Handout 3
Statistics Review*

Instructor: Vira Semenova Note author: Danylo Tavrov, Vira Semenova

1 Definitions and Language

1.1 Population and Sample

Recall that, broadly defined, a *population* is the complete data collection to be studied, containing all the objects of interest, and a *sample* is a part of the population of interest, a sub-collection selected from a population.

To put things in perspective, suppose we are interested in measuring heights of all people on Earth. The population here would be the heights of all the people on Earth, whereas the sample would be the list of heights measured for some group of people.

Let X denote a random variable[1] standing for the height of a person. Naturally, it is characterized by some probability distribution. For the sake of an argument, let X follow a normal distribution with mean μ and variance σ^2, i.e. $X \sim N\left(\mu, \sigma^2\right)$.

Now, suppose we want to find out μ, i.e. the *average height* of a person. We cannot observe μ directly so we need to *estimate* it using the sample. For example, we independently draw n realizations of X to form a sample of size n: $\{x_1, x_2, \dots, x_n\}$. *Equivalently*, we can think of $\{x_1, x_2, \dots, x_n\}$ as being drawn from n *independent and identically distributed* (i.i.d.) random variables $\{X_1, X_2, \dots, X_n\}$. Here, independent means that the realization of one specific X_i does not affect the realization of any other X_j (knowledge of observation X_i tells us nothing about X_j), $i \neq j$; identical means that the values are drawn from the same distribution.

Note that if X_i are i.i.d., or that $X_1, X_2, \dots, X_n$ are *representative* of X, then

$$\prod_{i=1}^{n} \mathbb{P}_{X_i}(X_i = x_i) = \prod_{i=1}^{n} \mathbb{P}_X(X_i = x_i) \,,$$

where $\mathbb{P}_{X_i}(X_i = x) = \mathbb{P}_X(X_i = x)\ \forall x$ is the same for all i.

1.2 Estimands, Statistics, Estimators

Last time, we learned that a *parameter* is a numerical measurement that describes a characteristic of a population[2]. These parameters, called *estimands*, describe a population property and are fixed constants.

Example. Some examples of estimands are:

- the population mean of random variable X, $\mu_X = \mathbb{E}[X]$;
- the population variance of random variable X, $\sigma_X^2 = \mathrm{Var}(X)$;
- the covariance of random variables X and Y, $\sigma_{XY} = \mathrm{Cov}(X, Y)$.

*We thank Prof. Anna Mikusheva and numerous former ECON 140 and 141 GSIs for their course notes.

[1] We typically denote random variables $X_1, X_2, \dots, X_n$ using capital letters, and use small letters $x_1, x_2, \dots, x_n$ for *realizations* of these random variables, the ones that we actually *observe*.

[2] Speaking more formally, a parameter is an index for a family of probability distributions, which characterizes a random variable.

We also learned that a *statistic* is a numerical measurement that describes a characteristic of a sample. A statistic $g(X_1, \ldots, X_n)$ is a *function of the sample*: as sample changes, the statistic changes along with it. An expression that depends on an unknown quantity cannot be a statistic.

Example. Some examples of valid statistics are:

- the sample mean:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \; ; \tag{1.1}$$
- the sample variance:
$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \; ; \tag{1.2}$$
- the first observation, $g(X_1, X_2, \ldots, X_n) = X_1$;
- the constant zero, $g(X_1, X_2, \ldots, X_n) = 0$;
- the median of the sample, i.e. the value such that 50% of the observations are smaller and other 50% are greater than this value;
- the sample mean excluding 10% smallest and 10% largest values in the sample.

Example. Examples of quantities that are *not* statistics:

- demeaned sample mean, $\overline{X} - \mu_X$, as it depends on unknown μ_X;
- Z-score (standardized mean), $Z = \dfrac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}}$, as it depends on unknown μ_X and σ_X.

An *estimator* is a statistic that is used to estimate a parameter. An estimator that is used to estimate the parameter is a *random variable* (because of the randomness involved in selecting the sample)! An *estimate* is the realization of an estimator given the data from a specific sample. It is a *constant*.

Typically, we denote an estimator of some population parameter θ by $\hat{\theta}$.

2 Properties of Estimators

Properties of estimators can be roughly divided into *small-sample properties* (finite-sample properties), which describe estimator's behavior for a specific sample size n, and *large-sample properties* (asymptotic properties), which describe the behavior as $n \to \infty$. Small-sample properties we will discuss include unbiasedness and efficiency, and a leading asymptotic property we are interested in is consistency.

2.1 Unbiasedness

Typically, we prefer our estimators to be "accurate," in a sense that they reflect the population parameter as closely as possible. Informally, it means that if we collect many samples and compute estimates from each of them, then the average of those estimates will be close to the true parameter.

Formally, we say that estimator $\hat{\theta}$ is an *unbiased* estimator of some parameter θ if

$$\mathbb{E}\left[\hat{\theta}\right] = \theta \; .$$

The *bias* then is defined as the average gap between the estimator and the parameter:

$$\text{Bias} = \mathbb{E}\left[\hat{\theta}\right] - \theta \; .$$

Example. Let $U_1, U_2, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0, b]$, where b is an unknown parameter. Consider an estimator

$$\hat{b} = \max\{U_1, U_2, \ldots, U_n\} .$$

First, observe that the CDF of this statistic is as follows:

$$\begin{aligned} F_{\hat{b}}(x) &= \mathbb{P}\left(U_1 \leq x \, , \ldots \, , U_n \leq x\right) \\ &= \mathbb{P}\left(U_1 \leq x\right) \cdot \ldots \cdot \mathbb{P}\left(U_n \leq x\right) \\ &= F_U(x)^n = \begin{cases} 0 \, , & x < 0 \\ \left(\dfrac{x}{b}\right)^n \, , & 0 \leq x \leq b \\ 1 \, , & b < x \end{cases} . \end{aligned}$$

where we used independence and identical distribution of $U_1, \ldots, U_n$ and applied the formula of CDF for uniformly distributed random variables.
Then, the PDF of the statistic is given by the following derivative:

$$f_{\hat{b}}(x) = \frac{d}{dx} F_{\hat{b}}(x) = \begin{cases} \dfrac{n x^{n-1}}{b^n} \, , & 0 \leq x \leq b \\ 0 \, , & \text{otherwise} \end{cases} .$$

Then,

$$\mathbb{E}\left[\hat{b}\right] = \int_0^b x \cdot \frac{n x^{n-1}}{b^n} \, dx = \frac{n}{b^n} \cdot \frac{x^{n+1}}{n+1}\bigg|_0^b = \frac{n}{n+1} \cdot b \, .$$

Thus, we conclude that $\hat{b}$ is a *biased* estimator of parameter b:

$$\text{Bias} = \frac{n}{n+1} \cdot b - b = -\frac{b}{n+1} \neq 0 \, .$$

We can easily introduce an *unbiased* estimator by correcting the bias as follows:

$$\hat{b}_{\text{corrected}} = \frac{n+1}{n} \hat{b} \, .$$

Note also that another unbiased estimator of b is given by

$$\tilde{b} = 2\overline{U} \equiv \frac{2}{n} \sum_{i=1}^{n} U_i \, .$$

We encourage the reader to prove this fact on their own.

2.2 Efficiency

When selecting an estimator, we want it to be not only accurate (unbiased) but also as concentrated around the true mean as possible, i.e. we want an estimator with the *smallest variance possible*. Efficiency is an important and desirable property, because in practice, we usually only have one sample, i.e., we can only get one realization of the estimator. If the variance of an estimator is very large, that realization could be very far away from the mean of the estimator (which, for unbiased estimators, equals to the parameter we are estimating).

As is, however, the question does not make much sense. For example, the constant estimator, $\hat{\theta}_c \equiv 0$ has zero variance. Therefore, we focus on the *unbiased* estimators going forward.

We say that estimator $\hat{\theta}$ is *more efficient* than estimator $\tilde{\theta}$ if $\text{Var}\left(\hat{\theta}\right) \leq \text{Var}\left(\tilde{\theta}\right)$ for all θ, and if $\text{Var}\left(\hat{\theta}\right) < \text{Var}\left(\tilde{\theta}\right)$ at least for some θ (otherwise the variances would be the same). We say that an estimator $\hat{\theta}$ is *the most efficient*, or simply *efficient*, in some class of estimators if it has the smallest variance among all the estimators in this class.

Example. An estimator $\tilde{\theta} = X_1$ of the population mean, μ_X, of some random variable with finite nonzero variance σ_X^2, is less efficient than estimator $\check{\theta} = (X_1 + X_2)/2$, because they are both unbiased ($\mathbb{E}\left[\tilde{\theta}\right] = \mathbb{E}\left[\check{\theta}\right] = \theta$) but $\check{\theta}$ has lower variance than $\tilde{\theta}$:

$$\text{Var}\left(\check{\theta}\right) = \frac{\sigma_X^2}{2} < \text{Var}\left(\tilde{\theta}\right) = \sigma_X^2 \ , \quad \forall \theta \in \mathbb{R} \ .$$

2.3 Mean Squared Error

To measure how far $\hat{\theta}$ is from θ on average, we use least squares distance. Define the *mean squared error* (MSE) as follows:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \ . \tag{2.1}$$

Decomposing $\hat{\theta} - \theta$ into bias and variance gives

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + \mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right) \cdot \left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)\right] + \left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2 \\
&= \text{Var}\left(\hat{\theta}\right) + 2\mathbb{E}\left[\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right] \cdot \left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right) + \text{Bias}^2\left(\hat{\theta}\right) \\
&= \text{Var}\left(\hat{\theta}\right) + 0 + \text{Bias}^2\left(\hat{\theta}\right) \\
&= \text{Var}\left(\hat{\theta}\right) + \text{Bias}^2\left(\hat{\theta}\right) \ .
\end{aligned}$$

Note that for unbiased estimators, MSE coincides with the variance.

2.4 Consistency

Speaking informally, an estimator is *consistent* if its distribution around the mean gets narrower and narrower as sample size n increases indefinitely. Formally, we say that an estimator is consistent if it *converges in probability* to the parameter it estimates as n goes to infinity:

$$\hat{\theta}_n \xrightarrow{p} \theta \ ,$$

where $\hat{\theta}_n$ is the estimator of θ for the sample of size n. Or, using the definition of convergence in probability:

$$\forall \varepsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0 \ .$$

One special case of consistency follows from the *Chebyshev inequality*, which states the following:

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(|X - \mu| \geq \varepsilon\right) \leq \frac{\text{Var}\left(X\right)}{\varepsilon^2} \ ,$$

where X is a random variable with mean μ. If estimator $\hat{\theta}$ is unbiased, then the Chebyshev inequality tells us that

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(\left|\hat{\theta} - \theta\right| \geq \varepsilon\right) \leq \frac{\text{Var}\left(\hat{\theta}\right)}{\varepsilon^2} \ .$$

Therefore, if we know that an estimator is unbiased, and we can show that its variance tends to zero as n tends to infinity, we will be able to immediately conclude that it is consistent.

3 Sample Mean and Sample Variance

3.1 Sample Mean Estimator

Assume that X_i, $i = 1, \ldots, n$, are random variables that are i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$ for each i. More concisely, we will often write $X_i \sim \text{i.i.d.}(\mu, \sigma^2)$.

Consider the estimator (1.1) of μ. We can show that it possesses several useful properties.

3.1.1 Properties of Sample Mean

Unbiasedness. To show that $\overline{X}$ is an unbiased estimator of μ, observe:

$$\begin{aligned}\mathbb{E}\left[\overline{X}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu\,,\end{aligned}$$

where we have used the property of linearity of expectation and the fact that X_i are identically distributed.

Consistency. To show that $\overline{X}$ is a consistent estimator of μ, we will make use of the observation from Sect. 2.4 that an unbiased estimator is consistent if its variance tends to zero as n tends to infinity. To that end, let's first calculate the variance of the sample mean:

$$\begin{aligned}\text{Var}\left(\overline{X}\right) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\left(\text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) + 2\text{Cov}(X_1, X_2) + 2\text{Cov}(X_1, X_3) + \ldots\right) \\ &= [\text{by independence}] \\ &= \frac{1}{n^2}\left(\text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) + 0 + 0 + \ldots\right) \\ &= [\text{identically distributed}] \\ &= \frac{1}{n^2}\left(\text{Var}(X_1) + \text{Var}(X_1) + \ldots + \text{Var}(X_1)\right) \\ &= \frac{1}{n^2}\left(n\text{Var}(X_1)\right) \\ &= \frac{1}{n^2}\left(n\sigma^2\right) \\ &= \frac{\sigma^2}{n}\,.\end{aligned}$$

Indeed, this last expression goes to zero as n goes to infinity.

We have just proven the *(weak) law of large numbers* (WLLN), whose formulation is as follows. Let all X_i be i.i.d. random variables with $\mathbb{E}[X_i] = \mu_X$ and $\text{Var}(X_i) = \sigma_X^2$ for all i[3]. Then, $\overline{X}$ converges in probability to the population mean μ_X:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu_X\,.$$

[3] The law of large numbers requires that $\text{Var}(X)$ is finite, which is always the case in this class.

Efficiency. An estimator $\hat{\theta}$ is linear in sample vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$ if it can be represented as a linear combination of X_i:

$$\hat{\theta} = \sum_{i=1}^{n} w_i X_i \equiv \mathbf{w}^\top \mathbf{X}, \tag{3.1}$$

where $\mathbf{w}^\top = (w_1, w_2, \ldots, w_n)$ is a *weighting* vector of fixed weights. Expression (3.1) is also called a *weighted average* of the random variables.

Common examples of linear estimators conclude:

- the *first observation* $\hat{\theta} = X_1$, with the weights $\mathbf{w}^\top = (1, 0, \ldots, 0)$;
- the *sample mean* $\hat{\theta} = \overline{X}$, with the weights $\mathbf{w}^\top = (1/n, 1/n, \ldots, 1/n)$.

Let us consider a class of estimators of the population mean that are weighted averages, $\mathcal{F} = \{\sum_{i=1}^{n} w_i X_i \ , \ \mathbf{w} \in \mathbb{R}^n\}$, and investigate properties of the estimators within this class.

First of all, a linear estimator $\hat{\theta}$ is an unbiased estimator of the population mean μ_X if and only if

$$\mathbb{E}\left[\hat{\theta}\right] = \mathbb{E}\left[\sum_{i=1}^{n} w_i X_i\right] = \left(\sum_{i=1}^{n} w_i\right) \mu_X = \mu_X \ .$$

Therefore, $\hat{\theta}$ is unbiased if and only if[4]

$$\sum_{i=1}^{n} w_i = 1 \ .$$

Note that by this logic, both the first observation and the sample mean are unbiased estimators of the population mean.

A linear estimator $\hat{\theta}$ has the following variance:

$$\mathrm{Var}\left(\hat{\theta}\right) = \mathrm{Var}\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i^2 \mathrm{Var}\left(X_i\right) = \left(\sum_{i=1}^{n} w_i^2\right) \sigma_X^2 \ ,$$

where we have used independence and the fact that X_i are identically distributed.

We are now seeking to prove that the sample mean $\overline{X}$ is (the most) efficient in the class $\mathcal{F}$. Note that a linear unbiased estimator $\hat{\theta}$ has the *smallest* variance in $\mathcal{F}$ when expression for $\mathrm{Var}\left(\hat{\theta}\right)$ is minimized over all $\mathbf{w}$, subject to our unbiasedness condition:

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{i=1}^{n} w_i^2 \\ &\text{s.t.} \sum_{i=1}^{n} w_i = 1 \ , \end{aligned}$$

where the optimal solution is[5]

$$\mathbf{w}^* = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \ .$$

This vector of weights exactly corresponds to the sample mean. Therefore, the sample mean estimator is the *best linear unbiased estimator* (BLUE) of the population mean. In other words, $\overline{X}$ is efficient in the class of linear unbiased estimators.

[4] Unless $\mu_X = 0$, in which case the weights can sum up to any constant, but we will *normalize* them by requiring that they sum up to 1.

[5] The Lagrangian for this problem is $\mathcal{L} = \sum_{i=1}^{n} w_i^2 - \lambda\left(\sum_{i=1}^{n} w_i - 1\right)$, and the first order conditions are $2w_j = \lambda$ for each $j = 1, \ldots, n$. Noting the constraint, we observe that $\sum_{i=1}^{n} w_i = \frac{n\lambda}{2} = 1$, from where it follows that $\lambda = \frac{n}{2}$, and thus $w_j = \frac{1}{n}$ for each j.

Note that if $\text{Var}(X_i) = \sigma_i^2$, i.e. if the variances are not required to be the same for all X_i, then the minimization program becomes

$$\min_{\mathbf{w}} \sum_{i=1}^{n} w_i^2 \sigma_i^2$$
$$\text{s.t.} \sum_{i=1}^{n} w_i = 1 \ ,$$

and its solution can be shown to be equal to $w_i = \frac{1}{C\sigma_i^2}$ for each i, where $C = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$.

3.1.2 Sample Mean as a Least Squares Estimator

Suppose that we observe a sample $(x_1, \ldots, x_n)$, and our goal is to find a value m^*, such that the sum of the distances from each x_i, $i = 1, \ldots, n$, to $\hat{\theta}$ is the smallest.
Formally, we have the following minimization problem[6]:

$$m^* = \arg\min_{m} \sum_{i=1}^{n} (x_i - m)^2 \ .$$

We can rearrange the minimand as follows:

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - m)^2 &= \sum_{i=1}^{n} \left(x_i - \overline{X} + \overline{X} - m\right)^2 \\
&= \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 + 2\sum_{i=1}^{n} \left(x_i - \overline{X}\right)\left(\overline{X} - m\right) + \sum_{i=1}^{n} \left(\overline{X} - m\right)^2 \\
&= \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 + 2\left(\overline{X} - m\right)\sum_{i=1}^{n} \left(x_i - \overline{X}\right) + n\left(\overline{X} - m\right)^2 \ .
\end{aligned}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n} X_i - \overline{X} = 0 \ ,$$

and therefore,

$$\sum_{i=1}^{n} (x_i - m)^2 = \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 + n\left(\overline{X} - m\right)^2 \ .$$

Since $\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2$ does not depend on m, the whole sum is minimized when the second term is zero, i.e. when $m = \overline{X}$.
Therefore, the sample mean $\overline{X}$ is the *least squares* estimator of the population mean, i.e. it is such a value that the sum of squared distances from each value in the sample to $\overline{X}$ is minimal.

3.2 Illustration of Sample Mean Properties in R

Consider a random variable X distributed as a Gamma random variable with a shape parameter $k = 1$ and a scale parameter $\theta = 2$[7]. The mean of this random variable is given by $\mathbb{E}[X] = k\theta = 2$. We assume that we don't know this and want to estimate the mean using the sample we have.
Consider a sample of size n drawn from the corresponding population. We will use the sample mean $\overline{X}$ as our estimate of the population mean μ_X. Let us demonstrate the properties of this estimator (unbiasedness, consistency, efficiency) using *Monte Carlo simulations*. A Monte Carlo simulation is a repeated simulation of random data:

[6] Note that we could have used $|x_i - c|$ as the definition of distance, but using the squared difference is more analytically tractable.
[7] Refer to https://en.wikipedia.org/wiki/Gamma_distribution for visual depiction of the PDF of this distribution.

- for each simulation, we generate n random data points from an assumed *data generating process* (DGP), which in our case is given above as a Gamma random variable with specified parameters;
- we repeat the same simulation T times (each time with new random data) and thus generate many estimates of the statistics of interest;
- in the end, we can characterize the distribution of the statistics of interest that we generated using some summary. Here, we will plot appropriate histograms.

Let us perform a Monte Carlo simulation $T = 10{,}000$ times for three different sample sizes, $n = 10, 100, 1000$. Histograms for $\overline{X}$ are given in Fig. 3.1. In this picture, we see two properties of the sample mean:

- the sample mean is unbiased, because for each sample size the distribution of the estimator is center around the true mean;
- the sample mean is consistent, because as the sample size gets bigger, the distribution becomes narrower and narrower and is almost concentrated on the true mean;

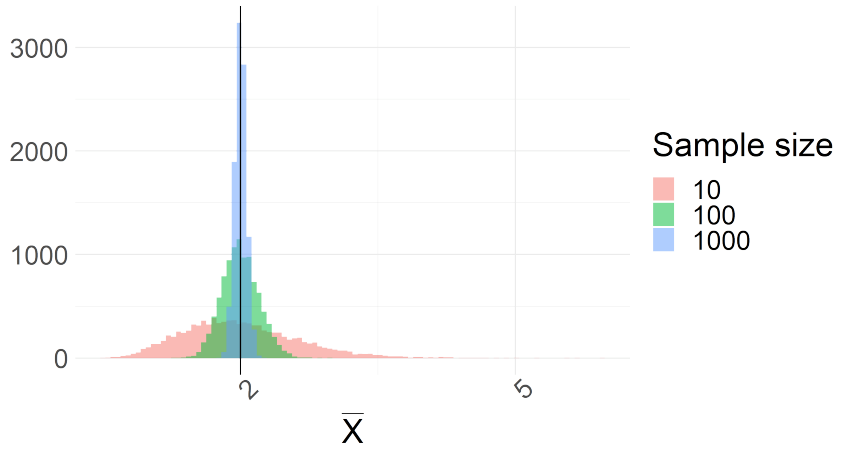

Figure 3.1: Histograms of the distribution of $\overline{X}$ for the Monte Carlo simulations (true parameter is indicated by a vertical bar)

To illustrate efficiency of the sample mean, we perform the same exercise, but for sample size $n = 10$ and two estimators, $\hat{\mu}_X = \overline{X}$ and $\tilde{\mu}_X = X_1$. We know from theory that both of them are unbiased but the sample mean is efficient. From Fig. 3.2, we see that, indeed, the sample mean has lower variance than the first observation.

The R code for this exercise is given in Listing 3.2[8].

Listing 1: R code used to perform Monte Carlo simulations for sample mean properties

```
library(ggplot2)
library(latex2exp)


# Helper function to plot one histogram
plot.histogram <- function(data, true.value, sample.sizes){
  # plot mean vs. first
  for (n in sample.sizes){
    # for each sample size

```

[8] Actually, when the script is run, more files are produced than given here. In particular, the file `hist_mean_first_1000.png` provides visual illustration to the fact that X_1 is not consistent.

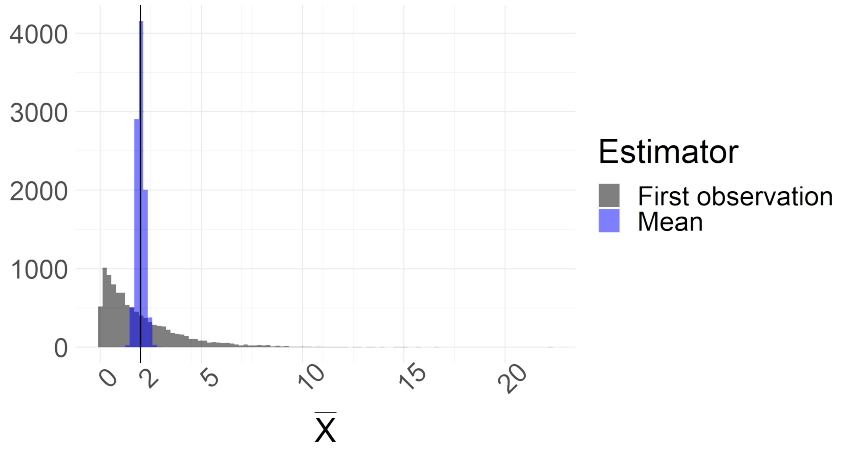

Figure 3.2: Histograms of the distribution of $\overline{X}$ and X_1 for the Monte Carlo simulations with sample size $n = 10$ (true parameter is indicated by a vertical bar)

```
    ggplot(
      data[data$sample.size == n, ],
      aes(x = coef, fill = type)
    ) +
      geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
      geom_vline(xintercept = true.value) +
      scale_fill_manual(
        name = "Estimator",
        values = c("black", "blue")
      ) +
      labs(x = TeX("$\\bar{X}$"), y = "") +
      scale_x_continuous(breaks = c(pretty(data$coef), true.value)) +
      theme_minimal() +
      theme(
        text = element_text(size = 25),
        plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
        axis.text.x = element_text(angle = 45)
      )

    ggsave(
      paste0("hist_mean_first_", n, ".png"),
      width = 3000, height = 1500, units = "px"
    )
  }

  # plot means for different sample sizes
  ggplot(data[data$type == "Mean", ], aes(x = coef, fill = as.factor(sample.size))) +
    geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
    geom_vline(xintercept = true.value) +
    guides(fill = guide_legend(title = "Sample size")) +
    labs(x = TeX("$\\bar{X}$"), y = "") +
    scale_x_continuous(breaks = c(pretty(data$coef), true.value)) +
    theme_minimal() +
    theme(
      text = element_text(size = 25),
      plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
      axis.text.x = element_text(angle = 45)
    )

  ggsave(
    paste0("hist_mean.png"),
    width = 3000, height = 1500, units = "px"
```

```
  )
}


# Function performing a Monte Carlo simulation
# for a data sample of size n with T repetitions
monte.carlo <- function(sample.sizes, T){
  # allocate memory
  X.mean <- rep(0, T)
  X.first <- rep(0, T)
  df <- NULL
  for (n in sample.sizes){
    # for each sample size

    for (i in 1:T){
      # generate a dataset
      X <- rgamma(n, shape = 1, scale = 2)

      # compute the mean
      X.mean[i] <- mean(X)
      X.first[i] <- X[1]
    }

    # add these means to the dataframe
    df <- rbind(
      df,
      data.frame(
        coef = c(X.mean, X.first),
        sample.size = rep(n, length(X.mean) + length(X.first)),
        type = c(
          rep("Mean", length(X.mean)),
          rep("First observation", length(X.first))
        )
      )
    )
  }

  # plot the histograms
  plot.histogram(df, true.value = 2, sample.sizes = sample.sizes)
}


# set random number generator seed for reproducibility
set.seed(100)

# set working directory to the current file
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# run Monte Carlo simulation
monte.carlo(
  sample.sizes = c(10, 100, 1000), T = 10000
)
```

3.3 Sample Variance Estimator

We know that the variance of a random variable X can be calculated by the expectation operator:

$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] .$$

We know from Sect. 3.1 that $\mathbb{E}[X]$ can be estimated by the sample mean (1.1), and that it possesses several nice properties when X_i are i.i.d., $i = 1, \ldots, n$.

A natural first attempt at selecting an estimator for a variance would be to use the following quantity:

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2 \ . \tag{3.2}$$

However, it is not a statistic since μ_X is an unknown population parameter. However, we can estimate μ_X by $\overline{X}$. If we plug this into (3.2), we get:

$$\hat{\sigma}_X^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \ .$$

Let us rearrange the terms as follows:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X+\mu_X-\overline{X}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2+\frac{2}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)\left(\mu_X-\overline{X}\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\mu_X-\overline{X}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2+\frac{2}{n}\left(n\overline{X}-n\mu_X\right)\left(\mu_X-\overline{X}\right)+\left(\mu_X-\overline{X}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2-2\left(\mu_X-\overline{X}\right)+\left(\mu_X-\overline{X}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2-\left(\mu_X-\overline{X}\right)^2 \ .
\end{aligned}$$

Expectation of this expression is as follows:

$$\mathbb{E}\left[\hat{\sigma}_X^2\right]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2\right]-\mathbb{E}\left[\left(\mu_X-\overline{X}\right)^2\right]=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right)-\mathrm{Var}\left(\overline{X}\right)=\sigma_X^2-\frac{\sigma_X^2}{n}=\frac{n-1}{n}\sigma_X^2 \ ,$$

where we have used the formula for the variance of the sample mean. Therefore, $\hat{\sigma}_X^2$ is *biased*. An unbiased estimator for a population variance is called *sample variance* and defined as follows:

$$s_X^2 \equiv \frac{n}{n-1}\hat{\sigma}_X^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \ . \tag{3.3}$$

We can show that $\hat{\sigma}_X^2$ is a consistent estimator of σ_X^2. Indeed, in the expression above, the first term, $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2$, is a sum of i.i.d. random variables[9], with expectation equal to σ_X^2 and variance

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2\right)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\left(X_i-\mu_X\right)^2\right)=\frac{1}{n^2}\sum_{i=1}^{n}\left(\mathrm{Var}\left(X_i^2\right)+4\mu^2\mathrm{Var}\left(X_i\right)\right) \ ,$$

which converges to 0 as $n\to\infty$. Therefore, the first term is consistent for $\mathrm{Var}\left(X_i\right)$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_X\right)^2 \xrightarrow{p} \mathrm{Var}\left(X_i\right) \ .$$

The second term, $\left(\mu_X-\overline{X}\right)^2$, converges in probability to 0, which establishes the fact that

$$\hat{\sigma}_X^2 \xrightarrow{p} \sigma_X^2 \ .$$

Having shown consistency $\hat{\sigma}_X^2$, we immediately conclude that $s_X^2 \xrightarrow{p} \sigma_X^2$ because $\frac{n}{n-1}\to 1$ as $n\to\infty$.

[9] Note that $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ is *not* a sum of i.i.d. random variables, therefore we had to go at some length to transform this into a more tractable expression.

4 Limit Theorems

4.1 Central Limit Theorem

We already encountered one limit theorem, WLLN, in the previous section. Another important theorem is *central limit theorem* (CLT), which is stated as follows.

Let all X_i be i.i.d. random variables[10] with $\mathbb{E}[X_i] = \mu_X$ and $\text{Var}(X_i) = \sigma_X^2$ for all i[11]. Denote by

$$Z_n = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \tag{4.1}$$

the *standardized mean*, or the *Z-score*[12]. Note that Z_n is *not* a statistic, in the sense that we cannot compute it without knowledge of population parameters μ_X and σ_X^2. The name "standardized" comes from the fact that (as is easy to calculate) $\mathbb{E}[Z_n] = 0$ and $\text{Var}(Z_n) = 1$.

Then, the following holds:

$$Z_n = \sqrt{n}\left(\frac{\overline{X} - \mu_X}{\sigma_X}\right) \xrightarrow{d} N(0,1) \ .$$

This means that as n increases, the sequence $\{Z_n\}$ of random variables *converges in distribution* to the standard normal distribution, in other words, $\lim_{n\to\infty} F_{Z_n}(x) = \Phi(x)$, where $\Phi(x)$ is the standard normal CDF[13]. It follows that

$$\sqrt{n}\left(\overline{X} - \mu_X\right) \xrightarrow{d} N(0, \sigma_X^2) \ .$$

A (somewhat informal) interpretation of this expression is as follows:

$$\overline{X} \overset{\text{a}}{\sim} N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \ , \tag{4.2}$$

meaning that asymptotically (for large n), $\overline{X}$ is approximately distributed with mean μ_X and variance $\frac{\sigma_X^2}{n}$.

4.2 Continuous Mapping Theorem

A very useful theorem that allows us to establish properties of estimators is the *continuous mapping theorem* (CMT). There are two versions of the theorem, one for convergence in probability, and another one for convergence in distribution.

More precisely, if $X_n \xrightarrow{p} c$, where $c \in \mathbb{R}$, and $g(\cdot)$ is a function continuous at c, then $g(X_n) \xrightarrow{p} g(c)$.

Likewise, if $X_n \xrightarrow{d} X$, and $g(\cdot)$ is discontinuous on a set of points D_g such that $\mathbb{P}(X \in D_g) = 0$[14], then $g(X_n) \xrightarrow{d} g(X)$.

CMT can be immediately applied to generalize the LLN, i.e. to prove that all sample moments of i.i.d. data converge to the corresponding population moments (so the sample mean converges to the population mean, the sample variance converges to the population variance, etc.)[15].

[10] This is the so called Lindeberg-Lévy CLT. There are other versions of CLT, for random variables that are not necessarily independent or identically distributed.

[11] The central limit theorem requires that $\text{Var}(X)$ is finite, which is always the case in this class.

[12] Technically speaking, a Z-score, for a random variable X with mean μ_X and variance σ_X^2, is defined as $Z = \frac{X - \mu_X}{\sigma_X}$. We slightly abuse the terminology and call (4.1) a Z-score when we should understand that it is a Z-score for a sample mean $\overline{X}$.

[13] In general, convergence in distribution tells us that a sequence of numbers $\{F_n(x)\}$, for some sequence of random variables distributed according to F_n, converges to the limiting distribution $F(x)$ for each x where $F(x)$ is continuous. Since $\Phi(x)$ is continuous everywhere, we omit this qualifier from the statement of the CLT.

[14] In other words, $g(\cdot)$ is basically "continuous at X."

[15] Assuming the expectation of the squared random variable in question is finite and so on.

4.3 Slutsky Theorem

Yet another powerful limit theorem is the *Slutsky Theorem*. It claims that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n Y_n \xrightarrow{d} Xc$;
- $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ if $c \neq 0$.

Example. Consider a random variable X with mean μ_X and variance σ_X^2. We know that, by CLT, $Z \xrightarrow{d} N(0,1)$, where Z is defined as in (4.1). Suppose we don't know the true value of σ_X^2 but estimate it using the sample variance s_X^2. We obtain the following *t-score*:

$$t = \frac{\overline{X} - \mu_X}{\sqrt{s_X^2/n}} .$$

We can show that $t \xrightarrow{d} N(0,1)$. Since we know that $s_X^2 \xrightarrow{p} \sigma_X^2$, by CMT, $\sqrt{\frac{\sigma_X^2}{s_X^2}} \xrightarrow{p} \sqrt{\frac{\sigma_X^2}{\sigma_X^2}} = 1$.

Then, by the Slutsky theorem,

$$t = Z \cdot \sqrt{\frac{\sigma_X^2}{s_X^2}} \xrightarrow{d} N(0,1) \cdot 1 = N(0,1) .$$

4.4 Different Types of Convergence

It is important to clarify the distinction between two different types of convergence:

- convergence *in probability* happens when our statistic converges to a number. For instance, in LLN, $\overline{X} \xrightarrow{p} \mu_X$, which is a number;
- convergence *in distribution* happens when our statistic converges to a *random variable*. For instance, in CLT, $\frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}}$ converges to a standard normal random variable as n gets large. It does *not* converge to an exact number, but we know the distribution that it converges to.

4.5 Standard Deviation and Standard Error

Above, we discussed the estimator of the population variance, which we called sample variance and denoted by s_X^2. Its square root,

$$s_X = \sqrt{s_X^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} ,$$

is an estimator of the true population standard deviation.

One shouldn't mistake it for the *standard error*. The standard error is the estimator of the standard deviation of the asymptotic distribution and is used to estimate the spread of an estimator, whose asymptotic distribution is characterized by the CLT. For instance, the standard error for the sample mean is $\text{s.e.}\left(\overline{X}\right) = \frac{s_X}{\sqrt{n}}$, because $\text{s.e.}\left(\overline{X}\right) \xrightarrow{p} \frac{\sigma_X}{\sqrt{n}} = \sqrt{\text{Var}\left(\overline{X}\right)}$.

4.6 Illustration of Limit Theorems in R

To illustrate the law of large numbers and the central limit theorem in practice, we will discuss a simple example in R that uses Monte Carlo simulations.

For our illustration, we will generate n independent exponentially distributed random variables with $\lambda = 4$, i.e. their pdf is as follows:

$$f(x; \lambda = 4) = \begin{cases} 4e^{-4x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

From the properties of this distribution, we know that $\mu_X = \frac{1}{\lambda} = \frac{1}{4}$ and $\sigma_X^2 = \frac{1}{\lambda^2} = \frac{1}{16}$. For a sample of n values, we calculate the following statistics:

- the sample mean $\overline{X}$;
- the sample variance s_X^2;
- the t-score;
- the Z-score.

We will perform four Monte Carlo simulations with $T = 10{,}000$ and different sample sizes: $n = 10$, 100, 1000, and 10,000. In other words, for each of the four simulations, we will have 10,000 different values of our four statistics.

Histograms for the sample mean for different n are given in Fig. 4.1. Each graph plots the histogram of the $T = 10{,}000$ simulated values of $\overline{X}$ for the given sample size. Notice that the spread of the histogram gets thinner and thinner as n grows and centers around 0.25, the true mean of X_i. This means that as n increases, $\overline{X}$ is converging to the population value of 0.25. In other words, we see an illustration of the LLN: the sample means converges to the population mean as $n \to \infty$, or $\overline{X} \xrightarrow{p} \mathbb{E}[X]$.

Also, we can see that the histogram looks "more normal" as n increases. This is an illustration of the approximation using the CLT given by (4.2).

The same can be said about the sample variance. The respective histograms are given in Fig. 4.2. Again, each graph plots the histogram of the $T = 10{,}000$ simulated values of the sample variance s_X^2 for the given sample size. Again, notice that the spread of the histogram gets thinner and thinner as n increases and centers around 0.0625, the true variance of X_i. This also illustrates the LLN (and CMT): the sample variance converges to the population variance as $n \to \infty$, or $s_X^2 \xrightarrow{p} \text{Var}(X)$.

Let us now turn to the illustration of the CLT. Figure 4.3 shows the four histograms for Z-scores. As n increases, the histogram of Z-scores looks more and more like a standard normal distribution.

Additionally, we can see that the t-score histograms follow the same pattern. The only difference is that we estimate σ_X^2 with s_X^2, but for large n, as we saw with the LLN, we know that s_X^2 approaches σ_X^2, which means that for n large enough, the distributions of Z and t are the same.

The R code for this exercise is given in Listing 4.6.

Listing 2: R code used to perform Monte Carlo simulations for limit theorems

```
library(ggplot2)


# Function performing a Monte Carlo simulation
# for a data sample of size n with T repetitions
monte.carlo <- function(n, T, lambda) {
  # allocate memory for our statistics
  sample.mean <- rep(0, T)
  sample.variance <- rep(0, T)
  t.score <- rep(0, T)
  Z.score <- rep(0, T)

  # pre-compute mean and variance of the distribution
```

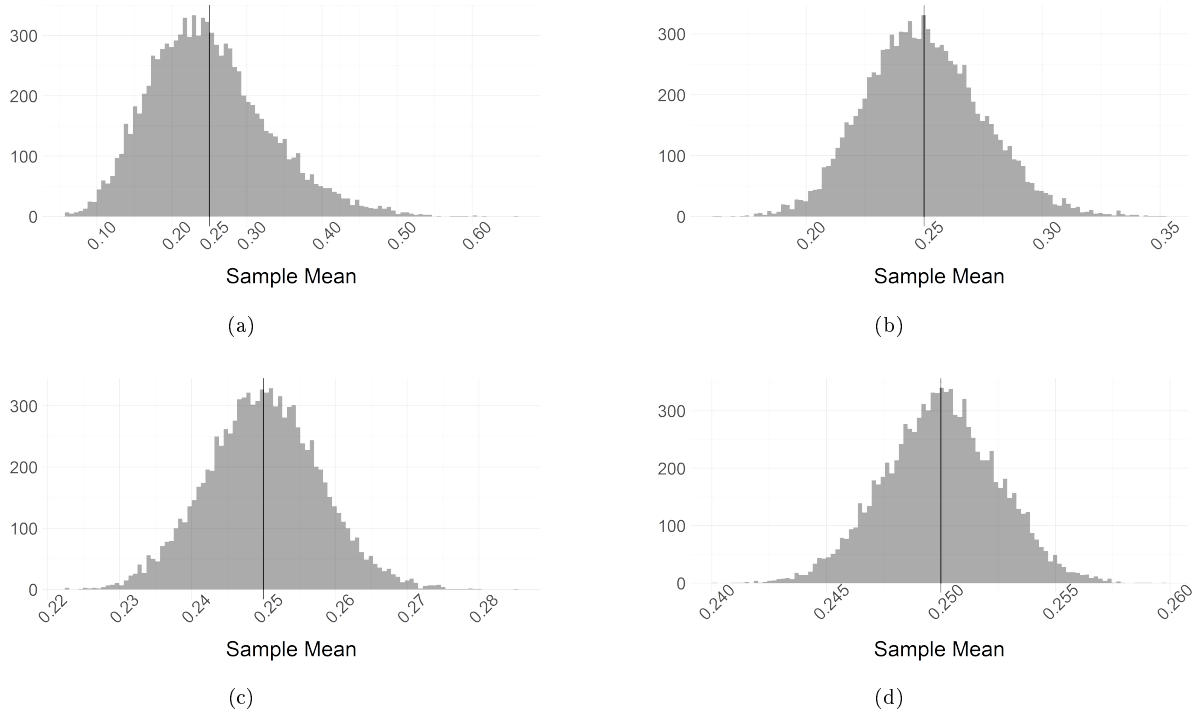

Figure 4.1: Histograms of the distribution of the sample means for the Monte Carlo simulations with sample size n (true mean is indicated by a vertical bar): (a) $n = 10$; (b) $n = 100$; (c) $n = 1000$; (d) $n = 10{,}000$

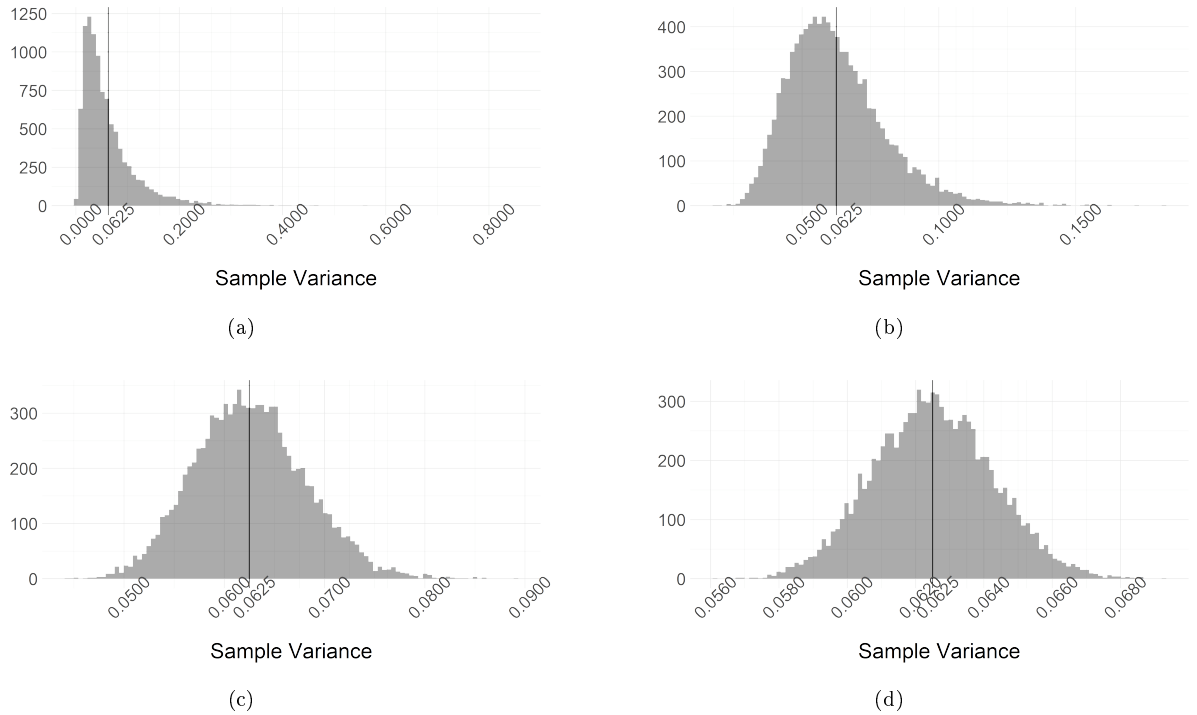

Figure 4.2: Histograms of the distribution of the sample variances for the Monte Carlo simulations with sample size n (true variance is indicated by a vertical bar): (a) $n = 10$; (b) $n = 100$; (c) $n = 1000$; (d) $n = 10{,}000$

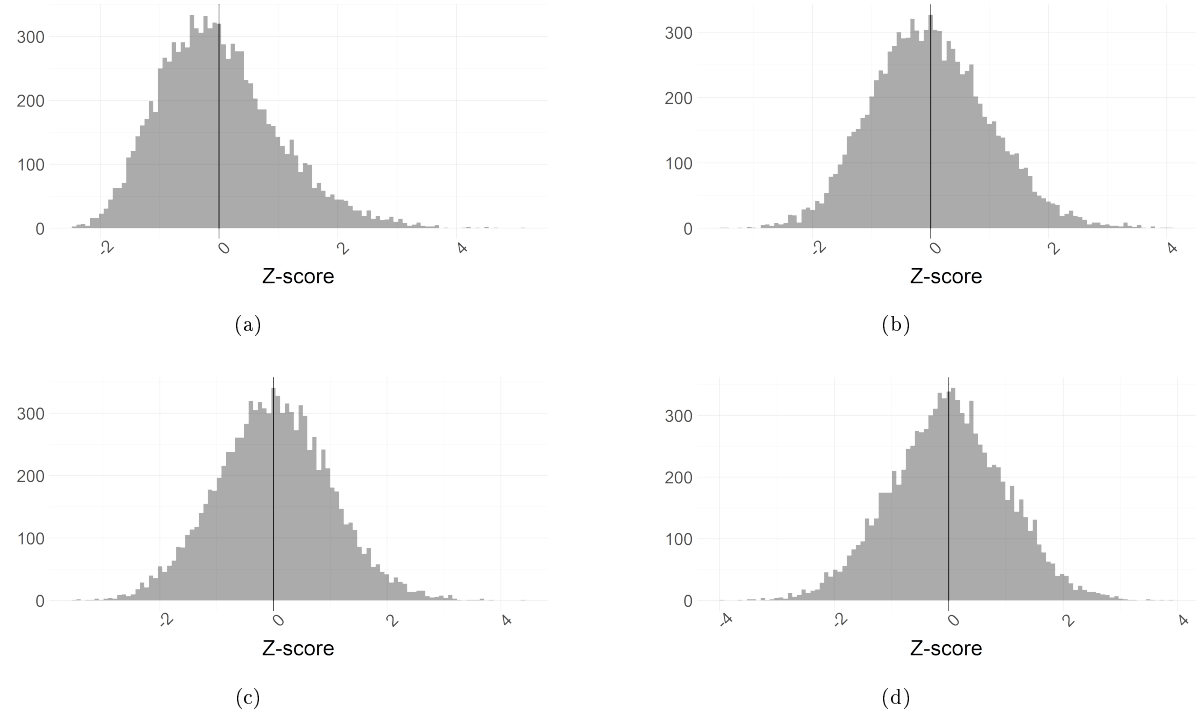

Figure 4.3: Histograms of the distribution of the Z-scores for the Monte Carlo simulations with sample size n: (a) $n = 10$; (b) $n = 100$; (c) $n = 1000$; (d) $n = 10{,}000$

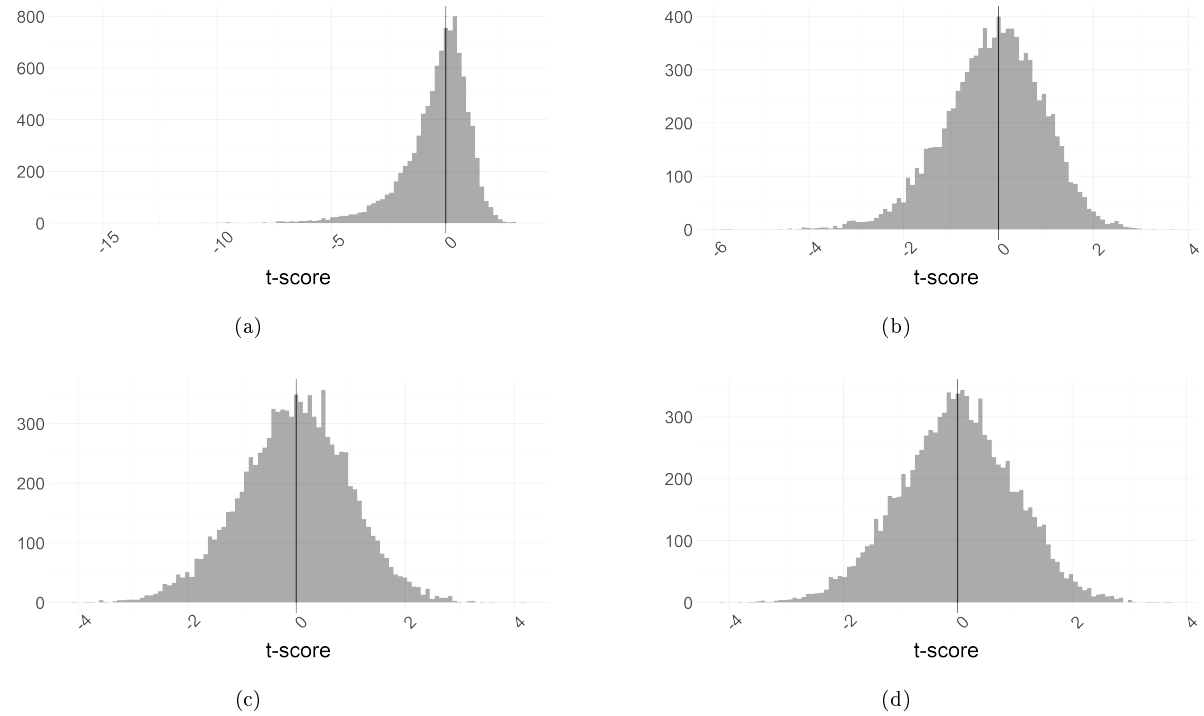

Figure 4.4: Histograms of the distribution of the t-scores for the Monte Carlo simulations with sample size n: (a) $n = 10$; (b) $n = 100$; (c) $n = 1000$; (d) $n = 10{,}000$

```
exp.mean <- 1/lambda
exp.variance <- 1/lambda^2
for (i in 1:T){
  # generate n iid exponential random variables
  sample <- rexp(n, lambda)

  # compute statistics
  sample.mean[i] <- mean(sample)
  sample.variance[i] <- var(sample)
  t.score[i] <- (sample.mean[i] - exp.mean) / (sqrt(sample.variance[i]/n))
  Z.score[i] <- (sample.mean[i] - exp.mean) / (sqrt(exp.variance/n))
}

# create a dataframe out of all these values
df <- data.frame(
  mean = sample.mean,
  variance = sample.variance,
  t = t.score,
  Z = Z.score
)

# plot histograms
ggplot(df, aes(x = mean)) +
  geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
  geom_vline(xintercept = 1/lambda) +
  labs(x = "Sample Mean", y = "") +
  scale_x_continuous(breaks = c(pretty(df$mean), 1/lambda)) +
  theme_minimal() +
  theme(
    text = element_text(size = 25),
    plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
    axis.text.x = element_text(angle = 45)
  )
ggsave(
  paste0("hist_limit_mean_", n, ".png"),
  width = 3000, height = 1500, units = "px"
)

ggplot(df, aes(x = variance)) +
  geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
  geom_vline(xintercept = 1/lambda^2) +
  labs(x = "Sample Variance", y = "") +
  scale_x_continuous(breaks = c(pretty(df$variance), 1/lambda^2)) +
  theme_minimal() +
  theme(
    text = element_text(size = 25),
    plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
    axis.text.x = element_text(angle = 45)
  )
ggsave(
  paste0("hist_limit_variance_", n, ".png"),
  width = 3000, height = 1500, units = "px"
)

ggplot(df, aes(x = t)) +
  geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
  geom_vline(xintercept = 0) +
  labs(x = "t-score", y = "") +
  scale_x_continuous(breaks = c(pretty(df$t), 0)) +
  theme_minimal() +
  theme(
    text = element_text(size = 25),
    plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
    axis.text.x = element_text(angle = 45)
  )
ggsave(
  paste0("hist_limit_t_", n, ".png"),
```

```
    width = 3000, height = 1500, units = "px"
  )

  ggplot(df, aes(x = Z)) +
    geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
    geom_vline(xintercept = 0) +
    labs(x = "Z-score", y = "") +
    scale_x_continuous(breaks = c(pretty(df$Z), 0)) +
    theme_minimal() +
    theme(
      text = element_text(size = 25),
      plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
      axis.text.x = element_text(angle = 45)
    )
  ggsave(
    paste0("hist_limit_Z_", n, ".png"),
    width = 3000, height = 1500, units = "px"
  )
}

# set random number generator seed for reproducibility
set.seed(100)

# set working directory to the current file
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# run the Monte Carlo simulations for different sample sizes
monte.carlo(n = 10, T = 10000, lambda = 4)
monte.carlo(n = 100, T = 10000, lambda = 4)
monte.carlo(n = 1000, T = 10000, lambda = 4)
monte.carlo(n = 10000, T = 10000, lambda = 4)
```