# Chapter 1: Estimators (Pages 4–12)

ETH Zürich Lecture Notes (Translated)
V. Tassion (based on a previous version from M. Schweizer)
D-INFK Spring 2022 (Updated: May 24, 2022)

## Introduction and Basic Ideas

**Basic Ideas:** One has observed data and wishes to draw inferences about the underlying mechanism generating these data.

A common first step is to present the data graphically. This is often useful to form an initial impression and to generate ideas. (Such methods belong to descriptive statistics, which are not treated here.)

We focus on *inductive statistics* in the following. The basic idea is simple: The data

$$x_1, x_2, \ldots, x_n$$

are viewed as realisations of random variables

$$X_1, X_2, \ldots, X_n,$$

and (under suitable assumptions) one seeks statements about the distribution of $X_1, \ldots, X_n$.

**Important:** Always clearly distinguish between the data

$$x_1, \ldots, x_n \quad \text{(lowercase, usually numbers)}$$

and the generating mechanism

$$X_1, \ldots, X_n \quad \text{(uppercase, i.e., random variables on } \Omega\text{)}.$$

As William James (1842–1910) remarked, "We must be careful not to confuse data with the abstractions we use to analyze them."

The collection of observations (or random variables) is called a *sample*, and the number $n$ is the *sample size*.

A typical analysis proceeds by finding a suitable model for the data. The model is described by a (possibly high-dimensional) parameter $\theta \in \Theta$, and to use the concepts and

notation precisely, one must specify exactly how probability-theoretic statements depend on $\theta$. Usually, one considers a family of probability spaces; that is, one has a fixed base space $(\Omega, \mathcal{F})$ and for each $\theta \in \Theta$ a probability measure $\mathbb{P}_\theta$. One may imagine that 'nature' chooses a parameter $\theta \in \Theta$ along with a stochastic mechanism $\mathbb{P}_\theta$. As statisticians we do not know which $\theta$ has been chosen; we thus treat the data $x_1, \ldots, x_n$ as outcomes of the random variables $X_1, \ldots, X_n$ under the unknown mechanism $\mathbb{P}_\theta$, and we attempt to draw conclusions about $\theta$.

A typical parametric statistical analysis comprises the following steps:

1. **Descriptive Statistics:** Graphical methods are used to form an initial idea for choosing an appropriate model. (We do not discuss this step further.)

2. **Choice of a (Parametric) Model:** Specify the parameter space $\Theta$ and the family $(\mathbb{P}_\theta)_{\theta\in\Theta}$ of models.

3. **Parameter Estimation:** Based on the data, choose the best-fitting model. For this, an *estimator* is used—a function mapping the data $x_1, \ldots, x_n$ to a parameter value $\theta \in \Theta$.

4. **Goodness-of-Fit Testing:** Test whether the chosen parameter $\theta$ or model $\mathbb{P}_\theta$ fits the data well using an appropriate statistical test.

5. **Assessing Reliability of the Estimates:** Instead of a single parameter value, one may specify a region in $\Theta$ (a confidence interval) such that, with a certain probability, all the models within that region are in agreement with the data.

# 1 Basic Concepts of Estimation

We wish to estimate an unknown parameter $\theta$ based on a sample

$$X_1, X_2, \ldots, X_n.$$

## 1.1 Definition of an Estimator

**Definition 1.1** (Estimator)**.** *An* estimator *is a random variable*

$$T : \Omega \longrightarrow \mathbb{R}$$

*of the form*

$$T = t(X_1, X_2, \ldots, X_n),$$

*where $t : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a measurable function. Inserting the observed data*

$$x_1, x_2, \ldots, x_n \quad (\textit{with } x_i = X_i(\omega))$$

*yields the* estimate

$$t(x_1, \ldots, x_n)$$

*for* $\theta$.

**Remark 1.1.** *It is essential to distinguish between the* estimator (*a random variable*) *and the* estimate (*the realised value when data are plugged in*).

**Example 1.1** (Tea Tasting Lady)**.** *An English lady claims that when drinking tea with milk she can tell by taste whether milk or tea was poured first into the cup. To test her claim, we ask her over* $n$ *days to classify two cups (one of type 1 and one of type 2) by indicating in which cup the milk was poured first. We record the outcomes as*

$$x_1, x_2, \ldots, x_n \in \{0, 1\} \quad \text{(with 1 indicating a correct and 0 an incorrect classification).}$$

*These data are treated as realisations of the random variables*

$$X_1, X_2, \ldots, X_n.$$

*Let*

$$S_n = \sum_{i=1}^{n} X_i,$$

*be the random number of correct classifications, and denote the observed sum by* $s_n$*. Assuming the* $X_i$ *are i.i.d.* $\sim \mathrm{Ber}(\theta)$ *with unknown success probability* $\theta \in [0, 1]$*, we have*

$$S_n \sim \mathrm{Bin}(n, \theta).$$

*Natural choices for estimators are:*

1. ***Last Observation Estimator:*** $T^{(1)} = X_n$.
2. ***Sample Mean Estimator:*** $T^{(2)} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$.

*For the observed data, these yield the estimates:*

$$t^{(1)}(x_1, \ldots, x_n) = x_n \quad \textit{and} \quad t^{(2)}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

## 1.2 Bias and Mean Squared Error

The estimator $T$ is a random variable whose distribution (under $\mathbb{P}_\theta$) depends on the unknown parameter $\theta$.

**Definition 1.2** (Unbiased Estimator)**.** *An estimator* $T$ *is called* unbiased *for* $\theta$ *if, for all* $\theta \in \Theta$*,*

$$\mathbb{E}_\theta[T] = \theta.$$

**Definition 1.3** (Bias and MSE). *For $\theta \in \Theta$, the* bias *of an estimator $T$ is defined as*

$$\text{Bias}_\theta(T) = \mathbb{E}_\theta[T] - \theta.$$

*The* mean squared error *(MSE) of $T$ is defined as*

$$\text{MSE}_\theta(T) = \mathbb{E}_\theta\big[(T-\theta)^2\big].$$

*In fact,*

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) + \Big(\text{Bias}_\theta(T)\Big)^2.$$

*For an unbiased estimator, the MSE equals the variance.*

**Example 1.2** (Tea Tasting Lady (Continued)). *Both estimators $T^{(1)} = X_n$ and $T^{(2)} = \frac{1}{n}\sum_{i=1}^n X_i$ are unbiased:*

$$\mathbb{E}_\theta\big[T^{(1)}\big] = \mathbb{E}_\theta[X_n] = \theta,$$

*and*

$$\mathbb{E}_\theta\big[T^{(2)}\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] = \theta.$$

*However, their variances are:*

$$\text{Var}_\theta\big[T^{(1)}\big] = \theta(1-\theta),$$

$$\text{Var}_\theta\big[T^{(2)}\big] = \frac{1}{n}\theta(1-\theta).$$

*Thus, the sample mean $T^{(2)}$ has a smaller variance since it uses the full sample information.*

$\diamondsuit$

## 1.3 Maximum Likelihood Estimation (MLE)

In this section, we introduce a systematic method for determining estimators. This method yields results that are both plausible and have good properties.

Assume we know the joint distribution of $X_1, \ldots, X_n$ under $\mathbb{P}_\theta$. For i.i.d. data, the joint probability (or density) is given by the product of the individual probabilities (or densities). For a given sample $(x_1, \ldots, x_n)$, the function

$$L(x_1, \ldots, x_n; \theta)$$

is called the *likelihood function* for $\theta$.

**Definition 1.4** (Likelihood Function). *For the observed sample $(x_1, \ldots, x_n)$, the likelihood*

*function is defined by*

$$L(x_1, \ldots, x_n; \theta) = \begin{cases} \prod_{i=1}^{n} p_{X_i}(x_i; \theta), & \text{if } X_i \text{ are discrete,} \\ \prod_{i=1}^{n} f_{X_i}(x_i; \theta), & \text{if } X_i \text{ are continuous.} \end{cases}$$

**Definition 1.5** (Maximum Likelihood Estimator (MLE))**.** *The* maximum likelihood estimator *of $\theta$ is defined as*

$$\hat{\theta}_{\mathrm{ML}}(x_1, \ldots, x_n) \in \arg\max_{\theta \in \Theta} L(x_1, \ldots, x_n; \theta).$$

*In practice, one maximises the log-likelihood function,*

$$\ell(\theta; x_1, \ldots, x_n) = \log L(x_1, \ldots, x_n; \theta),$$

*and then obtains the estimator by replacing the data with the random variables:*

$$T_{\mathrm{ML}} = t_{\mathrm{ML}}(X_1, \ldots, X_n).$$

**Example 1.3** (Bernoulli Distribution)**.** *Let $X_1, \ldots, X_n$ be i.i.d. $\sim \mathrm{Ber}(p)$ with unknown $p \in (0, 1)$. The probability mass function is*

$$p_X(x; p) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}.$$

*Hence, the likelihood function is*

$$L(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

*Taking the logarithm,*

$$\ell(p; x_1, \ldots, x_n) = \left( \sum_{i=1}^{n} x_i \right) \log p + \left( n - \sum_{i=1}^{n} x_i \right) \log(1-p).$$

*Differentiating with respect to $p$:*

$$\frac{\mathrm{d}}{\mathrm{d}p} \ell(p; x_1, \ldots, x_n) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0.$$

*Solving,*

$$\frac{\sum x_i}{p} = \frac{n - \sum x_i}{1-p} \quad \Longrightarrow \quad \sum_{i=1}^{n} x_i (1-p) = \left( n - \sum_{i=1}^{n} x_i \right) p.$$

*Thus,*

$$\sum_{i=1}^{n} x_i = np \quad \Longrightarrow \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

*Therefore, the MLE for $p$ is given by*

$$T_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

$\diamondsuit$

## 1.4 Models with Multiple Parameters

Thus far, our discussion has focused on models with a single parameter $\theta \in \mathbb{R}$. However, many situations require models with multiple parameters $\theta_1, \theta_2, \ldots, \theta_m$, where $m \geq 2$. We now develop a general theory for such cases.

Consider the parameter space

$$\Theta \subset \mathbb{R}^m,$$

where $m$ is the number of parameters. The stochastic model is given by a family of probability measures $(P_\theta)_{\theta\in\Theta}$, and our goal is to estimate the vector

$$\theta = (\theta_1, \theta_2, \ldots, \theta_m).$$

All previous definitions (of estimator, bias, MSE, MLE, etc.) extend naturally to this setting.

**Example 1.4** (Normal Distribution)**.** *Let $X_1, \ldots, X_n$ be i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Here, the unknown parameter is*

$$\theta = (\mu, \sigma^2),$$

*so that $m = 2$. We wish to estimate both $\mu$ and $\sigma^2$.*
*The density function for $X_i$ is*

$$f_{X_i}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

*Thus, the likelihood function is:*

$$L(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big).$$

*Taking the logarithm,*

$$\log L(x_1, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

***Differentiation with Respect to $\mu$:***

*Differentiate with respect to $\mu$:*

$$\frac{\partial}{\partial \mu} \log L(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0.$$

*Hence,*

$$\hat{\mu} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

***Differentiation with Respect to $\sigma^2$:***

*Differentiate with respect to $\sigma^2$:*

$$\frac{\partial}{\partial \sigma^2} \log L(x_1, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = 0.$$

*Solving for $\sigma^2$ gives:*

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

*By expanding the square, we may also write:*

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x}_n)^2.$$

**Remark 1.2.** *The maximum likelihood estimator $\hat{\sigma}^2$ is not unbiased since*

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2 < \sigma^2.$$

*A commonly used unbiased estimator for $\sigma^2$ is*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

*Furthermore, the estimator*

$$T = (T_1, T_2)$$

*with*

$$T_1 = \hat{\mu} = \bar{X}_n \quad \textit{and} \quad T_2 = \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2,$$

*is, in general, also the so-called* moment estimator *for* $(\mathbb{E}[X], \mathrm{Var}[X])$ *in any model $P_\theta$ where $X_1, \ldots, X_n$ are i.i.d. However, this estimator has the general drawback that it is*

*not unbiased for* $(\mathbb{E}[X], \mathrm{Var}[X])$. *In fact, while*

$$\mathbb{E}_\theta[T_1] = \mathbb{E}_\theta[X_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[X_i] = \mathbb{E}_\theta[X],$$

*we obtain*

$$\mathbb{E}_\theta[(X_n)^2] = \frac{1}{n^2}\sum_{i,k=1}^{n} \mathbb{E}_\theta[X_i X_k] = \frac{1}{n^2}\left( n\sum_{i=1}^{n} \mathbb{E}_\theta[X_i^2] + \sum_{i\neq k} \mathbb{E}_\theta[X_i]\mathbb{E}_\theta[X_k] \right),$$

*and, due to independence (for* $i \neq k$*),*

$$\mathbb{E}_\theta[X_i X_k] = \mathbb{E}_\theta[X_i]\mathbb{E}_\theta[X_k] = \big(\mathbb{E}_\theta[X]\big)^2.$$

*Thus, one can show that*

$$\mathbb{E}_\theta[T_2] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[X_i^2] - \mathbb{E}_\theta[(X_n)^2] = \frac{n-1}{n}\,\mathrm{Var}_\theta[X].$$

*To obtain an unbiased estimator for* $(\mathbb{E}[X], \mathrm{Var}[X])$*, one typically uses*

$$T_1' = T_1 = \bar{X}_n, \quad T_2' = \frac{n}{n-1} T_2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

*The estimator* $T_2'$ *is often denoted by* $S^2$ *and is called the* empirical sample variance.

♢

# 2 Confidence Intervals

## 2.1 Definition

In the preceding chapter we introduced methods for estimating unknown parameters using formulas. A natural question is: How reliable are these estimators? For example, suppose we toss a coin $n$ times without knowing the probability $p$ of heads. If we observe, say, 70 heads, then the maximum likelihood estimator is $T_{\mathrm{ML}} = 0.7$. But how far can $T_{\mathrm{ML}}$ deviate from the true value of $p$? To answer such questions, we introduce the concept of a confidence interval.

**Definition 2.1** (Confidence Interval)**.** *Let* $\alpha \in [0, 1]$*. A* confidence interval *for* $\theta$ *with confidence level* $1 - \alpha$ *is a random interval*

$$I = [A, B],$$

*with endpoints*

$$A = a(X_1, \ldots, X_n), \quad B = b(X_1, \ldots, X_n),$$

*where* $a$, $b : \mathbb{R}^n \to \mathbb{R}$, *such that for all* $\theta \in \Theta$

$$P_\theta[A \leq \theta \leq B] \geq 1 - \alpha.$$

**Remark 2.1.** *In the definition above, the parameter* $\theta$ *is nonrandom (fixed but unknown), while the endpoints* $A$ *and* $B$ *are random variables (functions of the data).*

**Example 2.1** (Confidence Interval for a Normal Model with Known Variance)**.** *Assume we have i.i.d. random variables*

$$X_1, \ldots, X_n \sim N(m, 1),$$

*i.e. a normal model with known variance* $\sigma^2 = 1$ *but with unknown mean* $m$*. One may show that the maximum likelihood estimator is the sample mean*

$$T = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

*We now seek a confidence interval for* $m$ *of the form*

$$I = \Big[T - \frac{c}{\sqrt{n}}, T + \frac{c}{\sqrt{n}}\Big],$$

*where* $c > 0$ *is a constant independent of* $n$*. Note that*

$$P_\theta\Big[T - \frac{c}{\sqrt{n}} \leq m \leq T + \frac{c}{\sqrt{n}}\Big] = P_\theta\Big[-c \leq \sqrt{n}(T - m) \leq c\Big].$$

*Since*

$$Z = \sqrt{n}(T - m) \sim N(0, 1),$$

*it follows that*

$$P_\theta[-c \leq Z \leq c] = 2\Phi(c) - 1.$$

*Consulting the standard normal table shows that* $2\Phi(1.96) - 1 \geq 0.95$*, so by choosing* $c = 1.96$ *we obtain a 95%-confidence interval:*

$$I = \Big[T - \frac{1.96}{\sqrt{n}}, T + \frac{1.96}{\sqrt{n}}\Big].$$

**What does this mean exactly?**

Imagine that we perform $n$ measurements of a physical quantity. For example, suppose we wish to determine, at room temperature, the temperature at which water begins to boil. The characteristics of the thermometer suggest that each measurement can be modeled

by a normally distributed random variable $N(m, 1)$, where $m$ is the (unknown) boiling temperature. We perform a series of measurements—say, $x_1 = 99.2$, $x_2 = 98.7$, .... After $n = 100$ successive trials, we calculate the empirical average

$$\hat{m}(x) = \frac{x_1 + \cdots + x_n}{n} = 99.106.$$

The confidence interval obtained above thus indicates that, assuming the stochastic model is correct, the true value $m$ lies (with 95% probability) in the interval

$$\Big[99.106 - 0.196,\ 99.106 + 0.196\Big] = [98.910,\ 99.302].$$

**What are the key points?**
In the above example the most important observation is that the random variable

$$Z = \sqrt{n}(T - m)$$

is normally distributed for every value of the unknown parameter $\theta$.
In general, one may attempt to obtain a confidence interval for a parameter $\theta$ by first determining an estimator $T$ for $\theta$. Next, one seeks to find a random variable of the form

$$Z = f(T, \theta)$$

whose distribution can be explicitly determined and that does not depend on $\theta$. This is generally easier when the random variables $X_1, \ldots, X_n$ are normally distributed since operations on normally distributed random variables are well understood—for instance, we have used above that the sum of independent normally distributed random variables is itself normally distributed.

In the next section we will introduce new distributions that arise from operations on normally distributed random variables.

## 2.2 Distribution Statements

In many situations it is useful or necessary to know the distribution of an estimator (or a function thereof) under $P_\theta$, for every $\theta \in \Theta$ or for certain values of $\theta$. There are only a few exact general results; for the normal distribution, precise results are available.

**Definition 2.2** (Chi-Squared Distribution)**.** *A continuous random variable $X$ is said to be* chi-squared distributed *with $m$ degrees of freedom if its density is given by*

$$f_X(y) = \frac{1}{2^{m/2}\Gamma(m/2)}\, y^{\frac{m}{2}-1} e^{-y/2}, \quad y \geq 0.$$

*We write*

$$X \sim \chi^2_m.$$

*Here the Gamma function is defined as*

$$\Gamma(v) = \int_0^\infty t^{v-1} e^{-t}\, dt,$$

*and for* $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$.

**Remark 2.2.** *The chi-squared distribution with* $m$ *degrees of freedom is a special case of the Gamma*$(\alpha, \lambda)$*-distribution with* $\alpha = \frac{m}{2}$ *and* $\lambda = \frac{1}{2}$. *For* $m = 2$, *it corresponds to an exponential distribution with parameter* $\frac{1}{2}$.

**Theorem 2.1** (Sum of Squares Theorem)**.** *If* $X_1, X_2, \ldots, X_m$ *are i.i.d.* $\sim N(0,1)$, *then*

$$Y = \sum_{i=1}^{m} X_i^2 \sim \chi_m^2.$$

**Definition 2.3** (t-Distribution)**.** *A continuous random variable* $X$ *is said to be* t-distributed *with* $m$ *degrees of freedom if its density is given by*

$$f_X(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad x \in \mathbb{R}.$$

*We write*

$$X \sim t_m.$$

**Remark 2.3.** *For* $m = 1$ *the t-distribution is equivalent to a Cauchy distribution, and as* $m \to \infty$ *it converges asymptotically to the standard normal distribution* $N(0,1)$. *Like the* $N(0,1)$ *distribution, the t-distribution is symmetric about 0; however, it is heavy-tailed (i.e. its density decays more slowly to 0 as* $|x| \to \infty$*), and this effect is more pronounced the smaller* $m$ *is.*

**Theorem 2.2** (Satz 2.6)**.** *Let* $X$ *and* $Y$ *be independent random variables with*

$$X \sim N(0,1) \quad \textit{and} \quad Y \sim \chi_m^2.$$

*Then the quotient*

$$Z := \frac{X}{\sqrt{Y/m}}$$

*is t-distributed with* $m$ *degrees of freedom.*

## 2.3 Normal Model with Unknown Variance and Mean

For normally distributed samples, we have exact distributional statements that we will also use in hypothesis testing later. Recall the definitions of the sample mean and the

sample variance:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

**Theorem 2.3.** *If $X_1, \ldots, X_n$ are i.i.d. $\sim N(m, \sigma^2)$, then $\bar{X}_n$ and $S^2$ are independent.*

**Remark 2.4.** *(Proof: See Rice, Section 6.3.)*

**Example 2.2** (Ostrich Eggs)**.** *Mr. Smith and Dr. Thurston, two Australian researchers, are debating the average weight of ostrich eggs. They agree that the weights can be approximately modeled as normally distributed. Mr. Smith claims that the mean weight is 1100 g, while Dr. Thurston contends that it is 1200 g. To settle their dispute, they collect $n = 8$ eggs with the following weights (in grams):*

$$1090,\ 1150,\ 1170,\ 1080,\ 1210,\ 1230,\ 1180,\ 1140.$$

*These data are viewed as realizations of i.i.d. random variables $X_1, \ldots, X_8 \sim N(m, \sigma^2)$. The natural estimators for $m$ and $\sigma^2$ are the sample mean*

$$\bar{X}_8 = \frac{1}{8}\sum_{i=1}^{8} X_i,$$

*and the sample variance*

$$S^2 = \frac{1}{7}\sum_{i=1}^{8}(X_i - \bar{X}_8)^2.$$

*A confidence interval for $m$ is constructed in the form*

$$C(X_1, \ldots, X_8) = \Big[\bar{X}_8 - t_{7,1-\alpha/2}\frac{S}{\sqrt{8}},\ \bar{X}_8 + t_{7,1-\alpha/2}\frac{S}{\sqrt{8}}\Big],$$

*where $t_{7,1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the $t$-distribution with 7 degrees of freedom. For example, if $1 - \alpha = 99\%$, and the observed values are*

$$\bar{x}_8 = 1156.25, \quad s = 52.90, \quad t_{7,0.995} = 3.499,$$

*the resulting confidence interval for $m$ is approximately*

$$[1090.81,\ 1221.69].$$

*Thus, both the claims of 1100 g and 1200 g are plausible with these data.*
*To construct a confidence interval for $\sigma^2$, we use the fact that*

$$\frac{(8-1)S^2}{\sigma^2} \sim \chi^2_7.$$

*If $\chi^2_{7,\gamma}$ denotes the $\gamma$-quantile of a $\chi^2_7$ distribution, then a $1-\alpha$ confidence interval for $\sigma^2$*

*is given by*

$$C(X_1, \ldots, X_8) = \left[ \frac{7S^2}{\chi^2_{7,1-\alpha/2}},\ \frac{7S^2}{\chi^2_{7,\alpha/2}} \right].$$

*For instance, with* $1 - \alpha = 95\%$*, if*

$$s^2 = 2798.21, \quad \chi^2_{7,0.025} = 1.69, \quad \chi^2_{7,0.975} = 16.01,$$

*one obtains the confidence interval for* $\sigma^2$*, and by taking square roots, the interval for* $\sigma$ *is approximately*

$$[34.98,\ 107.66].$$

**Remark 2.5.** *In this example, the confidence intervals are exact because precise distributional results are available. In many other situations, one can only obtain approximate confidence intervals using the central limit theorem.*

## 2.4 Approximate Confidence Intervals

A general approximate approach is provided by the central limit theorem (CLT). Often, an estimator $T$ is a function of a sum, say,

$$T = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

By the CLT, for large $n$,

$$\sum_{i=1}^{n} Y_i \approx N\Big(n\,\mathbb{E}[Y_i],\ n\ \mathrm{Var}[Y_i]\Big),$$

which can be used to approximate the distribution of $T$ and, hence, to construct approximate confidence intervals.

**Example 2.3** (Tea Tasting Lady (Approximate Confidence Intervals))**.** *Suppose that in* $n = 10$ *trials the tea tasting lady correctly classifies 6 pairs (i.e.,* $s = 6$*). In any model* $P_\theta$*, the number of successes* $S_{10}$ *is* $\mathrm{Bin}(10, \theta)$*. We wish to obtain a confidence interval for the unknown parameter* $\theta$*.*
*By the central limit theorem, for large* $n$*,*

$$S_{10} \approx N\Big(10\,\theta,\ 10\,\theta(1-\theta)\Big).$$

*Define the standardized statistic*

$$S^*_{10} = \frac{S_{10} - 10\theta}{\sqrt{10\,\theta(1-\theta)}} \approx N(0,1).$$

*For $\theta = \theta_0 = \frac{1}{2}$, this becomes*

$$T = \frac{S_{10} - 5}{\sqrt{10 \cdot \frac{1}{4}}} = \frac{2S_{10} - 10}{\sqrt{10}} \approx N(0, 1)$$

*under $P_{\theta_0}$.*
*To perform a two-sided test at a given level $\alpha$, choose the critical region*

$$K = (-\infty, -c) \cup (c, \infty),$$

*so that*

$$P_{\theta_0}(|T| > c) \approx \alpha.$$

*Since the distribution of $T$ under $P_{\theta_0}$ is symmetric, we have*

$$\alpha \approx 2\Big(1 - \Phi(c)\Big),$$

*and hence,*

$$c \approx \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big).$$

*For example, if $\alpha = 0.01$, then*

$$c \approx \Phi^{-1}(0.995) = 2.576.$$

*One can then derive an approximate confidence interval for $\theta$ by solving*

$$|S_{10} - 10\theta| \leq c\sqrt{10\,\theta(1-\theta)}.$$

*There are several methods to solve this inequality:*

***Method 1:*** *Assume $\theta(1-\theta) \approx \frac{1}{4}$ (its maximum value). Then the inequality simplifies to*

$$|S_{10} - 10\theta| \leq \frac{c}{2}\sqrt{10},$$

*yielding the approximate confidence interval*

$$\Big[S_{10} - \frac{c}{2}\sqrt{10},\ S_{10} + \frac{c}{2}\sqrt{10}\Big].$$

***Method 2:*** *Use the exact approximate distribution*

$$S_{10} \sim N\Big(10\theta,\ 10\,\theta(1-\theta)\Big),$$

*and write the approximate confidence interval for $\theta$ as*

$$\theta \approx \frac{S_{10}}{10} \pm \frac{c}{\sqrt{10}}\sqrt{\frac{\theta(1-\theta)}{10}}.$$

*Then substitute $\theta$ by its estimate $\frac{S_{10}}{10}$ to obtain a "double approximate" interval.*

*For instance, with $1-\alpha = 95\%$ (so that $c = 1.96$) and an observed $s = 6$, the estimated $\theta$ is* 0.6. *Using Method 1, one obtains an interval approximately* [0.290, 0.910]*, and Method 2 gives an interval about* [0.296, 0.904]. *(Solving the quadratic equation exactly might yield an interval such as* [0.3127, 0.8318]*.) For larger sample sizes (say, $n = 100$ with $s = 60$), the approximate intervals become much narrower and tend to agree closely across methods.*

# 3 Tests

## 3.1 Null and Alternative Hypotheses

The starting point is, as in the previous section, a sample $X_1, \ldots, X_n$. We again consider a family of probability measures $P_\theta$ with $\theta \in \Theta$ that describes our possible models. (Note that $\theta$ can be unidimensional or multidimensional.) We often have a prior belief about where in $\Theta$ the correct (but unknown) parameter $\theta$ might lie, and we wish to test this belief using the data. The basic problem is to decide between two competing classes of models – namely, the null hypothesis and the alternative hypothesis.
More precisely, one sets

$$\text{Null hypothesis } H_0 : \theta \in \Theta_0,$$

$$\text{Alternative hypothesis } H_A : \theta \in \Theta_A,$$

with $\Theta_0 \cap \Theta_A = \varnothing$. (If no explicit alternative is specified, one takes $\Theta_A = \Theta \setminus \Theta_0$.) When $\Theta_0$ or $\Theta_A$ consists of a single value $\theta_0$ or $\theta_A$, they are called *simple*; otherwise, they are called *composite*.
In explicit terms, the null hypothesis states:

$$H_0 : \text{ "the true (but unknown) parameter } \theta \text{ lies in the set } \Theta_0\text{."}$$

and the alternative hypothesis

$$H_A : \text{ "the true (but unknown) parameter } \theta \text{ lies in the set } \Theta_A\text{."}$$

**Example 3.1** (Tea Tasting Lady)**.** *An English lady claims that when drinking tea with milk she can, by taste alone, distinguish whether the milk or the tea was poured into the cup first. How can one verify whether this claim is true?*

*As in the previous section, we ask the lady over $n$ days to classify two cups (one of type 1 and one of type 2); that is, she is required to state in which cup the milk was poured first. We record the outcomes $x_1, x_2, \ldots, x_n \in \{0, 1\}$ (where 0 indicates an incorrect classification and 1 a correct classification) and, as usual, treat these data as realizations of the random variables $X_1, X_2, \ldots, X_n$. Let*

$$S_n = \sum_{i=1}^{n} X_i$$

*denote the (random) number of correctly classified pairs.*
*As our model, we again assume that the $X_i$ are independent and identically distributed according to a Bernoulli distribution, i.e.,*

$$X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$$

*with parameter $\theta \in \Theta = [0, 1]$. Then, naturally,*

$$S_n \sim \text{Bin}(n, \theta)$$

*under $P_\theta$; in other words, in the model $P_\theta$ (which corresponds to a given $\theta$), the number $S_n$ of successes is binomially distributed with parameters $n$ and $\theta$.*
*As skeptics, we doubt the lady's claimed ability. Therefore, we choose as our (simple) null hypothesis*

$$H_0 : \theta = \tfrac{1}{2},$$

*i.e. $\Theta_0 = \{\frac{1}{2}\}$ ("random guessing—anyone can do that"). The (composite) alternative hypothesis, asserting that the lady possesses special abilities, is then*

$$H_A : \theta > \tfrac{1}{2},$$

*i.e. $\Theta_A = \left(\frac{1}{2}, 1\right]$.*
*To proceed further, we now need to formalize the decision-making process based on the data. (We will return to the details of this example later.)* $\lozenge$

## 3.2 Tests and Decisions

**Definition 3.1** (Test)**.** *A* test *is a pair $(T, K)$, where*

- *$T$ is a statistic of the form $T = t(X_1, \ldots, X_n)$ (the* test statistic*), and*
- *$K \subset \mathbb{R}$ is a (deterministic) set, called the* critical region *(or* rejection region*).*

Given the observed data $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$, a statistical test enables us to systematically accept or reject the null hypothesis $H_0$. We first compute the test statistic

$T(\omega) = t(X_1(\omega), \ldots, X_n(\omega))$ and then follow the decision rule:

$$\text{Reject } H_0 \text{ if } T(\omega) \in K,$$

$$\text{Do not reject } H_0 \text{ if } T(\omega) \notin K.$$

Because $T$ is a random variable, the event $\{T \in K\}$ has a probability which we can consider under each model $P_\theta$.
There are two types of errors:

1. A *Type I error* occurs when the null hypothesis is wrongly rejected even though it is true. For $\theta \in \Theta_0$, this error has probability $P_\theta[T \in K]$.

2. A *Type II error* occurs when the null hypothesis is not rejected (or is accepted) even though it is false. For $\theta \in \Theta_A$, this error has probability $P_\theta[T \notin K] = 1 - P_\theta[T \in K]$.

**Example 3.2** (Tea Tasting Lady)**.** *In the tea tasting lady example, a Type I error occurs if we reject the null hypothesis of random guessing even though it is true (i.e. the lady has no special ability). Conversely, a Type II error occurs if we fail to reject the null hypothesis, thereby missing the lady's special ability when it is indeed present.*

## 3.3 Significance Level and Power

When choosing an appropriate test, minimizing the probability of a Type I error is crucial. A Type I error occurs if we reject $H_0$ (i.e. if $T \in K$) even though $H_0$ is true. We would like our test to have a low probability of a Type I error. To this end, we define the significance level of a test.

**Definition 3.2** (Significance Level)**.** *Let $\alpha \in (0, 1)$. A test $(T, K)$ is said to have significance level $\alpha$ if for all $\theta \in \Theta_0$*

$$P_\theta[T \in K] \leq \alpha.$$

Our second objective is to avoid a Type II error. This leads directly to the definition of the power of a test.

**Definition 3.3** (Power)**.** *The* power *of a test $(T, K)$ is defined as the function*

$$\beta : \Theta_A \to [0, 1], \quad \theta \mapsto \beta(\theta) := P_\theta[T \in K].$$

The primary goal is to minimize the probability of a Type I error (i.e., keep the test's significance level low). Having fixed a level $\alpha$, we design a test with significance level $\alpha$. Our secondary goal is to maximize the power (i.e., minimize the probability of a Type II error, which is $1 - \beta(\theta)$ for $\theta \in \Theta_A$). Notice that this asymmetry means it is inherently more difficult to reject $H_0$ than to fail to reject it. Therefore, a serious test often adopts as the null hypothesis the negation of the statement one actually wishes to prove. If one

can still reject $H_0$ under these stringent conditions, one can be more confident that an effect is truly present.
It follows that the decision of a test depends on how one defines the null and alternative hypotheses. In fact, the same question might lead to different decisions if the roles of $H_0$ and $H_A$ are interchanged. (We will illustrate this with an example later.)
**Important:** The decision in a test is never a proof; it is only an interpretation of how well the data agree with the presumed model. If $T(\omega) \in K$, we reject $H_0$ and thus disbelieve that $\theta \in \Theta_0$, which may (but need not) lead us to believe that $\theta$ is in $\Theta_A$. If $T(\omega) \notin K$, we do not reject $H_0$ and are reinforced in our belief that $\theta \in \Theta_0$. However, we know no more about the true value of $\theta$ than before — a test does not provide a proof.

**Example 3.3.** *Tea Testing Lady (Cont.)*
*Under $P_\theta$ the random variables $X_1, \ldots, X_n$ are again assumed to be i.i.d. distributed as* $\mathrm{Ber}(\theta)$*, and hence the total number of successes*

$$S_n = \sum_{i=1}^{n} X_i$$

*is distributed as* $\mathrm{Bin}(n, \theta)$*. In this example the null and alternative hypotheses are set as*

$$H_0 : \ \theta = \tfrac{1}{2} \quad \textit{and} \quad H_A : \ \theta > \tfrac{1}{2}.$$

*Because one would expect more ones (correct classifications) for $\theta > \frac{1}{2}$ than for $\theta = \frac{1}{2}$, a large value of $S_n$ supports $H_A$. A plausible test is to use the test statistic*

$$T := S_n$$

*and to choose a critical region of the form*

$$K = (c, \infty).$$

*In other words, we reject the hypothesis of random guessing (i.e. $H_0$) if the lady achieves many successes.*
*Since our null hypothesis is $\theta = \frac{1}{2}$ (i.e. "no special ability"), we deliberately set the test up so that even if the lady has some ability, a positive result is required to reject the skeptical null hypothesis.*
*In order to determine the critical value $c$ corresponding to a significance level $\alpha$, we need the probabilities*

$$P_{\frac{1}{2}}[S_n > c]$$

*for $\theta = \frac{1}{2}$, and for the power function we also need*

$$\beta(\theta) = P_\theta[S_n > c]$$

*for $\theta > \frac{1}{2}$. In general, one needs to know the distribution of the test statistic $T$ under every $P_\theta$ (or at least under the null hypothesis). In practice, this is usually not possible exactly; if the distribution under $H_0$ cannot be obtained exactly, an approximation must suffice.*
*Suppose we perform the test over $n = 10$ days. The following table shows the binomial probabilities $P_{\frac{1}{2}}[S_{10} > k]$ for various values of $k$:*

| $\theta$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$ |
|---|---|---|---|---|
| 0.7 | 0.3828 | 0.1493 | 0.0282 | 0 |
| 0.6 | 0.1673 | 0.0464 | 0.0060 | 0 |
| 0.5 | 0.0547 | 0.0107 | 0.0010 | 0 |

*To obtain a significance level $\alpha$ of approximately 5%, we require that*

$$P_{\frac{1}{2}}[S_{10} > c] \leq 0.05.$$

*Choosing $c = 7$ yields $P_{\frac{1}{2}}[S_{10} > 7] = 0.0547$, which is roughly 5%. At this level, we are willing to reject the null hypothesis (of random guessing) when 8 or more successes are observed.*
*The power of the test can also be derived from the table. For example, for the chosen $c = 7$ we have*

$$\beta(0.6) = P_{0.6}[S_{10} > 7] = 0.1673, \quad \textit{and} \quad \beta(0.7) = 0.3828.$$

*Thus, we see that*

$$1 - \beta(\theta) = P_\theta[S_{10} \leq 7]$$

*becomes rather large for $\theta$ in the alternative, indicating that the test has a significant probability of a Type II error (i.e. failing to detect a real ability) when the deviation from 0.5 is weak.*
$\Diamond$

**Remark 3.1** (Remark 3.4)**.** *Because the test statistic $T$ in the above example is discrete, it is generally impossible to achieve exactly the preassigned significance level $\alpha$. That is, one cannot usually find a critical region $K$ such that*

$$P_{\theta_0}[T \in K] = \alpha.$$

*(Indeed, even for a simple null hypothesis $\Theta_0 = \{\theta_0\}$ this is problematic in the discrete case.) A common workaround is to use a* randomized test*: One selects a number $\gamma \in [0, 1]$ such that*

$$\gamma\, P_{\theta_0}[T > c] + (1 - \gamma)\, P_{\theta_0}[T > c + 1] = \alpha,$$

*and then decides as follows: If $T > c$, the null hypothesis is rejected with probability $\gamma$; that*

*is, $H_0$ is rejected if (i) $T > c$ holds, and (ii) an independent $U(0,1)$-distributed random variable takes a value less than or equal to $\gamma$.*
*In the above example, using such a randomized test one can achieve the exact level $\alpha = 5\%$ by choosing $c = 7$ and setting*

$$\gamma = \frac{\alpha - P_{\theta_0}[T > c+1]}{P_{\theta_0}[T > c] - P_{\theta_0}[T > c+1]} \approx 0.893.$$

The above situation, with simple null and alternative hypotheses, is so specific that it rarely occurs in practice. However, the basic idea can be generalized and often leads to good or even optimal tests under less restrictive assumptions. In later sections, examples will illustrate that the resulting tests are often intuitively very plausible.

## 3.4 Construction of Tests

In this section, we explain a systematic approach to test construction which, in many situations, leads to an optimal test. The idea dates back to Neyman and Pearson.
Assume that $\theta_0 \neq \theta_A$ are two fixed numbers. In this section, we assume that both the null hypothesis and the alternative hypothesis are simple, i.e.,

$$H_0 : \theta = \theta_0, \qquad H_A : \theta = \theta_A.$$

Furthermore, we assume that the random variables $X_1, \ldots, X_n$ are either jointly discrete or jointly continuous under both $P_{\theta_0}$ and $P_{\theta_A}$. In particular, the likelihood function $L(x_1, \ldots, x_n; \theta)$ is well-defined for $\theta = \theta_0$ and $\theta = \theta_A$ (see Definition 1.4).

**Definition 3.4** (Likelihood Ratio)**.** *For every $x_1, \ldots, x_n$, define the likelihood ratio*

$$R(x_1, \ldots, x_n) := \frac{L(x_1, \ldots, x_n; \theta_A)}{L(x_1, \ldots, x_n; \theta_0)}.$$

*By convention, if $L(x_1, \ldots, x_n; \theta_0) = 0$, we set $R(x_1, \ldots, x_n) = +\infty$.*

Intuitively, a large ratio indicates that the observations $x_1, \ldots, x_n$ are far more likely under the alternative $P_{\theta_A}$ than under the null $P_{\theta_0}$. Hence, it makes sense to define the test statistic as

$$T := R(X_1, \ldots, X_n),$$

and to choose the critical region as

$$K := (c, \infty)$$

for some constant $c$.

**Definition 3.5** (Likelihood Ratio Test)**.** *Let $c \geq 0$. The* likelihood ratio test *with param-*

*eter $c$ is the test $(T, K)$ where*

$$T = R(X_1, \ldots, X_n) \quad \textit{and} \quad K = (c, \infty).$$

The likelihood ratio test is optimal in the following sense: Any other test with significance level no greater than the level of the likelihood ratio test will have lower power (i.e., a higher probability of a Type II error).

**Theorem 3.1** (Neyman–Pearson Lemma (Theorem 3.7))**.** *Let $c \geq 0$ and let $(T, K)$ be the likelihood ratio test with parameter $c$ and significance level $\alpha^* := P_{\theta_0}[T > c]$. If $(T', K')$ is any other test with significance level $\alpha \leq \alpha^*$, then*

$$P_{\theta_A}[T' \in K'] \leq P_{\theta_A}[T \in K].$$

**Remark 3.2.** *(Proof: See Krengel, Theorem 6.2.)*

The situation above with simple hypotheses is very special; in practice such cases occur rarely. However, the basic idea can be generalized to yield good or even optimal tests under less restrictive conditions. Often, for composite hypotheses the *generalized likelihood ratio*

$$R(x_1, \ldots, x_n) := \frac{\sup_{\theta \in \Theta_A} L(x_1, \ldots, x_n; \theta)}{\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta)}$$

(or an alternative version using the union $\Theta_A \cup \Theta_0$) is used, and one chooses the test statistic as $T := R(X_1, \ldots, X_n)$ with critical region $K = (c_0, \infty)$, where $c_0$ is chosen such that the test has the preassigned significance level.

**Example 3.4** (Tea Testing Lady via the Likelihood Ratio Method)**.** *Assume that in the coin-toss model underlying the tea testing lady experiment, the random variables*

$$X_1, \ldots, X_n$$

*are independent and identically distributed with*

$$X_i \sim \mathrm{Ber}(\theta),$$

*so that the probability mass function is*

$$p_X(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1 - x_i}, \quad \textit{for } x_i \in \{0, 1\}.$$

*Thus, the joint likelihood function for an observed sample $(x_1, \ldots, x_n)$ is given by*

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_X(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}.$$

*We now consider testing the hypothesis that the lady is merely guessing. Choose as the null hypothesis*

$$H_0 : \theta = \tfrac{1}{2},$$

*and as the alternative hypothesis*

$$H_A : \theta > \tfrac{1}{2}\,.$$

*For fixed data $x_1, \ldots, x_n$, define the likelihood ratio as*

$$R(x_1, \ldots, x_n; \theta_0, \theta_A) := \frac{L(x_1, \ldots, x_n; \theta_A)}{L(x_1, \ldots, x_n; \theta_0)}\,.$$

*Since under $H_0$ we have $\theta_0 = \frac{1}{2}$, it follows that*

$$L(x_1, \ldots, x_n; \tfrac{1}{2}) = \left(\frac{1}{2}\right)^n.$$

*Therefore, the likelihood ratio becomes*

$$R(x_1, \ldots, x_n; \tfrac{1}{2}, \theta_A) = \frac{L(x_1, \ldots, x_n; \theta_A)}{\left(\frac{1}{2}\right)^n} = \left(\frac{\theta_A}{\frac{1}{2}}\right)^{\sum_{i=1}^n x_i} \left(\frac{1-\theta_A}{\frac{1}{2}}\right)^{n-\sum_{i=1}^n x_i}.$$

*Because by assumption $\theta_A > \frac{1}{2}$, it follows that the ratio*

$$\frac{\theta_A(1-\theta_0)}{\theta_0(1-\theta_A)} = \frac{\theta_A\,(1-\frac{1}{2})}{(\frac{1}{2})(1-\theta_A)} = \frac{\theta_A}{1-\theta_A} > 1.$$

*Thus, $R(x_1, \ldots, x_n; \frac{1}{2}, \theta_A)$ is large exactly when the exponent $\sum_{i=1}^n x_i$ is large.*
*Instead of working with the full likelihood ratio, we note that it is equivalent (in terms of ordering the data) to use the total number of successes as the test statistic. Hence, we define*

$$T := \sum_{i=1}^n X_i = S_n,$$

*and choose the critical region*

$$K := (c, \infty),$$

*with the constant $c$ chosen to ensure that the test meets the specified significance level under $H_0$.*
*Thus, the Neyman–Pearson approach leads us to reject $H_0$ (i.e. the hypothesis of random guessing) if the observed sum $S_n$ is large. This procedure is exactly equivalent to the test procedure we earlier motivated by plausibility arguments. ◊*

**Example 3.5** (Example: Testing the Mean in a Normal Model with Known Variance)**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed under $P_\theta$ with*

$$X_i \sim N(\mu, \sigma^2)$$

*and with known variance $\sigma^2$; hence, the unknown parameter here is $\theta = \mu \in \mathbb{R}$. The probability density function of $X_i$ under $P_\theta$ is*

$$f_X(x_i;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) = \frac{1}{\sqrt{2\pi v}} \exp\Big(-\frac{(x_i-\mu)^2}{2v}\Big),$$

*where we have set $v = \sigma^2$. Since the $X_i$ are i.i.d., the joint likelihood function is*

$$L(x_1,\dots,x_n;\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big).$$

*The likelihood ratio is then defined as*

$$R(x_1,\dots,x_n;\theta_0,\theta_A) = \frac{L(x_1,\dots,x_n;\theta_A)}{L(x_1,\dots,x_n;\theta_0)}.$$

*In our application we wish to test the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_A : \theta = \theta_A.$$

*Because the likelihood under $H_0$ is*

$$L(x_1,\dots,x_n;\theta_0) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\Big),$$

*the ratio becomes*

$$R(x_1,\dots,x_n;\theta_0,\theta_A) = \exp\left(-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(x_i-\theta_A)^2 - \sum_{i=1}^{n}(x_i-\theta_0)^2\Big]\right).$$

*This can be rewritten (up to a multiplicative constant depending on $\sigma,\theta_0,\theta_A$) as*

$$R(x_1,\dots,x_n;\theta_0,\theta_A) = \textit{const.}(\sigma,\theta_0,\theta_A)\cdot \exp\Big(\frac{\theta_A-\theta_0}{\sigma^2}\sum_{i=1}^{n} x_i\Big).$$

*Thus, the likelihood ratio tends to be large when the exponent*

$$(\theta_A-\theta_0)\sum_{i=1}^{n} x_i$$

*is large. Note that the interpretation of "large" here depends on the sign of $\theta_A-\theta_0$. In any case, we choose as the test statistic*

$$T' := \sum_{i=1}^{n} X_i.$$

*If $\theta_A > \theta_0$ (so that $\theta_A - \theta_0 > 0$), the exponent is large when $T'$ is large; then we choose the critical region of the form*

$$K'_{(>)} := (c'_{(>)}, \infty),$$

*i.e. we reject $H_0$ when $T'$ is large. Conversely, if $\theta_A < \theta_0$, then the exponent is large when $T'$ is small (i.e. negative), and the critical region is of the form*

$$K'_{(<)} := (-\infty, c'_{(<)}).$$

*In both cases, the critical region must be chosen (i.e. the constants $c'_{(>)}$ or $c'_{(<)}$ determined) so that the test attains a preassigned significance level $\alpha$. That is, we wish to have*

$$P_{\theta_0}[T' \in K'] \leq \alpha,$$

*and for that we need the distribution of $T'$ under $P_{\theta_0}$, i.e. under $H_0$.*
*In the present case this is straightforward. Under any $P_\theta$ the $X_i$ are i.i.d. $\sim N(\theta, \sigma^2)$, hence the sum*

$$T' = \sum_{i=1}^{n} X_i \sim N(n\theta, n\sigma^2)$$

*under $P_\theta$. Equivalently, we may define*

$$T = \frac{X_n - \theta}{\sigma/\sqrt{n}} \sim N(0,1)$$

*under $P_\theta$, so that we can use $T$ in place of $T'$. One should note that $T$ is actually computable in the model $P_\theta$ for $\theta \in \Theta_0$, i.e. with $\theta = \theta_0$ (under $H_0$), because by assumption the variance $\sigma^2$ is known and the mean $\theta_0$ to be tested is also known. (The same applies to $T'$; however, the distribution of $T$ under $H_0$ is simpler than that of $T'$.)* $\Diamond$

## 3.5 Examples

In this section we illustrate the considerations above by a few examples. (We largely refrain from detailed derivations and present only the recipes.)

**Example 3.6** (Normal Distribution, Test for the Mean with Known Variance)**.** *This test is known as the* z-test*. Here the random variables*

$$X_1, \ldots, X_n \quad \text{are i.i.d.} \ \sim N(\theta, \sigma^2)$$

*under $P_\theta$ with known variance $\sigma^2$, and we wish to test the hypothesis*

$$H_0 : \quad \theta = \theta_0.$$

*Possible alternatives $H_A$ are either $\theta > \theta_0$ or $\theta < \theta_0$ (one-sided), or $\theta \neq \theta_0$ (two-sided);*

*which alternative is most appropriate depends on the concrete question.*
*In every case the test statistic is (see the last example in Section 3.4)*

$$T := \frac{X_n - \theta_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \textit{under } P_{\theta_0}.$$

*The critical region $K$ takes the form:*

- $K = (c_>, \infty)$ *for a one-sided test against* $H_A : \theta > \theta_0$,
- $K = (-\infty,\, c_<)$ *for a one-sided test against* $H_A : \theta < \theta_0$,
- $K = (-\infty, -c_=) \cup (c_=, \infty)$ *for a two-sided test against* $H_A : \theta \neq \theta_0$.

*For example, the condition*

$$\alpha = P_{\theta_0}[T \in (c_>, \infty)] = P_{\theta_0}[T > c_>] = 1 - P_{\theta_0}[T \leq c_>] = 1 - \Phi(c_>)$$

*implies that*

$$c_> = \Phi^{-1}(1-\alpha) \equiv z_{1-\alpha}.$$

*Thus, for $\theta > \theta_0$ we reject $H_0$ if*

$$X_n > \theta_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

*Analogous reasoning yields $c_< = z_\alpha = -z_{1-\alpha}$ and for the two-sided test $c_= = z_{1-\alpha/2}$.* $\Diamond$

**Example 3.7** (Ostrich Eggs (Known Variance))**.** *The Australians Mr. Smith and Dr. Thurston are still disputing the average weight of ostrich eggs. Both agree that the weights may be modeled as normally distributed; however, Mr. Smith claims that the mean is 1100 g while Dr. Thurston insists that the eggs are heavier (approximately 1200 g on average). To settle their dispute, they travel to Africa to search for ostrich eggs. Because these are usually well hidden, they find only eight eggs with the following weights (in grams):*

$$1090,\ 1150,\ 1170,\ 1080,\ 1210,\ 1230,\ 1180,\ 1140.$$

*Dr. Thurston proposes to test Mr. Smith's claim by taking the hypothesis*

$$H_0 : \quad \mu = \mu_0 = 1100$$

*against the alternative $H_A : \mu > 1100$ (or alternatively, against $\mu = 1200$) at the 5% level. The variance is known; in fact, $\sigma = 55$ g. Dr. Thurston computes the sample mean as*

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 1156.25,$$

*and from a standard normal table he finds* $z_{0.95} = 1.645$*. Hence, the test statistic is*

$$T_{\text{Th}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{8}} \approx \frac{1156.25 - 1100}{55/\sqrt{8}} \approx 2.89.$$

*Since* $T_{\text{Th}} > 1.645$*, the hypothesis* $\mu = 1100$ *is rejected at the 5% level.*
*Mr. Smith, however, feels that this procedure disadvantages him and suggests instead testing Dr. Thurston's claim with the hypothesis*

$$H_0 : \quad \mu = \mu_1 = 1200$$

*against the alternative* $H_A : \mu < 1200$ *(or alternatively,* $\mu = 1100$*). He computes the corresponding test statistic as*

$$T_{\text{Sm}} = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{8}} \approx \frac{1156.25 - 1200}{55/\sqrt{8}} \approx -2.25.$$

*Since* $-2.25 < -1.645$ *(with* $z_{0.05} = -1.645$*), the hypothesis* $\mu = 1200$ *is rejected at the 5% level.* $\Diamond$

**Example 3.8** (Normal Distribution, Test for the Mean with Unknown Variance (t-test))**.** *This test is known as the* t-test*. Here, the observations*

$$X_1, \ldots, X_n \quad \text{are i.i.d.} \sim N(\mu, \sigma^2)$$

*under* $P_\theta$*, but now the variance* $\sigma^2$ *is unknown. In this case the parameter is* $\theta = (\mu, \sigma^2)$ *(with* $\sigma^2$ *unknown). We wish to test*

$$H_0 : \quad \mu = \mu_0.$$

*Strictly speaking, this is a composite hypothesis since* $\sigma^2$ *is unspecified. Explicitly, the null set is*

$$\Theta_0 = \{\mu_0\} \times (0, \infty).$$

*The test statistic is defined as*

$$T := \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \text{under } P_{\theta_0},$$

*where*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

*Depending on whether the alternative is one-sided or two-sided, the critical region is chosen according to the appropriate* $t_{n-1}$ *quantiles.* $\Diamond$

**Example 3.9** (Ostrich Eggs (t-test Version))**.** *Now, Mr. Smith and Dr. Thurston won-*

*der whether, in their first experiment, they might have used an incorrect estimate of the variance of ostrich eggs. Therefore, they decide to perform the tests again without the assumption of known variance—that is, by using a t-test.*
*Dr. Thurston still insists on testing*

$$H_0: \quad \mu = 1100 \quad \textit{against} \quad H_A: \mu > 1100,$$

*at the 5% level. Since the variance is unknown, he calculates the sample variance as*

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 2798.21,$$

*so that* $s = 52.90$. *From a t-distribution with 7 degrees of freedom one finds* $t_{7,0.95} = 1.895$. *Then the test statistic is*

$$\tilde{T}_{\mathrm{Th}} = \frac{\bar{x} - 1100}{s/\sqrt{8}} \approx \frac{1156.25 - 1100}{52.90/\sqrt{8}} \approx 3.008.$$

*Since* $3.008 > 1.895$, *the hypothesis* $H_0: \mu = 1100$ *is rejected at the 5% level.*
*Not surprisingly, Mr. Smith remains unconvinced and suggests instead testing Dr. Thurston's claim with the alternative hypothesis reversed:*

$$H_0: \quad \mu = \mu_1 = 1200 \quad \textit{against} \quad H_A: \mu < 1200 \quad (\textit{or also } \mu = 1100).$$

*He computes the test statistic*

$$\tilde{T}_{\mathrm{Sm}} = \frac{\bar{x} - 1200}{s/\sqrt{8}} \approx \frac{1156.25 - 1200}{52.90/\sqrt{8}} \approx -2.339.$$

*Since for the 5% level we have* $t_{7,0.05} = -t_{7,0.95} = -1.895$ *and* $\tilde{T}_{\mathrm{Sm}} < -1.895$, *the hypothesis* $H_0: \mu = 1200$ *is rejected at the 5% level.* $\Diamond$

**Example 3.10** (Paired Two-Sample Test in a Normal Model)**.** *Suppose that a group of subjects is measured under two different conditions, yielding paired observations* $(X_1, Y_1), \ldots, (X_n, Y_n)$. *Assume that under* $P_\theta$ *the pairs are independent and that*

$$X_i \sim N(\mu_X, \sigma^2) \quad \textit{and} \quad Y_i \sim N(\mu_Y, \sigma^2),$$

*with identical variance* $\sigma^2$. *Such a situation occurs, for example, when the same subjects try two different treatments, yielding a natural pairing.*
*In this case one can reduce the two-sample problem to a one-sample problem by considering the differences*

$$Z_i := X_i - Y_i.$$

*Then the $Z_i$ are i.i.d. with*

$$Z_i \sim N(\mu_X - \mu_Y,\, 2\sigma^2).$$

*Thus, tests for the difference of the means can be performed by applying the one-sample test (using a z-test when $\sigma^2$ is known or a t-test when it is unknown).*
*These tests are known as the* paired two-sample z-test *(if $\sigma^2$ is known) or the* paired two-sample t-test *(if $\sigma^2$ is unknown).* $\Diamond$

**Example 3.11** (Unpaired Two-Sample Test in a Normal Model)**.** *Consider now two independent samples:*

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2) \quad \textit{and} \quad Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2),$$

*where the variance is assumed to be the same in both groups, but the sample sizes $n$ and $m$ may differ. In this unpaired situation, pairwise differences cannot be formed and one must use an unpaired test.*
***(a)*** *If $\sigma^2$ is known, the test statistic is given by*

$$T := \frac{(X_n - Y_m) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$$

*under every $P_\theta$. Here $\sigma$ is known and $\mu_X - \mu_Y$ is presumed known under $H_0$; this is the unpaired two-sample* z-test.
***(b)*** *If $\sigma^2$ is unknown, one first computes the sample variances*

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2.$$

*Then the pooled variance is defined as*

$$S^2 := \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

*The test statistic becomes*

$$T := \frac{(X_n - Y_m) - (\mu_X - \mu_Y)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

*under $P_\theta$. This test is known as the unpaired two-sample* t-test. $\Diamond$

Most of the tests presented above are based on the assumption that the samples are normally distributed; this situation is very convenient because then the distribution of the test statistic is available in explicit form. These tests work very well when the data are indeed normally distributed; however, if this assumption does not hold, they quickly lose a significant portion of their power. Therefore, it is advisable to also be familiar with

alternative tests that rely on less specific assumptions.

## 3.6 The p-value

Let $X_1, \ldots, X_n$ be a sample of size $n$. We wish to test a hypothesis

$$H_0 : \quad \theta = \theta_0$$

against an alternative

$$H_A : \quad \theta \in \Theta_A.$$

**Definition 3.6** (Ordered Family of Tests)**.** *A family of tests $(T, (K_t)_{t\geq 0})$ is said to be ordered with respect to the test statistic $T$ if for all $s, t \geq 0$*

$$s \leq t \quad \Longrightarrow \quad K_s \supset K_t.$$

*Typical examples are:*

$$K_t = (t, \infty) \quad \textit{(right-tailed test)}, \qquad K_t = (-\infty, -t) \quad \textit{(left-tailed test)},$$

*or*

$$K_t = (-\infty, -t) \cup (t, \infty) \quad \textit{(two-sided test)}.$$

**Definition 3.7.** *Let $H_0 : \theta = \theta_0$ be a simple null hypothesis and let $(T, (K_t)_{t\geq 0})$ be an ordered family of tests. The* p-value *is defined as the random variable*

$$\textit{p-value} = G(T),$$

*where the function $G : \mathbb{R}^+ \to [0, 1]$ is given by*

$$G(t) = P_{\theta_0}\big[T \in K_t\big].$$

**Remark 3.3.**

- *The p-value, as a function of the test statistic $T$, is itself a random variable.*
- *The p-value depends directly on the initial observations $X_1, \ldots, X_n$; repeating the test with new data yields a new (random) p-value.*
- *The p-value always lies in the interval $[0, 1]$. In the case where $T$ is continuous and $K_t = (t, \infty)$, it can be shown that under $P_{\theta_0}$ the p-value is uniformly distributed on $[0, 1]$.*

The p-value informs us which tests in our family $\{(T, K_t) : t \geq 0\}$ would lead to rejection of $H_0$. In fact, if the observed p-value is $p$, then every test with significance level $\alpha > p$ would reject $H_0$ and those with $\alpha \leq p$ would not. Notice that the p-value depends solely on the null hypothesis; the alternative hypothesis does not enter its definition.

**Example 3.12** (Coin Toss Example)**.** *Suppose we toss a coin 100 times and observe 60 heads. Our model assumes that*

$$X_1, \dots, X_{100} \text{ are i.i.d. } \sim \mathrm{Be}(\theta)$$

*with* $\theta \in [0,1]$*. Under the null hypothesis we have*

$$H_0: \quad \theta = \tfrac{1}{2},$$

*so that* $\Theta_0 = \{\frac{1}{2}\}$ *and the alternative is*

$$H_A: \quad \theta \neq \tfrac{1}{2},$$

*i.e.,* $\Theta_A = [0,1] \setminus \{\frac{1}{2}\}$*. The number of successes is*

$$S_{100} = \sum_{i=1}^{100} X_i,$$

*and under* $P_\theta$ *we have* $S_{100} \sim Bin(100, \theta)$*.*
*To simplify calculations, we approximate the binomial distribution by a normal distribution (by the central limit theorem). For every* $\theta$ *we have*

$$S_{100} \approx N(100\theta,\, 100\theta(1-\theta)).$$

*Thus, we may define*

$$T' := \frac{S_{100} - 100\theta}{\sqrt{100\theta(1-\theta)}} \approx N(0,1)$$

*under* $P_\theta$*. For* $\theta = \frac{1}{2}$*, this becomes*

$$T = \frac{S_{100} - 50}{\sqrt{25}} = \frac{S_{100} - 50}{5}.$$

*An equivalent form is*

$$T = \frac{2S_{100} - 100}{10} \approx N(0,1) \quad \text{under } P_{\theta_0}.$$

*For a two-sided test we choose the symmetric critical region*

$$K := (-\infty, -c) \cup (c, \infty),$$

*so that* $H_0$ *is rejected if* $|T| > c$*. To have a test of approximate level* $\alpha$*, we require*

$$\alpha = P_{\theta_0}[|T| > c] \approx 2\big(1 - \Phi(c)\big),$$

*which implies*

$$c \approx \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big).$$

*For example, if* $\alpha = 0.01$*, then* $c \approx \Phi^{-1}(0.995) = 2.576$*. In our coin toss example, this translates (after back-transformation) to rejection of* $H_0$ *if*

$$S_{100} > 62.88 \quad \textit{or} \quad S_{100} < 37.12.$$

*Finally, the realized p-value is computed as*

$$\textit{p-value}(\omega) = P_{\theta_0}\big[|T| > t_0\big]\Big|_{t_0 = T(\omega)} \approx 2\big(1 - \Phi(T(\omega))\big).$$

*For instance, if* $T(\omega) = 2$ *then*

$$\textit{p-value}(\omega) \approx 2\big(1 - \Phi(2)\big) = 2\big(1 - 0.97725\big) = 0.0455.$$

*Thus, with 60 successes we would reject the hypothesis of a fair coin at the 5% level, but not at the 1% level.* $\Diamond$

## Summary of Tests

To conclude this section, the general procedure for hypothesis testing is summarized as follows:

1. **Choice of the Model.** Specify the underlying probability model.
2. **Formulation of Hypotheses.** Clearly state the null hypothesis $H_0$ and the alternative hypothesis $H_A$.
3. **Test Statistic and Critical Region.** Determine the test statistic $T$ and the form of the critical region $K$ (this can be derived via a generalized likelihood ratio test or taken from standard statistical literature).
4. **Setting the Significance Level.** Choose the significance level $\alpha$ so that the critical region $K$ satisfies, approximately,

$$\sup_{\theta \in \Theta_0} P_\theta[T \in K] \leq \alpha.$$

5. **Decision Rule.** Compute the observed value $T(\omega)$ from the data. If $T(\omega) \in K$ reject $H_0$, otherwise do not reject $H_0$.
6. **(Optional) p-value.** Alternatively, compute the realized p-value. If it is less than or equal to $\alpha$, reject $H_0$.

**References:**

- Bronstein et al., *Taschenbuch der Mathematik*, 4th ed., Harri Deutsch (1999)
- Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 8th ed., Vieweg (2005)
- Lengler, Steger, and Welzl, *Algorithmen und Wahrscheinlichkeit* (2021)
- Lehn & Wegmann, *Einführung in die Statistik*, 4th ed., Teubner (2004)
- Rice, *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury Press (1995)
- Schweizer, *Wahrscheinlichkeit und Statistik* (2010)
- Stahel, *Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler*, 2nd ed., Vieweg (1999)
- Williams, *Weighing the Odds. A Course in Probability and Statistics*, Cambridge University Press (2001)