Econ 5023: Statistics for Decision Making

Univariate Statistics (X): Monte Carlo Simulation and Parametric Distributions: Continuous Distributions

Le Wang

November 28, 2018

Itinerary:

1. Uniform Distribution
   - 1.1 Applications of Uniform Distributions: Numerical Integration with Monte Carlo Simulation

2. Normal Distribution
   - 2.1 Standard Normal Distribution
   - 2.2 Manipulations of Standard Normal Distribution to Generate Other Normal Distributions: Multiplication and Addition
   - 2.3 Properties of Normal Distribution
   - 2.4 Applications of Normal Distributions
     - 2.4.1 Central Limit Theorem (CLT)
     - 2.4.2 Its Relationship to Z-Statistic
     - 2.4.3 Visualizing CLT

Things to pay attention when discussing a parametric distribution (To Answer Jordan's question earlier!):

1. What is the probability mass/density function? The relationship between a potential value and the (relative) probability
2. How does this density function look like?
3. What are the features of this distribution? Moments (mean, variance, skewness and kurtosis)
4. What kind of things can be characterized by this distribution?

Generally, we have NO idea what distribution data are drawn from!!

But surprisingly, we know the distribution from which some sample statistics (functions of data) are drawn. **Sample Distribution**

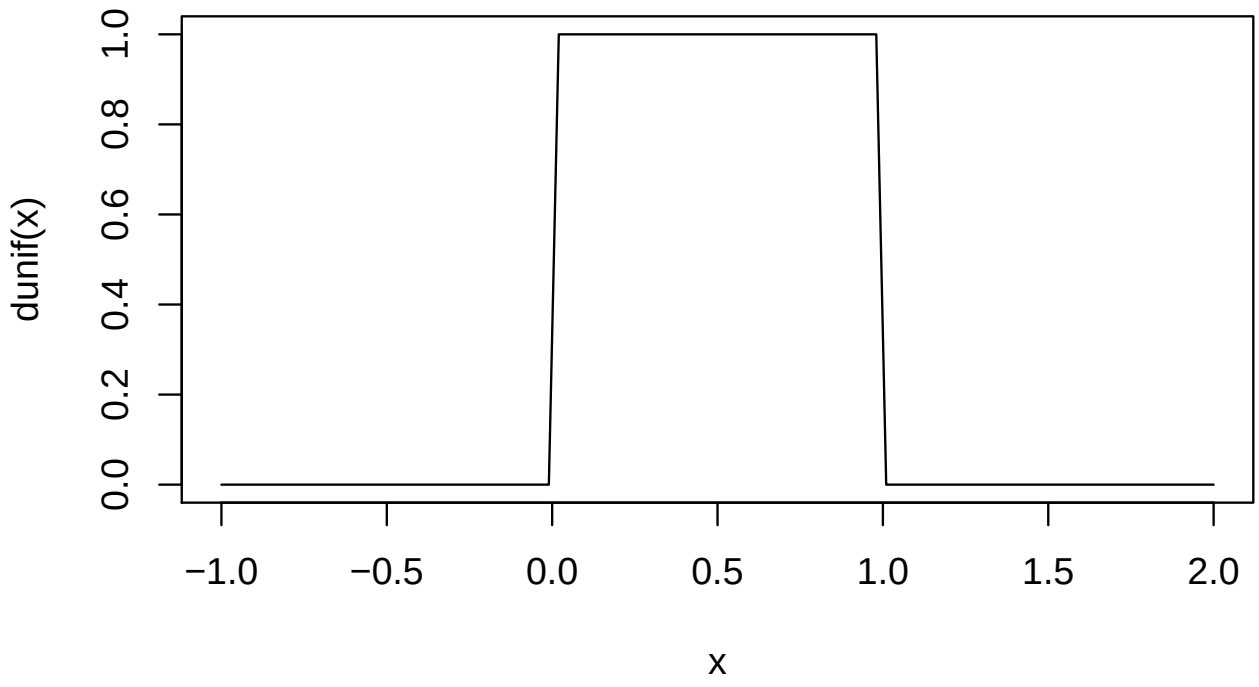
Continuous, Parametric Distribution (I): Uniform Distribution

1. Has a maximum value of b and a minimum value of a.

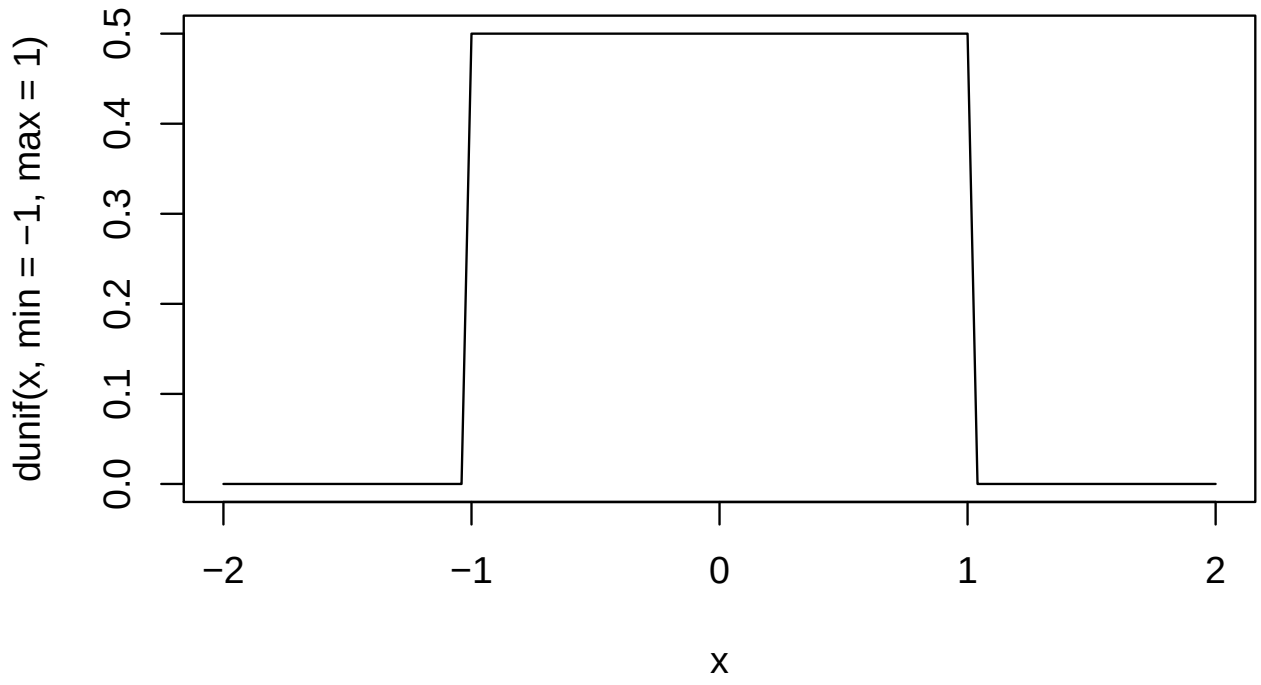
1. Has a maximum value of b and a minimum value of a.
2. Equal probability everywhere between $[a, b]$, zero otherwise. Mathematically,

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

For Example, $[a, b] = [0, 1] \implies \frac{1}{b-a} = \frac{1}{1-0} = 1$



For Example, $[a, b] = [-1, 1] \implies \frac{1}{b-a} = \frac{1}{1-(-1)} = .5$



We will these resutls in a minute

- $[a, b] = [-1, 1] \implies f(x) = .5$
- $[a, b] = [0, 1] \implies f(x) = 1$

CDF for Standard Uniform Variable with the support $[0, 1]$

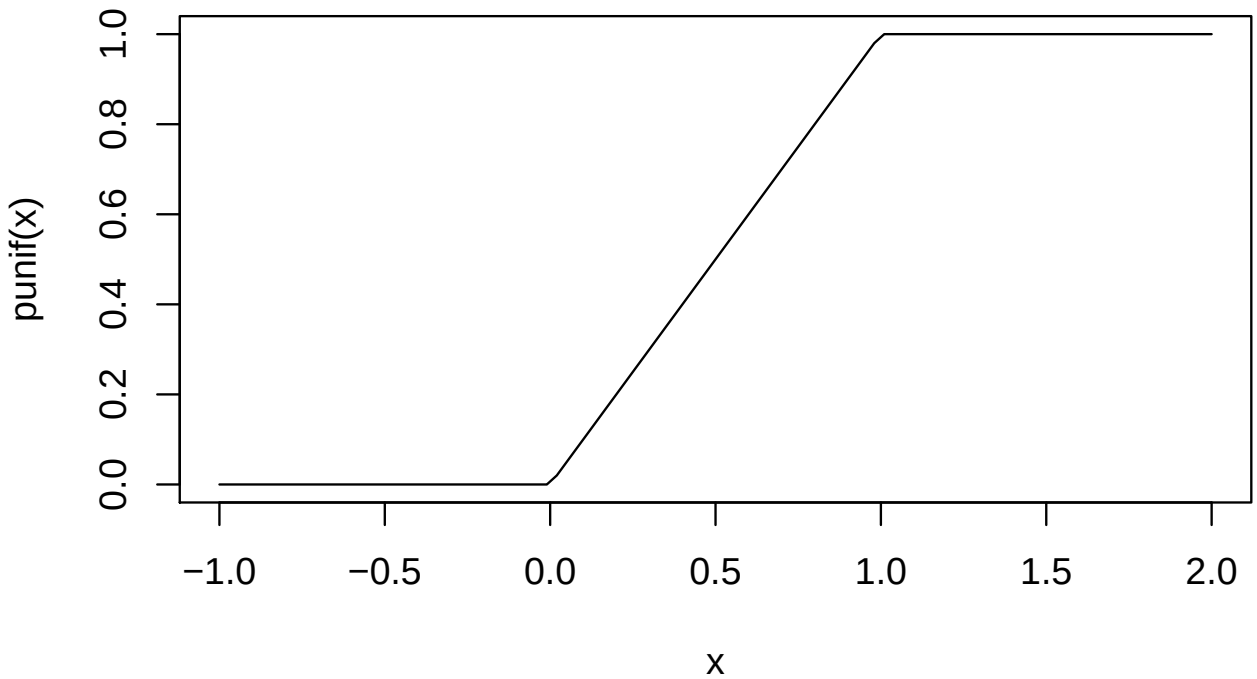
$$\Pr[X \leq x] = x \tag{2}$$

CDF for Standard Uniform Variable with the support $[0, 1]$

$$\Pr[X \leq x] = x \tag{2}$$

$$\Pr[x_2 \leq X \leq x_1] = x_1 - x_2 \tag{3}$$

For Example, $[a, b] = [0, 1] \implies \Pr[0 \leq X \leq 1] = 1 - 0 = 1$



Application I: One-Dimension Monte Carlo Integration

Remember that

$$\mathbb{E}[g(x)] = \int g(x)f(x)dx$$

Suppose that we want to compute

$$\theta = \int_0^1 g(x)dx$$

We cannot compute θ analytically, we can use numerical methods. The key is to realize that this expression is nothing but

$$\theta = \int_0^1 g(x) \cdot dx$$

Suppose that we want to compute

$$\theta = \int_0^1 g(x)dx$$

We cannot compute θ analytically, we can use numerical methods. The key is to realize that this expression is nothing but

$$\begin{aligned} \theta &= \int_0^1 g(x) \cdot 1 \cdot dx \\ &= \mathbb{E}[g(x)] \end{aligned}$$

Remember that 1 is the density function of the standard uniform.

$$\mathbb{E}[g(x)] = \int g(x) f(x) dx$$

$$\begin{aligned} \theta &= \int_0^1 g(x) \cdot 1 \cdot dx \\ &= \mathbb{E}[g(x)] \end{aligned}$$

If we know that

$$\begin{aligned} \theta &= \int_0^1 g(x)dx \\ &= \mathbb{E}[g(U)] \quad U \sim U(0,1) \end{aligned}$$

We can do the following

If we know that

$$\begin{aligned} \theta &= \int_0^1 g(x)dx \\ &= \mathbb{E}[g(U)] \quad U \sim U(0,1) \end{aligned}$$

We can do the following

1. Generate a random set of values $(U_1, U_2, U_3, \ldots, U_N)$ distributed from $U(0,1)$ and independent

If we know that

$$\begin{aligned}\theta &= \int_0^1 g(x)dx \\ &= \mathbb{E}[g(U)] \quad U \sim U(0,1)\end{aligned}$$

We can do the following

1. Generate a random set of values $(U_1, U_2, U_3, \ldots, U_N)$ distributed from $U(0,1)$ and independent
2. Estimate θ with

$$\widehat{\theta_n} = \frac{g(U_1) + g(U_2) + \cdots + g(U_N)}{N}$$

Example: Suppose that we wish to estimate

$$\int_0^1 x^3 dx = \frac{1}{4} = .25$$

```
# Set seed number to ensure reproducibility
set.seed(123456)

# Generate a set of random values drawn from the standard u
x<- runif(10)

# Generate function values
gx <- x^3

# Calculate sample means of the function values
mean(gx)

## [1] 0.2212977
```

Example: Suppose that we wish to estimate

$$\int_0^1 x^3 dx = \frac{1}{4} = .25$$

```
# Set seed number to ensure reproducibility
set.seed(123456)

# Generate a set of random values drawn from the standard uniform distributioin
x<- runif(100000)

# Generate function values
gx <- x^3

# Calculate sample means of the function values
mean(gx)


## [1] 0.2495327
```

Extensions

1. You can extend this to obtain integration over a suppose different from $[0, 1]$. For example,

$$\begin{aligned} \theta &= \int_1^3 (x^2 + x) dx \\ &= 2 \cdot \int_1^3 (x^2 + x) \frac{1}{2} dx \\ &= 2\mathbb{E}[X^2 + X] \end{aligned}$$

   Note that X is distributed from $U(1, 3)$ (Uniform distribution with support $[1, 3]$) whose density is $\frac{1}{3-2} = \frac{1}{2}$.

2. You can also simulate discrete variables based on the standard uniform distribution.

Simulating uniformly distributed variables (other than standard uniform)

1. Draw directly using R built-in command `runif(n,min=,max=)`
2. Draw indirectly using standard uniform first and then manipulate it (with addition and multiplication) into a uniform distribution with different support.

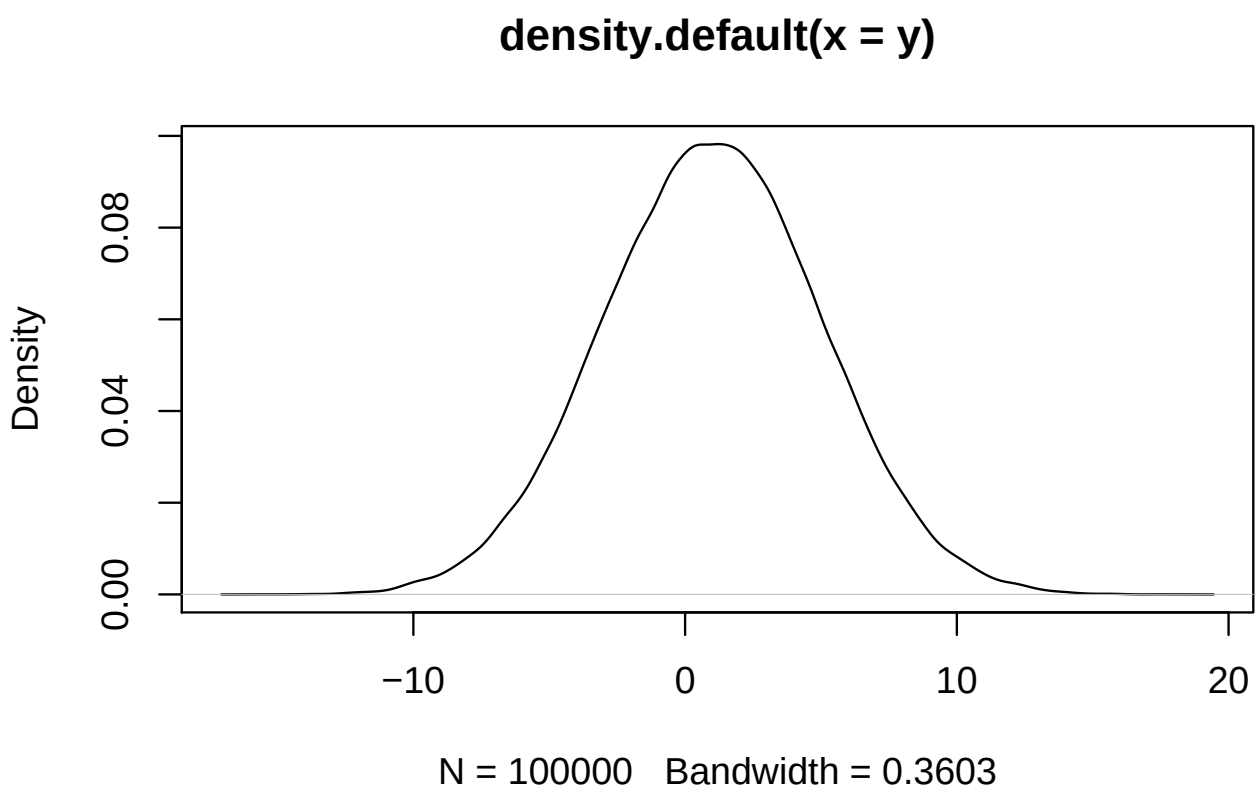
Continuous, Parametric Distribution (II)
Normal Distribution

For a normally distributed variable Y with mean μ and σ^2, the density function is given by

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad -\infty < y < \infty$$

```r
set.seed(123456)
y <- rnorm(100000, mean = 1,sd = 4)

plot(density(y))
```



```r
set.seed(123456)
y <- rnorm(100000, mean = 1,sd = 4)

mean(y)
```

```
## [1] 1.013107
```

```r
sd(y)
```

```
## [1] 4.00308
```

1. From Normal to Standard Normal
2. From Standard Normal to Normal

1. From Normal to Standard Normal

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

.

2. From Standard Normal to Normal

1. From Normal to Standard Normal

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

.

2. From Standard Normal to Normal

$$Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

Any normal variable is standardized by

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

.

This variable has a mean of zero and a standard deviation of one.

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

.

```
set.seed(123456)
y <- rnorm(100000, mean = 1,sd = 4)
z <- (y-mean(y))/sd(y)
head(cbind(y,mean(y),y-mean(y),sd(y),z))
```

```
##                y                                          z
## [1,]  4.3349327 1.013107  3.3218261 4.00308  0.82981765
## [2,] -0.1041911 1.013107 -1.1172977 4.00308 -0.27910955
## [3,] -0.4200074 1.013107 -1.4331140 4.00308 -0.35800287
## [4,]  1.3499497 1.013107  0.3368431 4.00308  0.08414598
## [5,] 10.0090229 1.013107  8.9959163 4.00308  2.24724894
## [6,]  4.3378405 1.013107  3.3247339 4.00308  0.83054404
```

```
mean(z)
```

```
## [1] -1.689621e-17
```

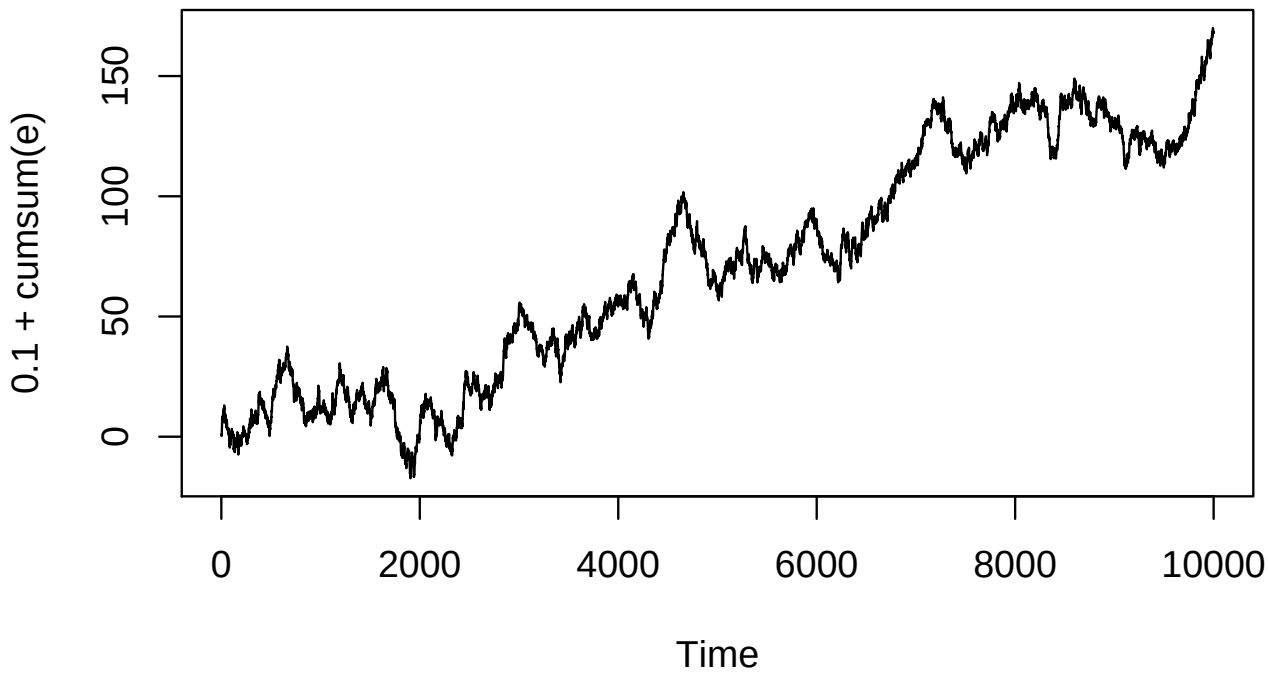
```
sd(z)
```

```
## [1] 1
```

Use normally distributed errors to simulate random walk variables.

$$\begin{aligned} x_t &= x_{t-1} + \epsilon_t \quad \epsilon_t \sim N(0,1) \\ &= x_0 + \sum_{i=1}^{t} \epsilon_i \end{aligned}$$

```
set.seed(123456)
e <- rnorm(10000)
plot.ts(0.1+cumsum(e))
```

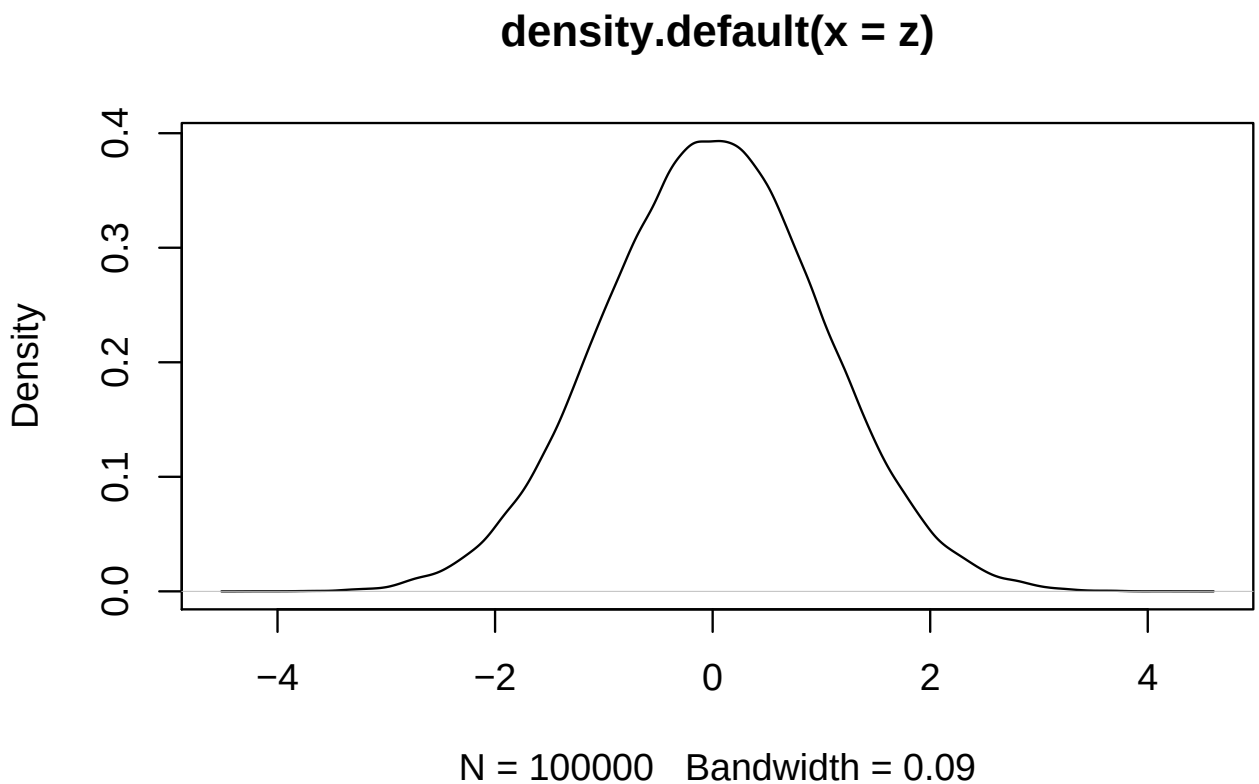


$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

.

```
set.seed(123456)
y <- rnorm(100000, mean = 1,sd = 4)
z <- (y-mean(y))/sd(y)

plot(density(z))
```



Data transformations are a core part of data engineering.

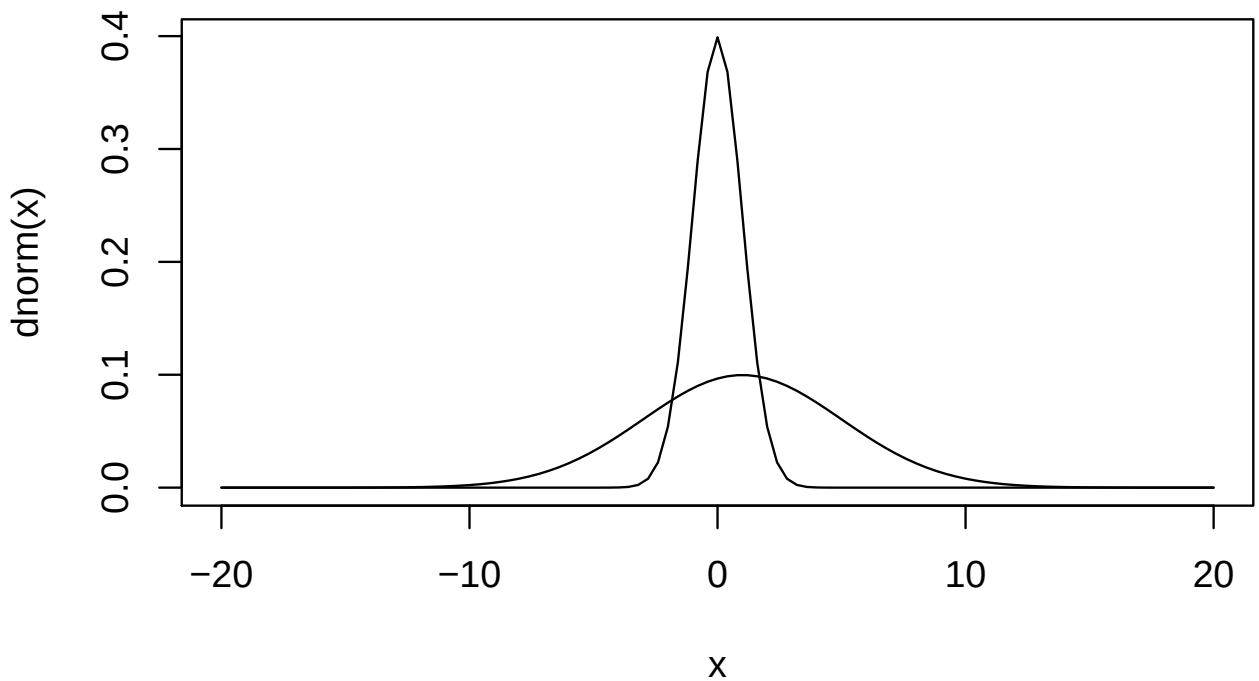
This type of transformation is also called **Normalization**. The idea is to transform a variable in such a way that its elements are not too large or too small compared to other variables' elements.

A good normalization will allow for the development of several new features (nothing but independent variables) that can significantly enhance the performance of basic (or raw) classifiers or regressors.

Drawbacks: It yields negative values, making the normalized variable incompatible with certain transformation like $\log(x)$. There are also no fixed boundaries for the normalized variable. (I don't have any preferences)

Also, if $Z \sim N(0,1)$, we can also generate a random variable distributed from $N(\mu, \sigma^2)$ by doing stretching and shifting the standard normal variable,

```
curve(dnorm(x),-20,20)
curve(dnorm(x,mean = 1,sd = 4),-20,20, add = T)
```



Mulitiplication: You blow up every value! And even larger values now have positive "probabilities". Cover a larger region.

Addition: Your center in this case does not change when you blow it up. It is still zero. How can I make my mean equal to one?

Add one to every value, then
$\frac{(x_1+1)+(x_2+1)+\cdots+(x_N+1)}{N} = \overline{x} + 1 = 0 + 1 = 1$

Also, if $Z \sim N(0,1)$, we can also generate a random variable distributed from $N(\mu, \sigma^2)$ by doing stretching and shifting the standard normal variable,

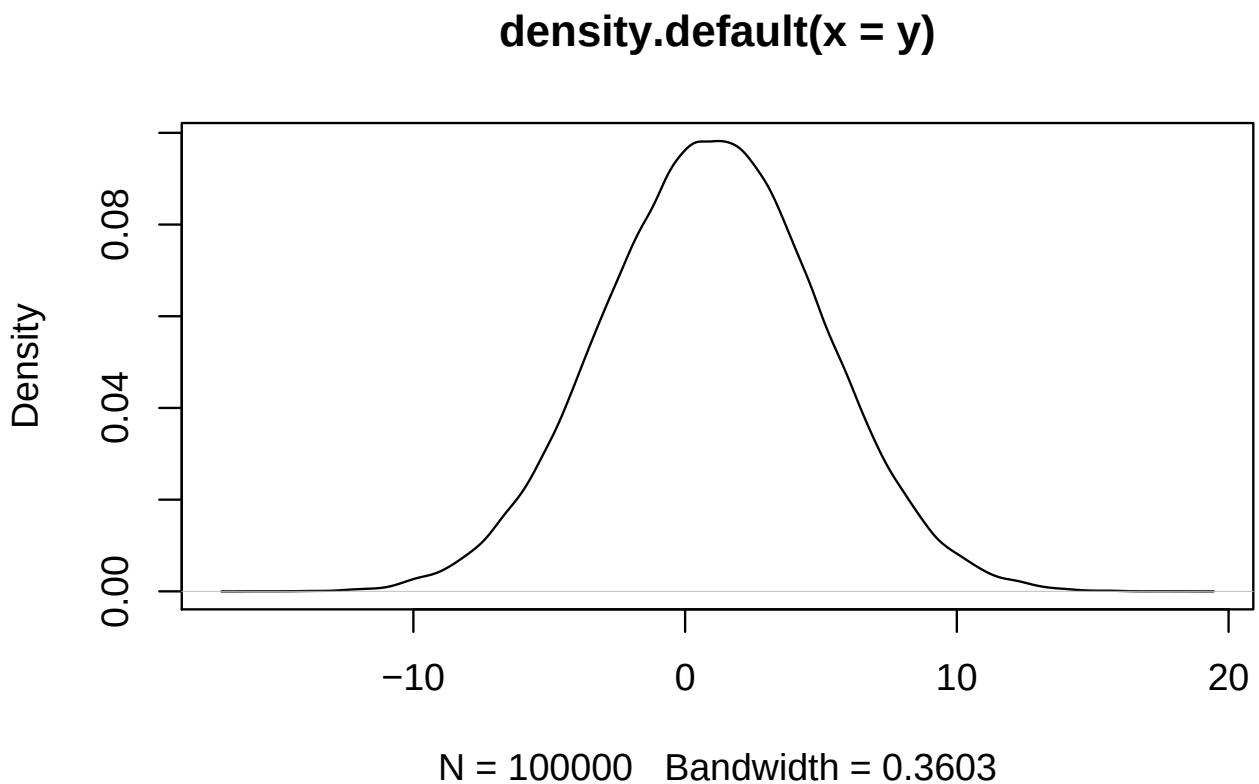
$$Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

$$Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

.

```
set.seed(123456)
z <- rnorm(100000, mean = 0,sd = 1)
y <- 1 + z*4

plot(density(y))
```



Useful Facts:

1. Skewness $= 0$ [symmetric]
2. Kurtosis $= 3$ [normal tail]
3. A very useful fact regarding the standard normal distribution is that the probability that an observation x is within 2 standard deviations from the mean is 0.95. In other words, a value greater 2 or less than -2 is very unlikely!

Remember these facts, later we will use these facts to test whether or not a random variable is indeed normal!

Alex: Symmetric Bell-shape distribution does not necessarily mean that a distribution is normal!

Lesson: Never believe the numbers you see from the data!!!!!!!!

Central Limit Theorem: An Important Theorem related to Normal Distribution

Sampling Distribution

Given a random sample of (fixed) size N, we have only one value for a particular statistic (estimator or a mathematical formula where you can plug in the data to obtain some value).

However, the value of a particular statistic is unknown beforehand and could vary with the random sample drawn from the underlying population.

A statistic should be considered as a random variable and follows a certain distribution: **sampling distribution**.

Interesting thing is that even though we do not know what each random sample would be, we can know the sampling distributions of many statistics.

In this case, sample average is a statistic (or estimator) that varies with random samples

And we know that its distribution is normal!

Mathematical Version (Lindeberg-Levy Central Limit Theorem):

Suppose $X_1, X_2, \ldots, X_N$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then, as n approaches infinity, the random variables $\sqrt{N}(\overline{X} - \mu)$ converges in distribution to a normal $N(0, \sigma^2)$:

$$\sqrt{(N)}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) - \mu\right] \sim N(0, \sigma^2)$$

Loose Translations: In the textbook, page 262.

Under certain conditions (see our notes on the law of large number for more explanations),

If all samples of particular size are selected from **any population**, the distribution of the sample mean (averages) is approximately a normal distribution. This approximation improves with larger samples.

Loosely speaking,

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{N})$$

Question: Does this look familiar?

How about now: Your Z-statistic when the σ is known!

$$\begin{aligned} \overline{X} &\sim N(\mu, \frac{\sigma^2}{N}) \\ \frac{\overline{X} - \mu}{\sqrt{\sigma^2/N}} &\sim N(0,1) \\ \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} &\sim N(0,1) \end{aligned}$$

Let's accumulate what we know!

1. $\frac{\overline{X}-\mu}{\sigma/\sqrt{N}} \sim N(0,1)$

How to Visualize the Central Limit Theorem?

```r
# Set seed number to ensure reproducibility
set.seed(123456)

# Define a vector to hold our sample means
means <- vector(,10000)

# Generate 1000 samples of size n and store their means
for (i in 1:10000){

  # Draw from the "fair-coin" distribution
  # Does not matter what distribution you draw
  # the data from
  sample   <- sample(0:1, 1000, replace = TRUE)
  means[i] <- mean(sample)

}
```

How to Visualize the Central Limit Theorem?

```
# Set seed number to ensure reproducibility

plot(density(means))
```