.657697e+00
#> 327    55.38571  1.081619e+00
#> 328    46.28893 -7.406669e-01
#> 329    50.85753  1.745237e-01
#> 330    53.12607  6.289630e-01
#> 331    44.30448 -1.138197e+00
#> 332    47.40936 -5.162209e-01
#> 333    60.19418  2.044862e+00
#> 334    45.67852 -8.629464e-01
#> 335    45.94438 -8.096877e-01
#> 336    53.64923  7.337633e-01
#> 337    43.08763 -1.381958e+00
#> 338    39.76107 -2.048342e+00
#> 339    51.20722  2.445749e-01
#> 340    46.75463 -6.473776e-01
#> 341    41.67283 -1.665373e+00
#> 342    50.40309  8.348982e-02
#> 343    44.67124 -1.064726e+00
#> 344    52.39211  4.819349e-01
#> 345    56.84487  1.373922e+00
#> 346    40.63786 -1.872702e+00
#> 347    52.52391  5.083364e-01
#> 348    53.87990  7.799717e-01
#> 349    54.20902  8.459018e-01
#> 350    52.67425  5.384534e-01
#> 351    46.61784 -6.747780e-01
#> 352    42.75268 -1.449056e+00
#> 353    52.97779  5.992593e-01
#> 354    53.89358  7.827133e-01
#> 355    49.13583 -1.703705e-01
#> 356    53.28542  6.608853e-01
#> 357    52.17630  4.387043e-01
#> 358    46.82345 -6.335900e-01
#> 359    59.92104  1.990145e+00
#> 360    53.30423  6.646531e-01
#> 361    49.70418 -5.651690e-02
#> 362    50.72426  1.478268e-01
#> 363    50.30320  6.348024e-02
#> 364    43.91452 -1.216314e+00
```

```
# mean of z-score variable
round(mean(df_generated_pop$z_norm_dist, na.rm = TRUE), digits = 4)
#> [1] 0
```

Plot the new z-score variable, which has:

- mean of about 0
- standard deviation of about 1

```
plot_distribution(df_generated_pop$z_norm_dist)
```

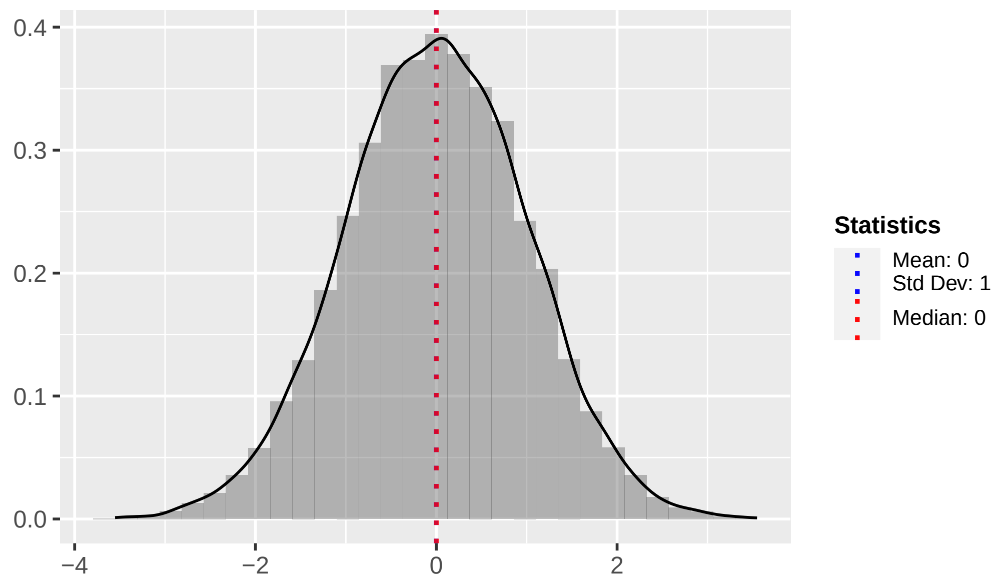


Standard normal distribution

Standard normal distribution is defined as a bell-shaped (i.e., normal) distribution that has a mean of 0 and a standard deviation of 1

Above, we created a variable `z_norm_dist` in the data frame `df_generated_pop` that has a standard normal distribution. Let's investigate and plot this variable:

```r
mean(df_generated_pop$stdnorm_dist, na.rm = TRUE)
#> [1] -0.002737966
sd(df_generated_pop$stdnorm_dist, na.rm = TRUE)
#> [1] 0.9983922

plot_distribution(df_generated_pop$stdnorm_dist)
```



Traits of standard normal distribution:

- The value of each observation is already in terms of z-scores. This means the value of each observation shows how many standard deviations it is from the mean.

Question: if the variable has a standard normal distribution, would it be likely to see an observation with a value of 3? - Answer: No. because a value of 3 would mean that the observation is three standard deviations greater than the mean. we know that for any variable with a normal distribution, less than 1% of observations have a value that is three standard deviations greater than the mean.