L4:Simulation and probability

Learning objectives

- Understand Point Estimates and Sampling Variability
- Visualize and Interpret Sampling Distributions
- Calculate and Interpret Standard Error

Can we verify this theorem??

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be iid random variables with finite mean μ and standard deviation σ. Then

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to Z \qquad \text{as } n \to \infty$$

where Z is a standard normal random variable.

1 - Different shapes of distributions

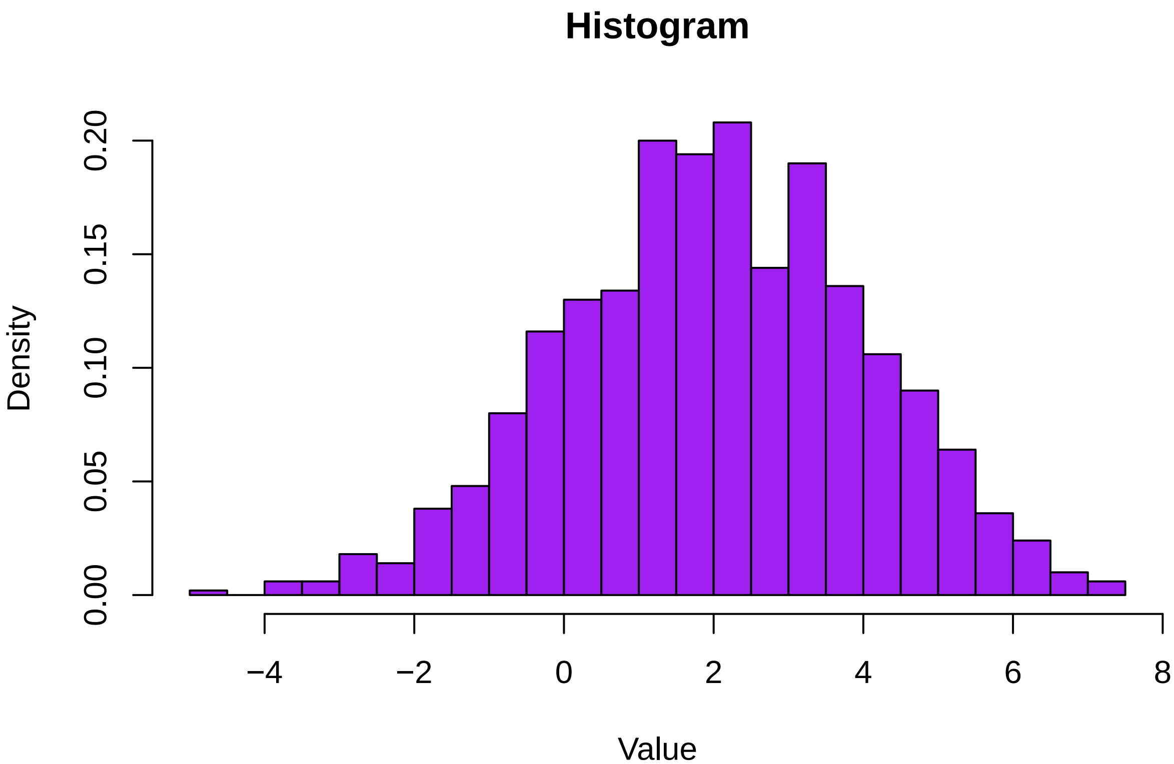
Some distribution functions in `R`.

- **Normal distribution** : `rnorm(n, mean, sd)`
  - `n`: Number of random values to generate
  - `mean`: Mean (center) of the distribution
  - `sd`: Standard deviation (spread) of the distribution
- **Exponential distribution** : `rexp(n, rate)`
  - `n`: Number of random values to generate
  - `rate`: Rate parameter (inverse of the mean)
  - The mean will be `1/rate`
- **Binomial distribution** : `rbinom(n, size, prob)`
  - `n`: Number of random values to generate
  - `size`: Number of trials
  - `prob`: Probability of success in each trial
  - The mean will be `n*prob`
- **Poisson distribution** : `rpois(n, lambda)`
  - `n`: Number of random values to generate
  - `lambda`: Expected number of occurrences in a fixed interval
  - The mean will be `lambda`
- **Gamma distribution** : `rgamma(n, shape, scale)`
  - `n`: Number of random values to generate

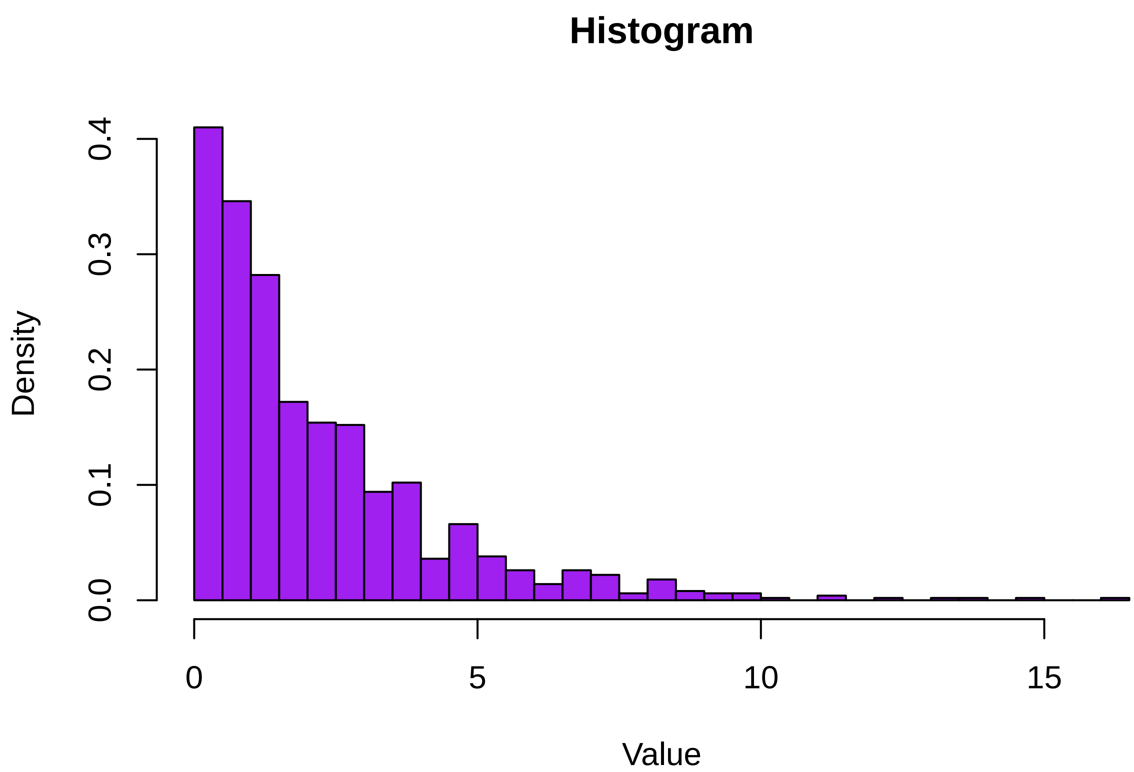
- `shape`: Shape parameter (determines the distribution's form)
- `scale`: Scale parameter (stretches or shrinks the distribution)
- The mean will be `shape*scale`

```
data <- rnorm(1000, mean=2, sd=2)

hist(data, breaks = 30, col = "purple", main = "Histogram", xlab = "Value", probability = TRUE)
```




```
data <- rexp(1000, rate = 1/2)
hist(data, breaks = 30, col = "purple", main = "Histogram", xlab = "Value", probability = TRUE)
```

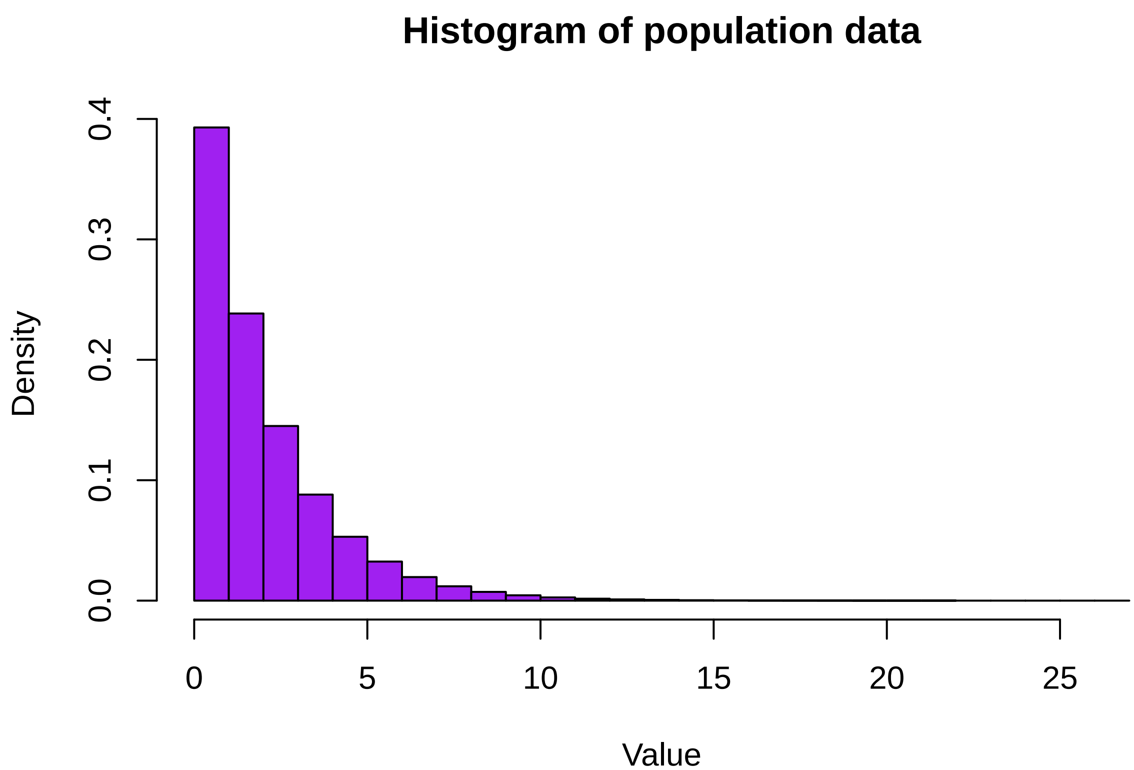


3 - Procedure

- 1. Obtain population data with some distribution.
- 2. Show the shape of the whole population.
- 3. Obtain $\overline{X}_i$ for $i = 1 \ldots n$ by repeating step 1.
- 4. Show the shape of $\overline{X}_i$ for $i = 1 \ldots n$.
- Change sample size n and do the whole step from 1 to 4.
- What can you conclude?

```
mean = 2
pop_data <- rexp(1000000, rate = 1/mean)

hist(pop_data, col = "purple", main = "Histogram of population data", xlab = "Value", probability = TRU
```

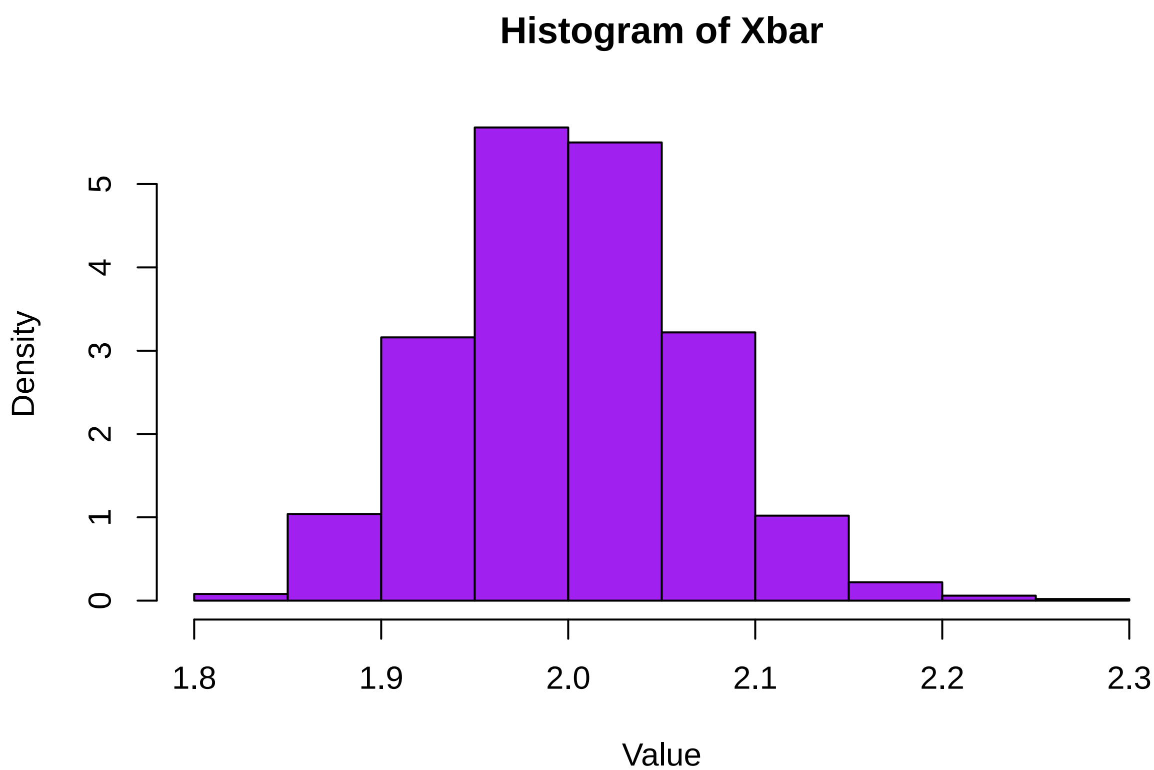



```r
sample_and_get_xbar <- function(n, pop_data){
    sample_data <- sample(pop_data, size=n, replace=TRUE)
    xbar = mean(sample_data)
    return(xbar)
}

n=1000
# Xbar1 = sample_and_get_xbar(n, pop_data)
Xbar_data <- replicate(n, sample_and_get_xbar(n, pop_data ))
head(Xbar_data)
```

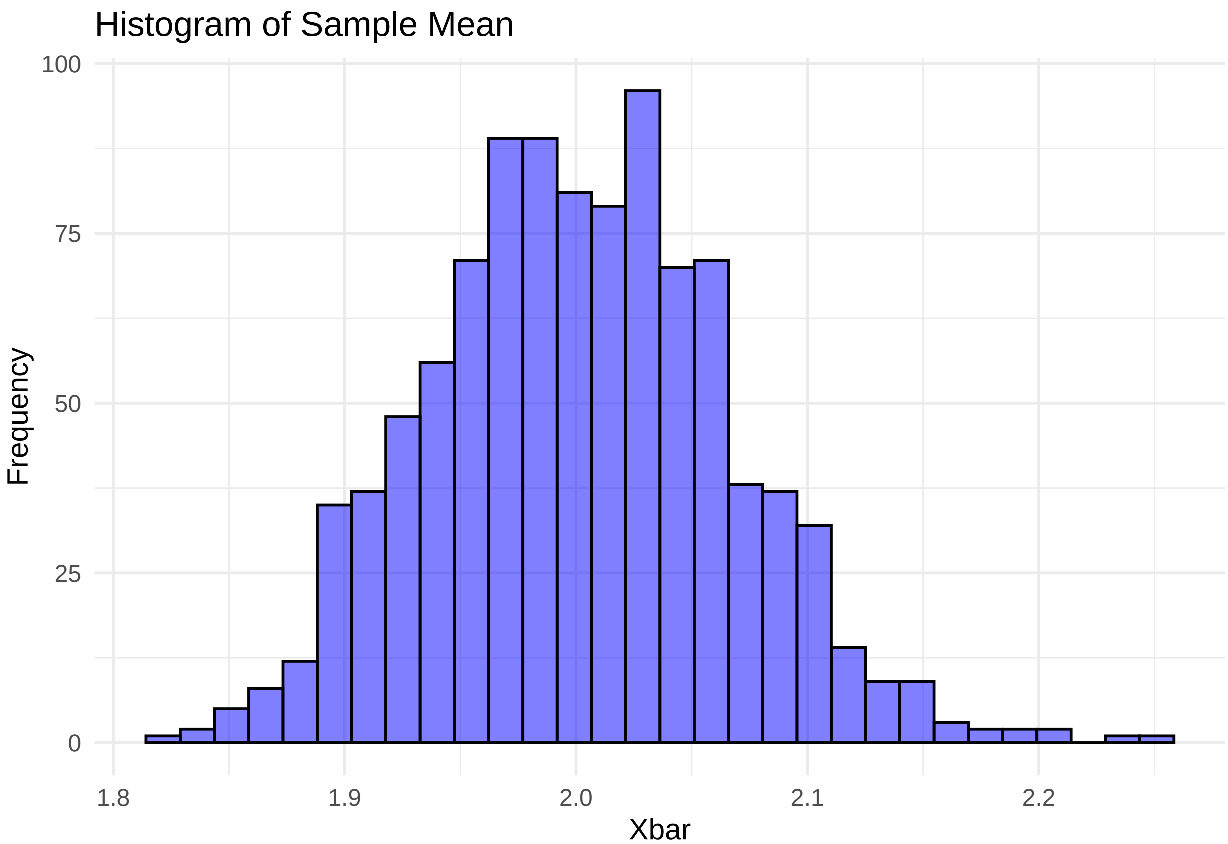
```
## [1] 2.065205 1.899304 2.067580 2.030113 1.919196 1.926914
```

```r
hist(Xbar_data, col = "purple", main = "Histogram of Xbar", xlab = "Value", probability = TRUE)
```




```r
# better version
ggplot(data = data.frame(Xbar_data),
             aes(x = Xbar_data)) +
        geom_histogram(fill = "blue", alpha = 0.5,
                       color = "black") +
        labs(title = "Histogram of Sample Mean",
             x = "Xbar",
             y = "Frequency") +
        theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Histogram of Sample Mean
Frequency
100
75
50
25
0
1.8
1.9
2.0
2.1
2.2
Xbar