Biostatistics
Week VII

Ege Ülgen, MD, PhD

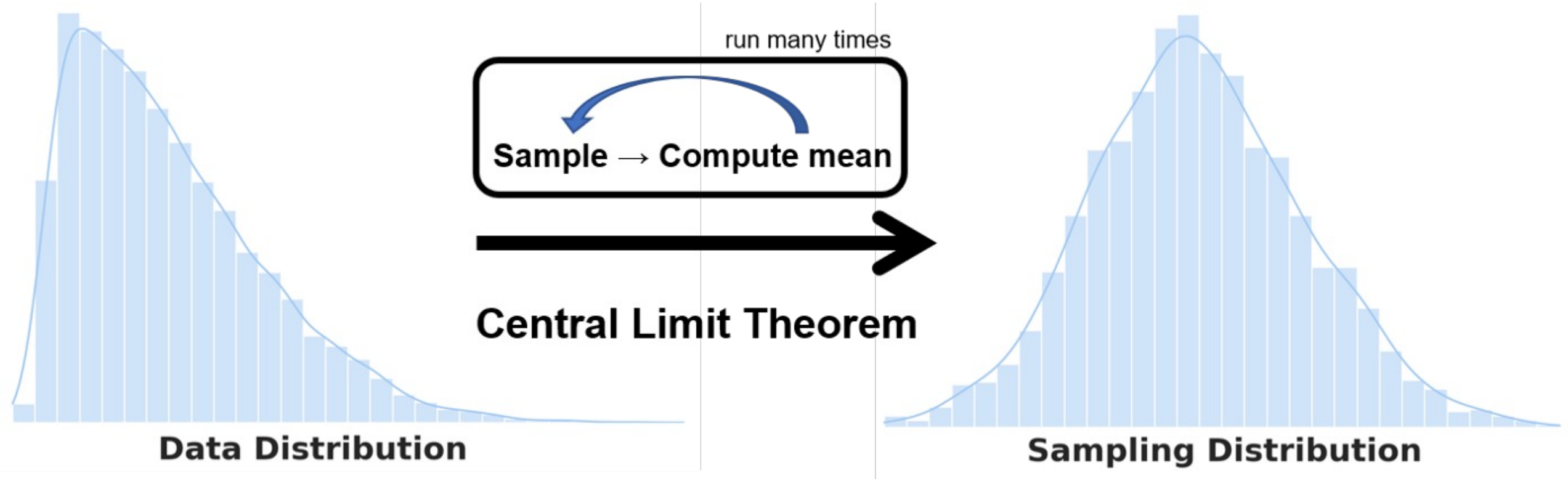
17 November 2022

ACIBADEM
MEHMET ALİ AYDINLAR ÜNİVERSİTESİ

Sampling Distribution

- Population Distribution
- Sample Distribution
- **Sampling Distribution**
  - theoretical probability distribution of a statistic obtained through a specific number of samples drawn from a specific population
  - if samples are randomly selected, the sample means will be somewhat different from the population mean (sampling error)

https://towardsdatascience.com/data-distribution-vs-sampling-distribution-what-you-need-to-know-294819109796



Sampling Distribution - cont.

- If sample size is large enough, the sampling distribution of the sample mean will be approximately normal

- the mean of the sample means will be the same as the population mean
- the standard deviation of the sample means = $\frac{\sigma}{\sqrt{n}}$

https://www.scribbr.com/statistics/confidence-interval/

Confidence Interval

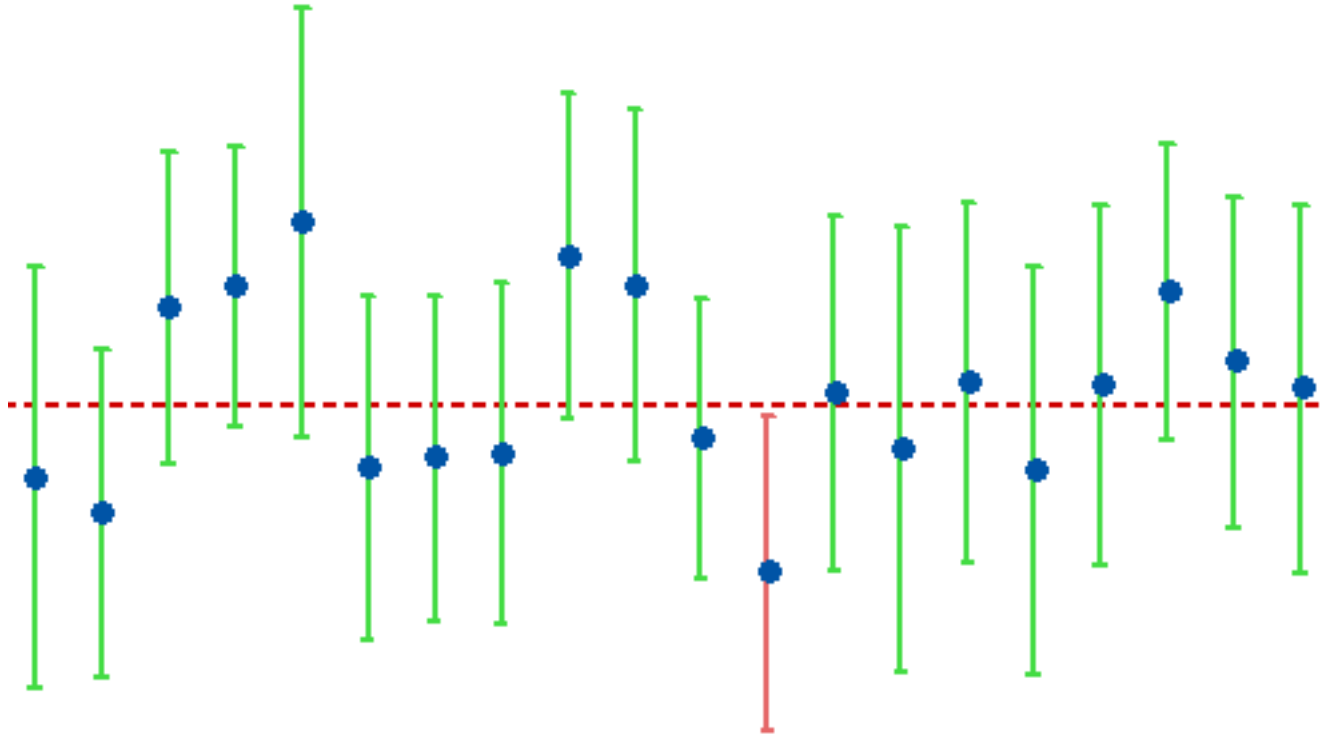
- When you make an estimate in statistics, there is always uncertainty around that estimate because the number is based on a single sample
- The confidence interval is the **range of values that you expect your estimate to fall between a certain percentage of the time** if you run your experiment again (re-sample the population in the same way)

https://www.scribbr.com/statistics/confidence-interval/

Confidence Interval

- The **confidence level** is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval
  - if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval

https://statisticsbyjim.com/hypothesis-testing/confidence-interval/



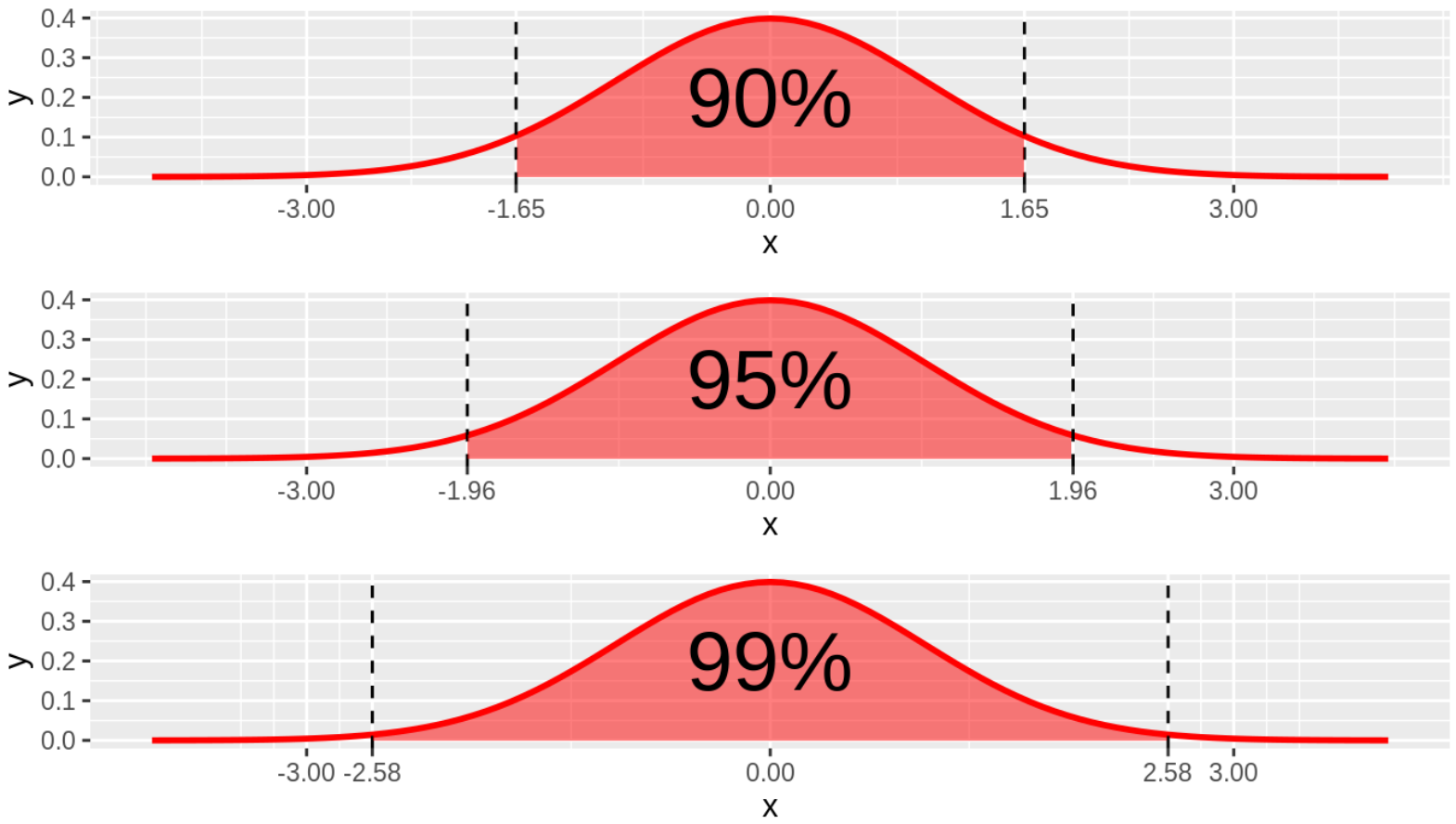
a 95% confidence interval [10 15] indicates that we can be 95% confident that the parameter is within that range

However, it does NOT indicate that 95% of the sample values occur in that range

Confidence Interval

$$CI = \bar{x} \pm Z * \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm t * \frac{s}{\sqrt{n}}$$



Confidence Interval - Example

id	week_1	cd4_1	week_2	cd4_2	perc_benefit
361	0	26	7.43	3	-11.905994
1017	0	13	7.00	10	-3.296703
519	0	3	8.14	5	8.190008
1147	0	65	33.00	97	1.491841
1216	0	36	8.00	31	-1.736111
52	0	16	9.43	31	9.941676
660	0	34	8.43	32	-0.697788
1145	0	41	8.00	71	9.146341
697	0	33	8.00	45	4.545455
560	0	21	8.00	27	3.571429

- Mean percentage benefit is 1.925015
- What is the 95% confidence interval of the mean percentage benefit?

Confidence Interval - Example (cont.)

Demo in R

- Mean percentage benefit is 1.925015
- Standard deviation is 6.702202
- Sample size is 10

$$95\%\ CI = [\bar{X} - t^* \frac{s}{\sqrt{n}}, \bar{X} + t^* \frac{s}{\sqrt{n}}] \qquad (t^* \sim t_{n-1} = t_9)$$

Maximum Likelihood Estimation

- MLE selects the set of values of the parameters that maximizes the likelihood function
  - maximizes the “agreement” of the selected model with the observed data

MLE of λ for Poisson Distribution

Recall that the pmf for Poisson distribution is:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

For observations $x_1, ..., x_N$ that are i.i.d. $\sim Pois(\lambda)$, the likelihood (L) of this observation is:

$$L = P((X_1 = x_1) \cap ... \cap (X_N = x_N)) = \prod_{i=1}^{N} P(X_i = x_i)$$

$$= \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-N\lambda}\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!}$$

MLE of λ for Poisson Distribution

Since the likelihood is monotonically increasing log-likelihood is then:

$$log(L) = -N\lambda + (\sum_{i=1}^{N} x_i) log(\lambda) - log(\prod_{i=1}^{N} x_i!)$$

MLE of λ for Poisson Distribution

$$log(L) = -N\lambda + (\sum_{i=1}^{N} x_i) log(\lambda) - log(\prod_{i=1}^{N} x_i!)$$

Under suitable regularity conditions, the maximum likelihood estimate (estimator) is defined as:

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}^+}{\operatorname{argmax}}\, log(L)$$

FOC:

$$\frac{\partial log(L)}{\partial \lambda}\Big|_{\hat{\lambda}} = -N + \frac{1}{\hat{\lambda}} \sum_{i=1}^{N} x_i = 0$$

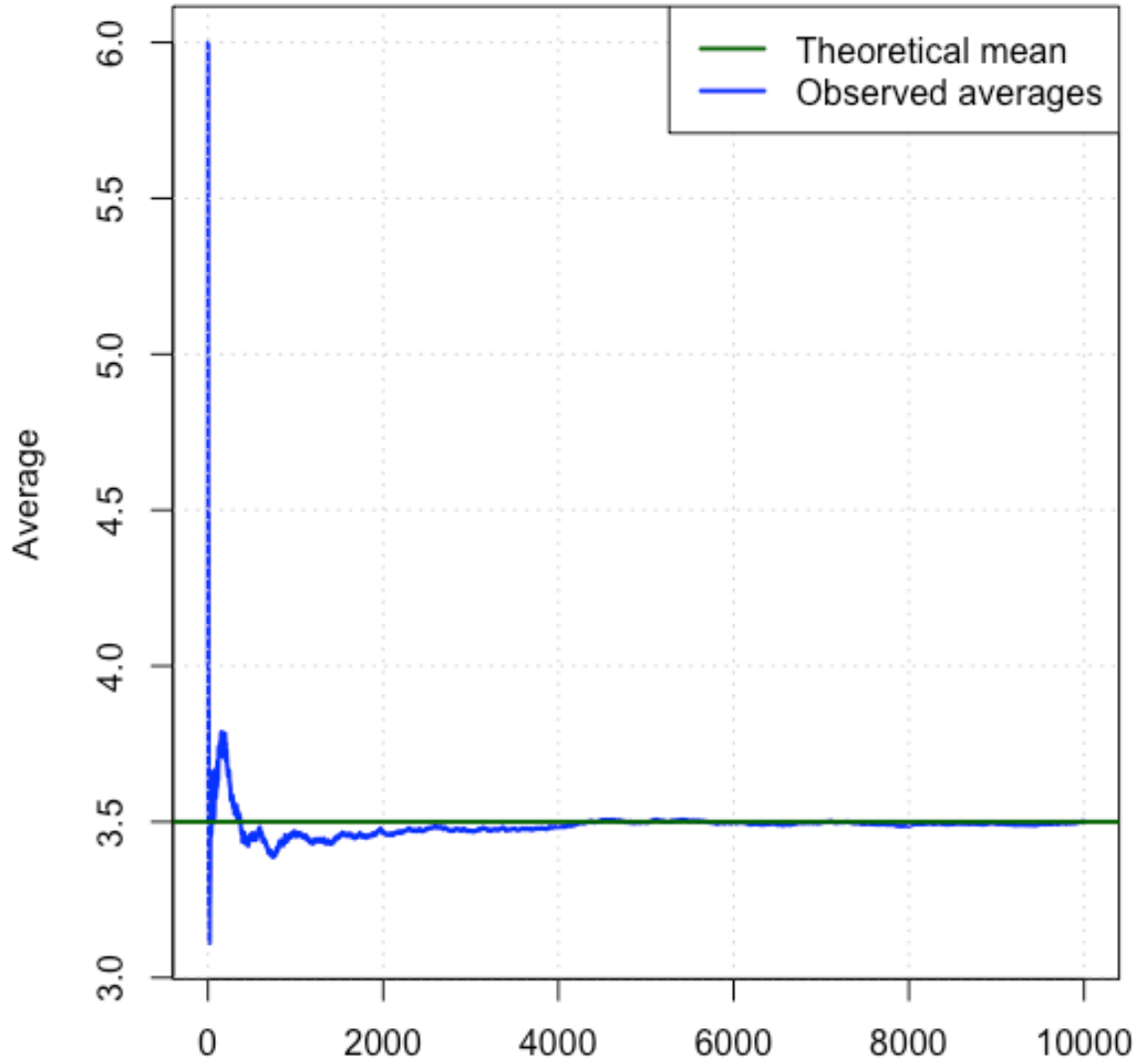
$$\iff \hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Law of Large Numbers

- the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed

- e.g., if $X_1, \ldots, X_n$ are i.i.d. normal variables with mean μ and variance σ^2, then $\bar{X}$ converges to μ as *n* increases

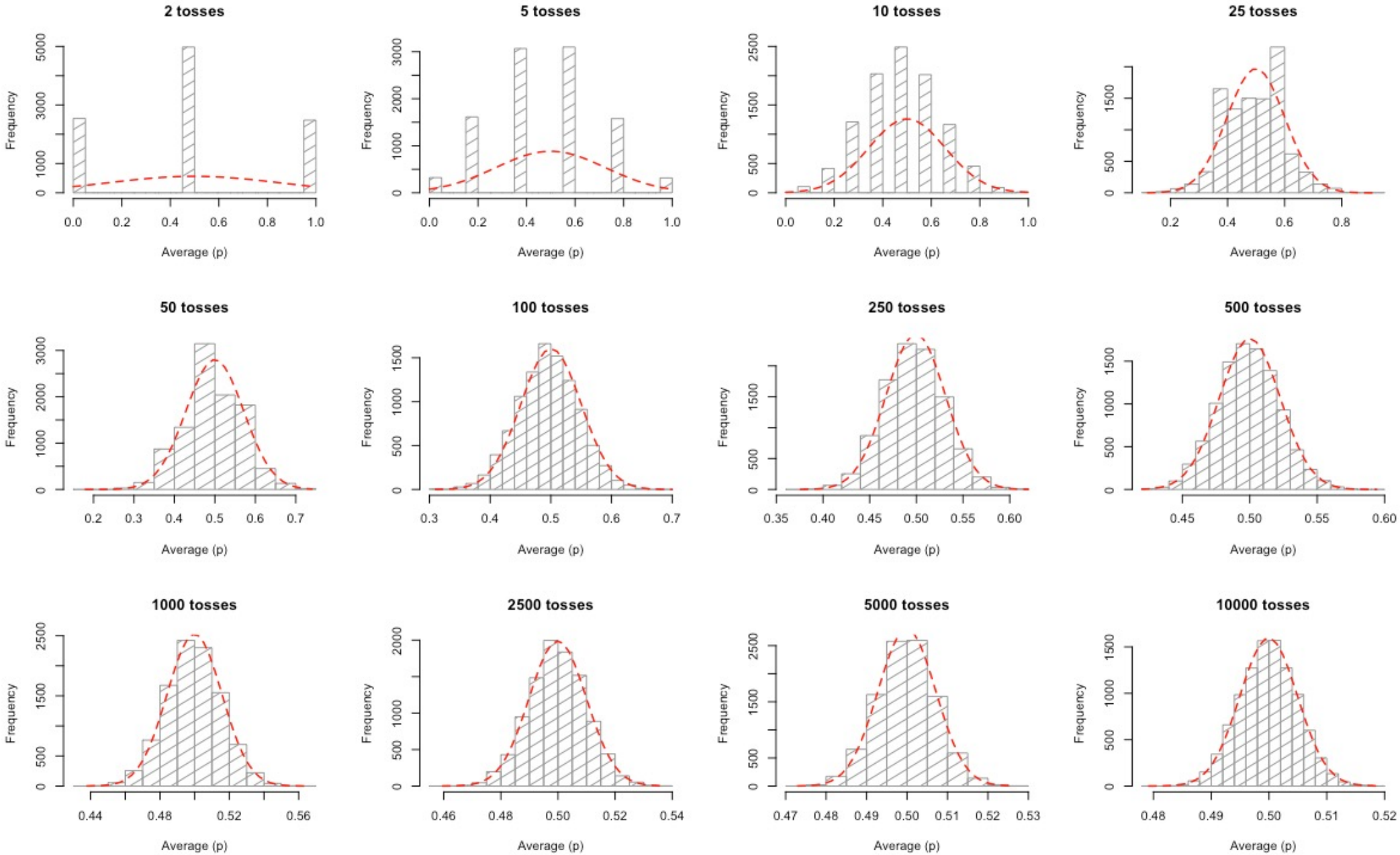
Law of Large Numbers



The Central Limit Theorem

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution

Tossing a coin n times (repeated for 10 000 times) – Distribution of sample means



The Central Limit Theorem

- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation
- A sufficiently large sample size can predict the characteristics of a population more accurately

Sampling Distributions of Mean and Variance

- $\bar{X}$ and s^2 are **point estimates**

- If $X_1, \ldots, X_n$ are i.i.d. RVs $\sim N(\mu, \sigma^2)$
  1. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
  2. $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(n-1)s^2}{\sigma^2}$
  3. $\bar{X}$ *and* s^2 *are independent*

Brief Summary

- MLE is a useful approach for finding an estimator that maximizes the "agreement" of the selected model with the observed data
- CLT states that "the distribution of sample means approximates a normal distribution as the sample size gets larger"
- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$