# ECE 581: Descriptive and Inferential Statistics

Henry D. Pfister
Duke University

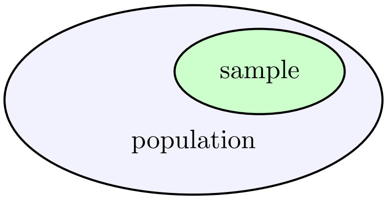
December 5, 2025

## Contents

## 1 Descriptive Statistics: Definitions and Examples

### 1.1 Introduction

**Definition 1.1** (Population, Sample, and Statistic)**.** A *population* is the complete collection of all possible observations of interest. A *sample* is a subset of observations drawn from the population. A *statistic* is a numerical summary computed from sample data and will be defined formally later.



Descriptive statistics help to summarize and visualize data without assuming any probabilistic model. They describe features such as the *center*, *spread*, and *shape* of the sample distribution. For a sample $x_1, x_2, \ldots, x_n \in \mathbb{R}$, consider the following statistics and visualizations.

*Example* 1.2 (Descriptive Summary). Consider the data set

$$\mathcal{D} = \{0, 1, 1, 2, 3, 3, 4, 4, 5, 5, 6, 8, 9, 10, 13\}$$

Then, we have the following statistics (see below)

$$\bar{x} \approx 4.93, \quad S^2 \approx 13.64, \quad Q_1 = 2, \quad Q_2 = 4, \quad Q_3 = 8, \quad \mathrm{IQR} = 6.$$

**Measures of Central Tendency.**

- **Sample Mean:**

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

  It represents the arithmetic average of the data points.

- **Median:** The value separating the higher half from the lower half when the data are sorted.

- **Mode:** The most frequently occurring value in the sample.

- **Trimmed Mean:** The mean computed after removing a fixed proportion (say 5%) of the smallest and largest values to reduce the effect of outliers.

**Measures of Variability.**

- **Range:** $\max(x_i) - \min(x_i)$.

- **Sample Variance:**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

  which estimates the population variance $\sigma^2$.

- **Standard Deviation:** $S = \sqrt{S^2}$.

- **Quartiles:** The first ($Q_1$), second ($Q_2$), and third ($Q_3$) quartiles are the 25th, 50th, and 75th percentiles respectively. The median is $Q_2$.

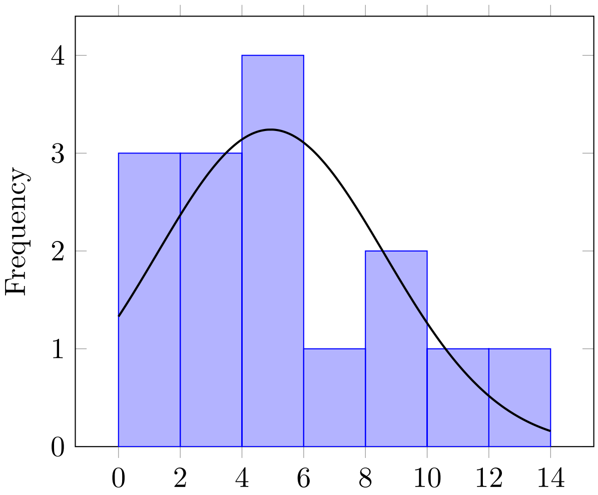
- **Interquartile Range (IQR):** $Q_3 - Q_1$, representing the spread of the middle 50% of data.

### 1.2 Visualization

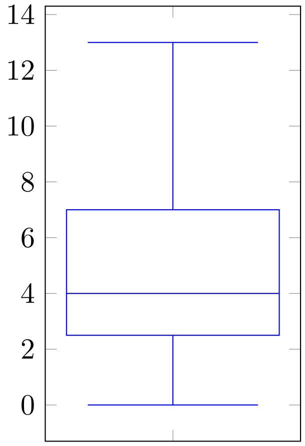
For visualization, we provide a histogram and a box plot. The histogram groups observations into contiguous bins along the horizontal axis and displays their frequencies as bars; with equal-width bins, the area of each bar is proportional to the number of observations in that bin. This reveals the sample's shape (modality, skew, and potential outliers). In the figure, we also overlay a normal curve scaled to the total count so you can compare the empirical shape to a Gaussian reference.

The box plot summarizes the distribution of $\mathcal{D}$ using quartiles and potential outliers. The box spans from the first quartile $Q_1 = 2$ to the third quartile $Q_3 = 8$ with interquartile range $\mathrm{IQR} = 6$, and the line inside the box marks the median $Q_2 = 4$. Whiskers extend to the most extreme points not beyond $1.5\,\mathrm{IQR}$ from the quartiles; any observations beyond the whiskers are flagged as outliers.

Histogram with Normal Fit



Boxplot of Data



## 2 Statistical Models and Sufficiency

A statistical model specifies a family of distributions that could generate the observed data. Here, we use uppercase $X_i$ for random variables and lowercase $x_i$ for their realized values.

**Definition 2.1** (Statistical Model)**.** A *statistical model* is a collection of distributions

$$\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$$

on $\mathcal{X} \subseteq \mathbb{R}^d$, where $\theta$ denotes a vector of unknown parameters and $p$ is a PMF if $\mathcal{X}$ is countable and a PDF if $\mathcal{X} \subseteq \mathbb{R}^d$ is a nice set such as $[a,b]^d$.

**Definition 2.2** (Likelihood Function)**.** Given an i.i.d. sample $x_1, \ldots, x_n$ drawn from $p(x;\theta)$, the *likelihood function* is

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

representing the joint probability mass (in the discrete case) or joint density (in the continuous case) of the observed data as a function of $\theta$. The *log-likelihood* is $\ell(\theta) = \ln L(\theta; x_1, \ldots, x_n)$.

**Definition 2.3** (Statistic)**.** Given a sample $X_1, \ldots, X_n$, a *statistic* is any function $T(X_1, \ldots, X_n)$.

**Definition 2.4** (Sufficient Statistic)**.** A statistic $T(X_1, \ldots, X_n)$ is *sufficient* for parameter $\theta$ if the conditional distribution $p(x_1, \ldots, x_n \mid T = t; \theta)$ does not depend on $\theta$.

**Theorem 2.5** (Factorization Theorem)**.** *A statistic $T(X_1, \ldots, X_n)$ is sufficient for $\theta$ if and only if the joint pdf/pmf can be factored as*

$$p(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n)\, g(T(x_1, \ldots, x_n); \theta),$$

*where $h$ does not depend on $\theta$.*

*Proof.* Assume a dominated model with joint density/pmf $p(x_1, \ldots, x_n; \theta)$. If there exist functions $h$ and $g$ such that $p(x; \theta) = h(x)\, g(T(x); \theta)$, then for any value $t$ in the range of $T$,

$$p(x \mid T = t, \theta) = \frac{p(x; \theta)}{\int_{T(y)=t} p(y; \theta)\, dy} = \frac{h(x)\, g(t; \theta)}{g(t; \theta) \int_{T(y)=t} h(y)\, dy} = \frac{h(x)}{\int_{T(y)=t} h(y)\, dy},$$

which does not depend on $\theta$. Hence $T$ is sufficient.

Conversely, suppose $T$ is sufficient, so that the conditional density $p(x \mid T = t, \theta)$ does not depend on $\theta$. Define

$$g(t; \theta) = \int_{T(y)=t} p(y; \theta)\, dy = \mathbb{P}_\theta(T = t), \qquad h(x) = p(x \mid T = T(x), \theta_0),$$

where $\theta_0$ is any fixed value in $\Theta$. Then, for all $\theta \in \Theta$,

$$p(x; \theta) = p(x \mid T = T(x), \theta)\, \mathbb{P}_\theta(T = T(x)) = h(x)\, g(T(x); \theta)$$

which yields the desired factorization. □

*Example* 2.6 (Sufficiency in Bernoulli and Poisson Families)*.* If $X_i \sim \text{Bernoulli}(p)$, then $T = \sum_i X_i$ is sufficient for $p$. Similarly, if $X_i \sim \text{Poisson}(\lambda)$ independently, then $T = \sum_i X_i$ is sufficient for $\lambda$.

The notion of sufficiency underlies data reduction and is fundamental for defining exponential families and efficient estimators.

*Proof of Sufficiency for Bernoulli Model.* Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$ with joint pmf

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

Let $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$. Then

$$p(x_1, \ldots, x_n; p) = h(x_1, \ldots, x_n)\, g(T(x_1, \ldots, x_n); p),$$

with $h(x_1, \ldots, x_n) = 1$ and $g(t; p) = p^t (1 - p)^{n - t}$, which satisfies the factorization theorem.

Alternatively, we can obtain the factorization by conditioning on $T = t$. For sequences with exactly $t$ ones, symmetry gives

$$\mathbb{P}\big((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \mid T = t\big) = \frac{1}{\binom{n}{t}},$$

which does not depend on $p$. Choosing $h(x) = 1/\binom{n}{T(x_1,\dots,x_n)}$ shows that

$$p(x_1,\dots,x_n;p) = \frac{1}{\binom{n}{T(x_1,\dots,x_n)}} \underbrace{\binom{n}{T(x_1,\dots,x_n)} p^{T(x_1,\dots,x_n)}(1-p)^{n-T(x_1,\dots,x_n)}}_{g(T(x_1,\dots,x_n);p)}.$$

Thus, this choice of $T$ is sufficient for $p$. □

*Proof of Sufficiency for Poisson Model.* Let $X_1,\dots,X_n \sim \text{Poisson}(\lambda)$. Then

$$p(x_1,\dots,x_n;\lambda) = \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = \left(\prod_{i=1}^n \frac{1}{x_i!}\right) e^{-n\lambda}\lambda^{\sum_i x_i}.$$

Define $T = \sum_i X_i$. We can write

$$p(x_1,\dots,x_n;\lambda) = h(x_1,\dots,x_n)g(T;\lambda),$$

where $h(x_1,\dots,x_n) = \prod_i (1/x_i!)$ and $g(T;\lambda) = e^{-n\lambda}\lambda^T$. Thus, by the factorization theorem, $T = \sum_i X_i$ is sufficient for $\lambda$. □

**Definition 2.7** (Minimal Sufficient Statistic)**.** A sufficient statistic $T(X_1,\dots,X_n)$ is *minimal sufficient* if, for any other sufficient statistic $S$, there exists a function $f$ such that $T = f(S)$ almost surely. Minimality captures the maximal reduction of the data without losing information about $\theta$.

**Definition 2.8** (Maximum Likelihood Estimator)**.** Given a statistical model $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ and data $x_1,\dots,x_n$, the *likelihood* $L$ and *log-likelihood* $\ell$ are

$$L(\theta;x_1,\dots,x_n) = \prod_{i=1}^n p(x_i;\theta), \quad \ell(\theta) = \ln L(\theta;x_1,\dots,x_n).$$

Any maximizer

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} L(\theta;x_1,\dots,x_n)$$

is called a *maximum likelihood estimator (MLE).*

# 3 Exponential Families

## 3.1 Definition and Structure

**Definition 3.1** (Exponential Family)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the sample space and $\Theta \subseteq \mathbb{R}^m$ the parameter space. A family of distributions is called an *exponential family* if its pdf/pmf can be expressed as

$$p(x;\theta) = h(x)\exp\Big(\eta(\theta)^\top T(x) - A(\theta)\Big), \text{ where}$$

- $h : \mathcal{X} \to [0,\infty)$ is the *base measure* that does not depend on $\theta$,
- $T : \mathcal{X} \to \mathbb{R}^k$ is the *sufficient statistic*,

- $\eta : \Theta \to \mathbb{R}^k$ is the *natural parameter map* where $\eta(\Theta) \subseteq \Omega$ with *natural parameter space*

$$\Omega = \Big\{ \eta \in \mathbb{R}^k : \int_{\mathcal{X}} h(x) \exp(\eta^\top T(x))\, dx < \infty \Big\},$$

- $A : \Theta \to \mathbb{R}$ is the *log partition function* ensuring normalization over $x$ for all $\theta \in \Theta$ via

$$A(\theta) = \ln \int_{\mathcal{X}} h(x) \exp\big(\eta(\theta)^\top T(x)\big)\, dx.$$

- Since $A(\theta)$ only depends on $\theta$ through $\eta(\theta)$, it is standard to reparameterize by $\eta$ and this gives

$$p(x;\eta) = h(x) \exp\Big(\eta^\top T(x) - A(\eta)\Big), \qquad A(\eta) = \ln \int_{\mathcal{X}} h(x) \exp\big(\eta^\top T(x)\big)\, dx.$$

**Examples.** The exponential family includes many common distributions: Bernoulli, Poisson, Gaussian (with known variance), and Exponential.

(a) **Bernoulli($p$):** $p(x;p) = p^x(1-p)^{1-x}$,

$$\eta(p) = \ln \frac{p}{1-p}, \quad T(x) = x, \quad h(x) = 1, \quad A(p) = \ln\left(1 + \frac{p}{1-p}\right).$$

(b) **Poisson($\lambda$):** $p(x;\lambda) = e^{-\lambda}\lambda^x/x!$, $\eta(\lambda) = \ln\lambda$, $T(x) = x$, $h(x) = 1/x!$, $A(\lambda) = \lambda$.

(c) **Exponential($\lambda$):** $p(x;\lambda) = \lambda e^{-\lambda x}\, \mathbf{1}_{\{x>0\}}$ for $\lambda > 0$, with

$$h(x) = \mathbf{1}_{\{x>0\}}, \quad T(x) = x, \quad \eta(\lambda) = -\lambda, \quad A(\lambda) = -\ln\lambda.$$

(d) **Normal($\theta = \mu$,$\sigma^2$ fixed):** $p(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$, $\eta(\mu) = \mu/\sigma^2$, $T(x) = x$, $A(\mu) = \frac{\mu^2}{2\sigma^2}$.

**Proposition 3.2** (Normalization implies the log-partition formula)**.** *For any $\theta$ such that $\eta(\theta) \in \Omega$ and*

$$p(x;\theta) = h(x) \exp\left(\eta(\theta)^\top T(x) - A(\theta)\right)$$

*is a valid pdf/pmf, the normalizing constant must be*

$$A(\theta) = \ln \int_{\mathcal{X}} h(x) \exp\big(\eta(\theta)^\top T(x)\big)\, dx.$$

*Proof.* Integrating $p(x;\theta)$ over $\mathcal{X}$ and using Fubini's theorem yields

$$1 = \int_{\mathcal{X}} p(x;\theta)\, dx = e^{-A(\theta)} \int_{\mathcal{X}} h(x) \exp\big(\eta(\theta)^\top T(x)\big)\, dx.$$

Solving for $A(\theta)$ gives the stated identity whenever the integral is finite. □

**Proposition 3.3** (Log-MGF of $T(X)$ in an exponential family)**.** *Let $X \sim p(x;\theta)$ in an exponential family and let $u \in \mathbb{R}^k$ be such that $\eta(\theta) + u \in \Omega$. The log moment generating function of $T(X)$ under parameter $\theta$ is*

$$K_T(u;\theta) := \ln \mathbb{E}_\theta\Big[e^{u^\top T(X)}\Big] = A\big(\eta(\theta)+u\big) - A\big(\eta(\theta)\big),$$

*where $A$ is treated here as a function of $\eta$. Equivalently, the MGF is $M_T(u;\theta) = \exp\big(A(\eta(\theta)+u) - A(\eta(\theta))\big)$.*

*Proof.* Write $p(x;\eta) = h(x)\exp\{\eta^\top T(x) - A(\eta)\}$ and let

$$Z(\eta) = \int h(x)\exp\{\eta^\top T(x)\}\,dx,$$

so that $A(\eta) = \ln Z(\eta)$ and

$$p(x;\eta) = \frac{h(x)\exp\{\eta^\top T(x)\}}{Z(\eta)}.$$

Then, for $\theta$ with natural parameter $\eta(\theta)$,

$$\begin{aligned}\mathbb{E}_\theta\Big[e^{u^\top T(X)}\Big] &= \int e^{u^\top T(x)}\,p(x;\eta(\theta))\,dx \\ &= \frac{1}{Z(\eta(\theta))}\int h(x)\exp\{(\eta(\theta)+u)^\top T(x)\}\,dx \\ &= \frac{Z(\eta(\theta)+u)}{Z(\eta(\theta))} = \exp\big(A(\eta(\theta)+u) - A(\eta(\theta))\big).\end{aligned}$$

Taking logarithms gives

$$K_T(u;\theta) = \ln \mathbb{E}_\theta\Big[e^{u^\top T(X)}\Big] = A(\eta(\theta)+u) - A(\eta(\theta)),$$

as claimed. □

**Definition 3.4** (Expectation and covariance notation)**.** For a parameter value $\theta$, we write $\mathbb{E}_\theta[\cdot]$ for expectation with respect to $p(x;\theta)$ and $\mathrm{Cov}_\theta(Y)$ for the covariance of a random vector $Y$ under $p(\cdot;\theta)$:

$$\mathrm{Cov}_\theta(Y) = \mathbb{E}_\theta\big[(Y - \mathbb{E}_\theta[Y])(Y - \mathbb{E}_\theta[Y])^\top\big], \qquad \mathrm{Var}_\theta(Y) = \mathrm{Cov}_\theta(Y) \text{ for scalar } Y.$$

Similarly, when the family is written in terms of the natural parameter $\eta$, we may use the subscript $\eta$. When unambiguous, we may omit the subscript and write $\mathbb{E}[\cdot]$, $\mathrm{Cov}(\cdot)$, and $\mathrm{Var}(\cdot)$.

**Lemma 3.5** (Log-Partition Function Generates Moments)**.** *For the exponential family $p(x;\eta) = h(x)\exp\{\eta^\top T(x) - A(\eta)\}$, the gradient and Hessian with respect to the natural parameter $\eta$ satisfy*

$$\nabla_\eta A(\eta) = \mathbb{E}_\eta[T(X)], \qquad \nabla^2_\eta A(\eta) = \mathrm{Cov}_\eta(T(X)).$$

*Proof.* Let $Z(\eta) = \int h(x)\exp\{\eta^\top T(x)\}\,dx$ so that $A(\eta) = \ln Z(\eta)$. Under standard regularity (allowing differentiation under the integral sign), for each component $j$ we have

$$\frac{\partial Z}{\partial \eta_j} = \int h(x)\,T_j(x)\,e^{\eta^\top T(x)}\,dx.$$

By the chain rule,

$$\frac{\partial A}{\partial \eta_j} = \frac{1}{Z(\eta)}\frac{\partial Z}{\partial \eta_j} = \int T_j(x)\,\frac{h(x)e^{\eta^\top T(x)}}{Z(\eta)}\,dx = \mathbb{E}_\eta[T_j(X)].$$

For the Hessian $H$, the $(j,k)$ entry is given by

$$\frac{\partial^2 A}{\partial \eta_j\,\partial \eta_k} = \frac{\partial}{\partial \eta_k}\mathbb{E}_\eta[T_j(X)] = \mathbb{E}_\eta[T_j(X)T_k(X)] - \mathbb{E}_\eta[T_j(X)]\,\mathbb{E}_\eta[T_k(X)] = \mathrm{Cov}_\eta\big(T_j(X), T_k(X)\big),$$

where we used that $\nabla_\eta \ln Z(\eta) = \frac{1}{Z}\nabla_\eta Z$ and the product rule. Stacking components yields the stated vector and matrix identities. □

### 3.2 Maximum-Likelihood Estimation for Exponential Families

Maximizing the likelihood is equivalent to maximizing the log-likelihood due to the monotonicity of the logarithm. In exponential families, first-order conditions reduce to matching the expectation of the statistics $T(X)$ to their empirical averages.

**Definition 3.6** (Log-Likelihood in an Exponential Family)**.** For i.i.d. samples $x_1, \ldots, x_n$ from $p(x;\theta) = h(x)\exp(\eta(\theta)^\top T(x) - A(\theta))$, the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n}\Big[\eta(\theta)^\top T(x_i) - A(\theta)\Big] + \sum_{i=1}^{n}\ln h(x_i).$$

**Definition 3.7** (Score Function)**.** The score is $U(\theta) = \nabla_\theta \ell(\theta)$, the gradient of the log-likelihood with respect to the parameter $\theta$.

**Lemma 3.8** (Score in Exponential Families)**.** *For an exponential family,*

$$U(\theta) = \nabla_\theta \ell(\theta) = \sum_{i=1}^{n} J_\eta(\theta)^\top\left(T(x_i) - \mathbb{E}_\theta[T(X)]\right),$$

*where $J_\eta(\theta) = \nabla_\theta \eta(\theta)$ is the Jacobian of the map $\eta\colon \Theta \to \mathbb{R}^k$ with elements $[J_\eta(\theta)]_{j,k} = \partial \eta_j/\partial \theta_k$.*

*Proof.* From the definition of $\ell(\theta)$ and linearity of differentiation,

$$\nabla_\theta \ell(\theta) = \sum_{i=1}^{n} J_\eta(\theta)^\top T(x_i) - n\,\nabla_\theta A(\theta).$$

By the chain rule and the moment identity for exponential families, $\nabla_\theta A(\theta) = J_\eta(\theta)^\top \nabla_\eta A(\theta) = J_\eta(\theta)^\top \mathbb{E}_\theta[T(X)]$. Substituting yields the stated form. □

**Lemma 3.9** (Score Vanishes at the MLE)**.** *If $\hat{\theta}$ is an interior maximizer of $\ell(\theta)$ and regularity conditions hold, then $U(\hat{\theta}) = \nabla_\theta \ell(\hat{\theta}) = 0$.*

*Proof.* First-order optimality for a differentiable function on an open set implies the gradient is zero at any interior maximizer. □

**Lemma 3.10** (Moment Matching at the MLE)**.** *Under mild conditions, the MLE $\hat{\theta}$ satisfies*

$$\mathbb{E}_{\hat{\theta}}[T(X)] = \frac{1}{n}\sum_{i=1}^{n} T(x_i).$$

*Proof.* Set the score to zero: $0 = \nabla_\theta \ell(\hat{\theta}) = J_\eta(\hat{\theta})^\top \sum_i \big(T(x_i) - \mathbb{E}_{\hat{\theta}}[T(X)]\big)$. If $J_\eta(\hat{\theta})$ has full column rank, then the bracketed sum must vanish, giving the stated equality. □

*Example* 3.11 (Exponential Distribution)*.* For $X_i \sim \text{Exp}(\lambda)$,

$$L(\lambda) = \lambda^n e^{-\lambda \sum_i x_i}, \quad \hat{\lambda} = \frac{n}{\sum_i x_i}.$$

The score equation predicts moment matching for $T(x) = x$: since $\mathbb{E}_\lambda[X] = 1/\lambda$, the condition $\mathbb{E}_{\hat{\lambda}}[X] = \bar{X}$ yields $\hat{\lambda} = 1/\bar{X} = n/\sum_i x_i$ as above.

*Example* 3.12 (Bernoulli Distribution)*.* If $X_i \sim \text{Bernoulli}(p)$, then $p(x;p) = p^x(1-p)^{1-x}$ is an exponential family with $T(x) = x$. The likelihood $L(p) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$ is maximized at

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

This agrees with moment matching since $\mathbb{E}_p[X] = p$.

# 4 Hypothesis Testing

## 4.1 Frequentist Approach

**Definition 4.1** (Null and Alternative Hypotheses)**.** A *null hypothesis* $H_0$ specifies a set of parameter values (e.g., $\theta = \theta_0$). An *alternative hypothesis* $H_1$ represents competing values (e.g., $\theta > \theta_0$ or $\theta \neq \theta_0$).

**Definition 4.2** (Test Statistic and Rejection Region)**.** A *test statistic* $T(X)$ is a function of the data used to decide whether to reject $H_0$. The rejection region $\mathcal{R}$ is the set of values for which $H_0$ is rejected.

*Example* 4.3 (Gaussian Mean Test)*.* Assume $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with $\sigma$ known. Under $H_0 : \mu = \mu_0$, the standardized statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

has the standard normal distribution.

One-sided upper-tail test ($H_1 : \mu > \mu_0$): reject $H_0$ at level $\alpha$ if $Z > z_{1-\alpha}$; the $p$-value is $p = 1 - \Phi(z_{\text{obs}})$.

One-sided lower-tail test ($H_1 : \mu < \mu_0$): reject $H_0$ at level $\alpha$ if $Z < z_\alpha$; the $p$-value is $p = \Phi(z_{\text{obs}})$.

Two-sided test ($H_1 : \mu \neq \mu_0$): reject $H_0$ at level $\alpha$ if $|Z| > z_{1-\alpha/2}$; the $p$-value is

$$p = 2\big(1 - \Phi(|z_{\text{obs}}|)\big).$$

Equivalently, the two-sided level-$\alpha$ test rejects $H_0$ iff $\mu_0$ lies outside the $(1-\alpha) \times 100\%$ confidence interval

$$\bar{X} \pm z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}.$$

**Definition 4.4** (Type-I and Type-II Errors)**.** A *Type-I error* occurs when $H_0$ is rejected although true. A *Type-II error* occurs when $H_0$ is not rejected although false. The probability of Type-I error is denoted by $\alpha$ and called the *significance* of the test. The probability of Type-II error is denoted by $\beta$ and the quantity $1-\beta$ is called the *power* of the test.

**Definition 4.5** (Likelihood Ratio Test (LRT))**.** Given hypotheses $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$, and an observed sample $x$, one rejects the hypothesis $H_0$ if the *likelihood ratio* (LR) statistic

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta_1} L(\theta; x)}$$

is less than some threshold $c$ (e.g., 1/10). In that case, the likelihood of the most likely explanation of $x$ under $H_0$ is smaller than $c$ times the likelihood of the most likely explanation of $x$ under $H_1$.

*Remark* 4.6. For a LRT, applying a strictly monotone function to both the likelihood-ratio $\Lambda(x)$ and the threshold $c$ gives an equivalent test where the direction of the inequality is flipped if the function is decreasing. Thus, any such function can be used to simplify the test expression in terms of $x$ without loss in performance.

**Definition 4.7** ($p$-value)**.** The *p-value* for observed data $x$ is the smallest level of significance $\alpha$ at which $H_0$ would be rejected by the test, equivalently the tail probability under $H_0$ of obtaining a test statistic that is more extreme than observed.

**Theorem 4.8** (Neyman–Pearson Lemma)**.** *For testing simple hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, the test that rejects $H_0$ when*

$$\frac{L(\theta_0; x)}{L(\theta_1; x)} < c_\alpha$$

*has the highest power among all tests of significance $\alpha$.*

**Lemma 4.9** (Gaussian Mean Test)**.** *For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with known $\sigma$, the two-sided test for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ using the likelihood ratio test reduces to computing the classical $Z$-statistic*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

*and rejecting if $|Z| > z_{1-\alpha/2}$, where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi$ is the standard Gaussian CDF.*

*Proof.* For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with known $\sigma$, the sample mean $\overline{X}$ is a sufficient statistic for $\mu$ and $\overline{X}$ is also Gaussian with mean $\mu$ and variance $\sigma^2/n$. Thus, the likelihood ratio test for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, in terms of the sample mean $x$, is given by

$$\Lambda(x) = \frac{L(\mu_0; x)}{\sup_{\mu_1 \neq \mu_0} L(\mu_1; x)} = \frac{e^{-n(x-\mu_0)^2/(2\sigma^2)}}{\sup_{\mu_1 \neq \mu_0} e^{-n(x-\mu_1)^2/(2\sigma^2)}} = e^{-n(x-\mu_0)^2/(2\sigma^2)}.$$

To find the classical $Z$-test, we apply the strictly decreasing function $\sqrt{-\ln(x)}$ to $\Lambda(x)$. This gives

$$\sqrt{-2\ln\Lambda(x)} = \sqrt{\frac{n(\bar{X}-\mu_0)^2}{\sigma^2}} = \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| = |Z|.$$

Since we applied a decreasing function, the test rejects if $|Z| > c$. Choosing the threshold $c$ to achieve significance $\alpha$ gives $c = \Phi^{-1}(1-\alpha/2) = z_{1-\alpha/2}$ and results in the classical two-sided $Z$-test that rejects if $|Z| > z_{1-\alpha/2}$. □

*Example* 4.10 (Practical Illustration). Suppose $\sigma = 10$, $n = 25$, and the observed mean is $\bar{x} = 53$ with $\mu_0 = 50$. Then $Z = (53-50)/(10/\sqrt{25}) = 1.5$. At $\alpha = 0.05$, $z_{1-\alpha/2} \approx 1.96$, so we do not reject $H_0$.

## 4.2 Bayesian Approach

While the frequentist approach makes no assumptions about $\theta$, the Bayesian approach treats $\theta$ as a random variable with prior $\pi(\theta)$. Consider testing $H_0$ (no disease) versus $H_1$ (disease) with prior probabilities $\Pr(H_0) = \pi_0$ and $\Pr(H_1) = \pi_1 = 1 - \pi_0$ where typically $\pi_0 \gg \pi_1$. Given a test outcome $X$, the *posterior odds* are

$$\frac{\Pr(H_1 \mid X)}{\Pr(H_0 \mid X)} = \underbrace{\frac{\pi_1}{\pi_0}}_{\text{prior odds}} \times \underbrace{\frac{p(X \mid H_1)}{p(X \mid H_0)}}_{\text{Bayes factor } B_{10}}.$$

Even with a positive test, a small prior $\pi_1$ can keep the posterior probability of disease low unless $B_{10}$ is large. Decision thresholds follow from comparing posterior odds to the loss-weighted threshold.

*Example* 4.11 (COVID testing example). Suppose the disease prevalence is $\pi_1 = 0.01$ (so $\pi_0 = 0.99$), a test has sensitivity $\Pr(+ \mid H_1) = 0.90$ and specificity $\Pr(- \mid H_0) = 0.95$ (so $\Pr(+ \mid H_0) = 0.05$). For a positive test,

$$B_{10} = \frac{\Pr(+ \mid H_1)}{\Pr(+ \mid H_0)} = \frac{0.90}{0.05} = 18.$$

The posterior odds are

$$\frac{\Pr(H_1 \mid +)}{\Pr(H_0 \mid +)} = \frac{\pi_1}{\pi_0} B_{10} = \frac{0.01}{0.99} \times 18 \approx 0.1818,$$

so $\Pr(H_1 \mid +) = \frac{0.1818}{1+0.1818} \approx 0.154$ is not even greater than 0.5. But, two independent positive tests would yield $B_{10} = 18^2 = 324$ and posterior probability $\approx \frac{(0.01/0.99)\times 324}{1+(0.01/0.99)\times 324} \approx 0.77$.

**Definition 4.12** (Posterior and Bayes Factor)**.** Let $P(H_j) = \pi_j$ denote the prior probability of hypothesis $H_j$ for $j = 0, 1$, and let $\pi_j(\theta)$ be the prior density of $\theta$ under $H_j$ on $\Theta_j$ (so $\int_{\Theta_j} \pi_j(\theta)\, d\theta = 1$). The posterior satisfies $p(\theta \mid x) \propto L(\theta; x)\, \pi(\theta)$. Then, the marginal likelihoods under $H_0$ and $H_1$ are

$$p(x \mid H_j) = \int_{\Theta_j} L(\theta; x)\, \pi_j(\theta)\, d\theta, \qquad j = 0, 1.$$

The *Bayes factor* in favor of $H_1$ over $H_0$ is

$$B_{10} = \frac{p(x \mid H_1)}{p(x \mid H_0)}.$$

The posterior odds equal prior odds times the Bayes factor:

$$\frac{\Pr(H_1 \mid x)}{\Pr(H_0 \mid x)} = \frac{\pi_1}{\pi_0} B_{10}.$$

*Example* 4.13 (Gaussian Mean with Conjugate Prior). Let $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ with known $\sigma^2$. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \sim N(\mu_0, \tau^2)$. Then under both hypotheses the joint density factors into a term in $\bar{X}$ and an ancillary term in the residuals. Thus, we can derive a closed-form Bayes factor by integrating out $\mu$ to get $\bar{X} \mid H_1 \sim N(\mu_0,\ \sigma^2/n + \tau^2)$. Evaluating the two Normal densities at the observed $\bar{x}$ and taking their ratio yields

$$B_{10} = \frac{p(\bar{x} \mid H_1)}{p(\bar{x} \mid H_0)} = \frac{\phi\big(\bar{x}; \mu_0,\ \sigma^2/n + \tau^2\big)}{\phi(\bar{x}; \mu_0,\ \sigma^2/n)},$$

where $\phi(\cdot; m, v)$ is the Normal pdf with mean $m$ and variance $v$.

# 5 Summary

Descriptive statistics summarize observed data, while inferential statistics use probabilistic models to draw conclusions about parameters. Exponential families unify many common models through sufficient statistics and log-partition functions, leading naturally to the likelihood principle and MLEs. Hypothesis testing formalizes decision-making under uncertainty through one-sided, two-sided, and likelihood ratio tests. Frequentist testing focuses on long-run error rates, while Bayesian testing incorporates prior beliefs via the Bayes factor.

# 6 Exercises

1. Unbiasedness of the sample variance. Let $X_1, \ldots, X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2 < \infty$. Show that $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ satisfies $\mathbb{E}[S^2] = \sigma^2$.

   **Solution:** We use the variance-decomposition identity

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

   Taking expectations on both terms and using independence,

$$\mathbb{E}\left[\sum_{i=1}^n (X_i - \mu)^2\right] = n \operatorname{Var}(X_1) = n\sigma^2, \qquad \mathbb{E}\big[n(\bar{X} - \mu)^2\big] = n \operatorname{Var}(\bar{X}) = n \cdot \frac{\sigma^2}{n} = \sigma^2.$$

   Hence

$$\mathbb{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2,$$

   and therefore

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \sigma^2.$$

For completeness, the identity follows by expanding and simplifying:

$$\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_i (X_i - \mu) + n(\bar{X} - \mu)^2,$$

and noting that the middle term vanishes because $\sum_i (X_i - \mu) = n(\bar{X} - \mu)$.

2. Exponential MLE and confidence interval. Let $X_i \sim \text{Exp}(\lambda)$ i.i.d. Given data $(0.7, 1.2, 0.4, 2.0, 1.5)$, compute the MLE $\hat{\lambda}$ and an approximate 95% confidence interval using the asymptotic normal approximation.

   **Solution:** MLE derivation:

$$\ell(\lambda) = \sum_{i=1}^{n} \big( \ln \lambda - \lambda x_i \big) = n \ln \lambda - \lambda \sum_i x_i \;\Rightarrow\; \frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_i x_i = 0 \;\Rightarrow\; \hat{\lambda} = \frac{n}{\sum_i x_i}.$$

   Numerics: $\sum x_i = 5.8$, $n = 5$, so $\hat{\lambda} = 5/5.8 \approx 0.8621$.

   Asymptotic standard error: the Fisher information for a single observation is $I_1(\lambda) = 1/\lambda^2$, so $I_n(\lambda) = n/\lambda^2$ and

$$\text{SE}(\hat{\lambda}) \approx \sqrt{\frac{1}{I_n(\hat{\lambda})}} = \frac{\hat{\lambda}}{\sqrt{n}} \approx \frac{0.8621}{\sqrt{5}} \approx 0.3855.$$

   A 95% Wald CI around an estimate equals the estimated value plus or minus 1.96 standard deviations because that would imply 95% confidence if the posterior distribution was Gaussian. Thus, we have

$$\hat{\lambda} \pm 1.96\,\text{SE} \approx 0.8621 \pm 1.96 \times 0.3855 \approx (0.1065,\ 1.6177).$$

   Exact CI (recommended for small $n$): using $2\lambda \sum_i X_i \sim \chi^2_{2n}$, invert to get

$$\left[ \frac{\chi^2_{0.025,\ 2n}}{2\sum_i x_i}\ ,\ \frac{\chi^2_{0.975,\ 2n}}{2\sum_i x_i} \right] = \left[ \frac{\chi^2_{0.025,10}}{11.6}\ ,\ \frac{\chi^2_{0.975,10}}{11.6} \right] \approx (0.279,\ 1.765),$$

   using $\chi^2_{0.025,10} \approx 3.246$ and $\chi^2_{0.975,10} \approx 20.483$.

3. Sufficiency in the Normal model with known variance. Let $X_i \sim N(\mu, \sigma^2)$ i.i.d. with known $\sigma^2$. Show $T = \sum_i X_i$ is sufficient for $\mu$ via factorization.

   **Solution:** Start from the joint density

$$p(x; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\!\left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

   Expand the square: $(x_i - \mu)^2 = x_i^2 - 2\mu x_i + \mu^2$. Then

$$p(x; \mu) = (2\pi\sigma^2)^{-n/2} \exp\!\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right) \exp\!\left( \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2} \right).$$

   This factors as $h(x)\, g(T; \mu)$ with

$$h(x) = (2\pi\sigma^2)^{-n/2} \exp\!\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right), \quad T(x) = \sum_{i=1}^{n} x_i, \quad g(T; \mu) = \exp\!\left( \frac{\mu}{\sigma^2} T - \frac{n\mu^2}{2\sigma^2} \right),$$

which does not depend on $\mu$ except through $T$. By the factorization theorem, $T = \sum_i X_i$ (equivalently $\bar{X}$) is sufficient for $\mu$.

4. Z-test with known variance (numerical). Suppose $\sigma = 12$, $n = 36$, $\bar{x} = 205$, and we test $H_0 : \mu = 200$ vs. $H_1 : \mu \neq 200$ at $\alpha = 0.05$. Compute the test statistic, $p$-value, and decision.

   **Solution:** Compute the standardized statistic

   $$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{205 - 200}{12/\sqrt{36}} = \frac{5}{2} = 2.5.$$

   Two-sided $p$-value:

   $$p = 2\big(1 - \Phi(2.5)\big) \approx 2(1 - 0.9938) = 0.0124.$$

   Decision at $\alpha = 0.05$: since $p < 0.05$ (equivalently, $|Z| = 2.5 > 1.96$), reject $H_0$.

   Cross-check via confidence interval: a 95% CI is

   $$\bar{x} \pm z_{0.975} \frac{\sigma}{\sqrt{n}} = 205 \pm 1.96 \cdot 2 = (201.08,\ 208.92),$$

   which does not contain $\mu_0 = 200$, so we again reject $H_0$.

5. Likelihood-ratio test for a Poisson mean. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ i.i.d. Test $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. Derive the LRT in terms of $T = \sum_i X_i$ and express the $p$-value in terms of a Poisson CDF. Evaluate the decision for $n = 10$, observed $T = 18$, and $\lambda_0 = 1.2$ at $\alpha = 0.05$.

   **Solution:** By sufficiency, the likelihood depends on the sample only through $T = \sum_i X_i$. Under $H_0$, $T \sim \text{Poisson}(n\lambda_0)$ with mean $n\lambda_0$. The LRT for a two-sided alternative rejects for values of $T$ in the tails (far from $n\lambda_0$). The corresponding two-sided $p$-value is

   $$p = 2\min\Big\{\mathbb{P}_{H_0}\big(T \le t_{\text{obs}}\big),\ \mathbb{P}_{H_0}\big(T \ge t_{\text{obs}}\big)\Big\} = 2\min\Big\{F_T(t_{\text{obs}}),\ 1 - F_T(t_{\text{obs}} - 1)\Big\},$$

   where $F_T(k) = \mathbb{P}_{H_0}(T \le k)$ and we used $\mathbb{P}(T \ge t) = 1 - F_T(t-1)$ for integer $t$.

   Numerics: with $n = 10$, $\lambda_0 = 1.2$, we have $T \sim \text{Poisson}(12)$ and $t_{\text{obs}} = 18$, so

   $$p = 2\min\big\{F_T(18),\ 1 - F_T(17)\big\}.$$

   Using the normal approximation with continuity correction gives

   $$Z = \frac{18.5 - 12}{\sqrt{12}} \approx 1.88, \quad \mathbb{P}(T \ge 18) \approx 0.030,$$

   so $p \approx 2 \times 0.030 = 0.060$. Exact computation yields $p \approx 0.067$. Decision: since $p > 0.05$, do not reject $H_0$.