# Normal Distribution - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

# Admin

- First quiz will be this week
- The first R-review session is available online

# Normal Distribution or Gaussian Distribution

- Normal Distribution or Gaussian distribution pdf:

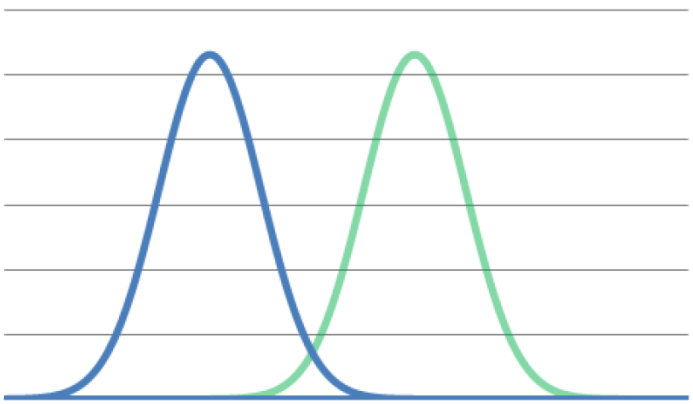
$$P_x[x] = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

- It has two parameters mean and standard deviation

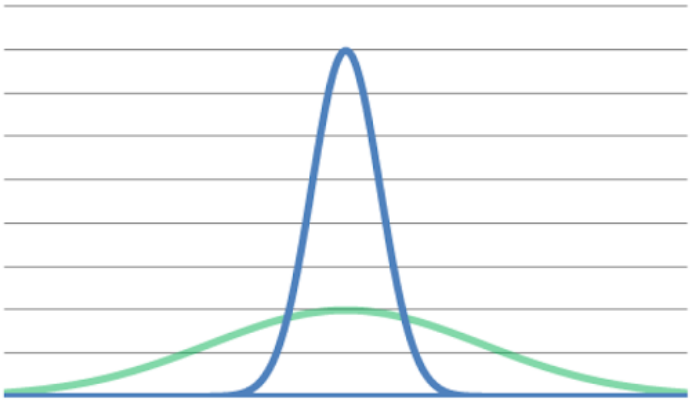
$$N(\mu, \sigma)$$

# Normal Distribution Examples

- Normal Distribution
- The mean is the center of the distribution and is the point that splits the area under the bell shaped curve in half.



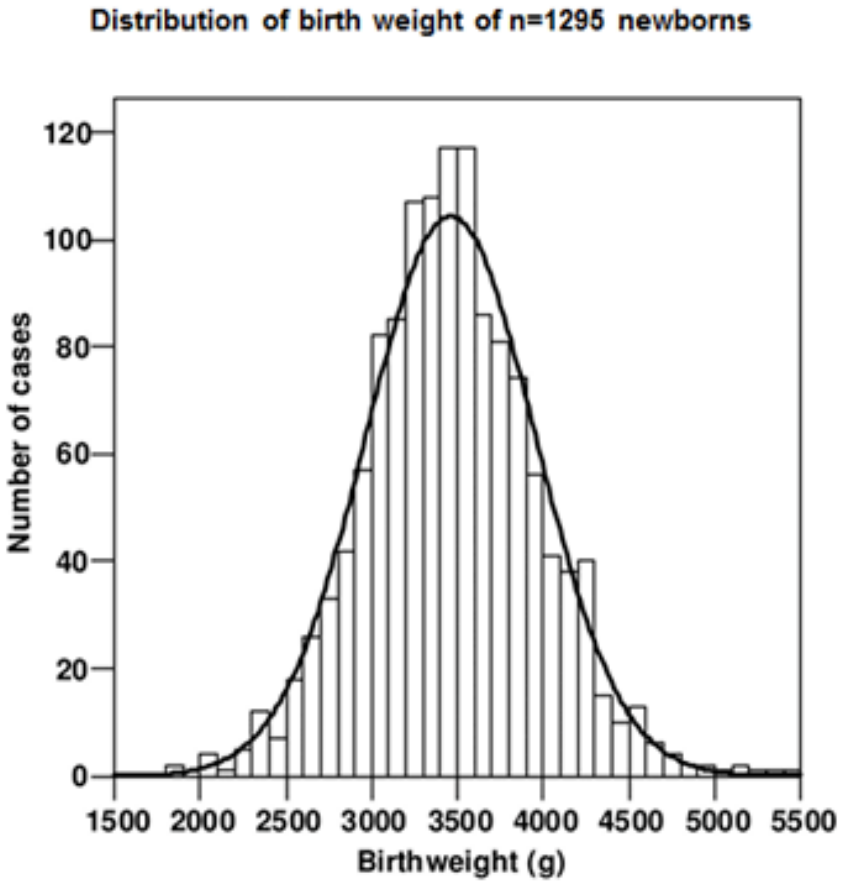
Different Means



Different Variance

# Normal Distribution – Example Curve

- A 2006 paper showed a histogram of the birth weights of newborns in their sample.



# Normal or Gaussian Distribution

- One of the most important distributions, since many natural phenomena follow normal distribution. E.g.
  - Tossing a Coin
  - Heights of people
  - Blood pressure
  - IQ scores
- Bell shape distribution – symmetric - mean=mode=median – Completely defined by $\mu$ and $\sigma$.

# Why Normal Distribution is Important? Poisson Distribution

- Poisson Distribution:
  - Expected number of occurrences of an event in a certain time interval is $\lambda$, the probability of occurring $n$ times in that interval is Poisson distribution

$$P(n,\lambda) = \frac{\lambda^n}{n!}\exp(-\lambda)$$

- Poisson distribution is everywhere, e.g. decay of radioactive, queueing dis.
- If $\lambda$ is large Poisson distribution gets approximated with Gaussian distribution

$$\mu = \lambda\,, \qquad \sigma^2 = \lambda$$

or

$$N\left(\lambda, \sqrt{\lambda}\right) = \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(x-\lambda)^2}{2\lambda}}$$

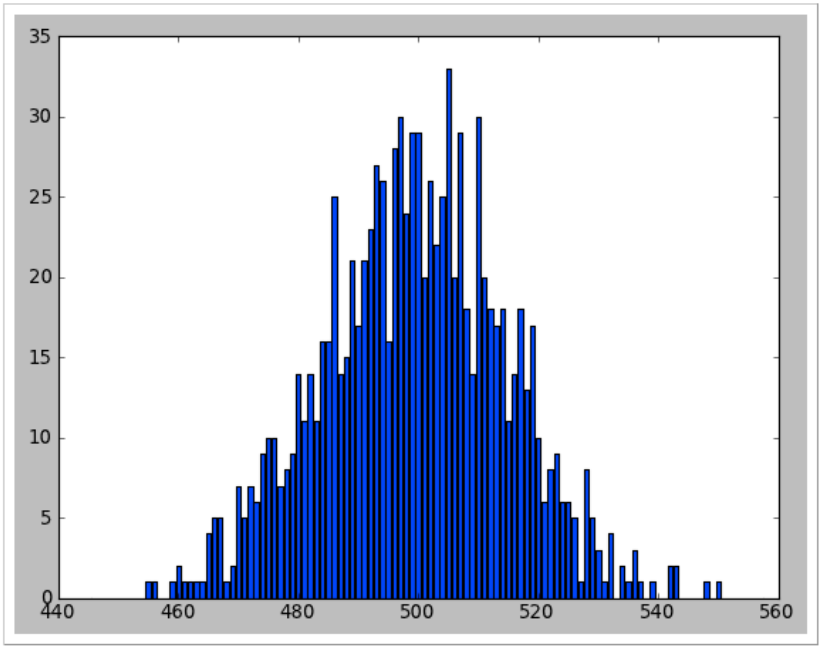
# Why Normal Distribution is Important? Binomial Distribution

- Binomial distribution:
  - n: number of total trials
  - p: probability of success on a single trial
  - $P_x = \binom{n}{x} p^x q^{(n-x)}$
- It can be shown that a Binomial Distribution with large enough trials, *n*, approaches normal distribution with

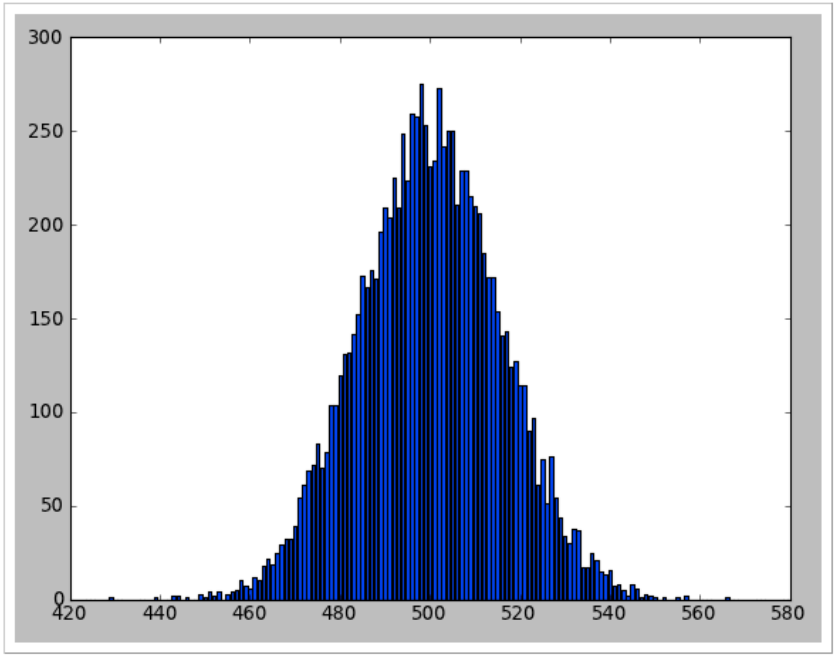
$$\mu = np\,, \qquad \sigma^2 = np(1-p)$$

- E.g. of Binomial: tossing a coin, winning slot machine.
- So, sum of a lot of independent binary random events is normal dis.

# Tossing a coin




1000 trials of tossing a coin



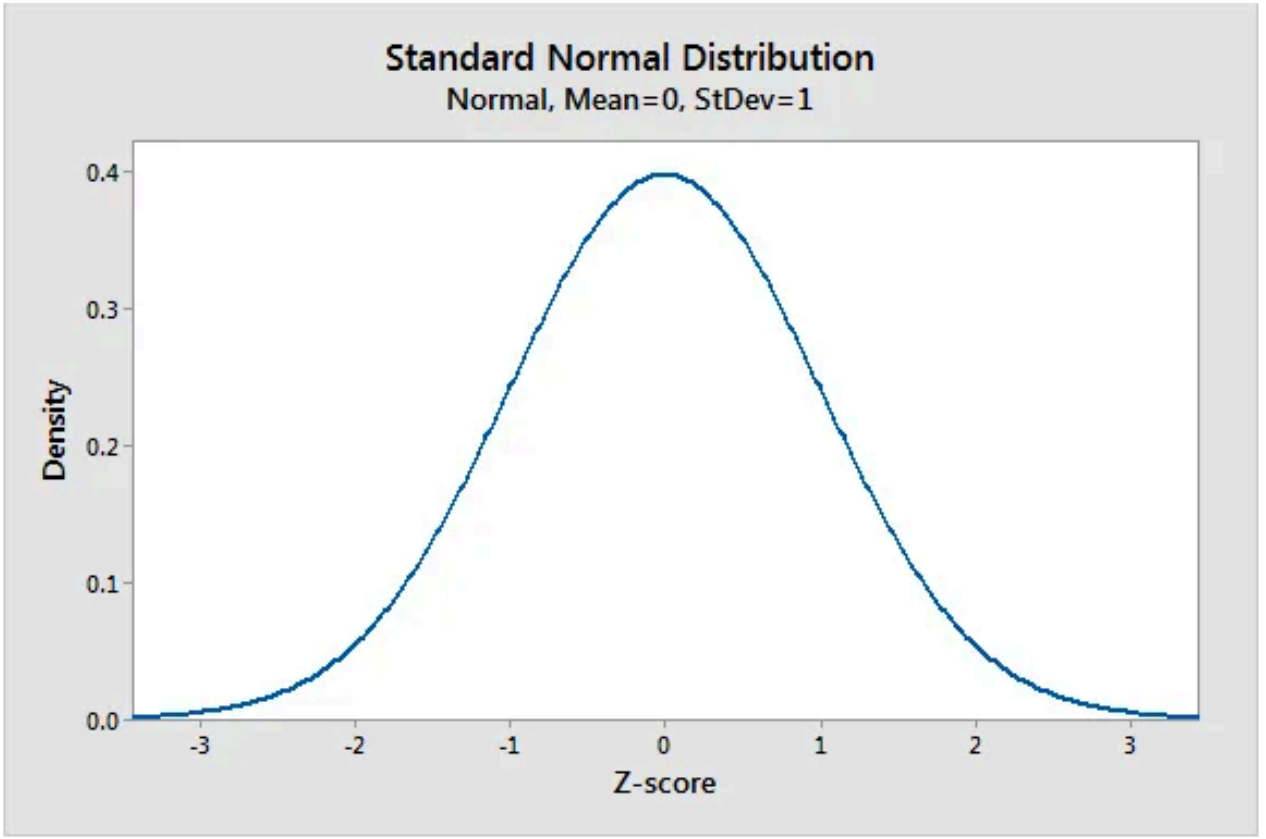

10,000 trials of tossing a coin

Reference – Pi-Cubed Programming challenge

# Standard Normal Distribution

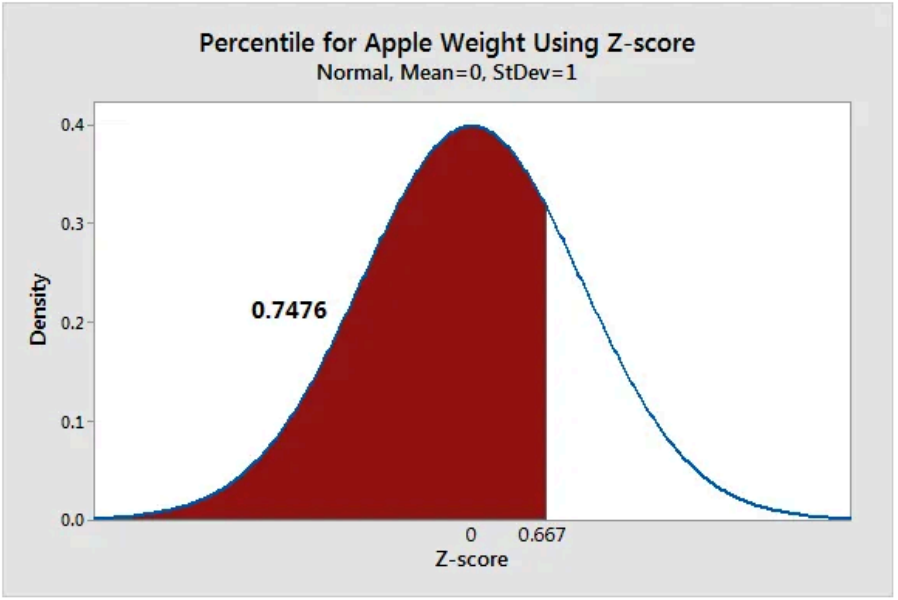
- Standard normal distribution

$$N(\mu = 0, \sigma = 1)$$



# Z-score

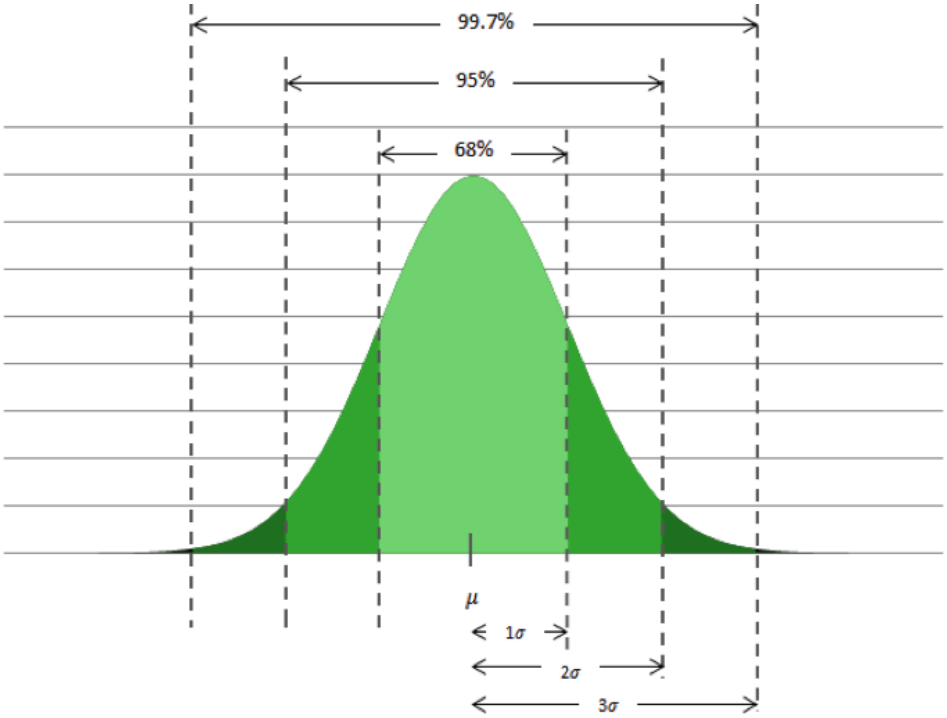
- Z-score and the area associated with the Z-score



# Area Under Normal Distribution

Assuming normal distribution $N(\mu, \sigma)$

- 68% of the observations fall within $\sigma$ of the $\mu$
- 95% of the observations fall within $2\sigma$ of the $\mu$
- 97.5% of the observations fall within $3\sigma$ of the $\mu$



# Exercise – SAT Score

The distribution of SAT scores for the verbal section for high school seniors is approximately a normal distribution with a **mean of 504** and a **standard deviation of 111**.

- What proportion of seniors score between 393 and 615?

# Exercise – SAT Score Answer

- $\mu=504, \sigma=111$
- $393 - 504 = -111$ & $615 - 504 = 111$
- 393 and 615 are one standard deviation away from the mean of 504

$$393 = 504-111 = \mu-\sigma$$

$$615 = 504+111 = \mu+\sigma$$

68% rule - 68% of scores are between 393 and 615.

# Example of Normal Distribution

The distribution of SAT scores for the verbal section for high school seniors is approximately a normal distribution with a **mean of 504** and a **standard deviation of 111**.

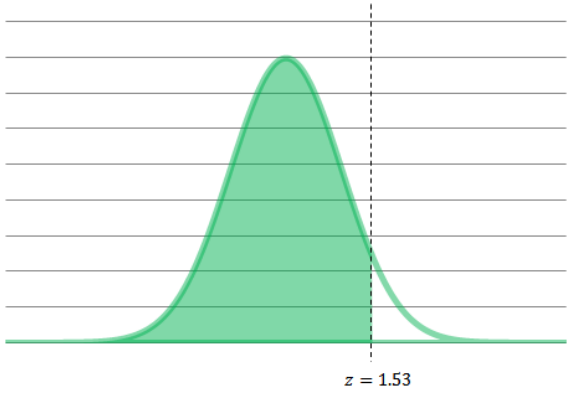
- What ration of students get SAT Score less than 674?
- What proportion of students get SAT score higher than 439.5

# Example of Normal Distribution

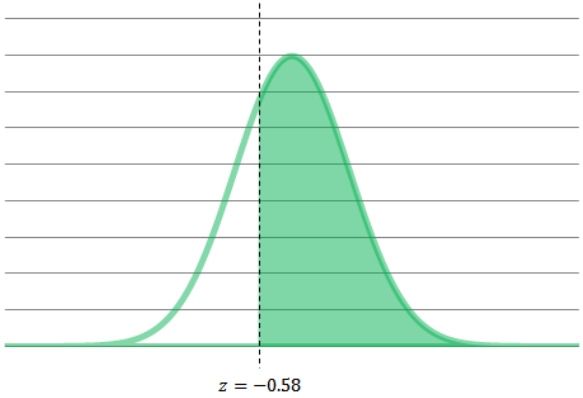
- Therefore, we find the area under the standard normal curve to the left of $z = 1.53$
- And the proportion of observations greater than $z = -0.58$

# Example of Normal Distribution

- $Z = \frac{x-\mu}{\sigma}$
- Therefore, we find the area under the standard normal curve to the left of z = 1.53
- And the proportion of observations greater than z = −0.58




The Area under the standard Normal curve to the left of z is 0.937




The Area under the standard Normal curve to the right of z is 1-0.281=0.719

## The Central Limit Theorem – An Example

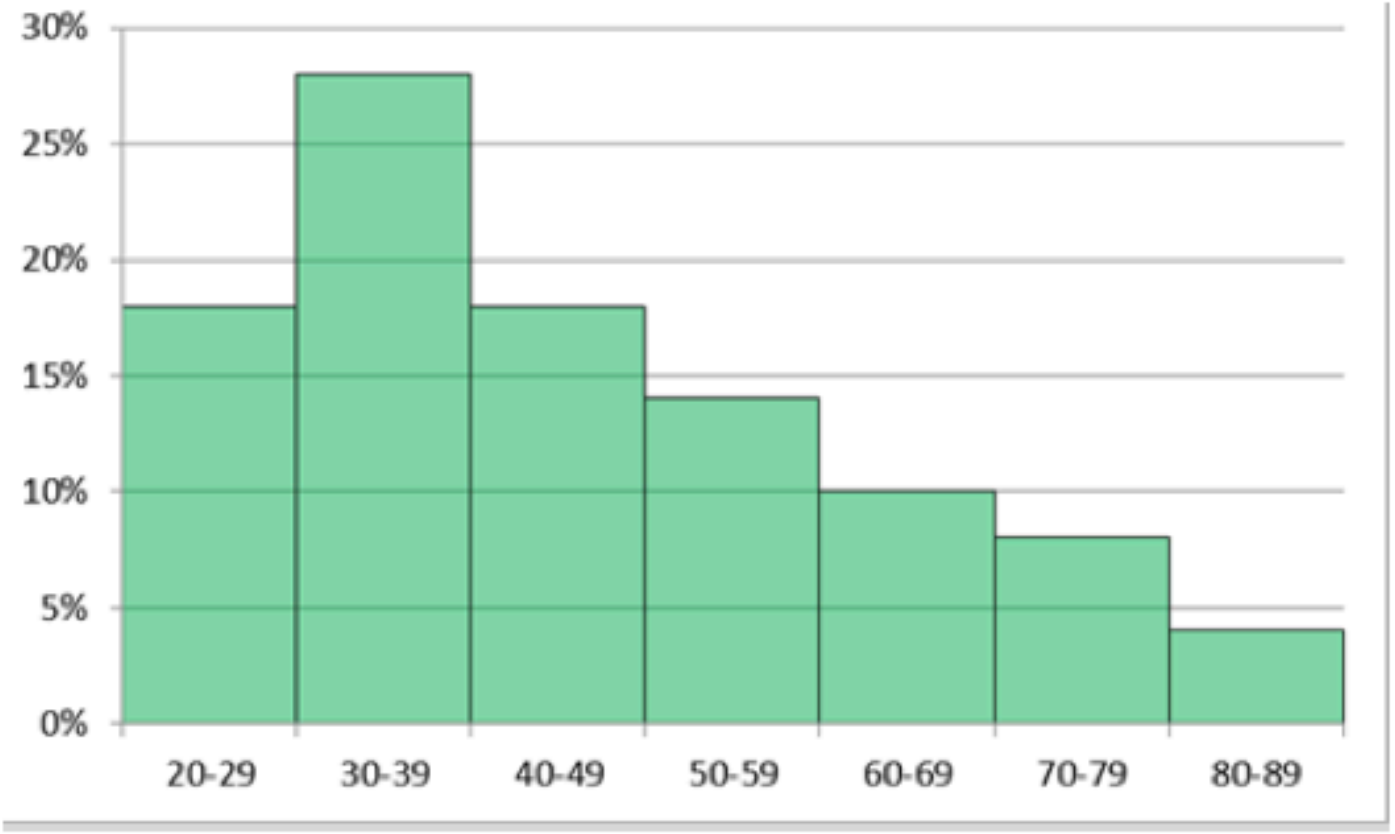
We would like to estimate the mean age of employees at company A. Let's assume that the population is made up of only 50 individuals with the following age distribution:

| | | | | |
|---|---|---|---|---|
| 42 | 20 | 32 | 47 | 31 |
| 66 | 25 | 64 | 25 | 46 |
| 76 | 56 | 32 | 20 | 50 |
| 60 | 58 | 31 | 83 | 51 |
| 22 | 32 | 64 | 49 | 75 |
| 40 | 43 | 54 | 44 | 62 |
| 46 | 27 | 32 | 49 | 37 |
| 38 | 59 | 33 | 59 | 73 |
| 26 | 26 | 83 | 71 | 39 |
| 35 | 33 | 35 | 28 | 35 |

The population mean of the ages of employees is 45.28 years

# The Central Limit Theorem – an example

We would like to estimate the mean age of employees at company A. Let's assume that the population is made up of **only 50 individuals** with the following age distribution:




The population mean of the ages of employees is 45.28 years

## The Central Limit Theorem – an example

Assume we take two random samples of size 5. The resulting ages in each of the two samples.

**Sample 1: 20,44,46,20,44**

**Sample 2: 83,32,31,50,32**

- The first sample has a sample mean of 34.6 and a sample median of 44.
- The second sample has a sample mean of 45.6 and a sample median of 32.

**Neither the sample mean nor the sample median** will always fall closer to the **population mean** in a given sample.

To evaluate each of these sample statistics and their ability to estimate the true population value, we must not rely on **just one example (one sample).**

We'd like to compare the distribution of the sample mean and sample median if we take hundreds of random samples of the same size.

# The Central Limit Theorem – An Example

- We want to evaluate **how well the sample mean and sample median perform as estimators of the population mean**.
- Using a computer to **randomly select 10,000 samples of size 5.**

| Sample | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $\bar{x}$ | $m$ |
|---|---|---|---|---|---|---|---|
| **1** | 20 | 44 | 46 | 20 | 44 | 34.6 | 43 |
| **2** | 83 | 32 | 31 | 50 | 32 | 45.6 | 32 |
| **3** | 58 | 49 | 31 | 32 | 50 | 44 | 49 |
| **4** | 49 | 38 | 31 | 71 | 32 | 44.2 | 38 |
| **5** | 40 | 27 | 76 | 47 | 62 | 50.4 | 47 |
| **6** | 59 | 27 | 32 | 66 | 46 | 46 | 46 |
| **7** | 31 | 64 | 38 | 42 | 62 | 47.4 | 42 |
| **8** | 83 | 49 | 50 | 39 | 22 | 48.6 | 49 |
| **9** | 40 | 26 | 22 | 49 | 54 | 38.2 | 40 |
| **10** | 27 | 47 | 64 | 39 | 54 | 46.2 | 47 |
| ... | | | | | | | |
| **10,000** | 25 | 26 | 33 | 60 | 71 | 43 | 33 |

# The Central Limit Theorem

When **the number of samples** taken from a population is sufficiently large, the **sampling distribution of the sample mean, will be approximately normally distributed** with an expected value of ($\mu$ and a standard deviation of $\sigma$ .

Say you take a random sample of size $n$ from a population with mean $\mu$ and standard deviation $\sigma$.

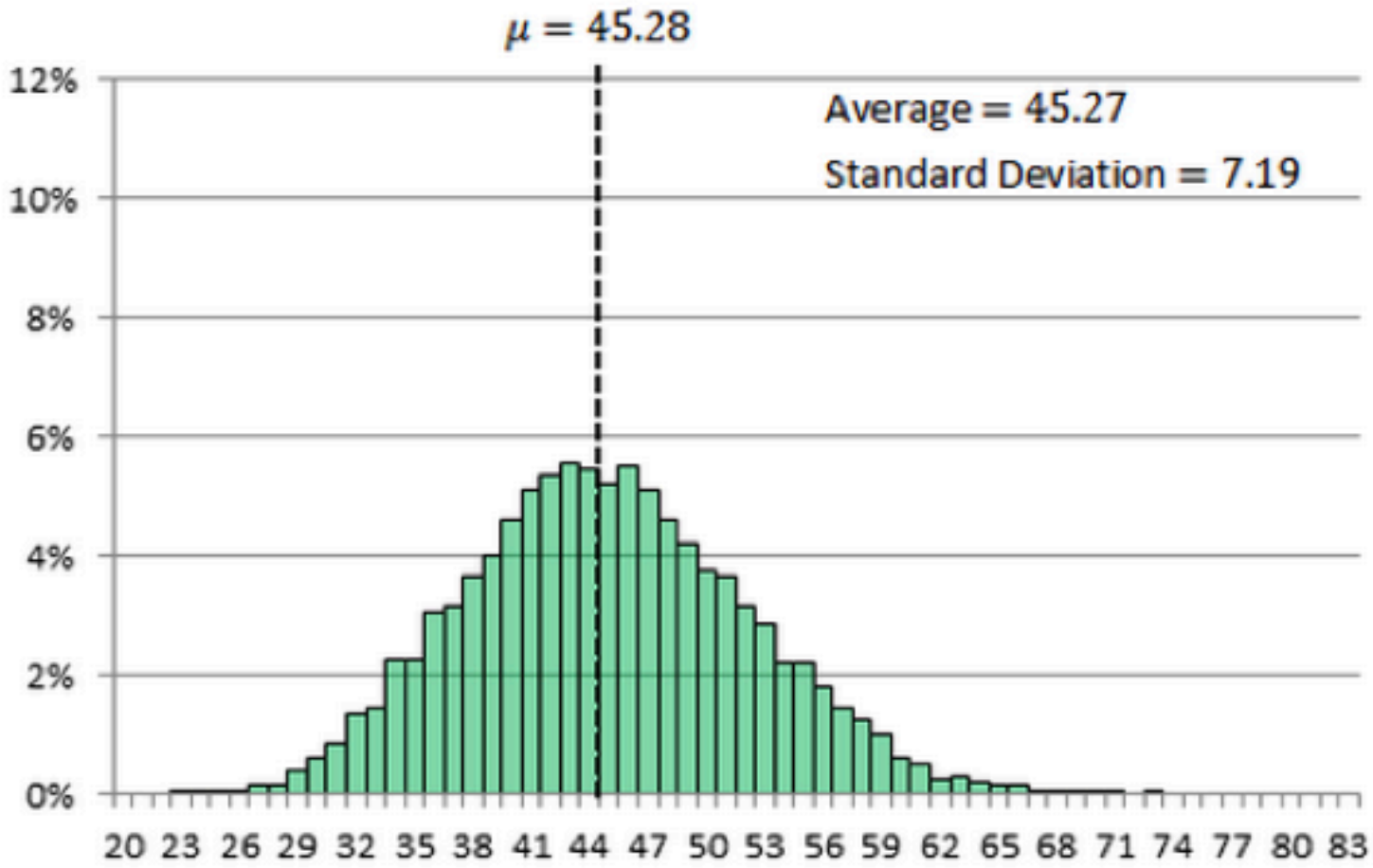
**The larger the sample size,**

1. the **closer** the sampling distribution of the sample means will be **to the normal distribution** and
2. the **smaller the variance** of the sample mean.

# The Central Limit Theorem – An Example

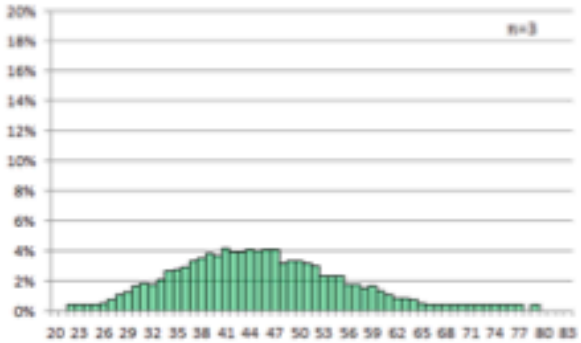
**The distribution of the sample mean $\bar{x}$ for 10000 samples of size 5.**

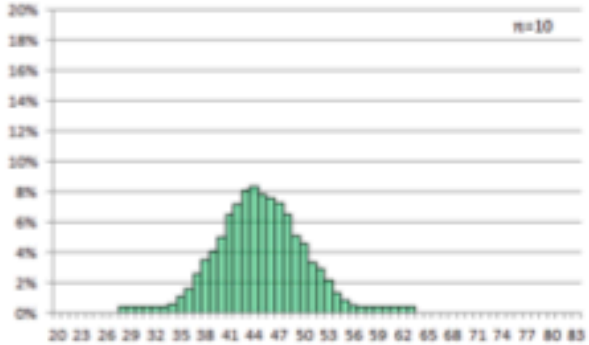
The distribution of the Sample mean is centered over the true value and has less spread or variability than the distribution of the sample mean.



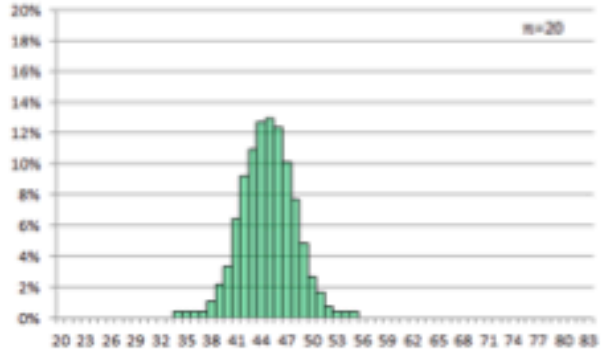
# The Central Limit Theorem - An example

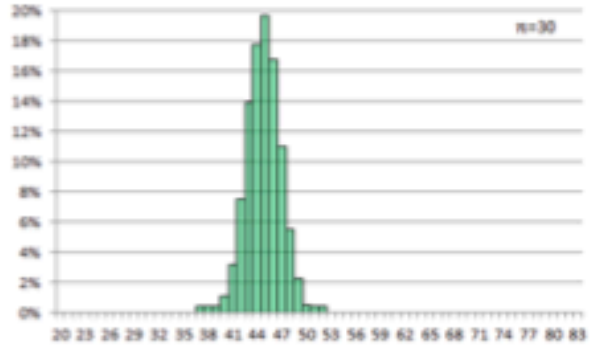
The population has a mean of 45.28 and a standard deviation of 17.20




| $n$ | $\bar{x}_{\bar{x}}$ | $\sigma_{\bar{x}}$ |
|---|---|---|
| 3 | 45.25 | 9.57 |
| 5 | 45.27 | 7.19 |
| 10 | 45.20 | 4.80 |
| 15 | 45.28 | 3.72 |
| 20 | 45.25 | 2.97 |
| 25 | 45.28 | 2.43 |
| 30 | 45.28 | 1.97 |



# Online Simulation of Central Limit Theorem

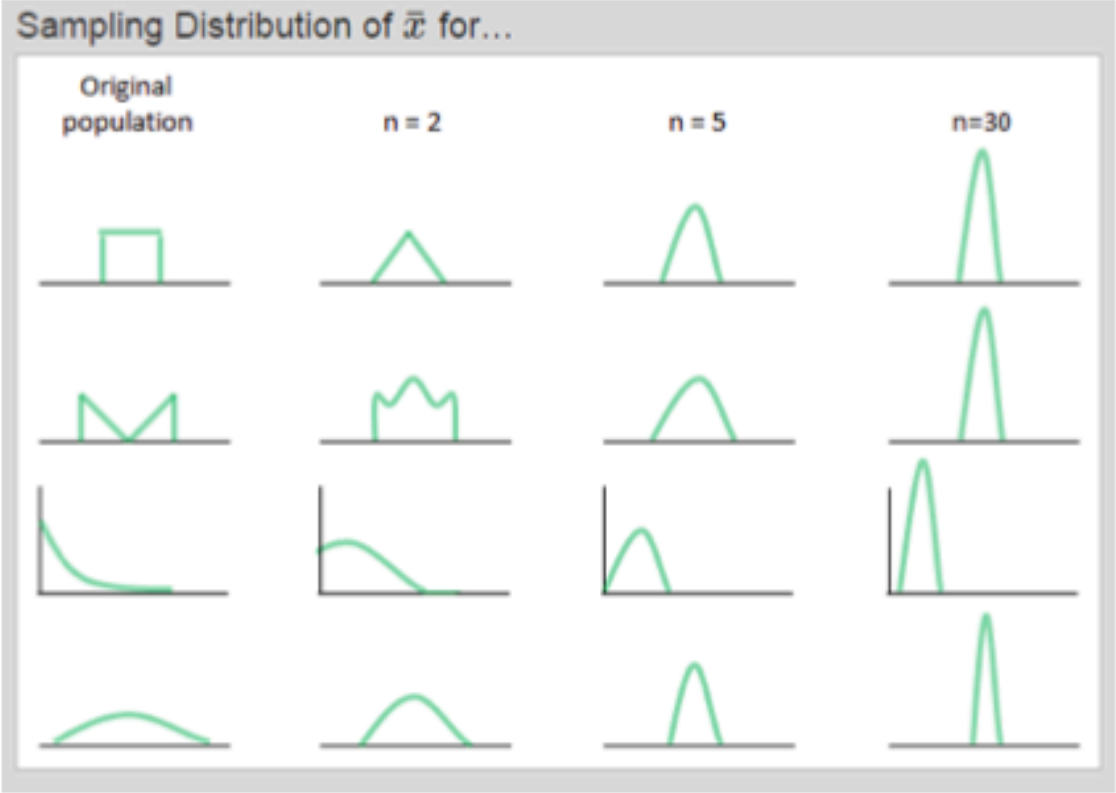
- Examples

  https://gallery.shinyapps.io/CLT_mean/

# How large should n be?

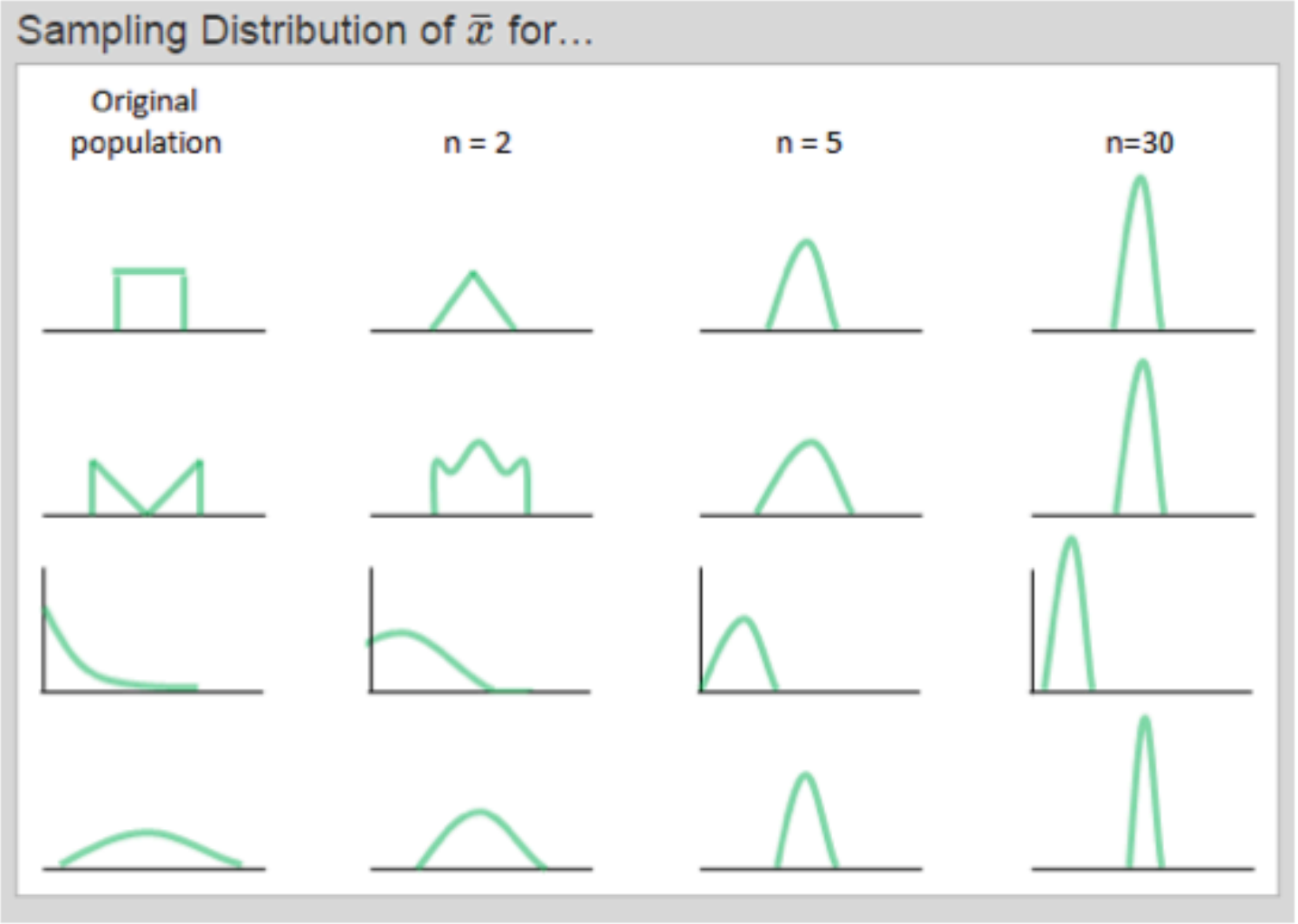
If the **underlying population is approximately normal,** then **even small values of $n$** will give a sampling distribution of sample means that are normally distributed.

Generally, the rule of thumb is that **$n$ should be $n \geq 30$** for the distribution of the sample means to be reasonably normally distributed.



# Skewed population distributions

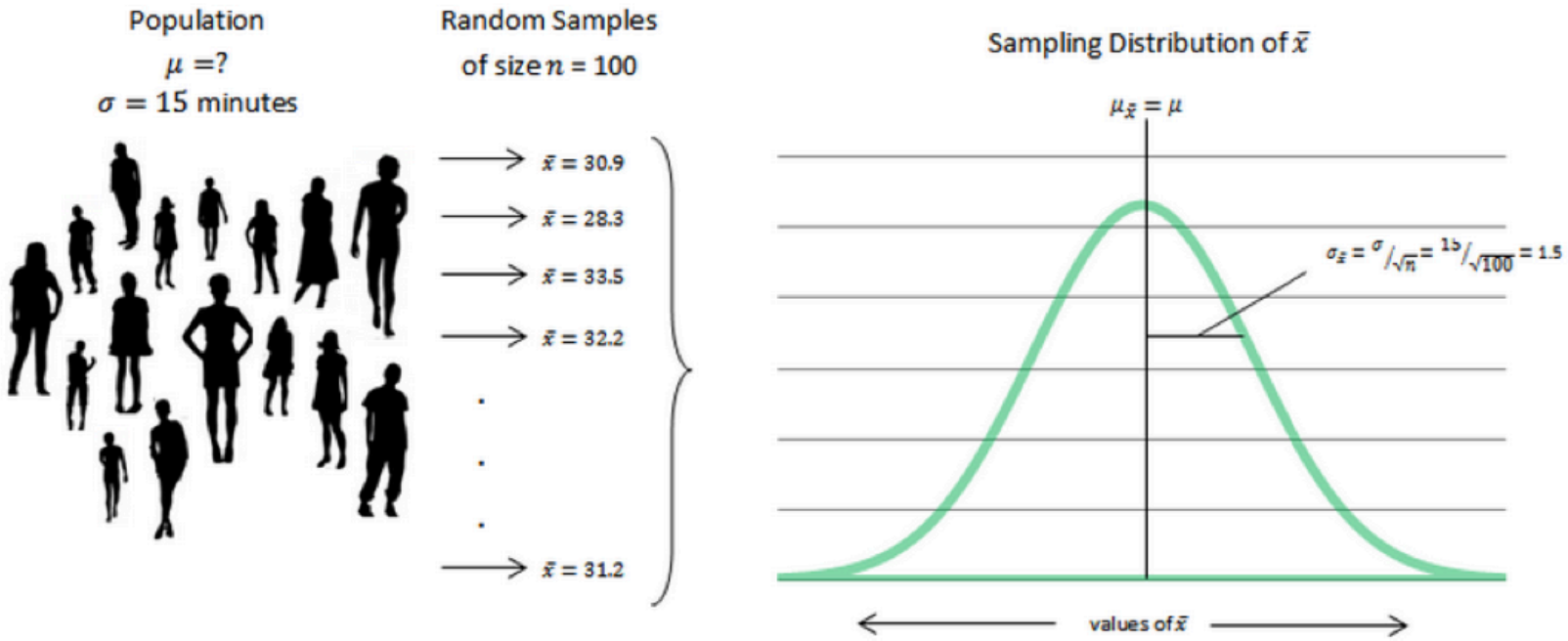
**For more skewed population distributions, n must be larger** before the sampling distribution is sufficiently normally distributed.



# Standard Deviation of Sample Mean

If $n \geq 30$ we know that the sample mean is approximately normally distributed with Standard Error(SE).

$$\mu = \mu_{\bar{x}} \qquad SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \text{Standard Error}$$



# Exercise

Suppose we have selected a random sample of **n=36** from a population with a mean of **80** and a sd of **6.**

**Find the probability that the sample mean will be between 79 and 81.**

## Answer

The Central Limit Theorem tells us that regardless of the shape of the underlying population, the sampling distribution of $\bar{x}$ is approximately normal when $n \geq 30$.

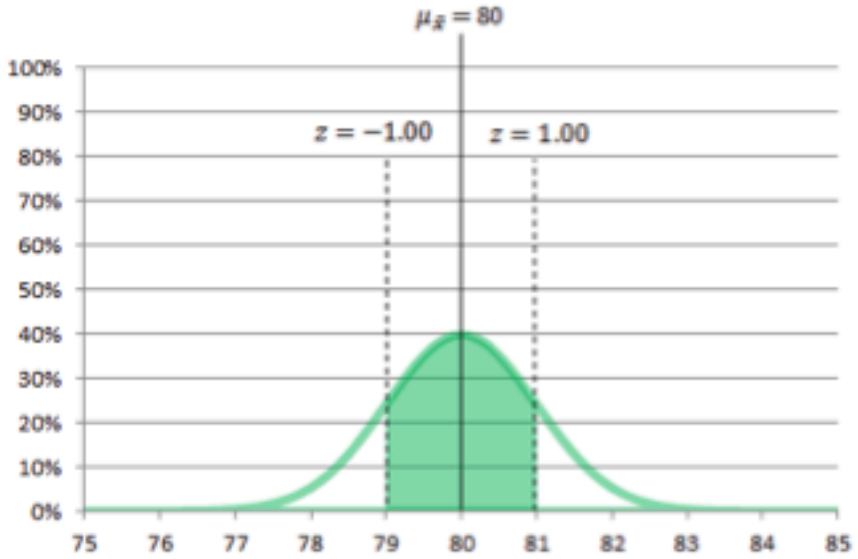
The sampling distribution will have a mean of $\mu_{\bar{x}} = 80$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{6} = 1$

We first standardize the distribution and then calculate the probabilities:

$$79 < x < 81$$

$$\frac{79-80}{1} < \frac{x-80}{1} < \frac{81-80}{1}$$

$$-1 < z < 1$$




0.8413 - 0.1587=0.6826 OR

0.8413 - 0.1587=0.6826

68.26%

# Answer - Using R

```
> 100 * (1 - 2 * pnorm(81, mean= 80, sd= 6/6 , lower.tail = F))
[1] 68.26895
```