Week 5: Confidence Intervals and Hypothesis Testing

Francisco Richter and Ernst Wit

Introduction to Data Science (MSc)

1 Introduction to Statistical Inference

Statistical inference represents the cornerstone of data science methodology, providing the mathematical framework for drawing conclusions about populations based on sample data while quantifying the uncertainty inherent in this process. The two fundamental paradigms of statistical inference—confidence intervals for estimation and hypothesis testing for decision making—form the foundation upon which modern data science applications are built.

The theoretical importance of statistical inference extends far beyond its computational applications. These methods embody fundamental principles of scientific reasoning, including the quantification of uncertainty, the control of error rates, and the systematic evaluation of evidence. Understanding these principles deeply provides essential insights into the reliability and limitations of data-driven conclusions, which is crucial for responsible data science practice.

From a mathematical perspective, statistical inference connects probability theory with practical decision making, demonstrating how abstract mathematical concepts translate into concrete tools for analyzing real-world data. The asymptotic theory underlying large-sample inference provides a bridge between finite-sample exact results and the approximate methods that are essential for analyzing complex, high-dimensional datasets common in modern applications.

2 Mathematical Foundations of Confidence Intervals

2.1 Theoretical Framework and Interpretation

The concept of confidence intervals requires careful mathematical formulation to avoid common misinterpretations while providing a rigorous foundation for uncertainty quantification. The frequentist interpretation of confidence intervals relies on the long-run behavior of the interval construction procedure rather than probability statements about fixed but unknown parameters.

Definition 2.1 (Confidence Interval)**.** *Let θ be an unknown parameter and let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution depending on θ. A $(1-\alpha)\times 100\%$ confidence interval for θ is a random interval $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ such that:*

$$P_\theta(L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n)) = 1 - \alpha$$

for all values of θ in the parameter space.

The subscript θ in P_θ emphasizes that the probability is computed under the assumption that θ is the true parameter value. This formulation clarifies that the randomness lies in the interval endpoints, which are functions of the random sample, rather than in the parameter θ, which is fixed but unknown.

The coverage probability property must hold for all possible values of θ, which is a strong requirement that distinguishes confidence intervals from other types of interval estimates. This uniform coverage property ensures that the confidence interval procedure is reliable regardless of the true parameter value, providing a robust foundation for statistical inference.

2.2 Pivotal Quantity Method and Exact Inference

The pivotal quantity method provides the most general approach to constructing confidence intervals with exact coverage probabilities. This method relies on finding functions of the data and parameter whose distributions are known and do not depend on unknown parameters.

Definition 2.2 (Pivotal Quantity)**.** *A pivotal quantity is a function $Q(X_1, \ldots, X_n, \theta)$ of the sample and the parameter such that the distribution of Q does not depend on θ or any other unknown parameters.*

The power of pivotal quantities lies in their ability to provide exact finite-sample inference without requiring asymptotic approximations. The distribution of a pivotal quantity is completely known, enabling precise probability calculations that form the basis for confidence interval construction.

Theorem 2.1 (Confidence Interval Construction via Pivotal Quantities)**.** *Let $Q(X_1, \ldots, X_n, \theta)$ be a pivotal quantity with known distribution. If there exist constants c_1 and c_2 such that $P(c_1 \leq Q \leq c_2) = 1 - \alpha$, then the set of values $\{\theta : c_1 \leq Q(X_1, \ldots, X_n, \theta) \leq c_2\}$ forms a $(1 - \alpha)$ confidence interval for θ.*

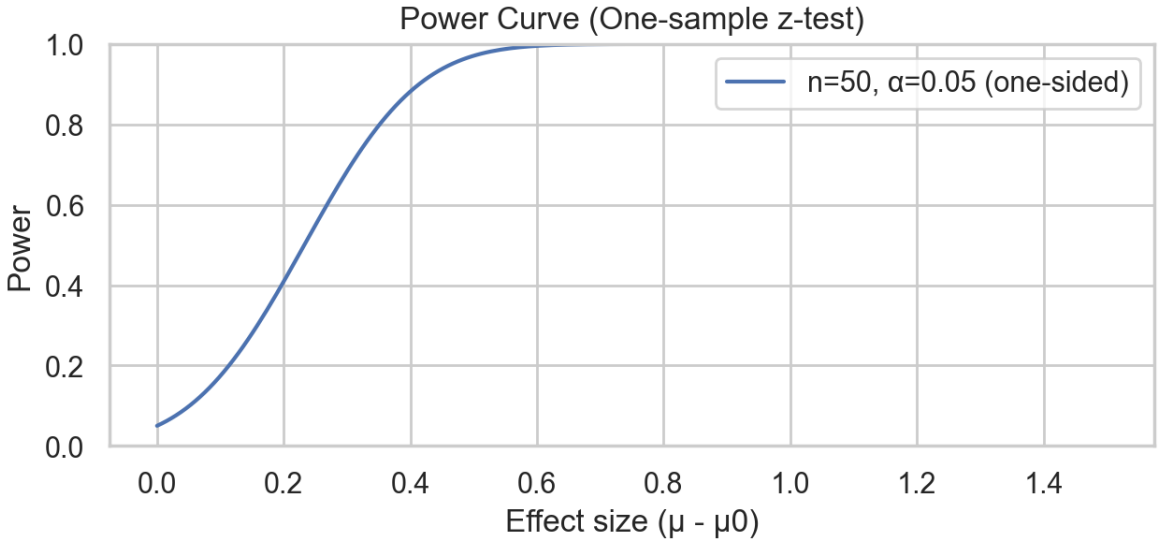

Figure 1: Power curve for a one-sample z-test (one-sided) as a function of effect size for fixed n.

Example 2.1 (Normal Mean with Known Variance)**.** *For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with known σ^2, the pivotal quantity is:*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Since $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$*, we have:*

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Solving the inequality for μ *yields the confidence interval:*

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

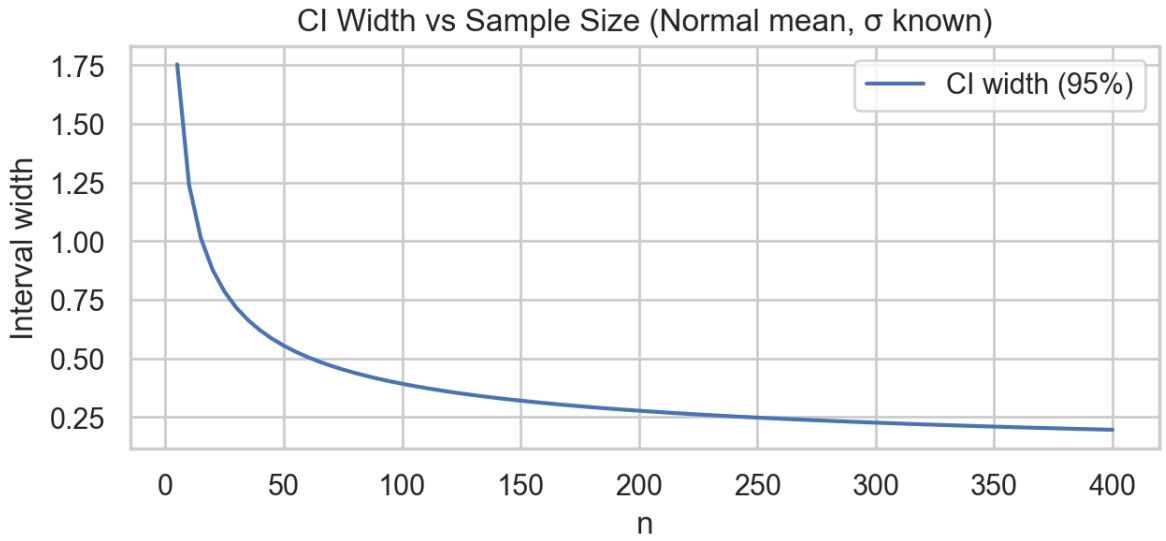

Figure 2: Width of a 95% confidence interval for μ with known σ versus sample size n.

The algebraic manipulation required to invert the pivotal quantity inequality demonstrates the mathematical precision underlying confidence interval construction. Each step in the inversion must preserve the direction of inequalities and maintain the probabilistic interpretation.

2.3 Asymptotic Theory and Large-Sample Methods

When exact pivotal quantities are not available, asymptotic theory provides a powerful framework for constructing approximate confidence intervals that become increasingly accurate as sample sizes grow. The foundation of large-sample inference rests on limit theorems that characterize the asymptotic behavior of estimators.

Theorem 2.2 (Asymptotic Normality of Maximum Likelihood Estimators)**.** *Under regularity conditions, the maximum likelihood estimator* $\hat{\theta}_n$ *satisfies:*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, I^{-1}(\theta))$$

where $I(\theta)$ *is the Fisher information matrix and* $\xrightarrow{D}$ *denotes convergence in distribution.*

The Fisher information matrix plays a central role in asymptotic inference, providing both the asymptotic variance of maximum likelihood estimators and a measure of the information content of the data about the parameter.

Definition 2.3 (Fisher Information)**.** *For a single parameter θ, the Fisher information is:*

$$I(\theta) = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right]$$

where $f(x;\theta)$ is the probability density function.

The equivalence of the two expressions for Fisher information follows from regularity conditions that allow differentiation and integration to be interchanged. The Fisher information quantifies how much information about θ is contained in a single observation, with larger values indicating more informative data.

Theorem 2.3 (Asymptotic Confidence Intervals)**.** *If $\hat{\theta}_n$ is asymptotically normal with $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$, then an approximate $(1-\alpha)$ confidence interval is:*

$$\left[\hat{\theta}_n - z_{\alpha/2}\frac{\hat{\sigma}(\hat{\theta}_n)}{\sqrt{n}}, \hat{\theta}_n + z_{\alpha/2}\frac{\hat{\sigma}(\hat{\theta}_n)}{\sqrt{n}}\right]$$

where $\hat{\sigma}^2(\hat{\theta}_n)$ is a consistent estimator of $\sigma^2(\theta)$.

The asymptotic approach requires careful attention to the rate of convergence and the quality of the normal approximation for finite samples. The accuracy of asymptotic confidence intervals depends on both the sample size and the underlying distribution, with some distributions requiring larger samples than others to achieve adequate approximation quality.

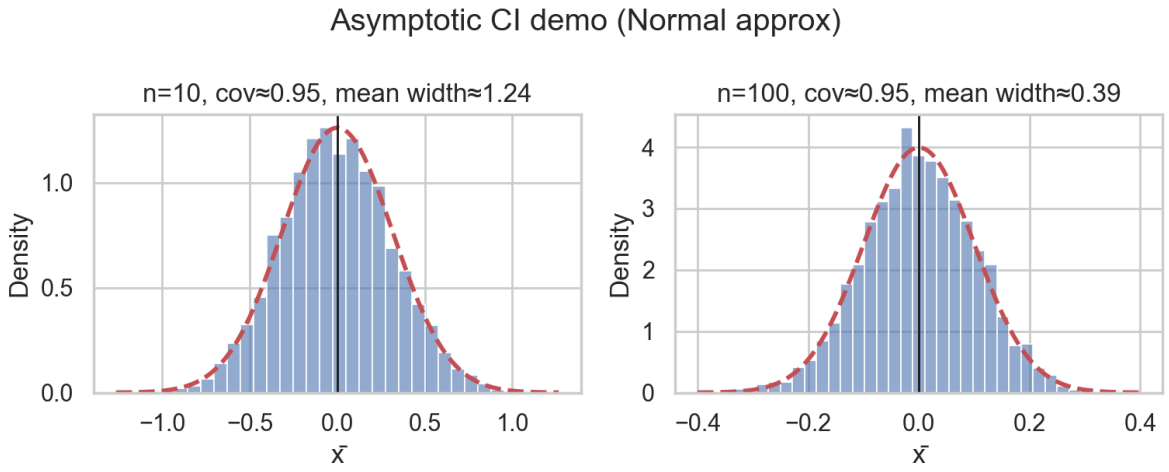

Figure 3: Asymptotic CI demonstration for small and large n: coverage and mean width.

2.4 Bootstrap Methods and Nonparametric Inference

Bootstrap methods provide a nonparametric approach to confidence interval construction that does not require distributional assumptions or asymptotic theory. The bootstrap principle uses resampling to approximate the sampling distribution of estimators, providing a flexible framework for inference in complex settings.

Definition 2.4 (Bootstrap Principle)**.** *Let $\hat{\theta}$ be an estimator of parameter θ based on sample $X_1, \ldots, X_n$. The bootstrap approximates the distribution of $\hat{\theta} - \theta$ by the distribution of $\hat{\theta}^* - \hat{\theta}$, where $\hat{\theta}^*$ is computed from a bootstrap sample $X_1^*, \ldots, X_n^*$ drawn with replacement from the original sample.*

The theoretical justification for the bootstrap relies on the consistency of the empirical distribution function as an estimator of the true distribution function. Under mild regularity conditions, the bootstrap distribution converges to the true sampling distribution, providing asymptotically valid inference.

Theorem 2.4 (Bootstrap Consistency)**.** *Under regularity conditions, if $\hat{\theta}_n$ is a consistent estimator of θ, then:*

$$\sup_x |P^*(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \leq x) - P(\sqrt{n}(\hat{\theta}_n - \theta) \leq x)| \xrightarrow{P} 0$$

where P^* denotes probability computed with respect to the bootstrap distribution.

Bootstrap confidence intervals can be constructed using several methods, each with different theoretical properties and practical performance characteristics. The percentile method uses quantiles of the bootstrap distribution directly, while the bias-corrected and accelerated (BCa) method adjusts for bias and skewness in the bootstrap distribution.

3 Hypothesis Testing: Mathematical Framework and Theory

3.1 Fundamental Concepts and Error Control

Hypothesis testing provides a formal framework for making decisions under uncertainty, with mathematical foundations rooted in decision theory and the control of error probabilities. The Neyman-Pearson framework establishes the theoretical foundation for modern hypothesis testing by formalizing the tradeoff between different types of errors.

Definition 3.1 (Statistical Hypothesis Testing Framework)**.** *A hypothesis test consists of:*

1. *A null hypothesis H_0 and alternative hypothesis H_1 that partition the parameter space*
2. *A test statistic $T(X_1, \ldots, X_n)$ that summarizes the evidence against H_0*
3. *A rejection region R such that H_0 is rejected if $T \in R$*
4. *A significance level α that controls the Type I error probability*

The mathematical formulation of hypothesis testing requires precise specification of the hypotheses in terms of parameter values or distributional properties. Simple hypotheses specify the parameter completely, while composite hypotheses specify only that the parameter lies in some subset of the parameter space.

Definition 3.2 (Type I and Type II Errors)**.** *For a hypothesis test with rejection region R:*

$$\textit{Type I Error:} \quad \alpha(\theta) = P_\theta(T \in R) \text{ when } \theta \in \Theta_0 \tag{1}$$

$$\textit{Type II Error:} \quad \beta(\theta) = P_\theta(T \notin R) \text{ when } \theta \in \Theta_1 \tag{2}$$

$$\textit{Power:} \quad \pi(\theta) = P_\theta(T \in R) \text{ when } \theta \in \Theta_1 = 1 - \beta(\theta) \tag{3}$$

The power function $\pi(\theta)$ characterizes the performance of a test across all possible parameter values, providing a complete description of the test's ability to detect departures from the null hypothesis. An ideal test would have $\pi(\theta) = \alpha$ for $\theta \in \Theta_0$ and $\pi(\theta) = 1$ for $\theta \in \Theta_1$, though such tests rarely exist in practice.

3.2 Neyman-Pearson Lemma and Optimal Tests

The Neyman-Pearson lemma provides the theoretical foundation for constructing optimal tests by characterizing the most powerful test for testing a simple null hypothesis against a simple alternative.

Theorem 3.1 (Neyman-Pearson Lemma)**.** *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. The most powerful test of size α has rejection region:*

$$R = \left\{ x : \frac{f(x|\theta_1)}{f(x|\theta_0)} > k \right\}$$

where k is chosen so that $P_{\theta_0}(X \in R) = \alpha$.

The likelihood ratio $\frac{f(x|\theta_1)}{f(x|\theta_0)}$ provides a natural measure of the evidence in favor of H_1 relative to H_0. Large values of this ratio indicate that the observed data are more likely under H_1 than under H_0, providing strong evidence against the null hypothesis.

The Neyman-Pearson lemma establishes that likelihood ratio tests are optimal in the sense of maximizing power for any given significance level. This optimality property provides theoretical justification for the widespread use of likelihood-based methods in statistical inference.

3.3 Likelihood Ratio Tests and Asymptotic Theory

For composite hypotheses, the generalized likelihood ratio test extends the Neyman-Pearson framework by comparing the maximum likelihood under the null and alternative hypotheses.

Definition 3.3 (Generalized Likelihood Ratio Test)**.** *For testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, the likelihood ratio test statistic is:*

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}$$

where $\Theta = \Theta_0 \cup \Theta_1$ is the entire parameter space.

The likelihood ratio Λ takes values between 0 and 1, with values close to 0 providing strong evidence against H_0 and values close to 1 providing little evidence against H_0. The test rejects H_0 when Λ is sufficiently small, or equivalently, when $-2 \log \Lambda$ is sufficiently large.

Theorem 3.2 (Wilks' Theorem)**.** *Under regularity conditions, if H_0 specifies r constraints on a p-dimensional parameter, then under H_0:*

$$-2 \log \Lambda \xrightarrow{D} \chi^2_r$$

as $n \to \infty$.

Wilks' theorem provides the asymptotic distribution of the likelihood ratio test statistic, enabling the construction of approximate tests for complex hypotheses. The degrees of freedom equal the difference in the number of free parameters between the null and alternative hypotheses, reflecting the dimensionality of the constraint imposed by H_0.

The regularity conditions required for Wilks' theorem include smoothness of the likelihood function, identifiability of parameters, and the assumption that the true parameter lies in the interior of the parameter space. Violations of these conditions can lead to non-standard asymptotic distributions that require specialized analysis.

4 Advanced Testing Procedures and Multiple Comparisons

4.1 Sequential Testing and Optional Stopping

Traditional hypothesis testing assumes a fixed sample size determined before data collection begins. Sequential testing procedures allow for data-dependent stopping rules that can improve efficiency and reduce expected sample sizes while maintaining error rate control.

Definition 4.1 (Sequential Probability Ratio Test)**.** *For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the sequential probability ratio test continues sampling until:*

$$\frac{L(\theta_1)}{L(\theta_0)} \leq A \quad \textit{(accept } H_0\textit{)} \quad \textit{or} \quad \frac{L(\theta_1)}{L(\theta_0)} \geq B \quad \textit{(reject } H_0\textit{)}$$

where A and B are chosen to achieve desired error probabilities.

Sequential tests can achieve the same error probabilities as fixed-sample tests with substantially smaller expected sample sizes, particularly when the true parameter is far from the boundary between H_0 and H_1. This efficiency gain makes sequential methods particularly valuable in applications where data collection is expensive or time-consuming.

4.2 Multiple Testing and Error Rate Control

Modern data science applications often involve testing thousands or millions of hypotheses simultaneously, creating a multiple testing problem where traditional error rate control becomes inadequate. The family-wise error rate and false discovery rate provide different frameworks for controlling errors in multiple testing scenarios.

Definition 4.2 (Family-Wise Error Rate)**.** *The family-wise error rate (FWER) is the probability of making at least one Type I error among all tests:*

$$FWER = P(\textit{at least one Type I error})$$

Theorem 4.1 (Bonferroni Correction)**.** *If each of m tests is conducted at level α/m, then $FWER \leq \alpha$.*

The Bonferroni correction provides a simple and widely applicable method for FWER control, but it can be overly conservative when the number of tests is large or when the tests are positively correlated. More sophisticated methods such as the Holm procedure provide uniformly more powerful alternatives while maintaining FWER control.

Definition 4.3 (False Discovery Rate)**.** *The false discovery rate (FDR) is the expected proportion of false discoveries among rejected hypotheses:*

$$FDR = E\left[\frac{V}{R \vee 1}\right]$$

where V is the number of false rejections and R is the total number of rejections.

Theorem 4.2 (Benjamini-Hochberg Procedure)**.** *Order the p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and let $k = \max\{i : p_{(i)} \leq \frac{i}{m}\alpha\}$. Reject hypotheses $H_{(1)}, \ldots, H_{(k)}$. This procedure controls FDR at level α when the test statistics are independent.*

The FDR framework is often more appropriate than FWER control in exploratory data analysis, where some false discoveries may be acceptable in exchange for increased power to detect true effects. The choice between FWER and FDR control depends on the relative costs of false discoveries and missed discoveries in the specific application context.

5 Bayesian Hypothesis Testing and Model Comparison

5.1 Bayes Factors and Evidence

Bayesian hypothesis testing provides an alternative framework that treats hypotheses as random quantities and uses probability to quantify evidence. Bayes factors provide a natural measure of the evidence in favor of one hypothesis relative to another.

Definition 5.1 (Bayes Factor)**.** *For hypotheses H_0 and H_1 with prior probabilities π_0 and π_1, the Bayes factor is:*

$$BF_{10} = \frac{P(Data|H_1)}{P(Data|H_0)} = \frac{m_1(x)}{m_0(x)}$$

where $m_i(x) = \int f(x|\theta_i)\pi_i(\theta_i)d\theta_i$ is the marginal likelihood under H_i.

The Bayes factor quantifies how much more likely the observed data are under H_1 compared to H_0, providing a direct measure of evidence that does not depend on arbitrary significance levels. Values of $BF_{10} > 1$ favor H_1, while values < 1 favor H_0.

Theorem 5.1 (Posterior Odds)**.** *The posterior odds in favor of H_1 are:*

$$\frac{P(H_1|Data)}{P(H_0|Data)} = BF_{10} \times \frac{\pi_1}{\pi_0}$$

This relationship shows how the Bayes factor updates prior odds to posterior odds, providing a coherent framework for incorporating both prior beliefs and data evidence in hypothesis evaluation.

5.2 Model Selection and Information Criteria

Bayesian model comparison extends hypothesis testing to the selection among multiple competing models, with information criteria providing approximate Bayesian solutions that are computationally tractable.

Definition 5.2 (Deviance Information Criterion)**.** *The DIC for model comparison is:*

$$DIC = -2\log f(y|\hat{\theta}) + 2p_D$$

where $\hat{\theta}$ is the posterior mean and p_D is the effective number of parameters.

Information criteria balance goodness of fit with model complexity, providing automatic penalties for overfitting that emerge naturally from Bayesian principles. The effective number of parameters p_D accounts for the reduction in effective dimensionality that occurs when informative priors constrain parameter estimates.

6 Applications in Modern Data Science

6.1 A/B Testing and Online Experimentation

A/B testing represents one of the most important applications of hypothesis testing in modern data science, enabling companies to make data-driven decisions about product changes and marketing strategies. The statistical framework must account for the unique challenges of online experimentation, including multiple testing, sequential monitoring, and heterogeneous treatment effects.

The traditional approach to A/B testing uses a two-sample test to compare conversion rates or other metrics between treatment and control groups. However, the online environment creates additional complications that require sophisticated statistical methods to address properly.

Sequential monitoring of A/B tests creates a multiple testing problem, as analysts may examine results multiple times during the experiment. Proper error rate control requires adjustment for this sequential monitoring, either through formal sequential testing procedures or through conservative significance level adjustments.

6.2 High-Dimensional Testing and Genomics

Modern genomics and other high-dimensional applications involve testing thousands or millions of hypotheses simultaneously, creating unprecedented multiple testing challenges. The statistical methods must balance the competing goals of discovering true effects while controlling false discoveries.

The two-groups model provides a framework for understanding the structure of high-dimensional testing problems. Under this model, a fraction π_0 of hypotheses are null (no effect), while the remaining fraction $1 - \pi_0$ are non-null (true effects). The goal is to identify the non-null hypotheses while controlling error rates.

Empirical Bayes methods provide a powerful approach to high-dimensional testing by estimating the proportion of null hypotheses and the distribution of effect sizes from the data. These methods can substantially improve power compared to traditional approaches by adapting to the specific characteristics of each dataset.

6.3 Machine Learning Model Validation

Statistical hypothesis testing plays a crucial role in machine learning model validation, providing formal frameworks for comparing model performance and assessing statistical significance of improvements. Cross-validation combined with appropriate statistical tests enables rigorous evaluation of model differences.

The McNemar test provides a framework for comparing the performance of two classifiers on the same dataset, accounting for the dependence between predictions. This test is particularly useful for comparing machine learning algorithms where traditional two-sample tests would be inappropriate due to the paired nature of the comparisons.

Permutation tests offer a nonparametric approach to model comparison that does not require distributional assumptions. By randomly permuting labels and recomputing performance metrics, these tests provide exact p-values for model comparison that are valid under minimal assumptions.

7 Summary

Confidence intervals and hypothesis testing form the mathematical foundation of statistical inference, providing rigorous frameworks for quantifying uncertainty and making decisions under uncertainty. The theoretical developments in this area, from the Neyman-Pearson lemma to modern multiple testing procedures, demonstrate the evolution of statistical thinking in response to increasingly complex data science applications.

The key insights from this material include the importance of error rate control, the power of asymptotic theory for large-sample inference, and the flexibility of bootstrap and other resampling methods for complex problems. These concepts are essential for understanding the reliability and limitations of statistical conclusions in data science applications.

The modern extensions to multiple testing, sequential analysis, and Bayesian methods demonstrate how classical statistical theory continues to evolve to address contemporary challenges. The ability to control error rates in high-dimensional settings, adapt to sequential data collection, and incorporate prior information makes these methods essential tools for modern data scientists.

The computational aspects of statistical inference, including bootstrap methods and permutation tests, illustrate how increased computational power has expanded the scope of statistical methods. Understanding these computational approaches is crucial for implementing robust and reliable statistical analyses in practice.

8 Exercises

1. **Confidence Interval Theory:** Derive the confidence interval for the difference of two normal means with unknown but equal variances. Show that the interval has exact coverage probability $(1 - \alpha)$ and explain the role of the pooled variance estimator.

2. **Fisher Information:** Calculate the Fisher information for the exponential distribution with parameter λ. Use this to derive the asymptotic variance of the maximum likelihood estimator and construct an asymptotic confidence interval.

3. **Bootstrap Consistency:** Explain why the bootstrap provides consistent estimates of sampling distributions. Discuss the conditions under which bootstrap confidence intervals have correct coverage probabilities.

4. **Neyman-Pearson Lemma:** Apply the Neyman-Pearson lemma to derive the most powerful test for $H_0 : \mu = 0$ versus $H_1 : \mu = 1$ when sampling from $N(\mu, 1)$. Calculate the power function for this test.

5. **Likelihood Ratio Tests:** Derive the likelihood ratio test for testing $H_0 : \sigma^2 = \sigma_0^2$ in a normal distribution with unknown mean. Show that the test statistic follows a chi-square distribution under H_0.

6. **Multiple Testing:** Compare the Bonferroni and Benjamini-Hochberg procedures for controlling error rates in multiple testing. Explain when each method is preferable and discuss their relative power properties.

7. **Bayes Factors:** Calculate the Bayes factor for comparing two normal distributions with different means but the same variance. Discuss how the choice of prior distributions affects the Bayes factor.

8. **Sequential Testing:** Design a sequential test for comparing two proportions in an A/B testing scenario. Explain how sequential monitoring affects the overall Type I error rate and discuss methods for controlling this inflation.