Statistical Thinking

Wei Lin

Lecture Notes 2025 Spring

Twist Shan, Xuan Chen, Quan Chen

June 15, 2025

Course Information

- Homework (30%), final paper (30%), final exam (40%)
- Three homework
- Final exam: 20250616

Boxes in Notes

Definition

Red Box (Definition) — This box is used to highlight **concepts**, **formal definitions**, and **specialized terminology** when they first appear. It serves as a glossary for precise understanding.

Theorem

Blue Box (Theorem) — This box is reserved for all mathematical results such as **Theorems**, **Propositions**, **Lemmas**, and other formal statements that require proof.

Example

Green Box (Example) — This box contains **worked examples**, **illustrative problems**, or small case studies that help concretely demonstrate abstract ideas.

Remark

Yellow Box (Remark) — This box is used for **comments**, **intuitions**, **warning points**, or things to keep in mind that aid deeper understanding but are not part of the formal derivation.

Contents

Lecture 1 Introduction

This lecture is based on the Chap. 1 of [Pol23].

Goals of statistics:

- Describe: The world is complex and we often need to describe it in a simplified way that we can understand.
- Decide: We often need to make decisions based on data, usually in the face of uncertainty. For example, inference and decision making.
- Predict: We often wish to make predictions about new situations based on our knowledge of previous situations. For example, machine learning and generalization.

Fundamental concepts of statistics:

- Learning from data.
- Aggregation: grouping, intercompostion (ANOVA) and regression (Cox).
- Uncertainty.
- Sampling.

Lecture 2 Data

This lecture is based on the Chap. 2-4 of [Pol23].

Definition

The word *data* is the plural of *datum*, coming from the Latin word *dare* which means give: give facts (quantity, quality) for future inference.

Types of data:

- qualitative (categorical) : describe a quality rather than a numeric quantity, i.e. nominal and ordinal.
- quantitative: discrete or continuous.

Definition

From data to information, we need aggregation. Idea of aggregation date is throwing away details of data to better summarize information:

$$\textbf{Data(full details)} = \textbf{Signal(deterministic patterns)} + \textbf{Noise(random)}.$$

Summarizing Data

To understand and visualize data, we often begin by summarizing it using frequency-based methods.

- **Absolute frequencies** can be converted into **relative frequencies**, providing a normalized view of the data distribution.
- **Histogram** is a common tool for displaying the frequency of data within specified intervals called **bins**. It shows the count (or relative frequency) of observations in each bin as vertical bars.
  - Narrow bins can produce **spikes** or jagged shapes.
  - Smoothing the histogram leads to **density estimation**, which approximates the underlying **distribution** (e.g., Normal, Gaussian, or mixture models).
- **Distribution shapes** provide key insights:
  - **Skewness**: Asymmetry in the data, often indicating different modes or shifts in the population. Skewed distributions can hint at underlying structure or subpopulations.
  - **Long-tailed distributions**: These have outliers or extreme values that stretch far from the center. They can influence statistical summaries like the mean or standard deviation.

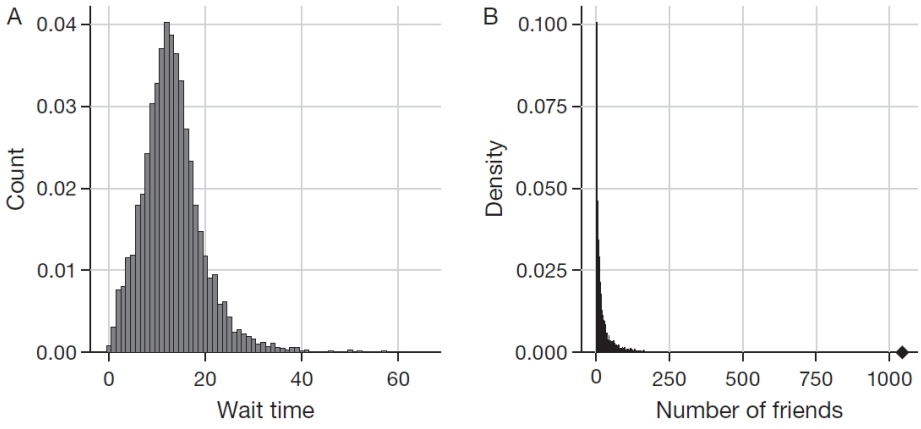

Figure 1: Examples of right-skewed and long-tailed distributions

Example

Example: Qualitative vs. Quantitative Variables
For each of the following variables, identify whether it is qualitative or quantitative:

- **The team memberships of a set of basketball players**: **Qualitative** — describes categorical groupings (teams).
- **The number of elections won by a politician**: **Quantitative** — a countable numerical value.
- **The version number of a car (for example, the BMW 330)**: **Qualitative** — despite having numbers, it acts as a label or identifier, not a measurement.
- **The proportion of students who have taken a statistics class**: **Quantitative** — a numerical value between 0 and 1, measurable.
- **The name of the textbook used for statistics classes across the country**: **Qualitative** — names are categorical variables.
- **Number of Twitter followers**: **Quantitative** — numerical count.
- **College major**: **Qualitative** — categories of academic fields.
- **Amount of daily rainfall**: **Quantitative** — measurable continuous quantity.
- **Presence/absence of disease antibodies**: **Qualitative** — categorical (binary: present/absent).

Data visualization

The way to group data:

- **Bar plot**: sample mean. Used when only one data point.
- **Beesuarm plot (bar plot with points)**: distribution, but hard to see due to the large number of data.
- **Violin plot**: distribution.
- **Box plot**: highlights the spread of the distribution along with any outliers.

Remark

It's better to make the visualization with more friendly color, and be friendly printed, photo-copy safe.

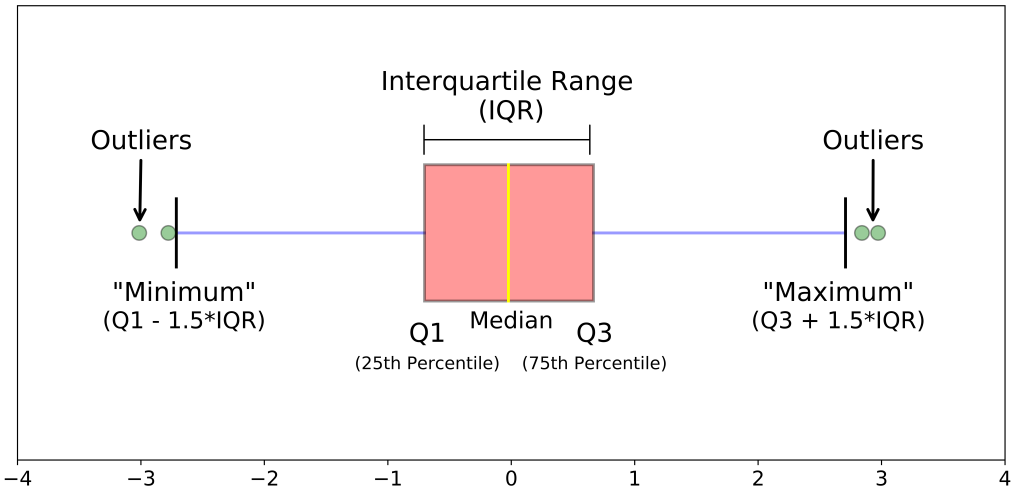

Figure 2: Understanding the boxplot

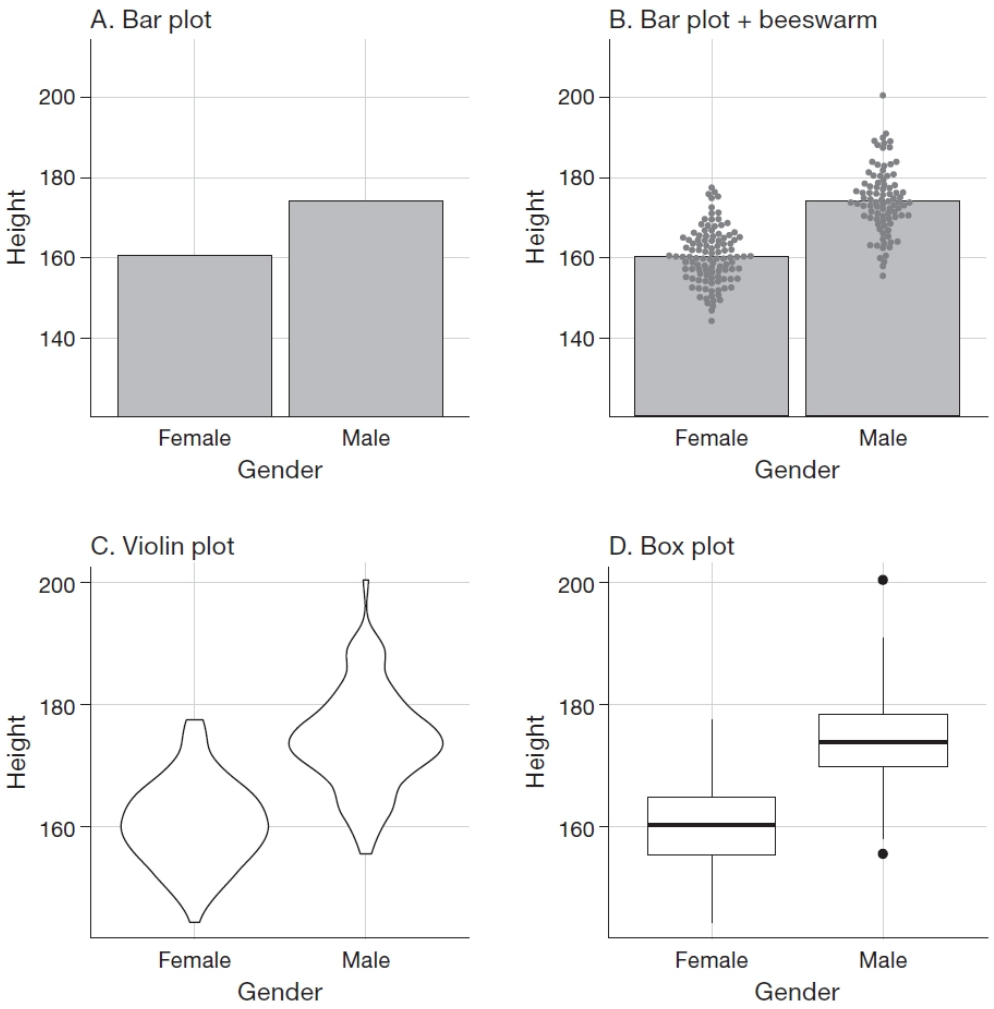

Figure 3: Four different ways of plotting

Principles of good visualization:

- Show the data and make them stand out.
- Maximize the $\frac{\text{data}}{\text{ink}}$ ratio.
- Avoid chartjunk and avoid distorting the data.
- Use the same font as possible: sans serif.

Example

Example: Why are pie charts particularly bad for data visualization?
Pie charts are often discouraged in data visualization due to the following reasons:

- They rely on our ability to perceive the **relative volume** (area of slices), which is cognitively difficult and inaccurate.
- They rely on our ability to **distinguish colors**, which can be problematic for colorblind viewers or when colors are too similar.
- They **do not allow correction for external factors** such as inflation or population size, limiting analytical depth.
- They may require the viewer to **remember color-to-category mappings** from a legend, which increases cognitive load and reduces readability.

Lecture 3 Benford's Law and Data Auditing

This lecture is based on [Hil95], [LSE00] and [TdCP17].

Benford's Law

In many naturally occurring collections of numbers, the leading digit is likely to be small.

Theorem

Benford's Law (First-Digit Law):
Let $D \in \{1, 2, \ldots, 9\}$ denote the first significant digit of a positive real number. Then the probability that the first digit equals d is given by

$$P(D = d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, \ldots, 9.$$

Then we can find Benford random variable c.d.f. of D:

$$F_D(d) = \sum_{k=1}^{d} \log_{10}(1 + \frac{1}{k}) = \log_{10}(d + 1).$$

This implies that lower digits occur much more frequently as the leading digit. For example,

$$P(D = 1) \approx 0.301, \quad P(D = 9) \approx 0.046.$$

Now that we have stated Benford's Law, how can we check whether a given distribution of T truly follows it? The key observation is to write

$$\log_{10} T = L + Z, \quad L = \lfloor \log_{10} T \rfloor \in \mathbb{Z}, \quad Z = \{\log_{10} T\} \in [0, 1).$$

In this form, the first digit D of T is simply $\lfloor 10^Z \rfloor$. Therefore, to verify that T satisfies Benford's Law, it suffices to show

$$Z = \{\log_{10} T\} \ \sim \ \text{Uniform}(0, 1).$$

Proof. Observe that $t' = 10^Z \in [1, 10)$ and $D = \lfloor t' \rfloor = \lfloor 10^Z \rfloor$. Hence for each $d = 1, \ldots, 9$,

$$\{D = d\} = \{\log_{10} d \le Z < \log_{10}(d + 1)\},$$

and if $Z \sim U(0, 1)$, then

$$P(D = d) = \log_{10}(d + 1) - \log_{10} d = \log_{10}\Big(1 + \frac{1}{d}\Big),$$

which completes the verification of Benford's Law. □

Example

Example: Triangular Distribution
Let $W \sim \text{Triangular}(0, 1, 2)$, with density

$$f_W(w) = \begin{cases} w, & 0 < w < 1, \\ 2 - w, & 1 \le w < 2. \end{cases}$$

Define $T = 10^W$, and consider the random variable

$$Z = \log_{10} T - \lfloor \log_{10} T \rfloor = W - \lfloor W \rfloor.$$

We want to show $Z \sim \text{Uniform}(0, 1)$, i.e., its CDF is $F_Z(z) = z$ on $[0, 1]$.

Note that $W \in (0,2)$ implies $T \in (1, 10^2)$. We split the interval into two cases:

$$\begin{aligned} F_Z(z) &= \sum_{\ell=-\infty}^{\infty} P\left(10^\ell \le T < 10^{\ell+1}\right) \cdot P\left(Z \le z \,\middle|\, 10^\ell \le T < 10^{\ell+1}\right) \\ &= P(1 \le T < 10) \cdot P(W \le z \mid 1 \le T < 10) \\ &\quad + P(10 \le T < 10^2) \cdot P(W - 1 \le z \mid 10 \le T < 10^2) \\ &= \frac{1}{2}\int_0^z w\, dw + \frac{1}{2}\int_z^1 (2-w)\, dw \\ &= \frac{1}{2} \cdot \frac{z^2}{2} + \frac{1}{2} \cdot \left[(2z - \frac{z^2}{2})\right] \\ &= \frac{1}{4}z^2 + z - \frac{1}{4}z^2 = z. \end{aligned}$$

Hence, $Z \sim \text{Uniform}(0,1)$, and so $T = 10^W$ satisfies Benford's Law.

Measures of Goodness

To assess whether an observed digit distribution $p = (p_1, \ldots, p_9)$ conforms to the expected Benford probabilities $q = (q_1, \ldots, q_9)$, we use goodness-of-fit measures. Two commonly used statistics are:

(1) Chi-Square Test:

$$\chi^2 = \sum_{j=1}^{9} \frac{(p_j - q_j)^2}{q_j}$$

This statistic measures the overall discrepancy between observed and expected frequencies, scaled by the expected count. It is often used with a null hypothesis $H_0 : p = q$, and the test statistic is compared to a chi-square distribution with 8 degrees of freedom.

(2) Maximum Norm (Uniform Distance):

$$\|p - q\|_\infty = \max_{1 \le j \le 9} |p_j - q_j|$$

This measures the largest single-digit deviation from the expected Benford probabilities. It is simple and interpretable, useful for identifying the most deviant digit.

Data Auditing

In practice, data auditing often involves checking whether observed categorical frequencies conform to expected distributions. Two widely used tools are:

(1) Chi-Square Test:

The chi-square goodness-of-fit test evaluates the overall deviation between observed counts x_j and expected counts np_j. The test statistic is

$$\chi^2 = \sum_{j=1}^{k} \frac{(x_j - np_j)^2}{np_j},$$

where:

- x_j is the observed count in category j,
- p_j is the expected probability of category j,
- $n = \sum_j x_j$ is the total number of observations.

Large values of χ^2 indicate significant deviation from the expected distribution.

(2) Pearson Residuals:

To identify which categories contribute most to the deviation, we examine the standardized residuals:

$$r_j = \frac{x_j - np_j}{\sqrt{np_j}}.$$

These are called Pearson residuals. A residual r_j far from zero (e.g., $|r_j| > 2$) suggests that the observed count in category j deviates significantly from expectation.

Lecture 4 Fitting Models to Data

This lecture is based on the Chap. 5 of [Pol23] and [Leh90].

Statistical Model

In the physical world, a **model** is a simplified representation of reality. Similarly, in statistics, a model provides a simplified description of how data are generated. Rather than capturing all complexity, a statistical model aims to capture the essential structure of the data-generating process.

Definition

A **statistical model** is a simplified description of how the data are generated. Formally, it is a family of probability distributions defined on a sample space.
Mathematically, a statistical model is written as:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where Θ is the parameter space and each P_θ is a probability distribution over the data space $\mathcal{X}$. The function

$$P : \Theta \to \mathcal{P}(\mathcal{A})$$

is called a **parametrization**, mapping parameters to distributions on the sample space $\mathcal{A}$.

Two extremes:

(i) **Completely known distribution:** $\Theta = \{\theta_0\}$. There is only one possible model; it is rigid but simple.

(ii) **Assumption-free model:** The set of all possible distributions $\mathcal{P}(\mathcal{X})$, which is highly flexible but difficult to work with (e.g., hard to normalize or estimate).

Remark

There is a tradeoff between flexibility and simplicity. Parametric models are easier to handle but may impose strong assumptions; nonparametric models are more flexible but often harder to estimate.

Model and Error Decomposition

The basic structure of a statistical model is:

$$\text{data} = \text{model} + \text{error}$$

This equation expresses the idea that any observed data can be decomposed into two parts:

- The portion explained by the model — representing our systematic understanding of the data-generating process.
- The error — representing randomness, noise, or aspects the model cannot explain.

For a specific observation i, we write:

$$\widehat{\text{data}}_i = \text{model}_i$$

where the hat denotes an estimate or prediction based on the model. The error is then:

$$\text{error}_i = \text{data}_i - \widehat{\text{data}}_i$$

This formalizes the idea that statistical modeling is about capturing the signal (model) and acknowledging the noise (error) inherent in real-world data.

Types of Models

- **Parametric models:** The parameter space $\Theta \subset \mathbb{R}^d$ is finite-dimensional. These models make strong assumptions about the data-generating distribution. Example: normal distribution with parameters μ, σ^2.
- **Nonparametric models:** The parameter space $\Theta \subset \mathcal{F}$ is infinite-dimensional. These models place minimal assumptions on the form of the distribution and are more flexible.
- **Semiparametric models:** The parameter $\theta = (\theta_1, \theta_2)$, where $\theta_1 \in \mathbb{R}^d$ is finite-dimensional and of primary interest, while $\theta_2 \in \Theta_2 \subset \mathcal{F}$ is infinite-dimensional and considered a **nuisance parameter**.

Definition

Properties of a Model (Identifiability)

A model is identifiable if

$$P_\theta = P_{\theta'} \quad \Rightarrow \quad \theta = \theta'.$$

That is, distinct parameter values should correspond to distinct probability distributions.

Example

Example: Identifiable Model (Normal Distribution) and Non-Identifiable Model (Symmetric Mixture)

- Consider the model where $X \sim \mathcal{N}(\mu, \sigma^2)$, and both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.
  Then the parameter (μ, σ^2) uniquely determines the distribution:

  $$\mathcal{N}(\mu, \sigma^2) \neq \mathcal{N}(\mu', \sigma'^2) \quad \text{if } (\mu, \sigma^2) \neq (\mu', \sigma'^2).$$

  Therefore, the model is **identifiable**.

- Consider the mixture model:

  $$f(x; \theta) = \frac{1}{2}\mathcal{N}(x; \theta, 1) + \frac{1}{2}\mathcal{N}(x; -\theta, 1),$$

  where $\theta \in \mathbb{R}$.

  Then for all θ,

  $$f(x; \theta) = f(x; -\theta),$$

  so the same distribution is produced by both θ and $-\theta$.

  Hence, the model is **not identifiable**, because we cannot distinguish between θ and $-\theta$ based on the data.

Performance Measures

To evaluate model fit, common error-based metrics include:

- **MSE (Mean Squared Error):**

  $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **SSE (Sum of Squared Errors):**

  $$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **RMSE (Root Mean Squared Error):**

  $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Example

Example: Children's Height Modeling
We want to model the height of children using NHANES data. The simplest model assumes all children have the same height:

$$y_i = \beta + \varepsilon_i$$

We estimate β using least squares, and the optimal choice is the sample mean $\bar{y}$. However, this model ignores age and gender. To improve the model, we gradually include more predictors:

- **Model 1 (Constant model):**

  $$\hat{y}_i = \hat{\beta}, \quad \text{same for all } i$$

  This gives a root mean squared error (RMSE) of about 39.0 cm.

- **Model 2 (Linear in age, no intercept):**

$$\hat{y}_i = \hat{\beta}_1 \cdot \text{Age}_i$$

This model underfits because when Age $= 0$, $\hat{y}_i = 0$, which is not realistic. RMSE is around 39.16 cm.

- **Model 3 (Linear in age with intercept):**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i$$

This significantly improves the fit. RMSE drops to around 27 cm.

- **Model 4 (Add gender):**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{Sex}_i$$

This yields slightly better performance (RMSE 26 cm), especially post-puberty.

This example shows how gradually adding relevant predictors improves model accuracy, moving from underfitting toward better generalization.

Model improvement techniques:

- Add more predictors (features)
- Allow interaction terms
- Use transformed or constructed features
- Avoid the **curse of dimensionality** by using additive models:

$$y_i = f(x_i) + \varepsilon_i \quad \Rightarrow \quad y_i = \beta_0 + \sum_{k=1}^{p} f_k(x_{ik}) + \varepsilon_i$$

- Consider varying coefficient models where effects change with conditions

What Is a Good Model?

A good model balances:

- **Underfitting** vs. **Overfitting**
- **Prediction error minimization**
- **Generalization** to new, unseen data

Overly simple models may fail to capture the structure (underfit), while overly complex models may capture noise (overfit). Good modeling practice involves choosing a model with low expected prediction error on future data.

Example

Example: Model Fit and Overfitting
Q1. The model that fits a particular dataset best (i.e., with the lowest sum of squared errors) is generally also the model that fits a new dataset best. True or false?

Answer: False

Explanation: A model that overfits may achieve a perfect fit on the training data but perform poorly on new data. The best training fit does not guarantee generalization.

Q2. Which of these concepts is most directly relevant to the previous question?

(A) Overfitting (B) Degrees of freedom (C) Variability (D) Standardized scores

Answer: (A)

Explanation: Overfitting refers to a model fitting noise rather than the underlying structure, explaining why perfect training performance may harm future predictive performance.

Formal Theory of Statistical Models

Statistical Experiment and Model Structure

A statistical experiment (either observational or controlled) can be formally described by a triplet:

$$(\mathcal{U}, \Omega, V)$$

where:

- $\mathcal{U}$: the set of statistical units (e.g., individuals in a population);
- Ω: covariate space (input features or design variables);
- V: response scale (possible output values, labels, or measurement space).

This decomposition forms the basis for formally reasoning about models, as it separates the structure of data generation from the particular statistical methods applied.

Definition

Model Logic

A key principle in formal statistical modeling is:

> A model is **logically sensible** if the meaning of its parameters does not depend on accidental design choices — such as the sample size, the layout of the data table, or the design of the experiment.

In other words, parameters should retain their interpretation under data transformations and should not be tied to specific datasets. This leads naturally into a category-theoretic formalization.

Category-Theoretic Perspective

Category theory provides a structural and abstract framework for understanding the relationships between data structures and models. A category $\mathcal{C}$ consists of:

- A collection of **objects** (e.g., sets $\Omega, \Omega', \ldots$);
- A collection of **morphisms** $\varphi : \Omega \to \Omega'$ (arrows), satisfying:
  (1) Identity: $\mathrm{id}_\Omega : \Omega \to \Omega$ is a morphism in $\mathcal{C}$;
  (2) Composition: If $\varphi : \Omega \to \Omega'$ and $\psi : \Omega' \to \Omega''$, then $\psi \circ \varphi : \Omega \to \Omega''$ is a morphism in $\mathcal{C}$.

Example

Example: Box–Cox Transformation as a Morphism
The Box–Cox transformation is a parametric family of functions used to stabilize variance and make data more Gaussian-like. It is defined as:

$$\varphi_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

This transformation maps the covariate space $\Omega \to \Omega'$, and under the category-theoretic view, each φ_λ is a morphism in the category of covariate spaces that preserves the modeling structure.

Example

Example: Z-score Standardization as a Morphism
Z-score normalization transforms a variable $x \in \Omega$ to:

$$Z(x) = \frac{x - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of the dataset.
This transformation centers and scales the data to have mean 0 and variance 1. As a mapping $\varphi : \Omega \to \Omega'$, it is a morphism that preserves relative distances and enables comparison across variables or datasets. In categorical terms, it is a structure-preserving arrow that maintains the interpretability of statistical parameters.

Functor: A functor $T : \mathcal{C} \to \mathcal{K}$ is a map between categories that preserves the morphism structure. It is either:

- **Covariant:** $T(\psi \circ \varphi) = T(\psi) \circ T(\varphi)$;
- **Contravariant:** $T(\psi \circ \varphi) = T(\varphi) \circ T(\psi)$.

A Statistical Model as a Diagram of Categories

A formal statistical model can be represented as a system of morphisms in three coordinated categories:

- **Category of Units:** $\mathcal{U}$, with injective maps preserving units (subsets, ID mappings);
- **Category of Covariates:** Ω, with structure-preserving injective transformations;
- **Category of Responses:** V, with typically surjective mappings capturing the outcome mechanisms.

These mappings induce a diagram of compatible morphisms representing the model:

$$\begin{array}{ccc} \mathcal{U} & \xrightarrow{\varphi^{\mathcal{U}}} & \mathcal{U}' \\ {\scriptstyle P(\cdot|\Omega)}\downarrow & & \downarrow{\scriptstyle P(\cdot|\Omega')} \\ V & \xrightarrow{\varphi^{V}} & V' \end{array}$$

Here $P(\cdot|\Omega)$ denotes the probability law of the model conditioned on the covariates Ω, and the diagram asserts compatibility under covariate and response transformations.

Remark

This abstract formulation emphasizes that the statistical model should respect the invariance and consistency of meaning across data transformations, enhancing robustness and interpretability.

Lecture 5 Bias Variance Trade-off

This lecture is based on the Secs. 7.2, 7.3 of [HTF09] and [Bre01].

Bias, Variance and Model Complexity

In supervised learning, we aim to learn a prediction function $\hat{f}(X)$ based on a training dataset $\mathcal{T}$, to predict a target variable Y. The accuracy of the prediction is measured by a loss function $L(Y, \hat{f}(X))$, such as:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{(squared error)} \\ |Y - \hat{f}(X)| & \text{(absolute error)} \end{cases}$$

Test error, or generalization error, is the expected loss on new, unseen data:

$$\mathrm{Err}_{\mathcal{T}} = \mathbb{E}[L(Y, \hat{f}(X)) \mid \mathcal{T}]$$

Averaging over all possible training sets gives the expected test error:

$$\mathrm{Err} = \mathbb{E}_{\mathcal{T}}[\mathrm{Err}_{\mathcal{T}}]$$

Training error is the average loss over the training data:

$$\overline{\mathrm{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

As model complexity increases, training error tends to decrease, but test error may increase due to overfitting. This leads to the bias–variance trade-off: more complex models reduce bias but increase variance. The goal is to find an optimal model complexity that minimizes the expected test error.

In practice, we often choose a model with tuning parameter α, and write $\hat{f}_\alpha(x)$. The aim is to find α that minimizes the expected test error.

There are two main tasks:

- **Model selection**: Compare different models to pick the best one.
- **Model assessment**: Estimate the generalization error of the final model.

If data is abundant, we can split it into training, validation, and test sets (e.g., 50%-25%-25%). Otherwise, we may use cross-validation or bootstrap methods.

Bias-Variance Decomposition

Suppose the data follows the model:

$$Y = f(X) + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$$

Let $\hat{f}(x)$ be the estimator trained on a dataset $\mathcal{T}$. The expected prediction error at a point x_0 under squared error loss is:

$$\mathrm{Err}(x_0) = \mathbb{E}[(Y - \hat{f}(x_0))^2 \mid X = x_0]$$

Theorem

Bias-Variance Decomposition:
The prediction error at x_0 can be decomposed as:

$$\mathrm{Err}(x_0) = \sigma_\varepsilon^2 + \Big(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\Big)^2 + \mathrm{Var}(\hat{f}(x_0))$$

That is:

$$\text{Prediction Error} = \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}$$

Proof. We compute the mean squared error:

$$\mathrm{MSE} = \mathbb{E}_{\mathcal{T},\varepsilon}[(Y - \hat{f}(x))^2] = \mathbb{E}[(f(x) + \varepsilon - \hat{f}(x))^2]$$

Expand the square:

$$\mathbb{E}[(f(x) - \hat{f}(x))^2] + 2\mathbb{E}[(f(x) - \hat{f}(x))\varepsilon] + \mathbb{E}[\varepsilon^2]$$

Since ε is independent of $\hat{f}(x)$ and has zero mean, the cross term vanishes, leaving:

$$= \mathbb{E}[(f(x) - \hat{f}(x))^2] + \sigma^2$$

Now decompose the first term:

$$\begin{aligned}
\mathbb{E}[(f(x) - \hat{f}(x))^2] &= \mathbb{E}\left[\left(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right] \\
&= \mathbb{E}\left[(f(x) - \mathbb{E}[\hat{f}(x)])^2 + 2(f(x) - \mathbb{E}[\hat{f}(x)])(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)) + (\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2\right] \\
&= (f(x) - \mathbb{E}[\hat{f}(x)])^2 + 2(f(x) - \mathbb{E}[\hat{f}(x)]) \cdot \mathbb{E}[\mathbb{E}[\hat{f}(x)] - \hat{f}(x)] + \mathbb{E}[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2]
\end{aligned}$$

Note that the second term vanishes:

$$\mathbb{E}[\mathbb{E}[\hat{f}(x)] - \hat{f}(x)] = \mathbb{E}[\mathbb{E}[\hat{f}(x)]] - \mathbb{E}[\hat{f}(x)] = 0$$

So we are left with:

$$= \underbrace{(f(x) - \mathbb{E}[\hat{f}(x)])^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]}_{\text{Variance}}$$

Thus,

$$\mathbb{E}[(f(x) - \hat{f}(x))^2] = \text{Bias}^2 + \text{Variance}$$

So:

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \sigma^2$$

□

Interpretation:

- **Bias:** How far the average model prediction is from the true function.
- **Variance:** How much the model prediction fluctuates around its average.
- σ^2: Irreducible error from noise in the data.

The bias-variance trade-off explains why overly complex models (low bias, high variance) or overly simple models (high bias, low variance) both perform poorly. The optimal model minimizes the total error.

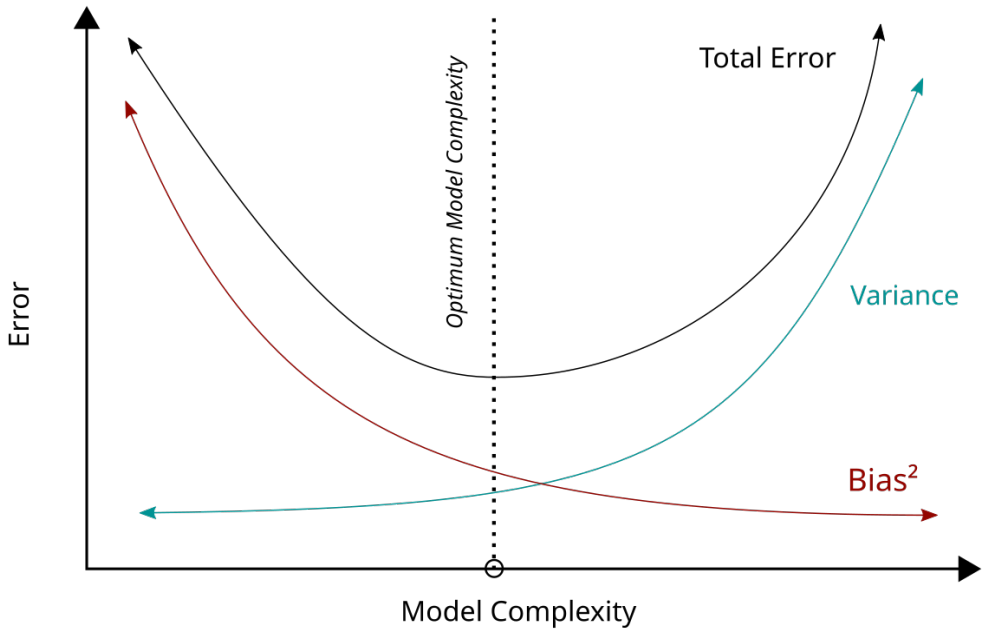

Figure 4: Bias and Variance Trade-off

Approximation-Estimation Decomposition

In practice, we often restrict ourselves to a function class $\mathcal{F}$ (e.g., linear functions), which might not contain the true function f. In this case, we define f^* as the best approximation to f within $\mathcal{F}$:

$$f^* = \arg\min_{g \in \mathcal{F}} \|f - g\|^2$$

That is, f^* is the projection of f onto the model space $\mathcal{F}$. Our learned estimator $\hat{f}$ is a random function in $\mathcal{F}$, depending on the training data.

Theorem

Approximation–Estimation Decomposition:
For any learned model $\hat{f} \in \mathcal{F}$, the squared bias decomposes as:

$$\mathbb{E}\left[f(x_0) - \hat{f}(x_0)\right]^2 = \underbrace{[f(x_0) - f^*(x_0)]^2}_{\text{Approximation Error}} + \underbrace{(f^*(x_0) - \mathbb{E}[\hat{f}(x_0)])^2}_{\text{Estimation Bias}^2} + \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{Variance}}$$

Accordingly, the total prediction error satisfies:

$$\text{Prediction Error} = \text{Approximation Error} + \text{Estimation Error} + \text{Variance} + \sigma^2$$

Proof. Decompose the squared term by inserting two intermediate terms $f^*(x_0)$ and $\mathbb{E}[\hat{f}(x_0)]$:

$$\begin{aligned}
\mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] &= \mathbb{E}\left[(f(x_0) - f^*(x_0)) + (f^*(x_0) - \mathbb{E}[\hat{f}(x_0)]) + (\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right]^2 \\
&= [f(x_0) - f^*(x_0)]^2 + [f^*(x_0) - \mathbb{E}[\hat{f}(x_0)]]^2 + \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] \\
&\quad + 2(f(x_0) - f^*(x_0))(f^*(x_0) - \mathbb{E}[\hat{f}(x_0)]) \\
&\quad + 2(f(x_0) - f^*(x_0)) \cdot \mathbb{E}[\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)] \\
&\quad + 2(f^*(x_0) - \mathbb{E}[\hat{f}(x_0)]) \cdot \mathbb{E}[\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)]
\end{aligned}$$

Next, we explain why each of the three cross terms vanishes under standard assumptions:

- **Cross Term 1:**
$$2(f(x_0) - f^*(x_0))(f^*(x_0) - \mathbb{E}[\hat{f}(x_0)])$$
This is a deterministic inner product. Since f^* is the orthogonal projection of f onto the function space $\mathcal{F}$, and $\mathbb{E}[\hat{f}] \in \mathcal{F}$, we have:
$$\langle f - f^*, f^* - \mathbb{E}[\hat{f}] \rangle = 0$$
So this cross term equals zero.

- **Cross Term 2 and 3:** We use the linearity of expectation:
$$\mathbb{E}[\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)] = \mathbb{E}[\mathbb{E}[\hat{f}(x_0)]] - \mathbb{E}[\hat{f}(x_0)] = 0$$
So this term is zero.

Hence,

$$\mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] = \underbrace{[f(x_0) - f^*(x_0)]^2}_{\text{Approximation Error}} + \underbrace{[f^*(x_0) - \mathbb{E}[\hat{f}(x_0)]]^2}_{\text{Estimation Bias}^2} + \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{Variance}}$$

□

Summary:

- **Approximation error:** The error from using a restricted function class.
- **Estimation error:** The error due to data randomness when estimating the best function in $\mathcal{F}$.
- These two sources of bias are orthogonal if $\mathcal{F}$ is a vector space.

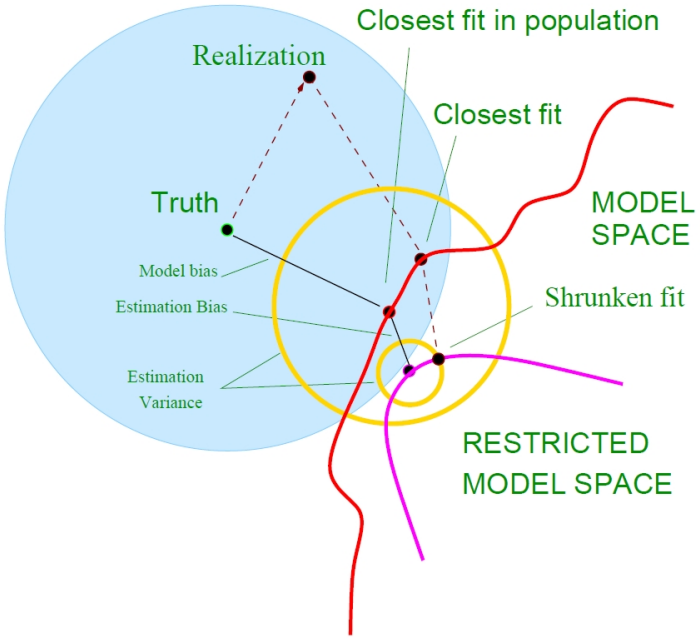

Figure 5: Schematic of the behavior of bias and variance.

Data Model versus Algorithmic Models

In statistical modeling, we distinguish between two main approaches to modeling the relationship between inputs x and outputs y:

- **Data models** assume a specific parametric form for the relationship, such as linear regression, logistic regression, or the Cox proportional hazards model. These models rely on statistical assumptions and focus on inference and interpretability.
- **Algorithmic models** treat the relationship $x \mapsto y$ as an unknown mapping, to be learned via flexible predictive algorithms such as decision trees, random forests, or neural networks. These models prioritize prediction over interpretability.

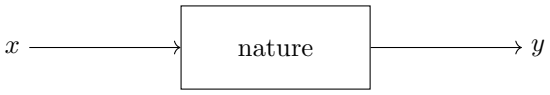

Figure 6: Real process: $x \to y$

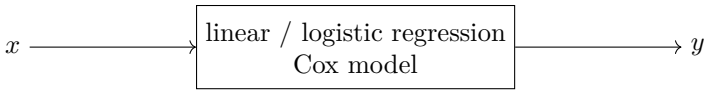

Figure 7: Data model: specified functional form $x \to y$

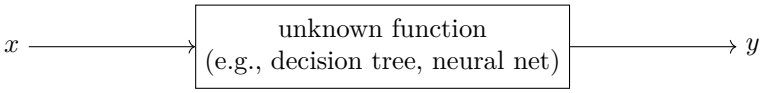

Figure 8: Algorithmic model: learned mapping $x \to y$

Model Validation
For data models, model validity is often assessed via:

- Goodness-of-fit tests
- Residual analysis (e.g., whether residuals are centered, homoscedastic, etc.)

Limitations of data models:

- Lack of power in goodness-of-fit tests when the model space is too large
- Fragility of statistical conclusions
- Curse of dimensionality

For algorithmic models, performance is typically assessed via prediction accuracy on unseen data:

- **Cross-validation:** involves training and testing on different subsets of the data, but may introduce bias due to data re-use
- **Independent test set:** preferred when available, as it gives an unbiased estimate of generalization error

Rashomon Effect

The Rashomon effect describes the situation where many different models achieve similarly good predictive performance, yet they may rely on entirely different mechanisms, structures, or feature combinations.

That is, even though the loss function or accuracy is nearly identical across models, their internal logic or interpretations can diverge significantly.

This presents a major challenge for scientific understanding and causal inference, where the goal is not just to predict well, but to explain how the system works.

Implications:

- There may not be a single "true" model—multiple plausible explanations may coexist.
- Interpretation becomes unstable: choosing one model may lead to completely different conclusions than choosing another.
- Model selection based solely on prediction accuracy may obscure deeper understanding.

This effect is especially pronounced in high-dimensional settings or with flexible model classes (e.g., ensembles, neural networks).

Occam's Razor

Occam's razor is the principle that, among competing models with similar performance, the simplest one should be preferred. Simplicity often leads to better interpretability and generalizability.

In practice, this gives rise to a trade-off between:

- **Predictive accuracy** — often favoring more complex models that better fit the data (e.g., chosen by AIC).
- **Model simplicity / interpretability** — favoring sparser models that are easier to explain (e.g., selected by BIC).

This is known as the AIC–BIC dilemma:

- AIC (Akaike Information Criterion) tends to select models with better predictive performance.
- BIC (Bayesian Information Criterion) tends to select simpler models that are more plausible from a generative perspective.

Conclusion: There is no universally best choice. The balance between prediction and interpretation should be guided by the goal of the analysis.

Lecture 6 Frequentist Inference

This lecture is based on the Chap. 2 of [EH16].

Frequentist Interpretation of Probability

The frequentist view defines probability as the long-run frequency of events in repeated trials. Statistical inference is about learning a fixed but unknown parameter θ of a population distribution F, using a random sample.

Suppose we observe data $x = (x_1, \ldots, x_n)$, modeled as i.i.d. draws from F:

$$X = (X_1, \ldots, X_n) \sim F, \quad \text{written as } F \to X.$$

We aim to infer a population property, e.g. the mean:

$$\theta = \mathbb{E}_F\{X\}.$$

A natural estimator is the sample average:

$$\hat{\theta} = \bar{x} = \frac{1}{n}\sum x_i = t(x).$$

To evaluate accuracy, we treat $\hat{\theta}$ as a realization of the random variable:

$$\hat{\Theta} = t(X),$$

whose behavior over repeated samples determines frequentist properties like:

$$\mu = \mathbb{E}_F[\hat{\Theta}], \quad \text{bias} = \mu - \theta, \quad \text{var} = \mathbb{E}_F[(\hat{\Theta} - \mu)^2].$$

Frequentism in practice

Q: How do we calculate population properties like θ, if F is unknown?

Frequentism evaluates statistical procedures by analyzing their behavior over hypothetical repeated samples. The key idea is: derive the probabilistic properties of an estimator $\hat{\Theta} = t(X)$, then apply them directly to the realized output $\hat{\theta} = t(x)$, even though the true distribution F is unknown.

Method 1: Plug-In Principle

Key idea: Replace the unknown population distribution F with the empirical distribution $\hat{F}_n$, and compute the quantity of interest by plugging in:

$$\hat{\theta}_{\text{plugin}} = T(\hat{F}_n).$$

Empirical distribution: Given a sample $X_1, \ldots, X_n$, the empirical distribution $\hat{F}_n$ assigns probability mass $1/n$ to each observed data point.

Example

Example: Plug-In Estimates

- If $\theta = \mathbb{E}[X]$ is the population mean, then the plug-in estimate is the sample mean:

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

- If $\theta = \text{Var}(X)$, then the plug-in estimate is the sample variance:

$$\widehat{\text{Var}}_F = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Why it works: By the law of large numbers, as $n \to \infty$, the empirical distribution $\hat{F}_n$ converges to the true distribution F, and thus $T(\hat{F}_n) \to T(F)$. That is, plug-in estimates become accurate for large n.

Method 2: Delta Method (Taylor Expansion)

Key idea: When we want to estimate the variance (or distribution) of a transformed statistic $g(\hat{\theta}_n)$, the Delta Method uses a first-order Taylor expansion to approximate it.

Taylor approximation: If $\hat{\theta}_n$ estimates θ, and $g(\cdot)$ is differentiable near θ, then:

$$g(\hat{\theta}_n) \approx g(\theta) + g'(\theta)(\hat{\theta}_n - \theta).$$

Variance approximation: If

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

then

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2\sigma^2).$$

Interpretation: A smooth transformation of an asymptotically normal estimator remains asymptotically normal. Its variance is scaled by the square of the derivative $g'(\theta)$. This method is commonly used to estimate standard errors of nonlinear functions of estimators and is widely used for constructing confidence intervals or standard errors for transformed parameters.

Example

Example: Delta method for squared sample mean
Let $\hat{\theta} = \bar{x}^2$, and recall $\frac{d}{d\bar{x}}(\bar{x}^2) = 2\bar{x}$. Using the Delta method:

$$\text{se}(\bar{x}^2) \approx 2|\bar{x}| \cdot \text{se}(\bar{x}).$$

This expresses the standard error of $\bar{x}^2$ in terms of the standard error of $\bar{x}$, using a local linear approximation.

Method 3: Parametric Models and MLE

Key idea: If we assume a parametric form F_θ for the distribution, then **maximum likelihood estimation (MLE)** is a standard frequentist approach for estimating θ.

Definition

Parametric model: Assume $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_\theta$, a family of distributions indexed by θ. For example, $F_\theta = \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$.
Maximum Likelihood Estimation: The MLE is defined as

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \prod_{i=1}^{n} p(X_i \mid \theta).$$

Properties (under regularity conditions):

- **Consistency:** $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_{\text{true}}$
- **Asymptotic normality:**
$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_{\text{true}}) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_{\text{true}})^{-1}\right),$$
where $I(\theta)$ is the Fisher information.

Method 4: Simulation and the Bootstrap

Key idea: The bootstrap is a resampling method. We approximate the unknown distribution F using the observed data (or a fitted parametric model), and simulate repeated samples from this approximation to assess estimator variability.

Basic nonparametric bootstrap procedure:

1. From the sample $\{X_1, \ldots, X_n\}$, draw a sample of size n with replacement.
2. Compute the statistic of interest $\hat{\theta}^*$ on this bootstrap sample.
3. Repeat steps 1–2 for B times to obtain $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$.

Bootstrap distribution: The empirical distribution of the $\hat{\theta}^*$'s estimates the variability of $\hat{\theta}$. This allows us to compute:

- Standard errors (e.g., via sample standard deviation of $\hat{\theta}^*$),
- Bias (difference between mean of $\hat{\theta}^*$ and $\hat{\theta}$),
- Confidence intervals (percentile or pivotal methods).

Parametric bootstrap: If a model F_θ is assumed, we can first compute $\hat{\theta}_{\text{MLE}}$, then simulate new samples from $F_{\hat{\theta}_{\text{MLE}}}$ instead of the empirical data.

Remark

Classical methods like plug-in, delta method, and MLE work best when $\hat{\theta} = t(x)$ is a smooth function. The bootstrap extends inference to more complex or irregular estimators, and is especially useful in modern computational statistics.

Method 5: Pivotal Statistics

Key idea: A pivotal statistic is a function of the sample and unknown parameter(s) whose distribution does not depend on any unknown quantities. Pivotal quantities are powerful tools for constructing exact (or asymptotically exact) confidence intervals and hypothesis tests.

Definition

Pivotal Statistic: A statistic $Q(X_1, \ldots, X_n; \theta)$ is called **pivotal** if its distribution does not depend on the unknown parameter θ. That is, the distribution of Q is the same for all values of θ.

Example

Example: Standardized Mean (Student's t-Statistic)
Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Then the statistic

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t-distribution with $n-1$ degrees of freedom, regardless of the values of μ and σ^2. This makes Z a pivotal quantity.

Confidence Interval Construction:
If the variance is known, we can directly invert a normal pivotal quantity:

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

which leads to an exact $1 - \alpha$ confidence interval for μ:

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

Example

Example: Two-Sample t-Test
We observe two samples: $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n_1})$, $\mathbf{x}_2 = (x_{21}, \ldots, x_{2n_2})$ and wish to test $H_0 : \mu_1 = \mu_2$.
We construct the test statistic

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\hat{s}_d}, \quad \text{where} \quad \hat{s}_d = \hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}$$

and pooled variance estimate

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

Then under H_0, the test statistic $t \sim t_{n_1+n_2-2}$, which is a pivotal distribution.

Advantages:

- Pivotal statistics do not depend on unknown parameters, so they allow construction of exact (finite-sample) confidence intervals or tests.
- They avoid the need for asymptotic approximations or plug-in variance estimation.

Good Inference Procedure

What makes a statistical inference procedure good?

- **Validity:** The procedure gives correct coverage or type I error rate under the assumed model. For example, a 95% confidence interval truly contains the parameter 95% of the time.
- **Optimality:** Among all valid procedures, the chosen one achieves the best performance — e.g., shortest confidence interval, lowest variance, most powerful test, etc.

Frequentist Optimization

Frequentist optimization refers to selecting an inference procedure (such as an estimator or test) that performs best according to a frequentist criterion, typically minimizing error or variance under repeated sampling from a model F.

The idea is to evaluate estimators or tests based on their long-run properties under a given model. Two classical forms of such optimality include:

- Maximum likelihood estimation (MLE), which minimizes (asymptotic) standard error;
- The Neyman–Pearson lemma, which provides the most powerful test between two simple hypotheses.

Example

Example: Neyman–Pearson Hypothesis Testing
Suppose we observe data x, and wish to choose between null $f_0(x)$ and alternative $f_1(x)$. A test decision rule is defined as $t(x) \in \{0, 1\}$, where:

- $t(x) = 1$: reject the null H_0;
- $t(x) = 0$: do not reject the null.

We define the two types of errors:

$$\alpha = \Pr_{f_0}\{t(x) = 1\} \quad \text{(Type I error)}, \quad \beta = \Pr_{f_1}\{t(x) = 0\} \quad \text{(Type II error)}.$$

Define the likelihood ratio:

$$L(x) = \frac{f_1(x)}{f_0(x)}.$$

The Neyman–Pearson lemma states that the most powerful test of size α is given by:

$$t_c(x) = \begin{cases} 1, & \text{if } \log L(x) \geq c \\ 0, & \text{if } \log L(x) < c \end{cases}$$

The cutoff c is determined to satisfy the constraint on Type I error:

$$\Pr_{f_0}(\log L(x) \geq c) = \alpha.$$

This test $t_c(x)$ minimizes the Type II error β among all tests with Type I error at most α, and is thus the most powerful level-α test.

Flaws in Frequentist Inference

While frequentist optimality theory has offered powerful tools for classical statistical problems, it faces limitations in more complex modern data settings. The framework often requires simplified assumptions and does not naturally adapt to flexible or data-driven algorithms.

Frequentist principles provide strong theoretical guarantees under idealized conditions (e.g., large-sample asymptotics, parametric models), but in real-world applications, the inference process may appear ad hoc or fragile. As a result, effective practice often involves supplementing frequentist reasoning with empirical tools, computational methods, and insights from Bayesian or algorithmic paradigms.

Lecture 7 Bayesian Inference

This lecture is based on the Chap. 3 of [EH16].

Bayesian Interpretation of Probability

Bayesian inference treats probability as a degree of belief. The core object is a **family of probability densities**:

$$\mathcal{F} = \{f_\mu(x) : x \in \mathcal{X}, \mu \in \Omega\}$$

where x is observed data in sample space $\mathcal{X}$, and μ is an unknown parameter in parameter space Ω.

In addition to $\mathcal{F}$, Bayesian inference requires a **prior density**:

$$g(\mu), \quad \mu \in \Omega$$

This represents knowledge or belief about μ before observing data.

Theorem

Bayes' Rule: Given prior $g(\mu)$ and likelihood $f_\mu(x)$, the posterior is:

$$g(\mu \mid x) = \frac{g(\mu) f_\mu(x)}{f(x)}, \quad \text{where} \quad f(x) = \int_\Omega f_\mu(x) g(\mu) d\mu$$

Bayes' Rule can be rewritten as:

$$g(\mu \mid x) = c_x \cdot L_x(\mu) \cdot g(\mu)$$

where $L_x(\mu) = f_\mu(x)$ is the likelihood function and c_x is a normalizing constant so that $g(\mu \mid x)$ integrates to 1.

For two parameter values $\mu_1, \mu_2 \in \Omega$, the posterior ratio is:

$$\frac{g(\mu_1 \mid x)}{g(\mu_2 \mid x)} = \frac{g(\mu_1)}{g(\mu_2)} \cdot \frac{f_{\mu_1}(x)}{f_{\mu_2}(x)}$$

i.e., *Posterior odds = Prior odds × Likelihood ratio*

Remark

Multiplying the likelihood by any constant depending only on x does not affect posterior shape since the constant is absorbed into normalization.

Uninformative Prior Distributions

Q: How to choose $g(\theta)$? Given $\theta \in (-1, 1)$

In practice, when prior information about parameters is scarce or nonexistent, we often seek an **uninformative prior**—a prior that minimizes influence on the posterior distribution.

Uniform (Flat) Prior

Assigns equal density across a bounded interval, e.g.,

$$g(\theta) = \frac{1}{2}, \quad \text{for } \theta \in [-1, 1]$$

This reflects ignorance but lacks invariance under reparameterization. For example, if $\theta = \log \gamma$, then

$$\Pr(\gamma > 1 \mid \hat{\theta}) \neq \Pr(\theta > 0 \mid \hat{\theta})$$

implying that flat priors can be sensitive to the chosen parameterization.

Jeffreys Prior (Invariant)

Based on the Fisher Information $\mathcal{I}_\theta$, this prior is defined as:

$$g(\theta) = \mathcal{I}_\theta^{1/2}$$

where

$$\mathcal{I}_\theta = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right)^2 \right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(x) \right]$$

This prior is invariant under reparameterizations $\theta \mapsto \gamma(\theta)$.

$$\tilde{\mathcal{I}}_\gamma = \left(\frac{\partial\theta}{\partial\gamma}\right)^2 \mathcal{I}_\theta \Rightarrow \tilde{g}(\gamma) = \left|\frac{\partial\theta}{\partial\gamma}\right| g(\theta)$$

Alternative view: Because $\mathcal{I}_\theta^{-1} \approx \text{Var}(\hat{\theta}_{\text{MLE}})$, another approximation is:

$$g(\theta) \propto \frac{1}{\sigma_\theta}$$

where σ_θ is the standard error of the MLE $\hat{\theta}$, e.g.,

$$\sqrt{n}(\theta - \hat{\theta}) \xrightarrow{d} N(0, (1-\theta^2)^2) = N(0, \sigma_\theta^2)$$

Other Examples

- **Triangular prior**: $g(\theta) = 1 - |\theta|$, for $\theta \in [-1, 1]$

Remark

Uniform priors can be misleading under reparameterizations. Jeffreys prior corrects this via Fisher information, yielding invariance and aligning with frequentist notions of precision.

Flaws in Bayesian Inference

While Bayesian methods offer coherence and robustness to sampling plans, they are sensitive to the choice of prior.

In high-dimensional settings, flat or uninformative priors (like the Jeffreys prior) can lead to poor performance, especially when applied jointly to multiple parameters.

Bayesian inference is immune to selection bias due to its reliance on the likelihood function, but it is not immune to prior misspecification.

As model complexity increases, the focus in Bayesian analysis shifts from posterior calculation to careful and possibly data-driven prior construction (e.g., empirical Bayes, hierarchical Bayes).

Frequentist versus Bayesian

- **Choice of Method/Prior:** Frequentists choose a test or method $t(x)$ for each specific problem; Bayesians choose a prior $g(\mu)$, which is fixed and updated with data.
- **All Questions vs. Specific One:** Bayesians can answer many questions simultaneously via the posterior; frequentists must tailor an estimator to each target.
- **Dynamic Updating:** Bayesian inference naturally supports sequential or online updating; frequentist methods may struggle with incrementality.
- **Objectivist vs. Subjectivist:** Frequentism aims at objectivity through repeated sampling logic; Bayesianism allows subjectivity via prior choice.
- **Regularization Analogy:** Bayesian priors act like regularization (e.g., Lasso, Ridge) in high-dimensional models, influencing posterior estimates.

Lecture 8 Fisherian Inference and MLE

This lecture is based on the Chap. 4 of [EH16].

Likelihood

The concept of **likelihood** lies at the heart of Fisherian inference. Suppose the data vector $x = (x_1, x_2, \ldots, x_n)$ is observed and fixed. For a parametric model $\{f_\theta(x)\}$, we define the likelihood function as:

$$L_x(\theta) = f_\theta(x)$$

It is often more convenient to work with the log-likelihood:

$$\ell_x(\theta) = \log L_x(\theta) = \log f_\theta(x)$$

The MLE of the parameter θ is defined as:

$$\hat{\theta} = \arg\max_{\theta \in \Omega} \ell_x(\theta)$$

MLEs are typically obtained by maximizing the log-likelihood over the parameter space. In many cases, the maximizer exists and is unique. When the parameter of interest is a function $\gamma = g(\theta)$, the MLE of γ is simply:

$$\hat{\gamma} = g(\hat{\theta})$$

This is known as the **invariance property** of MLEs.

The MLE procedure is widely used because of the following advantages:

- **Automatic:** A single numerical optimization routine typically suffices to compute $\hat{\theta}$. No special-case derivations are required, in contrast to unbiased estimators.
- **Easy to implement:** Especially for standard models, MLEs are readily computable using software or basic calculus.
- **Strong frequentist properties:** In large samples, MLEs are asymptotically unbiased and efficient — achieving the Cramér–Rao lower bound. Even in small samples, they often perform well.
- **Bayesian justification:** Under a flat prior, the MLE coincides with the mode of the posterior distribution. Specifically, from Bayes' rule:
  $$g(\theta \mid x) = c_x g(\theta) e^{\ell_x(\theta)}$$
  If $g(\theta)$ is constant, then $\hat{\theta}$ is also the MAP (maximum a portion) estimator.

Example

Example: Normal Distribution

Suppose $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$, where both θ and σ^2 are unknown. The joint likelihood function is:

$$L_x(\theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right)$$

Taking logarithms gives the log-likelihood:

$$\ell_x(\theta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2$$

Maximizing $\ell_x(\theta, \sigma^2)$ with respect to θ and σ^2 yields the following MLEs:

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

That is, the MLE of the mean is the sample average, and the MLE of the variance is the sample variance (without Bessel correction).

Fisher Information

Consider a one-parameter family of densities

$$\mathcal{F} = \{f_\theta(x) : \theta \in \Omega,\, x \in \mathcal{X}\}$$

where $\Omega \subseteq \mathbb{R}$ and $f_\theta(x)$ is differentiable in θ.

Definition

Score Function
Given a family of densities $f_\theta(x)$, the log-likelihood function is

$$\ell_x(\theta) = \log f_\theta(x)$$

The score function is the derivative of the log-likelihood with respect to the parameter θ:

$$\dot{\ell}_x(\theta) = \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{\dot{f}_\theta(x)}{f_\theta(x)}$$

Unbiasedness of Score: Under regularity conditions (permitting differentiation under the integral sign), the expectation of the score function is zero:

$$\mathbb{E}_\theta[\dot{\ell}_x(\theta)] = \int \dot{\ell}_x(\theta) f_\theta(x) dx = \frac{\partial}{\partial\theta} \int f_\theta(x) dx = \frac{\partial}{\partial\theta} 1 = 0$$

Definition

Fisher Information
The Fisher information at parameter value θ is defined as the variance of the score function:

$$\mathcal{I}_\theta = \mathrm{Var}_\theta[\dot{\ell}_x(\theta)] = \mathbb{E}_\theta\left[\left(\dot{\ell}_x(\theta)\right)^2\right] = \int \left(\dot{\ell}_x(\theta)\right)^2 f_\theta(x) dx$$

Equivalently, under regularity conditions (allowing differentiation under the integral sign), it can also be written as:

$$\mathcal{I}_\theta = -\mathbb{E}_\theta[\ddot{\ell}_x(\theta)] = -\int \ddot{\ell}_x(\theta) f_\theta(x) dx$$

Fisher information not only quantifies the amount of information that the data carries about the parameter θ, but also plays a central role in determining the precision of the MLE. Specifically, it appears in the asymptotic variance of the MLE, leading to the following fundamental result:

Theorem

Asymptotic Normality of the MLE
Let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of a scalar parameter θ. Under standard regularity conditions, we have:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

i.e.,

$$\hat{\theta} \xrightarrow{approx.} \mathcal{N}(\theta, \frac{1}{n\mathcal{I}_\theta})$$

Proof. Let the log-likelihood for a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x \mid \theta)$ be:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

By definition, the MLE $\hat{\theta}$ satisfies the score equation:

$$\ell_n'(\hat{\theta}) = 0$$

Now apply a Taylor expansion around the true value θ_0:

$$0 = \ell_n'(\hat{\theta}) = \ell_n'(\theta_0) + (\hat{\theta} - \theta_0)\ell_n''(\bar{\theta})$$

for some $\bar{\theta}$ between $\hat{\theta}$ and θ_0. Solve for $\hat{\theta} - \theta_0$:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left[\frac{1}{n}\ell_n''(\bar{\theta})\right]^{-1} \cdot \left[\frac{1}{\sqrt{n}}\ell_n'(\theta_0)\right]$$

Now use the following facts under regularity conditions:

- By the central limit theorem:
$$\frac{1}{\sqrt{n}}\ell_n'(\theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0))$$
- By the law of large numbers:
$$\frac{1}{n}\ell_n''(\bar{\theta}) \xrightarrow{p} \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta_0)\right] = -\mathcal{I}(\theta_0)$$

Combining these, we get:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

□

Example

Example: Normal Distribution with Known Variance
Suppose $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known. The log-likelihood function is:

$$\ell_x(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2 - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$

Taking derivatives, we obtain the score and observed information:

$$\dot{\ell}_x(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta), \quad \ddot{\ell}_x(\theta) = -\frac{n}{\sigma^2}$$

Solving the first-order condition $\dot{\ell}_x(\hat{\theta}) = 0$ yields the MLE:

$$\hat{\theta} = \bar{x}$$

The Fisher information is:

$$\mathcal{I}_\theta = \frac{n}{\sigma^2}$$

so by asymptotic normality:

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$$

Cramer-Rao Lower Bound

Let $\tilde{\theta} = t(x)$ be any unbiased estimator of a scalar parameter θ, based on an i.i.d. sample $x = (x_1, \ldots, x_n) \sim f_\theta(x)$. That is,

$$\mathbb{E}_\theta[\tilde{\theta}] = \theta$$

Theorem

Cramér–Rao Lower Bound
Under regularity conditions, the variance of any unbiased estimator $\tilde{\theta}$ satisfies

$$\mathrm{Var}_\theta(\tilde{\theta}) \geq \frac{1}{n\mathcal{I}_\theta}$$

where $\mathcal{I}_\theta$ is the Fisher information in one observation.
Equivalently, since the total Fisher information in the sample is $n\mathcal{I}_\theta$, the inequality can be written as:

$$\mathrm{Var}_\theta(\tilde{\theta}) \geq \left[\mathcal{I}_\theta^{(n)}\right]^{-1} = \frac{1}{n\mathcal{I}_\theta}$$

Proof. Let $\tilde{\theta} = t(x)$ be an unbiased estimator of θ, i.e., $\mathbb{E}_\theta[t(x)] = \theta$. Consider the identity:

$$\int t(x)\, \dot{\ell}_x(\theta)\, f_\theta(x)\, dx = \frac{d}{d\theta}\mathbb{E}_\theta[t(x)] = \frac{d\theta}{d\theta} = 1$$

By rewriting $t(x) = \theta + (t(x) - \theta)$, with $\int \dot{\ell}_x(\theta) f_\theta(x) dx = 0$, we obtain:

$$\int (t(x) - \theta)\, \dot{\ell}_x(\theta)\, f_\theta(x)\, dx = 1$$

Now, apply the Cauchy–Schwarz inequality:

$$\left(\int (t(x) - \theta)\, \dot{\ell}_x(\theta)\, f_\theta(x)\, dx\right)^2 \leq \left(\int (t(x) - \theta)^2 f_\theta(x) dx\right)\left(\int \left(\dot{\ell}_x(\theta)\right)^2 f_\theta(x) dx\right)$$

This simplifies to:

$$1 \leq \mathrm{Var}_\theta(\tilde{\theta}) \cdot \mathcal{I}_\theta^{(n)} \quad \Rightarrow \quad \mathrm{Var}_\theta(\tilde{\theta}) \geq \frac{1}{\mathcal{I}_\theta^{(n)}} = \frac{1}{n\mathcal{I}_\theta}$$

□

Remark

Interpretation: The bound tells us that no unbiased estimator can have smaller variance than $1/(n\mathcal{I}_\theta)$. While the MLE is not necessarily unbiased in finite samples, its bias is typically of order $1/n$, whereas its standard deviation is of order $1/\sqrt{n}$. Thus, the MLE is "asymptotically efficient," and comparison with the Cramér–Rao bound remains meaningful in large samples.

"Statistic Triangle"

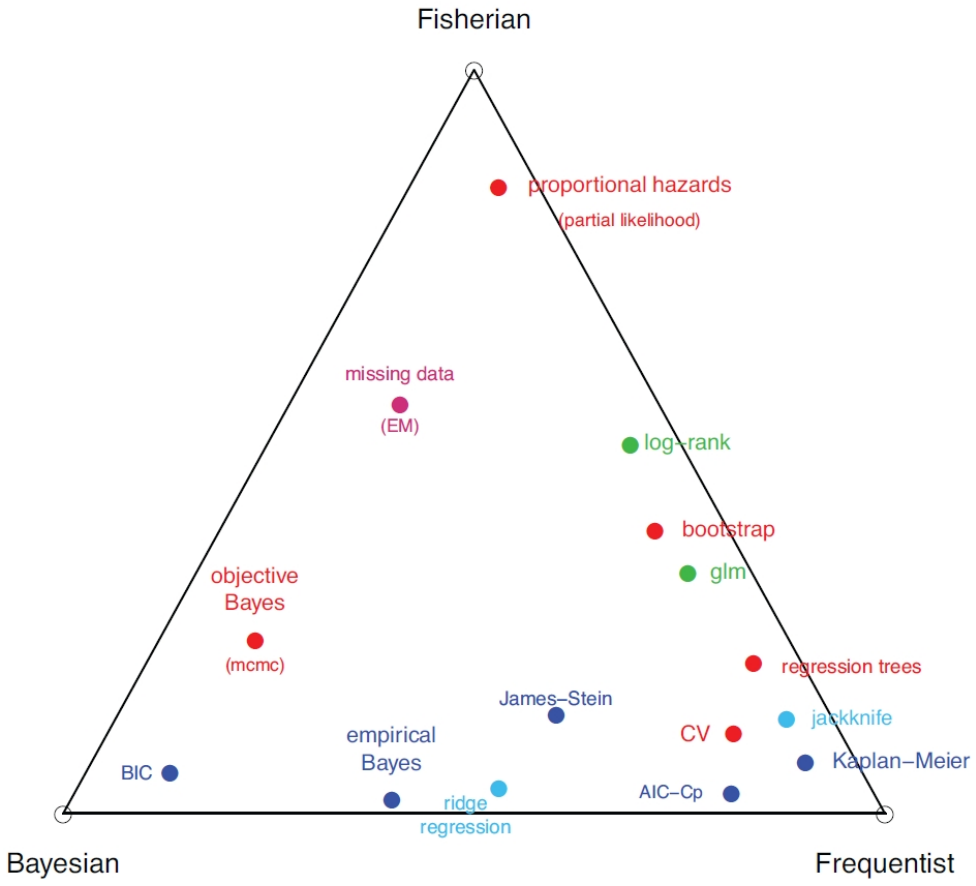

Figure 9: Bayesian, frequentist, and Fisherian influences, as described in the text, on 15 major topics, 1950s through 1990s. Colors indicate the importance of electronic computation in their development: red, crucial; violet, very important; green, important; light blue, less important; blue, negligible.

This triangle diagram illustrates the relationships and philosophical tensions among three major statistical traditions: **Fisherian**, **Frequentist**, and **Bayesian**. Each statistical method can be viewed as lying somewhere within this triangle, depending on the inferential logic it primarily relies on.

- **Fisherian** inference emphasizes likelihood-based reasoning, sufficiency, and information—the foundation of maximum likelihood estimation (MLE), Fisher information, and related methods like partial likelihood in proportional hazards models.
- **Frequentist** inference is grounded in repeated-sampling properties and error rates. Classical tools like the jackknife, cross-validation (CV), AIC, and regression trees are evaluated based on their performance across hypothetical repetitions.
- **Bayesian** inference incorporates prior beliefs and updates them via Bayes' theorem. Methods like MCMC, BIC, and both objective and empirical Bayes approaches embody this framework.

Logic of Inductive Inference

Conditional Inference

Fisher emphasized that statistical inference should condition on relevant ancillary information when available. This leads to the principle of **conditional inference**, which aims to provide conclusions that are more directly relevant to the observed data.

One of Fisher's key insights is that conditioning can lead to:

- **More relevant inferences**: Conditioning on observed features of the data-generating process often gives standard errors or confidence intervals that better reflect what was actually observed.
- **Simpler inferences**: In many cases, conditioning eliminates nuisance variation, leading to more interpretable and tractable calculations.

Example

Example: Conditioning on Sample Size Chosen by Coin Toss

Suppose we observe $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and estimate $\hat{\theta} = \bar{x}$. However, the sample size n was not fixed in advance—it was determined by flipping a fair coin:

$$n = \begin{cases} 25 & \text{with probability } 1/2 \\ 100 & \text{with probability } 1/2 \end{cases}$$

Assume we observe $n = 25$. What is the standard deviation of $\bar{x}$?

- The unconditional frequentist answer considers both possibilities: $\bar{x} \sim \mathcal{N}(0, 1/25)$ or $\mathcal{N}(0, 1/100)$ with equal probability, giving total variance

$$\frac{1}{2} \cdot \frac{1}{25} + \frac{1}{2} \cdot \frac{1}{100} = 0.03 \Rightarrow \text{SD} \approx 0.173$$

- The conditional inference answer, favored by Fisher, uses only the observed $n = 25$, leading to

$$\text{SD}(\bar{x}) = \frac{1}{\sqrt{25}} = 0.2$$

This illustrates Fisher's view: inference should condition on the observed sampling structure to ensure relevance.

In regression, the same logic applies: we typically treat the covariates x as fixed, and condition on them when computing standard errors for estimated coefficients. This is another form of conditional inference.

Ancillary Statistics: Fisher introduced the idea of ancillary statistics—statistics whose distribution does not depend on the parameter, and which can be used to condition inference. In the earlier example, the sample size n is ancillary. So are marginal totals in a contingency table and the design matrix X in regression.

Conditioning on such statistics is thought to:

- Increase relevance (by focusing inference on the situation actually observed),
- Increase simplicity (by removing irrelevant variation).

While some argue this discards information, Fisher contended that clarity and interpretability often outweigh this cost.

Permutation and Randomization

Fisherian methodology was often criticized for its reliance on normality assumptions. As an alternative, Fisher proposed a purely data-driven method: **permutation testing**.

Basic Idea: Given two groups (e.g., AML and ALL patients), we test whether the observed group difference is significant by comparing the original test statistic (e.g., two-sample t) to its distribution under all possible rearrangements of group labels.

Permutation Procedure: Suppose we observe $n = 72$ values, split into groups of size 47 and 25. The permutation test:

- Randomly permutes the group labels (many times, say $B = 10{,}000$);
- Recomputes the test statistic for each permutation;
- Estimates the two-sided p-value as:

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}(|t_i^*| \geq |t|)$$

Interpretation: Fisher provided two rationales for the permutation test:

- If the null hypothesis holds (i.e., all data come from the same distribution), then any label reassignment is equally likely;
- A small p-value implies the observed grouping (e.g., AML vs ALL) is unlikely under the null, providing evidence against it.

This exemplifies Fisher's *logic of inductive inference*: conclusions should follow directly from the data, without strong distributional assumptions.

Randomization: Fisher extended this idea to experimental design, advocating for random assignment of treatments. Randomization:

- Justifies permutation tests in experiments;
- Balances covariates (e.g., age, weight) across groups;
- Forms the foundation of modern RCTs.

Modern Perspective: Permutation methods have seen widespread resurgence, supported by computational power. They remain especially valuable when model assumptions (e.g., normality) are in doubt.

Lecture 9 Exponential Families

This lecture is based on the Chap. 5 of [EH16].

Univariate Distributions

We begin with five canonical univariate parametric families: Normal, Poisson, Binomial, Gamma, and Beta. Their basic forms, parameter domains, and first two moments are summarized below.

Name & Notation	**Density**	$\mathcal{X}$	Ω	**Expectation & Variance**
Normal $\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}^1$	$\mu \in \mathbb{R}^1,\ \sigma^2 > 0$	$\mu,\ \sigma^2$
Poisson $\mathrm{Poi}(\mu)$	$\frac{e^{-\mu}\mu^x}{x!}$	$\{0, 1, 2, \dots\}$	$\mu > 0$	$\mu,\ \mu$
Binomial $\mathrm{Bi}(n, \pi)$	$\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$	$\{0, 1, \dots, n\}$	$0 < \pi < 1$	$n\pi,\ n\pi(1-\pi)$
Gamma $\mathrm{Gam}(\nu, \sigma)$	$\frac{x^{\nu-1}e^{-x/\sigma}}{\sigma^\nu \Gamma(\nu)}$	$x \geq 0$	$\nu > 0,\ \sigma > 0$	$\nu\sigma,\ \nu\sigma^2$
Beta $\mathrm{Be}(\nu_1, \nu_2)$	$\frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1}(1-x)^{\nu_2-1}$	$0 \leq x \leq 1$	$\nu_1 > 0,\ \nu_2 > 0$	$\frac{\nu_1}{\nu_1+\nu_2}$, $\frac{\nu_1\nu_2}{(\nu_1+\nu_2)^2(\nu_1+\nu_2+1)}$

Table 1: Common univariate distributions: densities, support, parameter domains, and moments.

Additional Relationships.

- **Normal scaling:**
$$\mathcal{N}(\mu, \sigma^2) \sim \mu + \sigma Z, \quad Z \sim \mathcal{N}(0, 1)$$
- **Multivariate normal scaling:**
$$\mathcal{N}_p(\mu, \Sigma) \sim \mu + \Sigma^{1/2} Z, \quad Z \sim \mathcal{N}_p(0, I_p)$$
- **Chi-squared:**
$$\chi^2_\nu \sim \mathrm{Gam}\left(\frac{\nu}{2}, 2\right)$$
- **Uniform and Beta:**
$$U(0, 1) \sim \mathrm{Beta}(1, 1)$$
- **Beta–Gamma relationship:**
$$\mathrm{Be}(\nu_1, \nu_2) \sim \frac{\mathrm{Gam}(\nu_1, \sigma)}{\mathrm{Gam}(\nu_1, \sigma) + \mathrm{Gam}(\nu_2, \sigma)}$$
- **Poisson approximation to Binomial:**
$$\mathrm{Bi}(n, \pi) \approx \mathrm{Poi}(n\pi), \quad \text{for large } n,\ \text{small } \pi$$

Multivariate Normal Distribution

Definition

Multivariate Normal Distribution:
A p-dimensional random vector $x = (x_1, \dots, x_p)'$ has mean vector

$$\mu = \mathbb{E}[x] = (\mathbb{E}[x_1], \dots, \mathbb{E}[x_p])',$$

and covariance matrix

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)'\right], \quad \Sigma \in \mathbb{R}^{p\times p}.$$

We write

$$x \sim (\mu, \Sigma), \quad \text{or if normal, } x \sim \mathcal{N}_p(\mu, \Sigma).$$

The correlation between coordinates x_i and x_j is given by

$$\text{cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Multivariate Normal via Transformation: Let $z \sim \mathcal{N}_p(0, I_p)$, and let $T \in \mathbb{R}^{p\times p}$ be a non-singular matrix. Then the transformation

$$x = \mu + Tz$$

implies

$$x \sim \mathcal{N}_p(\mu, \Sigma), \quad \text{where } \Sigma = TT'.$$

The multivariate normal density is given by

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right).$$

Partitioned Form and Conditional Distribution: Suppose $x = (x'_{(1)}, x'_{(2)})'$ with corresponding partition of mean and covariance:

$$\begin{pmatrix} x_{(1)} \\ x_{(2)} \end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

Then the conditional distribution of $x_{(2)} \mid x_{(1)}$ is

$$x_{(2)} \mid x_{(1)} \sim \mathcal{N}_{p_2}\left(\mu_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(x_{(1)} - \mu_{(1)}),\ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right).$$

In the scalar case ($p_1 = p_2 = 1$), this reduces to:

$$x_2 \mid x_1 \sim \mathcal{N}\left(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x_1 - \mu_1),\ \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right).$$

Normal Scaling:

$$\mathcal{N}_p(\mu, \Sigma) \sim \mu + \Sigma^{1/2} Z, \quad Z \sim \mathcal{N}_p(0, I_p)$$

Fisher Information in Multi-dimension

Score Function: Let $\theta \in \mathbb{R}^p$, and suppose $x \sim f_\theta(x)$. The (vector-valued) score function is defined as the gradient of the log-likelihood:

$$\dot{\ell}_x(\theta) = D_\theta \log f_\theta(x) \in \mathbb{R}^{p\times 1}.$$

Fisher Information Matrix: The Fisher information is a $p \times p$ positive semi-definite matrix:

$$\mathcal{I}_\theta = \mathbb{E}_\theta\left[\dot{\ell}_x(\theta)\dot{\ell}_x(\theta)^\top\right] \in \mathbb{R}^{p\times p}.$$

Asymptotic Distribution of MLE: Under standard regularity conditions, the MLE $\hat{\theta}$ satisfies:

$$\hat{\theta} \overset{approx.}{\sim} \mathcal{N}_p\left(\theta, \mathcal{I}_\theta^{-1}\right).$$

That is, the MLE is asymptotically normal with mean θ and covariance matrix $\mathcal{I}_\theta^{-1}$.

Nuisance Parameter

Suppose the full parameter $\mu \in \mathbb{R}^p$ is partitioned as:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \text{with } \mu_1 \in \mathbb{R}^{p_1}, \mu_2 \in \mathbb{R}^{p_2}, \quad p_1 + p_2 = p$$

where μ_1 is the parameter of primary interest, and μ_2 is considered a **nuisance parameter**.

Assume the Fisher information matrix is similarly partitioned:

$$\mathcal{I}_\mu = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}$$

Theorem

Fisher Information with Nuisance Parameters
Let $\hat{\mu}_1$ denote the MLE of μ_1, marginalizing over μ_2. Then:

$$\hat{\mu}_1 \sim \mathcal{N}\left(\mu_1, \left(\mathcal{I}_\mu^{-1}\right)_{11}\right)$$

If μ_2 were known, then:

$$\tilde{\mu}_1 \sim \mathcal{N}\left(\mu_1, \left(\mathcal{I}_{11}\right)^{-1}\right)$$

and we always have:

$$\left(\mathcal{I}_\mu^{-1}\right)_{11} \succeq \left(\mathcal{I}_{11}\right)^{-1}$$

where $A \succeq B$ denotes that $A - B$ is positive semi-definite.

Proof. We begin with the partitioned Fisher information matrix:

$$\mathcal{I}_\mu = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}$$

From matrix inversion results (Schur complement formula), the upper-left block of the inverse satisfies:

$$\left(\mathcal{I}_\mu^{-1}\right)_{11} = \left(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}$$

Now compare with the case where μ_2 is known: then the Fisher information for μ_1 is $\mathcal{I}_{11}$, and the corresponding variance is $\left(\mathcal{I}_{11}\right)^{-1}$.

Note that the term $\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}$ is always positive semi-definite, so:

$$\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21} \preceq \mathcal{I}_{11} \quad \Rightarrow \quad \left(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1} \succeq \mathcal{I}_{11}^{-1}$$

Therefore:

$$\left(\mathcal{I}_\mu^{-1}\right)_{11} \succeq \left(\mathcal{I}_{11}\right)^{-1}$$

This proves that the marginal variance of $\hat{\mu}_1$ increases in the presence of nuisance parameters μ_2. □

Interpretation: When estimating μ_1, the presence of nuisance parameters μ_2 increases the asymptotic variance of the MLE. This illustrates a fundamental cost of nuisance parameters in classical inference—sometimes referred to as the "nuisance penalty".

Multinomial Distribution

Definition

Multinomial Distribution:
Consider a categorical outcome with L possible classes. Define the one-hot representation:

$$e_\ell = (0, 0, \ldots, \underset{\text{position } \ell}{1}, \ldots, 0)^\top, \quad \ell = 1, \ldots, L$$

Suppose we observe n independent categorical outcomes $x_1, \ldots, x_n$, with

$$x_i \in \{e_1, \ldots, e_L\}, \quad \mathbb{P}(x_i = e_\ell) = \pi_\ell$$

Let $X = \sum_{i=1}^n x_i = (X_1, \ldots, X_L)^\top \in \mathbb{Z}_{\geq 0}^L$ be the count vector, with $\sum_{\ell=1}^L X_\ell = n$. Then:

$$X \sim \text{Mult}_L(n, \boldsymbol{\pi})$$

The parameter space is the L-simplex:

$$\boldsymbol{\pi} \in \mathcal{S}_L = \left\{ \boldsymbol{\pi} \in \mathbb{R}^L : \pi_\ell \geq 0, \ \sum_{\ell=1}^L \pi_\ell = 1 \right\}$$

It has Probability Mass Function:

$$f_{\boldsymbol{\pi}}(x) = \frac{n!}{x_1! \cdots x_L!} \prod_{\ell=1}^L \pi_\ell^{x_\ell}, \quad x \in \mathbb{Z}_+^L, \ \sum x_\ell = n$$

Mean and Covariance: Let $x_i \sim \text{Categorical}(\boldsymbol{\pi})$, then

$$\mathbb{E}[x_i] = \sum_{\ell=1}^{L} \pi_\ell e_\ell = \boldsymbol{\pi}$$

$$\begin{aligned}
\text{Var}(x_i) &= \mathbb{E}[x_i x_i^\top] - \mathbb{E}[x_i]\mathbb{E}[x_i]^\top \\
&= \mathbb{E}\left[\left(\sum_{\ell=1}^{L} x_{i\ell} e_\ell\right)\left(\sum_{m=1}^{L} x_{im} e_m\right)^\top\right] - \left(\mathbb{E}[x_i]\right)\left(\mathbb{E}[x_i]\right)^\top \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L} x_{i\ell}^2 e_\ell e_\ell^\top\right] - \boldsymbol{\pi}\boldsymbol{\pi}^\top \\
&= \sum_{\ell=1}^{L} \mathbb{E}[x_{i\ell}]\, e_\ell e_\ell^\top - \boldsymbol{\pi}\boldsymbol{\pi}^\top \\
&= \sum_{\ell=1}^{L} \pi_\ell e_\ell e_\ell^\top - \boldsymbol{\pi}\boldsymbol{\pi}^\top \\
&= \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top
\end{aligned}$$

So for the sum:

$$X \sim \text{Mult}_L(n, \boldsymbol{\pi}) \quad \Rightarrow \quad \begin{aligned} \mathbb{E}[X] &= n\boldsymbol{\pi} \\ \text{Var}(X) &= n\left(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top\right) \end{aligned}$$

Sample Space:

$$\mathcal{X} = \left\{ x \in \mathbb{Z}_+^L : \sum_{\ell=1}^{L} x_\ell = n \right\} = n \cdot \mathcal{S}_L$$

Multinomial and Poisson

There is a useful relationship between the multinomial distribution and the Poisson distribution.

1. From Poisson to Multinomial (Conditional Distribution)
Let $\mathbf{S} = (S_1, S_2, \ldots, S_L)^\top \sim \text{Poi}(\boldsymbol{\mu})$, where $S_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_l)$. Let the total count be:

$$S_+ = \sum_{l=1}^{L} S_l = \mathbf{1}^\top \mathbf{S}$$

Then, conditional on the total sum S_+, the vector $\mathbf{S}$ follows a multinomial distribution:

$$\mathbf{S} \mid S_+ \sim \text{Mult}_L(S_+, \boldsymbol{\mu}/\mu_+), \quad \mu_+ = \sum_{l=1}^{L} \mu_l$$

2. From Multinomial to Poisson (Marginal Distribution)
Suppose $N \sim \text{Poi}(n)$, and conditional on N, we have:

$$\mathbf{x} \mid N \sim \text{Mult}_L(N, \boldsymbol{\pi})$$

Then the marginal distribution of $\mathbf{x}$ is:

$$\mathbf{x} \sim \text{Poi}(n\boldsymbol{\pi})$$

This result shows that:

$$\text{Mult}_L(N, \boldsymbol{\pi}) \sim \text{Poi}(n\boldsymbol{\pi}) \quad \text{if } N \sim \text{Poi}(n)$$

3. Interpretation and Application
The Poisson approximation is useful because it removes the negative correlations between components of the multinomial vector, which simplifies computation when n is large:

$$\mathbf{x} \sim \text{Mult}_L(n, \boldsymbol{\pi}) \approx \text{Poi}(n\boldsymbol{\pi})$$

This equivalence also justifies the multinomial as a foundation for nonparametric inference. Since it includes all discrete distributions over L categories, it serves as a model for general categorical data modeling. The multinomial model is heavily used in bootstrap methods, nonparametric statistics, and modern large-scale data analysis.

Exponential Families

Definition

Exponential Family: A probability density or mass function $f_\alpha(x)$ belongs to the exponential family if it can be written in the form:

$$f_\alpha(x) = \exp\left(\alpha^\top y - \psi(\alpha)\right) f_0(x), \quad \alpha \in A,$$

where:

- $\alpha \in \mathbb{R}^p$ is the **natural** (or **canonical**) parameter;
- $y = t(x) \in \mathbb{R}^p$ is the **sufficient statistic**;
- $f_0(x)$ is a base measure;
- $\psi(\alpha)$ is the **log-partition function** (or **cumulant generating function**), defined by:

$$\psi(\alpha) = \log \int_{\mathcal{X}} e^{\alpha^\top y} f_0(x) dx,$$

ensuring that $f_\alpha(x)$ integrates or sums to 1.

Mean and Variance:

$$\psi'(\alpha) = \mathbb{E}_\alpha[y], \qquad \psi''(\alpha) = \mathrm{Var}_\alpha(y)$$

More generally, for $\alpha \in \mathbb{R}^p$, the derivatives are:

$$\dot{\psi}(\alpha) = \nabla_\alpha \psi(\alpha) = \mathbb{E}_\alpha[y], \quad \ddot{\psi}(\alpha) = \nabla^2_\alpha \psi(\alpha) = \mathrm{Cov}_\alpha(y)$$

Example

Example: Gamma Distribution as Exponential Family
The Gamma distribution with parameters ν and σ has the density:

$$f(x) = \frac{1}{\sigma^\nu \Gamma(\nu)} x^{\nu-1} e^{-x/\sigma}$$

We write this in the exponential family form:

$$f(x) = \exp\left(\alpha_1 x + \alpha_2 \log x - \psi(\alpha)\right) \cdot \frac{1}{x}$$

with the following components:

- **Sufficient statistics:** $y(x) = (x, \log x)$
- **Natural parameters:** $\alpha = \left(-\frac{1}{\sigma},\ \nu\right)$
- **Log-partition function:**

$$\psi(\alpha) = \nu \log \sigma + \log \Gamma(\nu) = -\alpha_2 \log(-\alpha_1) + \log \Gamma(\alpha_2)$$

The natural parameter space is $\alpha_1 < 0, \alpha_2 > 0$, which is convex. Hence, the Gamma family is a two-parameter exponential family.

Generating Poisson Families

To construct the Poisson exponential family via exponential tilting, we follow three steps:

1. **Start with a base Poisson distribution.** Choose a fixed Poisson distribution $f_{\mu_0}(x)$, e.g., with $\mu_0 = 1$:

$$f_{\mu_0}(x) = \frac{e^{-\mu_0} \mu_0^x}{x!}, \quad x = 0, 1, 2, \ldots$$

2. **Apply exponential tilting.** For any $\mu > 0$, define the canonical parameter:

$$\alpha = \log\left(\frac{\mu}{\mu_0}\right)$$

Then construct the unnormalized tilted distribution:

$$\tilde{f}_\mu(x) = e^{\alpha x} f_{\mu_0}(x) = \left(\frac{\mu}{\mu_0}\right)^x f_{\mu_0}(x)$$

3. **Renormalize.** Define the normalizing function:

$$\psi(\alpha) = \log\left(\sum_{x=0}^{\infty} e^{\alpha x} f_{\mu_0}(x)\right)$$

Then the resulting probability mass function is:

$$f_\mu(x) = \frac{\tilde{f}_\mu(x)}{e^{\psi(\alpha)}} = \frac{e^{\alpha x - \psi(\alpha)} f_{\mu_0}(x)}{1}$$

This yields the canonical exponential family form:

$$f_\mu(x) = e^{\alpha x - \psi(\alpha)} f_{\mu_0}(x)$$

Why Exponential Tilting?

Exponential tilting plays a central role in exponential families because it preserves important properties under i.i.d. sampling. Specifically, it ensures that sufficient statistics remain additive and compress all relevant information into a fixed-dimensional summary, regardless of sample size.

Suppose we have $x = (x_1, \ldots, x_n)$ drawn i.i.d. from an exponential family:

$$f_\alpha(x_i) = e^{\alpha^\top y_i - \psi(\alpha)} f_0(x_i)$$

Then the joint density of the sample is:

$$f_\alpha(x) = \prod_{i=1}^{n} e^{\alpha^\top y_i - \psi(\alpha)} f_0(x_i) = e^{n\alpha^\top \bar{y} - n\psi(\alpha)} f_0(x)$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad f_0(x) = \prod_{i=1}^{n} f_0(x_i)$$

This shows that the sufficient statistic for inference remains $\bar{y}$, a p-dimensional average, which is independent of the sample size n. Only exponential tilting retains this property, making it essential for tractable inference and repeated sampling.

Cumulative Generating Function

Definition

The **moment generating function** (mgf) of a random variable X is defined as:

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The **cumulant generating function** (cgf) is the logarithm of the moment generating function:

$$K_X(t) = \log \mathbb{E}[e^{tX}] = \sum_{m=1}^{\infty} \kappa_m \frac{t^m}{m!}$$

where $\kappa_m = K_X^{(m)}(0)$ is the mth cumulant.

Examples of cumulants:

- $\kappa_1 = \mathbb{E}[X] = \mu$
- $\kappa_2 = \mathrm{Var}(X) = \sigma^2$
- $\kappa_3 = \mu_3 = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$ (skewness)
- $\kappa_4 = \mu_4 - \mu_2^2 = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4} - 3$ (kurtosis)

Properties:

- $K_Y(t) = \psi(t+\alpha) - \psi(\alpha)$, where $\psi(\alpha)$ is the log normalizing constant in the exponential family.
- The MLE of the expectation parameter $\mu = \mathbb{E}_\alpha[y]$ is $\hat{\mu} = \bar{y}$

Lecture 10 Information and Entropy

This lecture is based on the Chap. 1, Secs. 2.1–2.7, 8.1, 17.7 of [CT06].

Information Theory

In information theory, the amount of information carried by a distribution is measured by **entropy**. In contrast, in statistics, Fisher information quantifies the uncertainty about a parameter in a model.

Definition

The (Shannon) entropy of a discrete random variable X with probability mass function $p(x)$ is defined as

$$H(X) = -\mathbb{E}_p[\log_2 p(X)] = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x),$$

where the base-2 logarithm implies the unit of measurement is in bits.

Theorem

Proposition: For a discrete random variable X supported on a finite set of size m, the entropy $H(X)$ is maximized when X follows the uniform distribution:

$$p(x) = \frac{1}{m}, \quad \forall x \in \mathcal{X}.$$

Proof. Assume the values of X are $x_1, x_2, \ldots, x_m$, with corresponding probabilities $p_j = \mathbb{P}(X = x_j)$ for $j = 1, \ldots, m$, satisfying $\sum_{j=1}^m p_j = 1$.

Let $f(x) = -x\log x$. This function is concave on $(0, 1]$, and so by Jensen's inequality,

$$H(X) = -\sum_{j=1}^m p_j \log p_j \le -m \cdot \left(\frac{1}{m}\sum_{j=1}^m p_j\right) \log\left(\frac{1}{m}\sum_{j=1}^m p_j\right) = \log m.$$

The equality holds if and only if $p_1 = p_2 = \cdots = p_m = \frac{1}{m}$, i.e., when the distribution is uniform. □

We now extend the definition to a pair of random variables.

Definition

The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y),$$

which is equivalently

$$H(X, Y) = -\mathbb{E}[\log_2 p(X, Y)].$$

We also define the conditional entropy of one random variable given another as the expected entropy of the conditional distribution.

Definition

If $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X = x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|x) \\ &= -\mathbb{E}[\log_2 p(Y|X)]. \end{aligned}$$

Theorem

Chain Rule: The joint entropy satisfies the following identity:

$$H(X, Y) = H(X) + H(Y|X).$$

Proof. We start from the definition of joint entropy:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y).$$

Using the factorization $p(x,y) = p(x)p(y|x)$, we obtain:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log_2 [p(x)p(y|x)] \\ &= -\sum_{x,y} p(x,y) \left[\log_2 p(x) + \log_2 p(y|x)\right] \\ &= -\sum_{x,y} p(x,y) \log_2 p(x) - \sum_{x,y} p(x,y) \log_2 p(y|x) \\ &= -\sum_{x} p(x) \log_2 p(x) - \sum_{x,y} p(x,y) \log_2 p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

□

Relative Entropy and Mutual Information

Definition

The **relative entropy** or **Kullback–Leibler (KL) divergence** from distribution $p(x)$ to $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log_2 \frac{p(X)}{q(X)}\right],$$

where p represents the observed (true) distribution and q is the assumed (model) distribution.

Theorem

Non-negativity of KL divergence: For any two distributions p and q,

$$D(p \parallel q) \geq 0,$$

with equality if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. This follows from Gibbs' inequality. Define the function $f(t) = t \log t$, which is convex on $(0, \infty)$. Then

$$D(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} = \sum_{x} q(x) \cdot \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)}.$$

Applying Jensen's inequality to the convex function f with weights $q(x)$, we obtain:

$$\sum_{x} q(x) f\left(\frac{p(x)}{q(x)}\right) \geq f\left(\sum_{x} q(x) \cdot \frac{p(x)}{q(x)}\right) = f(1) = 0.$$

Thus $D(p \parallel q) \geq 0$, and equality holds if and only if $\frac{p(x)}{q(x)} = 1$ for all x, i.e., $p = q$. □

Definition

The **mutual information** between two random variables X and Y is defined as the KL divergence between the joint distribution and the product of the marginals:

$$I(X;Y) = D(p(x,y) \parallel p(x)p(y)) = \mathbb{E}_{p(x,y)} \left[\log_2 \frac{p(x,y)}{p(x)p(y)}\right].$$

It measures the reduction in uncertainty of one variable given knowledge of the other.

Theorem

Proposition: Mutual information satisfies the following properties:

1. **Symmetry and decomposition:**

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X).$$

2. **Self-information:**

$$I(X;X) = H(X).$$

Proof. **(1)** Start from the definition:

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log_2 \frac{p(y|x)}{p(y)}.$$

Thus,

$$I(X;Y) = \sum_x p(x) \sum_y p(y|x) \log_2 \frac{p(y|x)}{p(y)} = \sum_x p(x) D(p(y|x) \parallel p(y)) = H(Y) - H(Y|X).$$

Similarly, swapping the roles of X and Y gives $I(Y;X) = H(X) - H(X|Y)$.

(2) Since $p(x,x) = p(x)$ and $p(x)p(x)$ is the product of the marginal with itself, we compute:

$$I(X;X) = \sum_x p(x) \log_2 \frac{p(x)}{p(x)^2} = \sum_x p(x) \log_2 \frac{1}{p(x)} = H(X).$$

□

Definition

The **differential entropy** $h(X)$ of a continuous random variable X with probability density function $f(x)$ is defined as

$$h(X) = -\mathbb{E}_f[\log f(X)] = -\int f(x) \log f(x)\, dx.$$

Remark

Note that unlike discrete entropy, differential entropy can be negative and does not enjoy the same invariance under change of variables.

Entropy and Fisher Information

The **Fisher information** of a continuous random variable X with density f and location parameter θ is defined as

$$I(X) = \int_{-\infty}^{\infty} f(x-\theta) \left[\frac{\partial}{\partial \theta} \log f(x-\theta) \right]^2 dx.$$

In the location family case, we can shift the differentiation to the variable x, yielding the simplified expression:

$$I(X) = \int_{-\infty}^{\infty} f(x) \left[\frac{\partial}{\partial x} \log f(x) \right]^2 dx.$$

The following identity provides a fundamental link between entropy and Fisher information.

Theorem

De Bruijn's Identity: Let X be a continuous random variable with finite variance and density f, and let $Z \sim \mathcal{N}(0,1)$ be independent of X. Then, for $Y_t = X + \sqrt{t}Z$, the differential entropy h_e (to base e) satisfies:

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z) = \frac{1}{2} I(X + \sqrt{t}Z),$$

and in particular,

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z)\Big|_{t=0} = \frac{1}{2} I(X).$$

Proof. Let $Y_t = X + \sqrt{t}Z$. The density of Y_t is given by the convolution:

$$g_t(y) = \int_{-\infty}^{\infty} f(x) \cdot \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right) dx.$$

This is the convolution of f with the Gaussian kernel, i.e., $g_t = f * \phi_{\sqrt{t}}$.

It is known that:

$$\frac{\partial}{\partial t} g_t(y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} g_t(y).$$

Now compute the time derivative of the differential entropy:

$$\begin{aligned}
\frac{\partial}{\partial t} h_e(Y_t) &= -\frac{\partial}{\partial t} \int g_t(y) \log g_t(y)\, dy \\
&= -\int \left(\frac{\partial}{\partial t} g_t(y)\right) \log g_t(y)\, dy - \int \left(\frac{\partial}{\partial t} g_t(y)\right) dy \\
&= -\int \left(\frac{1}{2} \frac{\partial^2}{\partial y^2} g_t(y)\right) \log g_t(y)\, dy \quad (\text{since } \int \partial_t g_t(y)\, dy = \partial_t 1 = 0) \\
&= \frac{1}{2} \int \left(\frac{\partial g_t(y)}{\partial y}\right)^2 \frac{1}{g_t(y)}\, dy \\
&= \frac{1}{2} I(Y_t),
\end{aligned}$$

where we used integration by parts and the boundary term vanishes under regularity conditions. □

Lecture 11 Linear Regression

This lecture is based on the Chap. 14 of [Pol23] and Secs. 3.1–3.5 of [SL03].

Regression Problem Setup

In regression, we observe data pairs (x_i, y_i) for $i = 1, 2, \ldots, n$, where:

- y_i: the **outcome** or **dependent variable**, and
- x_i: the **predictors**, also called **independent variables** or **covariates**.

The goal of regression is twofold:

- **Inference**: Understand the relationship between x and y, often via estimated coefficients.
- **Prediction**: Given a new input x_0, predict the associated response $\hat{y}_0$.

The simplest form of a linear regression model assumes the response is a linear function of the predictors plus noise:

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 + \varepsilon,$$

where:

- $\mathbf{x} \in \mathbb{R}^p$ is the row vector of predictors,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the column vector of coefficients,
- $\beta_0 \in \mathbb{R}$ is the intercept,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is random noise.

The term **regression** comes from the phenomenon of regression to the mean, captured by the conditional expectation:

$$y = \mathbb{E}[y \mid x] + \varepsilon.$$

That is, even if an observed value of y is extreme, its conditional expectation given x may tend to be closer to the average. This reflects a general tendency of predictions to "regress" back toward the mean.

Least Squares Estimation

The method of least squares estimates the coefficient vector $\hat{\beta}$ by minimizing the residual sum of squares (RSS):

Definition

The least squares estimator is defined as

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 = \arg\min_{\beta} \|\boldsymbol{y} - X\beta\|_2^2,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $\beta \in \mathbb{R}^p$ is the coefficient vector. The objective function is:

$$\mathrm{RSS}(\beta) = \|\boldsymbol{y} - X\beta\|_2^2.$$

Taking derivative with respect to β and setting it to zero gives the **normal equation**:

$$\frac{\partial}{\partial \beta} \mathrm{RSS}(\beta) = -2X^\top (\boldsymbol{y} - X\beta) = 0.$$

Full Rank Case

If $\mathrm{rank}(X) = p \leq n$, then:

$$\mathrm{rank}(X^\top X) = p,$$

so $X^\top X$ is invertible, and the solution is:

$$\hat{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Geometric Interpretation

Let $\hat{\boldsymbol{y}} = X\hat{\beta}$ denote the fitted values. This vector is the orthogonal projection of $\boldsymbol{y}$ onto the column space $C(X)$. Let P denote the projection matrix onto $C(X)$, then:

$$\hat{\boldsymbol{y}} = P\boldsymbol{y}, \quad \text{where } P = X(X^\top X)^{-1}X^\top.$$

For any vector $\theta \in C(X)$, we have the orthogonality condition:

$$(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \theta) = 0.$$

Equivalently,

$$\boldsymbol{y}^\top(I - P)P(\boldsymbol{y} - \theta) = 0,$$

which implies that the residual $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to the column space of X:

$$\boldsymbol{y} - \hat{\boldsymbol{y}} \perp C(X).$$

Example

Example: Simple Linear Regression
In the simple linear model, we assume

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

This can be written in matrix form as

$$\boldsymbol{y} = \begin{pmatrix} \mathbf{1}_n & \mathbf{x} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \boldsymbol{\varepsilon}.$$

The least squares estimates are given by:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Theorem

Unbiasedness and Variance of $\hat{\beta}$:
Suppose $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$, $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2 I_n$, and X is full rank. Then:

$$\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1}X^\top \mathbb{E}[\boldsymbol{y}] = (X^\top X)^{-1}X^\top X\beta = \beta,$$
$$\mathrm{Var}[\hat{\beta}] = (X^\top X)^{-1}X^\top \mathrm{Var}[\boldsymbol{y}]X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

Proof. $\mathrm{Var}[\hat{\beta}] = \mathrm{Var}[(X^\top X)^{-1}X^\top \boldsymbol{y}] = (X^\top X)^{-1}X^\top \mathrm{Var}[\boldsymbol{y}]X(X^\top X)^{-1} = (X^\top X)^{-1}X^\top \mathrm{Var}[\boldsymbol{\varepsilon}]X(X^\top X)^{-1}$.
Under the assumption that $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2 I_n$, this simplifies to:

$$\mathrm{Var}[\hat{\beta}] = \sigma^2(X^\top X)^{-1}.$$

□

Remark

To achieve accurate estimation of $\hat{\beta}$, we need to control the noise level and $\mathbf{X}^\top\mathbf{X}$ needs to be away from singular case ($\mathbf{X}$ not too linear dependent).
The least squares estimator is optimal among all **linear unbiased estimators** (by Gauss–Markov theorem).

Additionally, we need an estimate of σ^2.

Theorem

If $\mathbb{E}[\boldsymbol{y}] = X\beta$, where $X \in \mathbb{R}^{n\times p}$ has rank $r \le p$, and $\mathrm{Var}[\boldsymbol{y}] = \sigma^2 I_n$, then the statistic

$$S^2 = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2}{n - r} = \frac{RSS}{n - r}$$

is an unbiased estimator of σ^2, i.e.,

$$\mathbb{E}[S^2] = \sigma^2.$$

Proof. Let $X_1 \in \mathbb{R}^{n\times r}$ be a full-rank representation of the column space $C(X)$, and suppose $\theta = X_1\alpha$, where $\alpha \in \mathbb{R}^r$. Then

$$\hat{\boldsymbol{y}} = X\hat{\beta} = P\boldsymbol{y}, \quad \text{where } P = X_1(X_1^\top X_1)^{-1}X_1^\top$$

is the projection matrix onto $C(X)$, and

$$\boldsymbol{y} - \hat{\boldsymbol{y}} = (I_n - P)\boldsymbol{y}.$$

Therefore, the residual sum of squares is

$$(n-r)S^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 = \boldsymbol{y}^\top(I_n - P)\boldsymbol{y}.$$

Since $\mathbb{E}[\boldsymbol{y}] = \theta = X_1\alpha \in C(X)$, we have $P\theta = \theta$, and by standard results:

$$\mathbb{E}\left[\boldsymbol{y}^\top(I_n - P)\boldsymbol{y}\right] = \sigma^2 \operatorname{tr}(I_n - P) + \theta^\top(I_n - P)\theta = \sigma^2(n-r).$$

Hence,

$$\mathbb{E}[S^2] = \frac{1}{n-r}\cdot\sigma^2(n-r) = \sigma^2.$$

□

Maximum Likelihood Estimation under Gaussian Noise

In the case of least squares estimation, assume the error term follows a Gaussian distribution:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n).$$

Under this assumption, the least squares estimator $\hat{\beta}$ coincides with the maximum likelihood estimator (MLE), and the MLE of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n}\left\|\boldsymbol{y} - X\hat{\beta}\right\|_2^2.$$

The log-likelihood function of the model is:

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - X\beta\|_2^2 + C,$$

where C is a constant independent of β and σ^2.

The Fisher information matrix is:

$$\mathcal{I}(\beta, \sigma^2) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial\theta\partial\theta^\top}\right] = \begin{pmatrix} \frac{1}{\sigma^2}X^\top X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

where $\theta = (\beta^\top, \sigma^2)^\top$.

Lecture 12 Generalized Linear Models

This lecture is based on the Chap. 8 of [EH16].

Why GLM?

In classical linear regression, we assume

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

which implies that $y \in \mathbb{R}$, and predicted values $\hat{y} = X\hat{\beta} \in \mathbb{R}$ as well.

However, this becomes inappropriate when:

- the response variable y is categorical or binary (e.g. success/failure),
- or when the mean response must lie in a restricted range (e.g. probabilities in $(0, 1)$).

To address these cases, we consider the framework of **Generalized Linear Models (GLM)**, which allows the response to come from a broader family of distributions (exponential family), and links the mean response to a linear predictor via a transformation (link function).

Example

Example: Logistic Regression
Suppose we observe binary responses:

$$y_i \sim \text{Bin}(n_i, \pi_i), \quad i = 1, \ldots, N,$$

where $\pi_i \in (0, 1)$ is the probability of success (e.g., true death rate), and n_i is the number of trials for group i.
We model the log-odds (logit) of the success probability as a linear function of covariates:

$$\lambda_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \alpha_1 x_i,$$

so that

$$\pi_i = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x_i)}}.$$

We estimate the coefficients via maximum likelihood:

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1)^\top,$$

and then obtain fitted probabilities:

$$\hat{\pi}(x) = \frac{1}{1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 x)}}.$$

The MLE does not have a closed-form solution and is typically obtained via iterative numerical optimization (e.g., Newton-Raphson or Fisher scoring).

Remark

GLMs generalize linear models by allowing:

- a non-Gaussian response distribution from the exponential family,
- a link function connecting the mean response to a linear predictor.

This allows us to model categorical, count, and proportion data more appropriately than standard linear regression.

General Form of Logistic Regression

We consider the binomial likelihood:

$$P(y_i \mid \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} = \binom{n_i}{y_i} \exp\left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) \right\}.$$

Define:

$$\lambda_i = \log\left(\frac{\pi_i}{1-\pi_i}\right), \quad \psi(\lambda_i) = \log(1+e^{\lambda_i}) = \log\left(\frac{1}{1-\pi_i}\right)$$

then the likelihood becomes:

$$P(y_i \mid \lambda_i) = \binom{n_i}{y_i} \exp\left(y_i\lambda_i - n_i\psi(\lambda_i)\right).$$

The joint likelihood (assuming independence) is:

$$f_\alpha(\boldsymbol{y}) = \prod_{i=1}^{N} \binom{n_i}{y_i} \exp\left(x_i^\top y_i \cdot \alpha - n_i\psi(x_i^\top \alpha)\right).$$

Define:

$$s_0 = \sum_{i=1}^{N} y_i, \quad s_1 = \sum_{i=1}^{N} x_i y_i,$$

as sufficient statistics for α.

We have the KL divergence:

$$D(p_i \parallel \hat{\pi}_i) = 2n_i\left[p_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) + (1-p_i)\log\left(\frac{1-p_i}{1-\hat{\pi}_i}\right)\right],$$

and the MLE minimizes the total deviance:

$$\hat{\alpha} = \arg\min_\alpha \sum_{i=1}^{N} D(p_i, \pi_\alpha(x_i)).$$

Logistic Regression with Interaction

In two-way designs with interaction (e.g., ratio i and day j), define:

$$\pi_{ij} = \text{probability of success for setting } (i,j),$$

$$\lambda_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + \alpha_i + \beta_j.$$

This is referred to as logistic ANOVA.

Problems:

- High dimensionality $\rightarrow$ instability.
- Zero or near-zero values $x_{ij} \approx 0$ lead to numerical issues with $\log(x_{ij}+\varepsilon)$.

Example

Example: Spam Data

Consider a binary classification task where we aim to predict whether an email is spam or not:

- $y_i \in \{0,1\}$: binary response indicating whether email i is spam ($y_i = 1$) or ham ($y_i = 0$);
- x_{ij}: relative frequency of word j in email i, normalized by message length.

We fit a logistic regression model of the form:

$$\lambda_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_{ij}, \quad \text{where } p = 57.$$

The goal is to estimate the coefficients α_j to identify influential keywords for spam detection.

Generalized Linear Models

We consider the 1-parameter exponential family:

$$f_\lambda(y) = \exp\left(y\lambda - \delta(\lambda)\right) f_0(y), \quad \lambda \in \Lambda.$$

Assume:

- $y = (y_1, \ldots, y_n)^\top$, with $y_i \overset{\text{i.i.d.}}{\sim} f_{\lambda_i}$,
- $\lambda_i = x_i^\top \alpha$, where $x_i \in \mathbb{R}^p$, $\alpha \in \mathbb{R}^p$, and $X \in \mathbb{R}^{n \times p}$.

The joint likelihood becomes:

$$f_\alpha(y) = \prod_{i=1}^n f_{\lambda_i}(y_i) = \exp\left(\alpha^\top X^\top y - \sum_{i=1}^n \delta(x_i^\top \alpha)\right) \cdot \prod_{i=1}^n f_0(y_i),$$

with cumulant function:

$$\psi(\alpha) = \sum_{i=1}^n \delta(x_i^\top \alpha).$$

This is a dimension reduction setting when $p \ll n$.

Properties of $\hat{\alpha}$

We denote $\mu(\alpha) = \mathbb{E}_\alpha[y] = (\mu_1, \ldots, \mu_n)^\top$, where $\mu_i = \delta'(x_i^\top \alpha)$, and

$$\Sigma(\alpha) = \mathrm{Var}(y) = \mathrm{diag}\left(\delta''(x_1^\top \alpha), \ldots, \delta''(x_n^\top \alpha)\right).$$

The MLE $\hat{\alpha}$ satisfies the estimating equation:

$$X^\top (y - \mu(\hat{\alpha})) = 0,$$

which is analogous to the normal equation $X^\top (y - X\hat{\beta}) = 0$ in least squares.

Using a Taylor expansion around the true value α, we obtain:

$$\hat{\alpha} \approx \alpha + V_\alpha^{-1} z, \quad z \sim \mathcal{N}(0, V_\alpha), \quad \text{with } V_\alpha = X^\top \Sigma(\alpha) X,$$

so that:

$$\hat{\alpha} \sim \mathcal{N}(\alpha, [X^\top \Sigma(\alpha) X]^{-1}).$$

In the Gaussian special case:

$$\Sigma(\alpha) = \sigma^2 I_n \Rightarrow \mathrm{Var}(\hat{\alpha}) = \sigma^2 (X^\top X)^{-1}.$$

Many common models used in GLMs—such as Normal, Poisson, Binomial, and Gamma—can be expressed in the one-parameter exponential family form:

$$f_\lambda(y) = \exp\left(y\lambda - \gamma(\lambda)\right) f_0(y),$$

with natural parameter λ, sufficient statistic y, and cumulant function $\gamma(\lambda)$. The deviance between two distributions in the family is:

$$D(f_1, f_2) = 2 \int f_1(y) \log\left(\frac{f_1(y)}{f_2(y)}\right) dy.$$

Distribution	λ	y	$D(f_1, f_2)$	$\gamma(\lambda)$
Normal $\mathcal{N}(\mu, \sigma^2)$	μ/σ^2	x	$\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2$	$\frac{\sigma^2 \lambda^2}{2}$
Poisson $\mathrm{Poi}(\mu)$	$\log \mu$	x	$2\mu_1 \left[\left(\frac{\mu_2}{\mu_1} - 1\right) - \log\left(\frac{\mu_2}{\mu_1}\right)\right]$	e^λ
Binomial $\mathrm{Bi}(n, \pi)$	$\log \frac{\pi}{1-\pi}$	x	$2n \left[\pi_1 \log \frac{\pi_1}{\pi_2} + (1 - \pi_1) \log \frac{1-\pi_1}{1-\pi_2}\right]$	$n \log(1 + e^\lambda)$
Gamma $\mathrm{Gam}(\nu, \sigma)$	$-1/\sigma$	x	$2\nu \left[\left(\frac{\sigma_1}{\sigma_2} - 1\right) - \log\left(\frac{\sigma_1}{\sigma_2}\right)\right]$	$-\nu \log(-\lambda)$

Table 2: Exponential family form and deviance for common GLMs

Theorem

Hoeffding's Lemma and Deviance Form
Let y follow a one-parameter exponential family with density

$$f_\lambda(y) = \exp(y\lambda - \gamma(\lambda))f_0(y),$$

and let $\hat{\lambda}$ be the MLE satisfying $\gamma'(\hat{\lambda}) = y$. Then the density under any λ satisfies:

$$f_\lambda(y) = f_y(y) \exp\left(-\frac{1}{2}D(y, \mu_\lambda)\right),$$

where $D(y, \mu)$ is the deviance between f_y and f_λ.

Proof. Let $\hat{\lambda}$ be the MLE given y, i.e., the maximizer of the log-likelihood:

$$\hat{\lambda} = \arg\max_\lambda \{y\lambda - \gamma(\lambda)\}.$$

Taking derivative:

$$\frac{d}{d\lambda}(y\lambda - \gamma(\lambda)) = y - \gamma'(\lambda),$$

which implies

$$\gamma'(\hat{\lambda}) = y \quad \Rightarrow \quad \mu_{\hat{\lambda}} = y.$$

Now define the deviance between two exponential family members as

$$D(\lambda_1, \lambda_2) = 2\int f_{\lambda_1}(y) \log\left(\frac{f_{\lambda_1}(y)}{f_{\lambda_2}(y)}\right) dy.$$

Applying the definition of exponential family, we compute

$$\log f_{\lambda_1}(y) - \log f_{\lambda_2}(y) = y(\lambda_1 - \lambda_2) - \gamma(\lambda_1) + \gamma(\lambda_2).$$

Therefore, we have:

$$\frac{1}{2}D(\lambda_1, \lambda_2) = \mathbb{E}_{\lambda_1}\left[(\lambda_1 - \lambda_2)y - \gamma(\lambda_1) + \gamma(\lambda_2)\right] = (\lambda_1 - \lambda_2)\mu_1 - [\gamma(\lambda_1) - \gamma(\lambda_2)].$$

Now fix $\lambda_1 = \hat{\lambda}$, and let $\mu = \mu_\lambda$, then we can rearrange:

$$f_\lambda(y) = f_{\hat{\lambda}}(y) \exp\left(-\frac{1}{2}D(\hat{\lambda}, \lambda)\right).$$

The log-likelihood is maximized at $\hat{\lambda}$, the MLE, which satisfies

$$\hat{\lambda} = \arg\max_\lambda \log f_\lambda(y).$$

Therefore, by definition,

$$f_y(y) := \sup_\lambda f_\lambda(y) = f_{\hat{\lambda}}(y),$$

since the MLE gives the density that maximizes the likelihood at y.

Since $f_{\hat{\lambda}}(y) = f_y(y)$, the data likelihood under the true observation, we obtain:

$$f_\lambda(y) = f_y(y) \exp\left(-\frac{1}{2}D(y, \mu_\lambda)\right),$$

which completes the proof. □

Probit Analysis

The probit model is a type of generalized linear model (GLM) for binary outcomes, where the link function is the inverse cumulative distribution function (cdf) of the standard normal distribution, denoted by Φ^{-1}.

In this model, we assume the probability of success (e.g., $y_i = 1$) is related to the covariate x_i through the following relationship:

$$\Phi^{-1}(\pi_i) = \alpha_0 + \alpha_1 x_i,$$

where $\pi_i = \mathbb{P}(y_i = 1 \mid x_i)$. Equivalently, solving for π_i, we obtain:

$$\pi_i = \Phi(\alpha_0 + \alpha_1 x_i),$$

where Φ is the standard normal cdf:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

This model arises naturally from a latent variable formulation:

- Introduce a latent (unobserved) continuous variable $\tilde{X}_i \sim \mathcal{N}(\alpha_0 + \alpha_1 x_i, 1)$;
- Define the observed binary outcome as:

$$y_i = \begin{cases} 1 & \text{if } \tilde{X}_i \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the probability of observing $y_i = 1$ is:

$$\mathbb{P}(y_i = 1 \mid x_i) = \mathbb{P}(\tilde{X}_i \leq 0 \mid x_i) = \mathbb{P}\left(Z \leq -(\alpha_0 + \alpha_1 x_i)\right) = \Phi\left(-(\alpha_0 + \alpha_1 x_i)\right).$$

Alternatively, if we define $y_i = 1$ when $\tilde{X}_i \geq 0$, then:

$$\pi_i = \mathbb{P}(y_i = 1) = \mathbb{P}(\tilde{X}_i \geq 0) = \Phi(\alpha_0 + \alpha_1 x_i).$$

Thus, the probit model is based on the assumption that a latent normal variable governs the binary outcome, and the observed binary variable is simply a thresholding (sign) of this latent variable.

Poisson Regression

In Poisson regression, the response variable is assumed to follow a Poisson distribution:

$$y_i \sim \text{Poi}(\mu_i),$$

where the mean μ_i is linked to covariates via a log-link function:

$$\log \mu_i = x_i^\top \alpha.$$

Poisson Deviance: To assess the fit of the model for each observation, we use the deviance residual. The standardized deviance residual is given by:

$$Z_i = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{D(y_i, \hat{\mu}_i)},$$

where $D(y_i, \hat{\mu}_i)$ is the deviance contribution from observation i, which measures the discrepancy between the observed count and the fitted value under the Poisson model.

Goodness-of-Fit Test: To evaluate the overall model adequacy, we consider the test statistic:

$$S = \sum_{j,k} Z_{jk}^2,$$

where Z_{jk} denotes the standardized deviance residual for the observation in row j and column k, in the case of tabulated or grouped data. Under the null hypothesis that the model is correctly specified, S approximately follows a chi-squared distribution:

$$S \sim \chi^2_{\text{df}},$$

with degrees of freedom $\text{df} = n - p$, where n is the number of observations and p is the number of estimated parameters in the model.

Lecture 13 Hypothesis Testing

This lecture is based on the Chap. 9 of [Pol23], Sec. 8.3.4 and 10.3.1 of [CB02], [HRD18] and Secs. 15.1–15.3, 15.5 of [EH16].

Statistical Inference

Statistical inference is the process of drawing conclusions about a population based on data. It encompasses several core tasks:

- **Estimation:** Determining unknown parameters, either as point estimates or interval estimates.
- **Testing:** Making formal decisions about hypotheses. This is typically framed as:

  $$H_0 \text{ (null hypothesis) vs. } H_1 \text{ (alternative hypothesis)},$$

  or equivalently, choosing between competing models.
- **Prediction:** Using observed data to forecast future or unobserved outcomes. Often operationalized via machine learning methods.
- **Attribution:** Identifying influential variables or selecting relevant models to explain the data, often in the context of causal inference or model selection.

The Process of Null Hypothesis Testing

We can break down the process of null hypothesis statistical testing into a number of steps:

1. Formulate a hypothesis that embodies our prediction (before seeing the data).
2. Specify null and alternative hypotheses.
3. Collect some data relevant to the hypothesis.
4. Fit a model to the data that represents the alternative hypothesis and compute a test statistic.
5. Compute the probability of the observed value of that statistic assuming that the null hypothesis is true.
6. Assess the "statistical significance" of the result.

Problems with Hypothesis Testing

While hypothesis testing is widely used, it also has notable limitations:

1. **Large-sample sensitivity:** When the sample size n is very large, even tiny, practically meaningless effects can lead to statistically significant results, causing H_0 to be rejected by default.
2. **Counterintuitive conclusions:** The logic of "rejecting" or "failing to reject" H_0 based on arbitrary thresholds (e.g., $\alpha = 0.05$) can feel unintuitive or misleading, especially in real-world decision-making contexts.

P-values

The concept of a p-value is central to hypothesis testing, but it admits multiple formalizations depending on perspective. We present three common definitions below.

Definition

Definition 1 (P-value as tail probability):
Let T be a test statistic and t_{obs} the observed value.

- For a two-sided test: $p = \mathbb{P}(|T| \geq |t_{\text{obs}}|)$
- For a one-sided test: $p = \mathbb{P}(T \geq t_{\text{obs}})$ or $\mathbb{P}(T \leq t_{\text{obs}})$

The p-value is the probability, under the null hypothesis, of observing a test statistic at least as extreme as the one actually observed.

An alternative definition, motivated by decision-theoretic principles, views the p-value as the smallest significance level under which a particular test would reject the null hypothesis:

Definition

Definition 2 (Shao):
Let $T(X) \in \{0,1\}$ be a test function that maps data X to a decision: $T(X) = 1$ means "reject H_0", and $T(X) = 0$ means "fail to reject H_0".
Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be the null and alternative models.

- Type I error (false alarm, reject H_0 when it is true): $\alpha_T(p) = \mathbb{P}_p(T(X) = 1)$, for $p \in \mathcal{P}_0$
- Type II error (miss, retain H_0 when it is false): $1 - \alpha_T(p) = \mathbb{P}_p(T(X) = 0)$, for $p \in \mathcal{P}_1$

The goal is to minimize Type II error over $\mathcal{P}_1$, subject to the constraint:

$$\sup_{p \in \mathcal{P}_0} \alpha_T(p) \leq \alpha \quad \text{(significance level)}.$$

This leads to the definition of a p-value as:

$$\text{p-value} = \inf\{\alpha \in (0,1) : T(X) = 1 \text{ under level } \alpha\}$$

A third definition formalizes p-values through their calibration properties. This definition is central to modern theoretical treatments, such as those in multiple testing or false discovery rate control:

Definition

Definition 3 (C & B):
A function $p(X)$ is a valid p-value function if for all $\theta \in \Theta_0$ and $\alpha \in [0,1]$,

$$\mathbb{P}_\theta\left(p(X) \leq \alpha\right) \leq \alpha.$$

Remark

Although all three definitions aim to capture the notion of statistical significance, they are not equivalent in general. The first definition is tied to tail probabilities of specific test statistics; the second arises from test function thresholds; and the third emphasizes validity under null distributions. One must be cautious not to treat them interchangeably without checking conditions under which they coincide.

Theorem

Theorem: Probability Integral Transformation (PIT)
Let $W(X)$ be a continuous random variable with cumulative distribution function F_θ under $\theta \in \Theta_0$. Then:

$$U := F_\theta(W(X)) \sim \text{Uniform}(0,1).$$

Proof. We want to compute the distribution of the transformed variable $U = F_\theta(W(X))$. For any $u \in [0,1]$, consider:

$$\mathbb{P}(U \leq u) = \mathbb{P}(F_\theta(W(X)) \leq u).$$

Since F_θ is continuous and strictly increasing, we can apply the inverse function F_θ^{-1}, and equivalently write:

$$\mathbb{P}(F_\theta(W(X)) \leq u) = \mathbb{P}(W(X) \leq F_\theta^{-1}(u)) = F_\theta(F_\theta^{-1}(u)) = u.$$

Hence, the CDF of U is the identity function on $[0,1]$, which means:

$$F_U(u) = u \quad \text{for all } u \in [0,1],$$

so $U \sim \text{Unif}(0,1)$. □

Theorem

Proposition: Let $W(X)$ be a test statistic such that larger values of W provide stronger evidence against H_0. Define:

$$p(X) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(X) \geq W(x)).$$

Then $p(X)$ is a valid p-value.

Proof. Fix any $\theta \in \Theta_0$, and let F_θ denote the cumulative distribution function (CDF) of $W(X)$ under θ.

We define the pointwise (non-supremum) p-value under θ as:

$$p_\theta(x) := \mathbb{P}_\theta(W(X) \geq W(x)).$$

Now, using the standard transformation trick:

$$p_\theta(x) = \mathbb{P}_\theta(-W(X) \leq -W(x)) = F_\theta(-W(x)),$$

where the second equality follows from the definition of CDF for $-W(X)$.

By the Probability Integral Transform (PIT), the random variable $F_\theta(W(X)) \sim \text{Unif}(0,1)$, so $p_\theta(X)$ is stochastically larger than or equal to $\text{Unif}(0,1)$, which gives:

$$\mathbb{P}_\theta(p_\theta(X) \leq \alpha) \leq \alpha \quad \text{for all } \alpha \in [0,1].$$

Since

$$p(X) = \sup_{\theta \in \Theta_0} p_\theta(X),$$

we conclude:

$$\mathbb{P}_{\theta_0}(p(X) \leq \alpha) \leq \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p_\theta(X) \leq \alpha) \leq \alpha,$$

which verifies that $p(X)$ is a valid p-value. □

Remark

Intuitive Explanation: The probability that a random variable is less than or equal to itself under its own distribution is uniformly distributed.

Likelihood Ratio Tests

The likelihood ratio test (LRT) is a general-purpose method for hypothesis testing based on the ratio of maximized likelihoods.

- Suppose $X \sim P_\theta$ with $\theta \in \Theta$.
- We want to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1,$$

where $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$.

Definition

Likelihood Function:
Let $\mathcal{L}(\theta \mid x)$ be the likelihood function. The likelihood ratio statistic is defined as:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta \mid x)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta \mid x)} = \frac{\mathcal{L}(\hat{\theta}_0 \mid x)}{\mathcal{L}(\hat{\theta} \mid x)},$$

where $\hat{\theta}_0$ is the MLE under H_0, and $\hat{\theta}$ is the unconstrained MLE.

Smaller $\lambda(x)$ indicates that the observed data are less compatible with H_0. Hence, we reject H_0 for small values of $\lambda(x)$.

To perform a level-α test, we find a critical value c such that:

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(\lambda(X) \leq c\right) = \alpha.$$

Theorem

Neyman–Pearson Lemma (Simple vs. Simple Hypotheses):
Let X be a random variable with density $f_\theta(x)$, and consider testing

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1,$$

for two fixed parameter values $\theta_0 \neq \theta_1$.

Then the most powerful test of level α rejects H_0 when the likelihood ratio satisfies:

$$\Lambda(x) := \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \geq c, \text{ or } \frac{f_{\theta_0}(x)}{\max\{f_{\theta_0}(x), f_{\theta_1}(x)\}} \leq c$$

for some constant $c > 0$ chosen so that the size constraint is met:

$$\mathbb{P}_{\theta_0}(\Lambda(X) \geq c) = \alpha.$$

Equivalently, the test rejects H_0 in the region:

$$\mathcal{R} = \left\{ x : \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \geq c \right\}.$$

Proof. Let $\phi_{NP}(x)$ be the Neyman-Pearson (NP) test, which is a function that equals 1 if we reject H_0 and 0 otherwise. By definition, $\phi_{NP}(x) = 1$ if $f_{\theta_1}(x) \geq c f_{\theta_0}(x)$ and $\phi_{NP}(x) = 0$ otherwise. The size of this test is $\mathbb{E}_{\theta_0}[\phi_{NP}(X)] = \alpha$.

Let $\phi_A(x)$ be any other test with size less than or equal to α. That is, $\mathbb{E}_{\theta_0}[\phi_A(X)] \leq \alpha$. We want to show that the power of the NP test is greater than or equal to the power of test ϕ_A.

$$\text{Power}(\phi_{NP}) = \mathbb{E}_{\theta_1}[\phi_{NP}(X)] \geq \mathbb{E}_{\theta_1}[\phi_A(X)] = \text{Power}(\phi_A).$$

Consider the key inequality which holds for all x by the definition of the NP test:

$$(\phi_{NP}(x) - \phi_A(x))(f_{\theta_1}(x) - c f_{\theta_0}(x)) \geq 0.$$

To see why this holds, consider two cases:

- If $\phi_{NP}(x) = 1$, then $f_{\theta_1}(x) \geq c f_{\theta_0}(x)$. The second term is non-negative. Since $\phi_A(x) \leq 1$, the first term $(\phi_{NP}(x) - \phi_A(x))$ is also non-negative. Thus, the product is non-negative.
- If $\phi_{NP}(x) = 0$, then $f_{\theta_1}(x) < c f_{\theta_0}(x)$. The second term is negative. Since $\phi_A(x) \geq 0$, the first term $(\phi_{NP}(x) - \phi_A(x))$ is non-positive. Thus, the product is non-negative.

Since the inequality holds for all x, its integral over the sample space must also be non-negative:

$$\int (\phi_{NP}(x) - \phi_A(x))(f_{\theta_1}(x) - c f_{\theta_0}(x)) dx \geq 0.$$

Expanding this integral gives:

$$\int (\phi_{NP}(x) - \phi_A(x)) f_{\theta_1}(x) dx - c \int (\phi_{NP}(x) - \phi_A(x)) f_{\theta_0}(x) dx \geq 0.$$

Rewriting this in terms of expectations:

$$(\mathbb{E}_{\theta_1}[\phi_{NP}] - \mathbb{E}_{\theta_1}[\phi_A]) - c(\mathbb{E}_{\theta_0}[\phi_{NP}] - \mathbb{E}_{\theta_0}[\phi_A]) \geq 0.$$

By the size constraints, we know $\mathbb{E}_{\theta_0}[\phi_{NP}] = \alpha$ and $\mathbb{E}_{\theta_0}[\phi_A] \leq \alpha$. Therefore, the difference $(\mathbb{E}_{\theta_0}[\phi_{NP}] - \mathbb{E}_{\theta_0}[\phi_A]) \geq 0$. Since $c > 0$, the entire second term $c(\mathbb{E}_{\theta_0}[\phi_{NP}] - \mathbb{E}_{\theta_0}[\phi_A])$ is non-negative.

For the whole expression to be greater than or equal to zero, the first term must also be non-negative:

$$\mathbb{E}_{\theta_1}[\phi_{NP}] - \mathbb{E}_{\theta_1}[\phi_A] \geq 0.$$

This implies $\mathbb{E}_{\theta_1}[\phi_{NP}] \geq \mathbb{E}_{\theta_1}[\phi_A]$, which means the NP test is the most powerful test of level α. □

Theorem

Simple Wilk's Theorem:
Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x \mid \theta)$, and we test:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0,$$

where $\theta \in \mathbb{R}$. Let $\hat{\theta}$ be the MLE and define the likelihood ratio statistic:

$$\lambda(X) = \frac{\mathcal{L}(\theta_0 \mid X)}{\mathcal{L}(\hat{\theta} \mid X)}.$$

Then under regularity conditions, under H_0,

$$-2\log\lambda(X) \xrightarrow{d} \chi^2_1 \quad \text{as } n \to \infty.$$

Proof. Let $\ell(\theta) = \log \mathcal{L}(\theta \mid X)$. Perform Taylor expansion around $\hat{\theta}$:

$$\ell(\theta_0) \approx \ell(\hat{\theta}) + \dot{\ell}(\hat{\theta})(\theta_0 - \hat{\theta}) + \frac{1}{2}\ddot{\ell}(\hat{\theta})(\theta_0 - \hat{\theta})^2.$$

Since $\hat{\theta}$ is the MLE, $\dot{\ell}(\hat{\theta}) = 0$. Then:

$$-2\log\lambda(X) = -2\ell(\theta_0) + 2\ell(\hat{\theta}) \approx (\theta_0 - \hat{\theta})^2 \cdot (-\ddot{\ell}(\hat{\theta})).$$

Recall that under H_0,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0)), \quad -\ddot{\ell}(\hat{\theta}) \xrightarrow{p} nI(\theta_0).$$

Combining these and applying **Slutsky's theorem**, we conclude:

$$-2\log\lambda(X) \xrightarrow{d} \chi^2_1.$$

□

Theorem

Slutsky's Theorem
Let $X_n \xrightarrow{d} X$ (converges in distribution), and $Y_n \xrightarrow{p} c$ (converges in probability to constant c).
Then:

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n \cdot Y_n \xrightarrow{d} cX$
- If $c \neq 0$, then $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$

In words: If one sequence of random variables converges in distribution and another converges in probability to a constant, then their sum, product, or ratio (if denominator nonzero) converges in distribution to the corresponding combination.

Proof. We prove the result for the ratio case: Suppose $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c$, and $c \neq 0$. We want to show:

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}.$$

Let $f(x, y) = x/y$, which is continuous at all $y \neq 0$. By the Continuous Mapping Theorem, if $(X_n, Y_n) \xrightarrow{d} (X, c)$, then:

$$f(X_n, Y_n) = \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}.$$

So it suffices to show joint convergence: $(X_n, Y_n) \xrightarrow{d} (X, c)$. This follows from:

$$X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{p} c \quad \Rightarrow \quad (X_n, Y_n) \xrightarrow{d} (X, c)$$

by a well-known result: convergence in distribution + convergence in probability ⇒ joint convergence.

Hence, $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.

The same logic applies to $X_n + Y_n$ and $X_n Y_n$ using $f(x, y) = x + y$ or $f(x, y) = xy$. □

Theorem

General Wilk's Theorem:
Suppose $\theta \in \Theta \subseteq \mathbb{R}^d$, and we test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta \setminus \Theta_0,$$

where $\dim(\Theta) = d$, $\dim(\Theta_0) = d_0$. Then the likelihood ratio statistic

$$\lambda(X) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta \mid X)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta \mid X)}$$

satisfies, under regularity conditions and H_0,

$$-2 \log \lambda(X) \xrightarrow{d} \chi^2_r, \quad \text{with } r = d - d_0 = \dim(\Theta) - \dim(\Theta_0)$$

Proof. Let $\ell(\theta) = \log \mathcal{L}(\theta \mid X)$ be the log-likelihood function. We denote the unrestricted and restricted Maximum Likelihood Estimators (MLEs) as:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta), \quad \text{and} \quad \tilde{\theta} = \arg\max_{\theta \in \Theta_0} \ell(\theta).$$

The log-likelihood ratio statistic is $W = -2 \log \lambda(X) = 2 \left[\ell(\hat{\theta}) - \ell(\tilde{\theta}) \right]$.

The core of the proof is to analyze the difference in log-likelihoods using a Taylor series expansion of $\ell(\tilde{\theta})$ around the unrestricted MLE $\hat{\theta}$:

$$\ell(\tilde{\theta}) \approx \ell(\hat{\theta}) + \nabla \ell(\hat{\theta})^T (\tilde{\theta} - \hat{\theta}) + \frac{1}{2} (\tilde{\theta} - \hat{\theta})^T \nabla^2 \ell(\hat{\theta}) (\tilde{\theta} - \hat{\theta}).$$

By the definition of the unrestricted MLE $\hat{\theta}$, its score (gradient) is zero: $\nabla \ell(\hat{\theta}) = \mathbf{0}$:

$$\ell(\hat{\theta}) - \ell(\tilde{\theta}) \approx -\frac{1}{2} (\tilde{\theta} - \hat{\theta})^T \nabla^2 \ell(\hat{\theta}) (\tilde{\theta} - \hat{\theta}) = \frac{1}{2} (\hat{\theta} - \tilde{\theta})^T \left(-\nabla^2 \ell(\hat{\theta}) \right) (\hat{\theta} - \tilde{\theta}).$$

Therefore, the test statistic is approximately:

$$W = 2[\ell(\hat{\theta}) - \ell(\tilde{\theta})] \approx (\hat{\theta} - \tilde{\theta})^T \left(-\nabla^2 \ell(\hat{\theta}) \right) (\hat{\theta} - \tilde{\theta}).$$

Under standard regularity conditions and as $n \to \infty$, the observed Hessian matrix, when scaled, converges to the Fisher Information matrix $I(\theta_0)$ evaluated at the true parameter θ_0:

$$-\frac{1}{n} \nabla^2 \ell(\hat{\theta}) \xrightarrow{p} I(\theta_0).$$

Substituting this into our expression for W, we get:

$$W \approx (\hat{\theta} - \tilde{\theta})^T (nI(\theta_0)) (\hat{\theta} - \tilde{\theta}) = \left(\sqrt{n}(\hat{\theta} - \tilde{\theta}) \right)^T I(\theta_0) \left(\sqrt{n}(\hat{\theta} - \tilde{\theta}) \right).$$

The final step is to characterize the asymptotic distribution of this quadratic form. From the asymptotic theory of MLE, we know that under H_0 (where the true parameter is θ_0):

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1}).$$

More directly, $\sqrt{n}(\tilde{\theta} - \theta_0)$ can be seen as an asymptotic projection of $\sqrt{n}(\hat{\theta} - \theta_0)$ onto the $(d-r)$-dimensional tangent space of Θ_0. Consequently, the difference vector $\sqrt{n}(\hat{\theta} - \tilde{\theta})$ lies approximately in the r-dimensional orthogonal complement.

This vector, let's call it $V = \sqrt{n}(\hat{\theta} - \tilde{\theta})$, is asymptotically normal, and the quadratic form $V^T I(\theta_0) V$ is its squared Mahalanobis length. For a normal vector projected onto an r-dimensional space, this quadratic form follows a chi-squared distribution with r degrees of freedom.

Thus, we conclude that:

$$W = -2 \log \lambda(X) \xrightarrow{d} \chi^2_r,$$

where $r = \dim(\Theta) - \dim(\Theta_0)$ is the number of constraints imposed by the null hypothesis. □

Remark

The degrees of freedom r equal the number of constraints imposed by H_0, i.e., the difference in dimension between Θ and Θ_0. It reflects the number of parameters being tested.

Meta Analysis: Analysis of Analogues

Meta-analysis refers to the analysis of data from multiple independent studies addressing a common scientific question. A central task is to combine the individual p-values $p_1, \ldots, p_n$ from each study into a single test statistic.

Theorem

P-value distribution under H_0
If each null hypothesis H_0 is simple and each test yields a valid p-value p_i, and the null distribution is continuous, then:

$$p_i \sim \text{Uniform}(0,1), \quad \text{for } i = 1, \ldots, n.$$

Common Combination Methods

We assume $p_1, \ldots, p_n \overset{i.i.d.}{\sim} \text{Uniform}(0,1)$ under H_0, and that under H_1, the p-values tend to be small.

- **1. Fisher's Method:**
$$S_F = 2\sum_{i=1}^{n} \log p_i \quad \Rightarrow \quad S_F \sim -\chi^2_{2n} \text{ under } H_0.$$
- **2. Pearson's Method:**
$$S_P = -2\sum_{i=1}^{n} \log(1-p_i) \quad \Rightarrow \quad S_P \sim \chi^2_{2n} \text{ under } H_0.$$
- **3. George's Method:**
$$S_G = S_F + S_P = \sum_{i=1}^{n} \log \frac{p_i}{1-p_i}.$$
- **4. Edgington's Method:**
$$S_E = \sum_{i=1}^{n} p_i.$$
- **5. Stouffer's Method:**
$$S_S = \sum_{i=1}^{n} \Phi^{-1}(p_i),$$
where Φ^{-1} is the inverse standard normal CDF.
- **6. Tippett's Method:**
$$S_T = \min\{p_1, \ldots, p_n\}.$$

Remark

Different methods emphasize different alternative scenarios: Fisher's is sensitive to many moderately small p-values, Tippett's focuses on the minimum, and Stouffer's assumes Gaussianity of transformed p-values.

Theorem

Proposition: Assume $H_1 : p_i \sim \text{Beta}(a,b)$, for $a \in (0,1]$, $b \in [1,\infty)$, and $a < b$.
Then the UMP test statistic for combining $p_1, \ldots, p_n$ is:

$$W = \sum_{i=1}^{n} \left[w \log p_i + (1-w)\log(1-p_i) \right], \quad \text{where } w = \frac{b-1}{b-a}.$$

This corresponds to a weighted combination of Fisher's and Pearson's method:

$$W = wS_F + (1-w)S_P.$$

Proof. We want to construct the most powerful test for combining $p_1, \ldots, p_n$, assuming the following model:

- Under the null hypothesis H_0: $p_i \sim \text{Uniform}(0,1)$
- Under the alternative H_1: $p_i \sim \text{Beta}(a,b)$, with parameters $a < b$, and $a \in (0,1], b \in [1,\infty)$

The Neyman–Pearson Lemma tells us that the most powerful test rejects H_0 for large values of the likelihood ratio:

$$\prod_{i=1}^{n} \frac{f_1(p_i)}{f_0(p_i)},$$

or equivalently, rejects when:

$$\sum_{i=1}^{n} \log \frac{f_1(p_i)}{f_0(p_i)} \text{ is large.}$$

Now compute the density functions:
- Under H_0: $f_0(p) = 1$, since $p \sim \text{Unif}(0,1)$
- Under H_1: $f_1(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$, the Beta density
So the log-likelihood ratio becomes:

$$\log \frac{f_1(p)}{f_0(p)} = \log \left(\frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1} \right) = \log \frac{1}{B(a,b)} + (a-1)\log p + (b-1)\log(1-p).$$

Since $B(a,b)$ is a constant that does not depend on p, it can be ignored in the test statistic. Therefore, the relevant test statistic is:

$$W(p) = \sum_{i=1}^{n} \left[(a-1)\log p_i + (b-1)\log(1-p_i) \right].$$

We can express this as a convex combination of Fisher's and Pearson's statistics. Define:

$$w = \frac{b-1}{b-a}, \quad \text{then } 1-w = \frac{1-a}{b-a}.$$

Now factor out the scale:

$$W(p) = (b-a) \sum_{i=1}^{n} \left[w \log p_i + (1-w)\log(1-p_i) \right].$$

Since $b - a > 0$, the most powerful test again depends on the statistic:

$$\sum_{i=1}^{n} \left[w \log p_i + (1-w)\log(1-p_i) \right],$$

which is a weighted sum of $\log p_i$ and $\log(1-p_i)$, combining Fisher's and Pearson's methods.

□

Multiple Testing

The classical hypothesis testing problem concerns a single hypothesis H_0. However, in many modern applications (e.g., genomics, neuroscience), we often test thousands of hypotheses simultaneously. This introduces the problem of **multiple testing**, where we aim to control some global error metric—most commonly.

Definition

The **family-wise error rate** (FWER), defined as:

$$\text{FWER} = \mathbb{P}(\text{Reject at least one true } H_0^i).$$

Around 1995, the rise of **microarray data** triggered the development of multiple testing theory. In particular:

Example

Benjamini–Hochberg (1995): FDR Control

In their influential paper, Benjamini and Hochberg analyzed gene expression data where:

- $n = 102$: number of individuals (52 prostate cancer patients, 50 normal controls),
- $N = 6033$: number of genes measured for each individual,
- Data matrix: $X = (x_{ij}) \in \mathbb{R}^{6033 \times 102}$,
- Test: for each gene i, test H_{0i}: no difference between groups.

For each gene, a two-sample t-test is performed:

$$t_i \sim t_{100} \quad \Rightarrow \quad p_i = F_{100}(t_i), \quad z_i = \Phi^{-1}(p_i),$$

so that under H_0: $z_i \sim \text{Unif}(0,1)$, and under H_1: $z_i \sim \text{Unif}(\mu_i, 1)$, where $\mu_i \neq 0$.

Remark

Key problem: When $N \gg n$, even if all nulls are true and we choose $\alpha = 0.05$, we expect $N \cdot \alpha$ false discoveries simply by chance.

Remedy 1: FWER (Family-Wise Error Rate)

When conducting multiple hypothesis tests simultaneously, a major concern is the probability of making at least one false rejection.

Bonferroni's Procedure: A classical and conservative method to control FWER. Simply reject each individual null hypothesis H_{0i} if

$$p_i \leq \frac{\alpha}{N},$$

where N is the total number of hypotheses tested, and α is the desired overall error rate.

Let $I_0 = \{i : H_{0i} \text{ is true}\}$, and denote $N_0 = |I_0|$. Then we can bound the FWER as:

$$\begin{aligned}
\text{FWER} &= \mathbb{P}\left(\bigcup_{i \in I_0} \left\{p_i \leq \frac{\alpha}{N}\right\}\right) \\
&\leq \sum_{i \in I_0} \mathbb{P}\left(p_i \leq \frac{\alpha}{N}\right) \quad \text{(by union bound)} \\
&= N_0 \cdot \frac{\alpha}{N} \\
&\leq \alpha.
\end{aligned}$$

Holm's Step-Down Procedure: An improvement over Bonferroni's method that maintains FWER control but with more power. The steps are:

1. Order the p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$
2. Let $H_{0(1)}, \ldots, H_{0(N)}$ be the corresponding hypotheses
3. For each i, reject $H_{0(i)}$ if

$$p_{(j)} \leq \frac{\alpha}{N - j + 1} \quad \forall j \leq i.$$

Proof. Let $i_0 = N - N_0 + 1$, which marks the index of the first true null among the ordered hypotheses. Let $\hat{i}$ be the stopping index of Holm's procedure:

$$\hat{i} = \max\left\{i : p_{(j)} \leq \frac{\alpha}{N - j + 1}, \ \forall j \leq i\right\}.$$

We want to show that:

$$\mathbb{P}(\text{at least one true } H_{0(i)} \text{ is wrongly rejected}) \leq \alpha.$$

To see this, note that under the null,

$$\mathbb{P}(p_{(i)} > \alpha/N_0) \geq 1 - \alpha (\text{Same as in Bonferroni}).$$

But since i_0 is the index where true nulls begin, we have:

$$\mathbb{P}(p_{(i_0)} > \alpha/N_0 = \alpha/(N - i_0 + 1)) \leq \mathbb{P}(\hat{i} < i_0),$$

i.e., the probability that Holm stops before encountering any true nulls. Hence, we conclude:

$$\text{FWER} \leq \mathbb{P}(\hat{i} < i_0) \leq \alpha.$$

□

Remark

FWER requires *no false rejection* across all tests, which is very conservative in large-scale testing. In real-world applications, especially when N is large (e.g., genomics or microarray data), a small fraction of false discoveries is often acceptable — this motivates alternative criteria such as the False Discovery Rate (FDR).

Remedy 2: FDR (False-Discoveries Rate)

When dealing with large-scale multiple testing problems, controlling the FWER is often too conservative. An alternative is to control the **False Discovery Rate (FDR)** — the expected proportion of false positives among the rejected hypotheses.

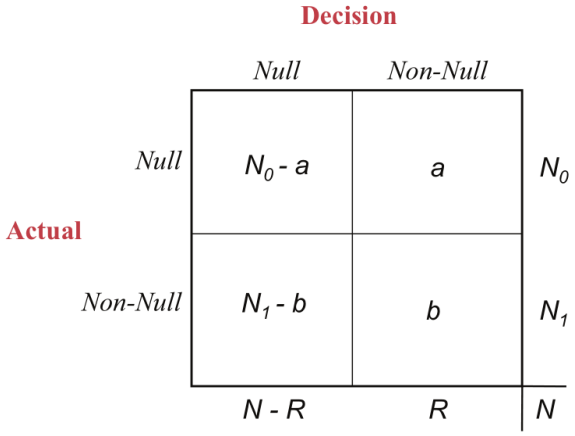

Figure 10: A decision rule D has rejected R out of N null hypotheses; a of these decisions were incorrect, i.e., they were "false discoveries," while b of them were "true discoveries".

Definition

FDR: We define the following:

$$\mathrm{FDP}(\mathbb{D}) = \frac{a}{R}, \quad \mathrm{FDR}(\mathbb{D}) = \mathbb{E}[\mathrm{FDP}(\mathbb{D})],$$

where FDP is the **false discovery proportion**, and FDR is its expectation.

Benjamini–Hochberg (BH) Procedure: We aim to control FDR at level q. Let the ordered p-values be $p_{(1)} \leq \cdots \leq p_{(N)}$. Define:

$$i_{\max} := \max\left\{ i : p_{(i)} \leq \frac{i}{N} \cdot q \right\},$$

then reject $H_{0(i)}$ for all $i \leq i_{\max}$.

Theorem

FDR Control under Independence: If the p-values are independent, and N_0 of them correspond to true nulls, then under the BH_q procedure:

$$\mathrm{FDR} = \pi_0 q = \frac{N_0}{N} q \leq q,$$

where $\pi_0 := \frac{N_0}{N}$ is the proportion of true nulls.

Proof. Let us define for any threshold $t \in (0, 1]$:

$$R(t) := \#\{i : p_i \leq t\}, \quad a(t) := \#\{i \in I_0 : p_i \leq t\},$$

where I_0 denotes the index set of true nulls.

Then define:

$$\mathrm{FDP}(t) := \frac{a(t)}{\max(R(t), 1)}, \quad Q(t) := \frac{Nt}{\max(R(t), 1)}.$$

Let:

$$t_q := \sup\{t \in (0,1] : Q(t) \le q\}.$$

Now, note that for the ordered p-value $p_{(i)}$, we have:

$$R(p_{(i)}) = i \quad \Rightarrow \quad Q(p_{(i)}) = \frac{Np_{(i)}}{i}.$$

Thus, the BH rejection rule becomes: reject all $H_{0(i)}$ such that $p_{(i)} \le t_q$.
Define:

$$A(t) := \frac{a(t)}{t}.$$

Assuming the null p-values are independent and uniformly distributed on $[0,1]$, it is known that $A(t)$ forms a nonnegative submartingale. Under certain regularity conditions (e.g., independence), $A(t)$ behaves like a martingale:

$$\mathbb{E}[A(s) \mid A(t)] = A(t), \quad \text{for all } s \le t.$$

By the Optional Stopping Theorem, we get:

$$\mathbb{E}[A(t_q)] = \mathbb{E}[A(1)] = \mathbb{E}[a(1)] = N_0.$$

Now relate this back to FDP:

$$\max(R(t_q), 1) = \frac{Nt_q}{Q(t_q)} = \frac{Nt_q}{q},$$

so:

$$\mathrm{FDP}(t_q) = \frac{a(t_q)}{\max(R(t_q),1)} = \frac{q}{N} \cdot \frac{a(t_q)}{t_q} = \frac{q}{N} A(t_q).$$

Taking expectation:

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}(t_q)] = \frac{q}{N} \cdot \mathbb{E}[A(t_q)] = \frac{qN_0}{N} = \pi_0 q.$$

□

Remark

Comparison Between the Three Methods:

- **Bonferroni vs. Holm:** Holm's step-down procedure uses thresholds $\frac{\alpha}{N-j+1}$, which are slightly looser than Bonferroni's uniform $\frac{\alpha}{N}$, especially when $N \gg j$.
- **Holm vs. Benjamini–Hochberg:** Consider the ratio of rejection thresholds:

$$\frac{\mathrm{Threshold}_{\mathrm{BH_q}}}{\mathrm{Threshold}_{\mathrm{Holm}}} = \frac{\frac{iq}{N}}{\frac{\alpha}{N-i+1}} = \frac{q}{\alpha}\left(1 - \frac{i-1}{N}\right) i.$$

  This shows that BH is more liberal, allowing more rejections and hence increasing power, but at the cost of allowing some false discoveries (in expectation).

Empirical Bayes Interpretation

The Benjamini–Hochberg (BH) procedure can also be viewed through the lens of an **empirical Bayes model**, which treats the observed test statistics as drawn from a mixture of null and non-null populations.

Assume the following mixture model:

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z),$$

where:

- $\pi_0 = \mathbb{P}(\text{null})$ is the prior probability a test is null,
- $f_0(z)$ is the density of z-scores under the null,
- $\pi_1 = \mathbb{P}(\text{nonnull}) = 1 - \pi_0$,
- $f_1(z)$ is the density of z-scores under the alternative.

Let us also define the corresponding survival functions (i.e., one minus CDFs):

$$S_0(z) = 1 - F_0(z), \quad S_1(z) = 1 - F_1(z), \quad S(z) = \pi_0 S_0(z) + \pi_1 S_1(z).$$

Then, the **Bayesian false discovery rate** at threshold z_0 is defined as:

$$\mathrm{Fdr}(z_0) = \mathbb{P}(\text{null} \mid z_i \geq z_0) = \frac{\pi_0 S_0(z_0)}{S(z_0)}.$$

This represents the probability that a test is actually null given that its statistic exceeds z_0.

Empirical estimation: To estimate $\mathrm{Fdr}(z_0)$ from data, assume $\pi_0 \approx 1$, and estimate the overall survival function $S(z_0)$ empirically as:

$$\hat{S}(z_0) = \frac{N(z_0)}{N}, \quad \text{where } N(z_0) := \#\{z_i \geq z_0\}.$$

Then the empirical estimator becomes:

$$\widehat{\mathrm{Fdr}}(z_0) = \frac{\pi_0 S_0(z_0)}{\hat{S}(z_0)}.$$

Connection to BH rule: Recall the BH rule rejects hypotheses with:

$$p_{(i)} \leq \frac{i}{N} q.$$

This is equivalent to saying:

$$S_0(z_{(i)}) \leq \hat{S}(z_{(i)}) q \quad \Leftrightarrow \quad \widehat{\mathrm{Fdr}}(z_{(i)}) \leq \pi_0 q.$$

In other words, the BH rule can be interpreted as rejecting all hypotheses where the estimated Bayesian FDR is below a target level $\pi_0 q$, assuming $\pi_0 = 1$. This gives a natural Bayesian justification to the BH procedure from the perspective of posterior probability control.

Remark

Choice of the Null Distribution: Something different happens in large-scale problems: with thousands of z-values to examine at once, the conventional theoretical null is inappropriate for the situation at hand. Put more positively, large-scale applications allow us to empirically determine a more realistic null distribution. An *MLE empirical null distribution* may be more appropriate.

Lecture 14 Survival Analysis

This lecture is based on the Chap. 9 of [EH16].

Grouped Data and Hazard Rates

Suppose X denotes a discrete lifetime variable. Let

$$f_i = \mathbb{P}(X = i)$$

be the probability of dying at age i, and

$$S_i = \mathbb{P}(X \geq i) = \sum_{j \geq i} f_j$$

be the probability of surviving to at least age i, also called **the survival function**.

Definition

Discrete Hazard Rate: The hazard rate at age i is defined as the conditional probability of dying at age i given survival until then:

$$h_i = \frac{f_i}{S_i} = \mathbb{P}(X = i \mid X \geq i).$$

The quantity h_i measures the risk of failure (death) at age i conditional on survival to that point. It can be estimated using grouped population data, where y_i is the number of deaths at age i, and n_i is the number at risk:

$$\hat{h}_i = \frac{y_i}{n_i}.$$

Definition

Conditional Survival Probability: Given survival past age i, the probability of surviving past age $j \geq i$ is

$$S_{ij} = \mathbb{P}(X > j \mid X \geq i) = \prod_{k=i}^{j} (1 - h_k).$$

This is the product of surviving each year between i and j.

In practice, we use plug-in estimators:

$$\hat{S}_{ij} = \prod_{k=i}^{j} (1 - \hat{h}_k).$$

This allows us to estimate the survival probability from life tables or grouped count data (e.g., from insurance datasets).

Continuous Time

In the continuous-time setting, we consider a non-negative random variable T, often referred to as "lifetime" or "time to event", with probability density function $f(t)$.

The survival function $S(t)$ gives the probability that the event has not occurred by time t:

$$S(t) = \mathbb{P}(T \geq t) = \int_t^\infty f(x)\, dx.$$

Definition

Hazard Rate: The hazard function $h(t)$ characterizes the instantaneous rate of failure at time t, given survival until time t:

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \delta t \mid T \geq t)}{\delta t}.$$

We can use the hazard function to describe the conditional survival probability from time t_0 to t_1. For $t_1 > t_0$, we have:

$$\mathbb{P}(T \geq t_1 \mid T \geq t_0) = \prod_{t=t_0}^{t_1} \left(1 - h(t)\, dt\right).$$

Using the approximation $\log(1 - h(t)\,dt) \approx -h(t)\,dt$, this becomes:

$$= \exp\left(\sum_{t=t_0}^{t_1} \log(1 - h(t)\,dt)\right) \approx \exp\left(-\sum_{t=t_0}^{t_1} h(t)\,dt\right) = \exp\left(-\int_{t_0}^{t_1} h(s)\,ds\right).$$

Theorem

Survival Function in Terms of Hazard: The survival function $S(t)$ is given by:

$$S(t) = \exp\left(-\int_0^t h(x)\,dx\right).$$

Proof. Since $f(t) = -\frac{d}{dt}S(t)$, we can write:

$$h(t) = -\frac{d}{dt}\log S(t).$$

Integrating both sides from 0 to t, we obtain:

$$\int_0^t h(x)\,dx = -\int_0^t \frac{d}{dx}\log S(x)\,dx = -\log S(t) + \log S(0).$$

Since $S(0) = 1$, we get:

$$\log S(t) = -\int_0^t h(x)\,dx,$$

and thus:

$$S(t) = \exp\left(-\int_0^t h(x)\,dx\right).$$

□

Example

Exponential Distribution. Let $f(t) = \frac{1}{c}e^{-t/c}$, for $t \geq 0$. Then:

- $S(t) = \int_t^\infty \frac{1}{c}e^{-x/c}\,dx = \exp(-t/c)$
- $h(t) = \frac{f(t)}{S(t)} = \frac{1}{c}$

The exponential distribution has a constant hazard rate and is memoryless, i.e.,

$$\mathbb{P}(T \geq t + \delta \mid T \geq t) = \mathbb{P}(T \geq \delta).$$

Censoring

In survival analysis, **right censoring** occurs when we only know that an individual's lifetime exceeds a certain value. In this case, we observe only the indicator $\mathbb{I}(T \geq t)$.

Each observed data point is recorded as $z_i = (t_i, d_i)$, where:

- t_i: the observed survival time (either time of death or censoring),
- $d_i \in \{0, 1\}$: event indicator, with:

$$d_i = \begin{cases} 1 & \text{if death is observed at time } t_i, \\ 0 & \text{if right-censored (i.e., death not observed.} \end{cases}$$

Kaplan-Meier Estimate

We consider n ordered observed survival times $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, with no ties, each associated with an indicator $d_{(k)} \in \{0, 1\}$ denoting whether a death was observed at $t_{(k)}$ (1 = death, 0 = censored).

Definition

Kaplan–Meier Estimator: The survival function estimate at time $t_{(j)}$ is:

$$\widehat{S}_{(j)} = \prod_{k \le j} \left(\frac{n-k}{n-k+1} \right)^{d_{(k)}},$$

where $d_{(k)} = 1$ if a death occurred at time $t_{(k)}$, and the denominator $n-k+1$ counts the number of individuals at risk just before $t_{(k)}$.

The estimator $\widehat{S}_{(j)}$ is a step function that only drops at observed death times $t_{(j)}$, and remains constant in between.

Variance Estimation

Theorem

Greenwood's Formula: Let $\widehat{S}_{(j)}$ denote the Kaplan–Meier estimate of the survival probability at ordered time $t_{(j)}$. Then its standard deviation is estimated by:

$$\mathrm{sd}(\widehat{S}_{(j)}) = \widehat{S}_{(j)} \cdot \left[\sum_{k \le j} \frac{y_k}{n_k(n_k - y_k)} \right]^{1/2},$$

where y_k is the number of deaths observed at time $t_{(k)}$, and n_k is the number of individuals at risk just before $t_{(k)}$.

Proof. The Kaplan–Meier estimate of the survival function at time $t_{(j)}$ is given by the product of conditional survival probabilities:

$$\widehat{S}_{(j)} = \prod_{k=1}^{j} \left(1 - \frac{y_k}{n_k} \right) = \prod_{k=1}^{j} \widehat{p}_k,$$

where $\widehat{p}_k = 1 - y_k/n_k$ is the estimated conditional probability of surviving past time $t_{(k)}$, given survival up to $t_{(k)}$.

To find the variance of a product, it is easier to work with the logarithm, which converts the product into a sum:

$$\log(\widehat{S}_{(j)}) = \log \left(\prod_{k=1}^{j} \widehat{p}_k \right) = \sum_{k=1}^{j} \log(\widehat{p}_k).$$

Assuming the terms $\log(\widehat{p}_k)$ are independent for different event times k, the variance of the sum is the sum of the variances:

$$\mathrm{Var}(\log(\widehat{S}_{(j)})) = \mathrm{Var} \left(\sum_{k=1}^{j} \log(\widehat{p}_k) \right) \approx \sum_{k=1}^{j} \mathrm{Var}(\log(\widehat{p}_k)).$$

We now need to find the variance of each term, $\mathrm{Var}(\log(\widehat{p}_k))$. Let $\widehat{q}_k = y_k/n_k$ be the estimated conditional probability of death at $t_{(k)}$. The number of deaths y_k among the n_k individuals at risk can be modeled as a binomial random variable, $y_k \sim \mathrm{Binomial}(n_k, q_k)$, where q_k is the true conditional probability of death.

The variance of the proportion $\widehat{q}_k$ is:

$$\mathrm{Var}(\widehat{q}_k) = \mathrm{Var} \left(\frac{y_k}{n_k} \right) = \frac{1}{n_k^2} \mathrm{Var}(y_k) = \frac{n_k q_k (1-q_k)}{n_k^2} = \frac{q_k(1-q_k)}{n_k}.$$

We use the Delta Method to find the variance of a function of $\widehat{q}_k$. We are interested in $g(\widehat{q}_k) = \log(\widehat{p}_k) = \log(1-\widehat{q}_k)$. The derivative is $g'(q_k) = -1/(1-q_k)$. By the Delta Method, the variance of $g(\widehat{q}_k)$ is approximately:

$$\begin{aligned} \mathrm{Var}(\log(1-\widehat{q}_k)) &\approx [g'(E(\widehat{q}_k))]^2 \, \mathrm{Var}(\widehat{q}_k) \\ &\approx \left(\frac{-1}{1-q_k} \right)^2 \frac{q_k(1-q_k)}{n_k} \\ &= \frac{1}{(1-q_k)^2} \frac{q_k(1-q_k)}{n_k} \\ &= \frac{q_k}{n_k(1-q_k)}. \end{aligned}$$

To obtain an estimate of this variance, we replace the true parameter q_k with its estimate $\widehat{q}_k = y_k/n_k$:

$$\widehat{\mathrm{Var}}(\log(\widehat{p}_k)) = \frac{\widehat{q}_k}{n_k(1-\widehat{q}_k)} = \frac{y_k/n_k}{n_k(1-y_k/n_k)} = \frac{y_k/n_k}{n_k\frac{n_k-y_k}{n_k}} = \frac{y_k}{n_k(n_k-y_k)}.$$

Summing these estimates gives the estimated variance of $\log(\widehat{S}_{(j)})$:

$$\widehat{\mathrm{Var}}(\log(\widehat{S}_{(j)})) = \sum_{k=1}^{j} \frac{y_k}{n_k(n_k-y_k)}.$$

Finally, we use the Delta Method again to relate the variance of $\log(\widehat{S}_{(j)})$ back to the variance of $\widehat{S}_{(j)}$. For a random variable X, we have $\mathrm{Var}(\log(X)) \approx \frac{\mathrm{Var}(X)}{[E(X)]^2}$. Rearranging and substituting our estimates, we get:

$$\widehat{\mathrm{Var}}(\widehat{S}_{(j)}) \approx [\widehat{S}_{(j)}]^2 \cdot \widehat{\mathrm{Var}}(\log(\widehat{S}_{(j)})).$$

Substituting the expression for $\widehat{\mathrm{Var}}(\log(\widehat{S}_{(j)}))$, we arrive at the estimated variance of the Kaplan–Meier estimator:

$$\widehat{\mathrm{Var}}(\widehat{S}_{(j)}) = [\widehat{S}_{(j)}]^2 \sum_{k=1}^{j} \frac{y_k}{n_k(n_k-y_k)}.$$

The standard deviation is the square root of the variance:

$$\mathrm{sd}(\widehat{S}_{(j)}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{S}_{(j)})} = \widehat{S}_{(j)} \left[\sum_{k=1}^{j} \frac{y_k}{n_k(n_k-y_k)}\right]^{1/2}.$$

This completes the proof. □

This variance estimate allows us to construct approximate confidence intervals for the survival function. In particular, vertical error bars in Kaplan–Meier plots typically correspond to $\widehat{S}_{(j)} \pm 1.96 \cdot \mathrm{sd}(\widehat{S}_{(j)})$, providing 95% confidence intervals.

Remark

In practice, $\widehat{S}(t)$ is calculated via a plug-in method, with stepwise updates at each death time. The variance estimate reflects accumulated uncertainty across observed timepoints.

Log-rank Test

The **log-rank test** (also known as the Mantel–Haenszel test) is a nonparametric method for comparing survival distributions between two groups, accounting properly for right censoring.

- It tests the null hypothesis that the hazard functions are equal across groups at all time points:

$$H_0 : h_{Ai} = h_{Bi} \quad \text{for all } i.$$

- This is done by comparing observed vs. expected deaths under H_0, across discrete time intervals (e.g., months).

Suppose in each time period i, we observe:

- n_{Ai}, n_{Bi}: numbers at risk in groups A and B;
- y_{Ai}, y_{Bi}: numbers of deaths in groups A and B;
- $n_i = n_{Ai} + n_{Bi}$, $y_i = y_{Ai} + y_{Bi}$: totals.

Under the null hypothesis, the expected number of deaths in group A at time i is:

$$E_i = \frac{n_{Ai}}{n_i} \cdot y_i,$$

with variance:

$$V_i = \frac{n_{Ai} n_{Bi} y_i (n_i - y_i)}{n_i^2 (n_i - 1)}.$$

Summing over time points gives the log-rank test statistic:

$$Z = \frac{\sum_{i=1}^{N}(y_{Ai} - E_i)}{\sqrt{\sum_{i=1}^{N} V_i}}.$$

Under H_0, the test statistic Z asymptotically follows a standard normal distribution:

$$Z \sim \mathcal{N}(0, 1).$$

Proportional Hazard Model

We consider a dataset $\{(t_i, d_i, \mathbf{Z}_i)\}$, where t_i is the observed survival time, $d_i \in \{0, 1\}$ is the censoring indicator (1 if death is observed), and $\mathbf{Z}_i \in \mathbb{R}^p$ is a vector of covariates or predictors.

Idea: We assume a baseline hazard $h_0(t)$, and that covariates act multiplicatively on the hazard. That is, under the Cox proportional hazard model:

$$h_i(t) = h_0(t) \cdot \exp(\mathbf{Z}_i^\top \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown coefficient vector. This is a semiparametric model (parametric in $\boldsymbol{\beta}$, nonparametric in $h_0(t)$).

From this model, the survival function becomes:

$$S_i(t) = \exp\left\{-\int_0^t h_i(s)\, ds\right\} = \exp\left\{-\exp(\mathbf{Z}_i^\top \boldsymbol{\beta}) \int_0^t h_0(s)\, ds\right\} = S_0(t)^{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta})},$$

where $S_0(t)$ is the baseline survival function.

Let $\theta_i = \exp(\mathbf{Z}_i^\top \boldsymbol{\beta})$, then $S_i(t) = S_0(t)^{\theta_i}$. Larger θ_i indicates faster hazard accumulation and thus shorter expected survival.

Partial Likelihood and Estimation

Let $T_{(1)} < T_{(2)} < \cdots < T_{(J)}$ be the ordered distinct event times, and for each j, let $\mathcal{R}_j = \{i : t_i \geq T_{(j)}\}$ be the risk set at time $T_{(j)}$.

Theorem

Lemma (Conditional Probability of Death under Proportional Hazards): Assume that only one individual dies at each observed failure time $T_{(j)}$, and let $\mathcal{R}_j = \{i : t_i \geq T_{(j)}\}$ denote the *risk set* at time $T_{(j)}$, i.e., the set of individuals still under observation just before time $T_{(j)}$. Let i_j be the index of the individual who dies at time $T_{(j)}$. Then under the proportional hazards model, the conditional probability that individual $i \in \mathcal{R}_j$ is the one who dies at time $T_{(j)}$ is:

$$\mathbb{P}(i_j = i \mid \mathcal{R}_j) = \frac{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta})}{\sum_{k \in \mathcal{R}_j} \exp(\mathbf{Z}_k^\top \boldsymbol{\beta})}.$$

Proof. Let $p_i = \mathbb{P}(\text{individual } i \text{ dies in } [T_{(j)}, T_{(j)} + \Delta t]) = h_i(T_{(j)})\Delta t$. Under the Cox proportional hazards model, this becomes:

$$p_i = h_0(T_{(j)}) \exp(\mathbf{Z}_i^\top \boldsymbol{\beta})\Delta t.$$

Let A_i denote the event that individual i dies at time $T_{(j)}$, while all other individuals in the risk set $\mathcal{R}_j \setminus \{i\}$ survive the small interval $[T_{(j)}, T_{(j)} + \Delta t]$. Then:

$$\mathbb{P}(A_i) = p_i \prod_{k \in \mathcal{R}_j \setminus \{i\}} (1 - p_k).$$

Since the events A_i for $i \in \mathcal{R}_j$ are disjoint, the probability that $i_j = i$ is:

$$\mathbb{P}(i_j = i \mid \mathcal{R}_j) = \frac{\mathbb{P}(A_i)}{\sum_{k \in \mathcal{R}_j} \mathbb{P}(A_k)} = \frac{p_i \prod_{k \in \mathcal{R}_j \setminus \{i\}} (1 - p_k)}{\sum_{k \in \mathcal{R}_j} p_l \prod_{l \in \mathcal{R}_j \setminus \{k\}} (1 - p_l)}.$$

Taking the limit $\Delta t \to 0$, we have $1 - p_k \approx 1$ for all k, and thus the products converge to 1. This simplifies the expression:

$$\mathbb{P}(i_j = i \mid \mathcal{R}_j) \approx \frac{p_i}{\sum_{k \in \mathcal{R}_j} p_k} = \frac{h_0(T_{(j)}) \exp(\mathbf{Z}_i^\top \boldsymbol{\beta})}{\sum_{k \in \mathcal{R}_j} h_0(T_{(j)}) \exp(\mathbf{Z}_k^\top \boldsymbol{\beta})} = \frac{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta})}{\sum_{k \in \mathcal{R}_j} \exp(\mathbf{Z}_k^\top \boldsymbol{\beta})}.$$

□

Thus, we have:

Definition

The **partial likelihood** (ignoring the baseline hazard) is:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\exp(\mathbf{Z}_{i_j}^\top \boldsymbol{\beta})}{\sum_{k \in \mathcal{R}_j} \exp(\mathbf{Z}_k^\top \boldsymbol{\beta})}.$$

Taking logs, the log-partial likelihood is:

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left[\mathbf{Z}_{i_j}^\top \boldsymbol{\beta} - \log \left(\sum_{k \in \mathcal{R}_j} \exp(\mathbf{Z}_k^\top \boldsymbol{\beta}) \right) \right].$$

We estimate $\boldsymbol{\beta}$ by maximizing the partial likelihood:

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}).$$

Asymptotic Inference

Let $\ddot{\ell}(\boldsymbol{\beta})$ be the observed information (Hessian matrix of the log-partial likelihood). Then under regularity conditions,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \ddot{\ell}(\boldsymbol{\beta})^{-1}),$$

and we can construct confidence intervals using the Wald approximation.

Lecture 15 Resampling Methods

This lecture is based on the Chap. 10-12 of [EH16].

Overview and Motivation

A fundamental question in statistical inference is: **How can we quantify uncertainty given only a single dataset?**

Classical approach:

- Assume a data generating model.
- Use asymptotic approximation (e.g., via Taylor expansion) to estimate variance and bias.

Modern idea: Perturb the data to simulate variability.

Resampling strategies:

- **Jackknife:** Leave-one-out (LOO) approach, systematically leaving out one observation at a time.
- **Bootstrap:** Resample the dataset with replacement to generate empirical distributions.
- **Cross-validation (CV):** Leave-one-fold-out; primarily used for model assessment and selection.

Inferential goals:

- **Jackknife/Bootstrap:** Estimate bias, standard error (SE), and construct confidence intervals (CIs).
- **CV:** Evaluate predictive performance, aid model selection.

Jackknife

Suppose the full dataset consists of i.i.d. observations:

$$X_i \overset{\text{i.i.d.}}{\sim} F, \quad i = 1, \ldots, n,$$

and let the estimator of interest be $\hat{\theta} = s(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

Leave-One-Out Estimation: Define the leave-one-out sample:

$$\mathbf{x}_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),$$

and compute the corresponding leave-one-out estimator:

$$\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)}).$$

Define the average of these leave-one-out estimates as:

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}.$$

Then the jackknife estimate of the standard error of $\hat{\theta}$ is:

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^2 \right]^{1/2}.$$

Example

Example: Jackknife Estimate for Sample Mean
Let $\hat{\theta} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the sample mean. For the leave-one-out sample $\mathbf{x}_{(i)}$, the jackknife estimate becomes:

$$\hat{\theta}_{(i)} = \frac{1}{n-1} \sum_{j \neq i} x_j = \frac{n\bar{x} - x_i}{n-1}.$$

Now compute the average of the jackknife estimates:

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} = \bar{x}.$$

So the deviation becomes:

$$\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} = \frac{n\bar{x} - x_i}{n-1} - \bar{x} = \frac{\bar{x} - x_i}{n-1}.$$

This shows that for the sample mean, the jackknife deviations are proportional to $\bar{x} - x_i$, and the jackknife standard error simplifies to:

$$\widehat{\text{se}}_{\text{jack}} = \left(\frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{\bar{x} - x_i}{n-1}\right)^2\right)^{1/2}.$$

Advantages:

- **Nonparametric:** Makes no distributional assumptions on F.
- **Automatic:** Can be applied mechanically to many estimators.

Limitations:

- **Sensitivity to smoothness:** The method may perform poorly if the statistic s is not smooth in sample size.
- **Biased upward:** The jackknife SE estimate is often biased upward. An alternative expression used for diagnostics is:

$$\hat{\text{se}}_{\text{jack}} = \left[\frac{1}{n^2}\sum_{i=1}^{n}D_i^2\right]^{1/2}, \quad \text{where } D_i = \frac{\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}}{c}, \quad c = \frac{1}{\sqrt{n(n-1)}}.$$

Bootstrap

The idea of bootstrap is to perform **resampling with replacement** in a more general manner, going beyond the leave-one-out deletion in jackknife.

We generate bootstrap samples:

$$x^{*b} = (x_1^{*b}, \cdots, x_n^{*b})^\top, \quad \text{for } b = 1, \ldots, B,$$

where each x_i^{*b} is sampled from the original data $x_1, \ldots, x_n$, with replacement.

On each bootstrap sample, we compute:

$$\hat{\theta}^{*b} = s(x^{*b})$$

Let $\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$ be the average.

Then, the bootstrap estimate of standard error is:

$$\widehat{\text{se}}_{\text{boot}} = \left[\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^*\right)^2\right]^{1/2}$$

This method simulates the sampling distribution of $\hat{\theta}$, under the assumption that the **empirical distribution** $\hat{F}$ is a good approximation to the true distribution F.

Bootstrap pipeline:

$$\text{Real World: } F \xrightarrow{\text{i.i.d.}} x \longrightarrow \hat{\theta}, \quad \text{vs.} \quad \text{Bootstrap: } \hat{F} \xrightarrow{\text{i.i.d.}} x^* \longrightarrow \hat{\theta}^* (\text{B times})$$

We replace the unknown distribution F by the empirical distribution $\hat{F}$, where $\hat{F}$ puts mass $1/n$ at each observed x_i. The use of sampling with replacement mimics i.i.d. draws from F.

This is called the **Substitution Principle**: when $n, B \to \infty$, the empirical bootstrap mimics the population behavior.

Example

Example: Bootstrap Implementation
Let $B = 200$. For constructing confidence intervals (CIs), typically we choose $B = 1000 \sim 2000$ to ensure stable variance estimates.

Comparison to MLE: Under MLE theory, the standard error is based on the Fisher Information:

$$\widehat{\text{se}}^2_{\text{Fisher}} = \left(nI(\hat{\theta})\right)^{-1}, \quad \text{(observed Fisher information)}$$

whereas the bootstrap directly replaces F by $\hat{F}$, without requiring any model specification.

Summary:

- Bootstrap is fully nonparametric and data-driven.
- It captures variability in $\hat{\theta}$ by empirically simulating its sampling distribution.
- It is widely used for standard errors, bias correction, and confidence interval construction.

Parametric Bootstrap

The **parametric bootstrap** assumes that the data come from a parametric family $\mathcal{F} = \{f_\mu(x) : \mu \in \Lambda\}$, and instead of resampling from the empirical distribution $\hat{F}$, we simulate from $f_{\hat{\mu}}$, where $\hat{\mu}$ is typically the MLE of the parameter μ.

Pipeline:

$$f_{\hat{\mu}} \longrightarrow x^* \longrightarrow \hat{\theta}^*$$

The key difference from the nonparametric bootstrap is that we **replace the empirical distribution** $\hat{F}$ with the estimated parametric model $f_{\hat{\mu}}$.

Example

Example 1: Normal Model
Assume $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, 1)$, and we estimate μ via $\hat{\mu} = \bar{X}$.
Then the parametric bootstrap sample is:

$$X_i^* \overset{i.i.d.}{\sim} \mathcal{N}(\bar{X}, 1)$$

Example

Example 2: Poisson Model for Binned Counts
Let Y_k be the number of data points in bin k, i.e., $Y_k = \#\{x_i \in \text{bin}_k\}$. Suppose

$$Y_k \overset{i.i.d.}{\sim} \text{Poisson}(\mu_k), \quad \text{with log-linear model} \quad \log \mu_k = \sum_{j=0}^{r} \beta_j x_k^j$$

We use MLE to obtain $\hat{\mu}_k$, then bootstrap:

$$Y_k^* \overset{i.i.d.}{\sim} \text{Poisson}(\hat{\mu}_k)$$

Remark

- The parametric bootstrap is more efficient when the model is correct.
- It requires stronger assumptions about the data-generating process.
- It is often used when the empirical distribution is too rough (e.g., small samples).

Resampling Plans

We can represent a general resampling scheme using a **resampling weight vector**

$$P = (P_1, \ldots, P_n)^\top \in \Delta_n,$$

where $\Delta_n = \{P \in \mathbb{R}^n : P_i \geq 0, \sum_{i=1}^n P_i = 1\}$ is the probability simplex.

Given a statistic $\hat{\theta} = s(\mathbf{x})$, we write it as a function of the data and the weight:

$$\hat{\theta} = s(\mathbf{x}) = s(P_0),$$

where $P_0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)^\top$ is the uniform weight used in the original empirical distribution.

Jackknife: The jackknife corresponds to resampling weights where one observation is left out:

$$P_{(i)} = \left(\frac{1}{n-1}, \ldots, \cancel{\frac{1}{n-1}}, \ldots, \frac{1}{n-1}\right)^\top \in \Delta_n,$$

with the i-th element set to zero. Alternatively written as:

$$P_{(i)} = \frac{1}{n-1}(1, \ldots, 1, 0, 1, \ldots, 1)^\top.$$

Bootstrap: The bootstrap samples correspond to resampling with replacement. Let $N = (N_1, \ldots, N_n)^\top$ be the counts of how many times each observation is drawn in a sample of size n. Then:

$$P^* = \frac{N}{n}, \quad \text{where } N \sim \text{Multinomial}(n, P_0).$$

Confidence Intervals

Definition

Confidence Interval: A $(1-\alpha)$-confidence interval for an unknown parameter θ is a random interval $C(X) = [L(X), U(X)]$, constructed from sample data X, such that:

$$\mathbb{P}_\theta\left(\theta \in C(X)\right) \geq 1 - \alpha.$$

This means that, over repeated samples from the distribution, the constructed interval will contain the true parameter at least $(1-\alpha) \times 100\%$ of the time.

Neyman's Construction and Pivot

To construct a $(1-\alpha)$-confidence interval for a parameter ϕ, the classical Neyman construction requires identifying a pivot.

Definition

Pivot: A pivot is a function $T(X, \theta)$ of the observed data X and the parameter θ such that the distribution of T is known and does not depend on θ. That is, $T(X, \theta) \sim F$ for some known distribution F, regardless of the true value of θ.

Suppose $\hat{\phi} \sim f_{\hat{\phi}|\phi}(r)$. We find interval bounds ϕ_ℓ and ϕ_u such that:

$$\int_{-\infty}^{\phi_\ell} f_{\hat{\phi}|\phi}(r)\, dr = \frac{\alpha}{2}, \quad \int_{\phi_u}^{\infty} f_{\hat{\phi}|\phi}(r)\, dr = \frac{\alpha}{2}.$$

Then the interval $C(\hat{\phi}) = [\phi_\ell, \phi_u]$ satisfies:

$$\mathbb{P}_\phi\left(\phi \in C(\hat{\phi})\right) = 1 - \alpha,$$

which defines a valid $(1-\alpha)$-level confidence interval for ϕ.

Once a pivot is constructed, one can invert its cumulative distribution function to construct an interval $C(\hat{\theta})$ satisfying:

$$\mathbb{P}_\theta\left(\hat{\theta}_\ell < \theta < \hat{\theta}_u\right) = 1 - \alpha,$$

which is the essence of the Neyman confidence interval framework.

Transformation Invariance

Suppose $\phi = m(\theta)$ is a monotonic transformation of θ. Then:

$$C^{\phi} = \left\{\phi = m(\theta) : \theta \in C(\hat{\theta})\right\}$$

is a $(1-\alpha)$-confidence interval for ϕ.

Example

Example: Fisher's Transformation for correlation coefficient.
Let

$$\phi = m(\theta) = \frac{1}{2}\log\left(\frac{1+\theta}{1-\theta}\right), \quad \text{then } \hat{\phi} \sim \mathcal{N}\left(\phi, \frac{1}{n-3}\right).$$

So the confidence interval for ϕ is:

$$C(\hat{\phi}) = \hat{\phi} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{1}{n-3}},$$

then inverse transform gives:

$$\theta = \frac{e^{2\phi}-1}{e^{2\phi}+1}.$$

Percentile Method

This is one of the most intuitive and widely used methods for constructing confidence intervals using the bootstrap.

Given a large number of bootstrap estimates $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$, we define the **empirical cumulative distribution function (c.d.f.)** as:

$$\hat{G}(t) = \frac{1}{B}\#\left\{\hat{\theta}^{*b} \le t\right\},$$

which gives the proportion of bootstrap samples less than or equal to t.

Then define the **empirical quantile function** as the inverse:

$$\hat{\theta}^*(\alpha) := \hat{G}^{-1}(\alpha),$$

which gives the value such that a proportion α of the bootstrap estimates are below it.

In other words, if you sort the $\hat{\theta}^{*b}$'s from smallest to largest:

$$\hat{\theta}^{*(1)} \le \hat{\theta}^{*(2)} \le \cdots \le \hat{\theta}^{*(B)},$$

then the α-quantile estimate is approximately:

$$\hat{\theta}^*(\alpha) \approx \hat{\theta}^{*(\lceil \alpha B \rceil)}.$$

Therefore, the **percentile bootstrap confidence interval** at level $1-\alpha$ is defined as:

$$CI_{1-\alpha}^{\text{percentile}} = \left[\hat{\theta}^*\left(\frac{\alpha}{2}\right), \hat{\theta}^*\left(1-\frac{\alpha}{2}\right)\right],$$

which simply takes the middle $100(1-\alpha)\%$ of the bootstrap distribution.

Prediction Error

Let the training dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and define a prediction rule:

$$\hat{y} = r_{\mathcal{D}}(x), \quad x \in \mathcal{X}.$$

Define the loss function $D(y, \hat{y})$, such as squared error $(y-\hat{y})^2$ or misclassification loss $\mathbb{I}(y \neq \hat{y})$. Assume training data $(x_i, y_i) \overset{i.i.d.}{\sim} F$, and that test point $(x_0, y_0) \sim F$ is independent of $\mathcal{D}$.

Definition

True error (risk):

$$\text{Err}_{\mathcal{D}} := \mathbb{E}_F[D(y_0, \hat{y}_0)], \quad \text{with } \mathcal{D} \text{ fixed and } (x_0, y_0) \text{ varying.}$$

Apparent error:

$$\text{err} := \frac{1}{n}\sum_{i=1}^{n} D(y_i, \hat{y}_i),$$

where $\hat{y}_i = r_{\mathcal{D}}(x_i)$, i.e. evaluating on training data itself.

Cross-Validation (CV)

Cross-validation is a general strategy for estimating prediction error by mimicking the process of applying a model to unseen data. The core idea is partitioning the data into two disjoint sets: one for training the model, and one for validating its performance.

Let $\mathcal{D}_{\text{val}} = \{(x_j, y_j)\}_{j=1}^{n_{\text{val}}}$ be the validation set, and let $\mathcal{D}_{\text{train}}$ be the training set (the rest of the data). Train the prediction rule $\hat{y}_j = r_{\mathcal{D}_{\text{train}}}(x_j)$, and compute the validation error:

$$\widehat{\text{Err}}_{\text{val}} = \frac{1}{n_{\text{val}}} \sum_{j=1}^{n_{\text{val}}} D(y_j, \hat{y}_j),$$

where $D(y, \hat{y})$ is a loss function such as squared error or 0–1 loss.

Leave-One-Out Cross-Validation (LOOCV)

LOOCV is a special case where we use one observation at a time for validation and the remaining $n - 1$ for training. For each $j \in \{1, \ldots, n\}$, define:

$$\hat{y}_{(j)} = r_{\mathcal{D}_{-j}}(x_j),$$

where $\mathcal{D}_{-j}$ denotes the training set with the j-th observation removed.

Then the LOOCV estimate of prediction error is:

$$\widehat{\text{Err}}_{\text{CV}} = \frac{1}{n} \sum_{j=1}^{n} D(y_j, \hat{y}_{(j)}).$$

Remark

LOOCV provides a nearly unbiased estimate of test error, since each model is trained on $n - 1$ data points. However, it may have high variance due to sensitivity to single observations.

Lecture 16 Stein's Phenomenon, Shrinkage and Ridge Regression

This lecture is based on the Chap. 7 of [EH16].

MLE and UMVUE

MLE (Maximum Likelihood Estimator): The MLE is typically nearly unbiased and has nearly minimum variance asymptotically. Under regularity conditions,

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

where $I(\theta)$ is the Fisher information.

Definition

UMVUE (Uniformly Minimum Variance Unbiased Estimator):
An estimator $\hat{\theta}$ is called the UMVUE of a parameter θ if it is unbiased for θ and has the smallest variance among all unbiased estimators of θ, uniformly over the parameter space.
For any estimator, the mean squared error (MSE) decomposes as:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta}).$$

Among unbiased estimators (i.e., $\mathrm{Bias}(\hat{\theta}) = 0$), minimizing MSE reduces to minimizing variance.

James–Stein Estimation

Bayesian Model

Assume a prior $\mu \sim \mathcal{N}(\mu_0, A)$, and likelihood $X \mid \mu \sim \mathcal{N}(\mu, 1)$. Then the posterior is:

$$\mu \mid X \sim \mathcal{N}(\hat{\mu}_{\text{Bayes}}, B), \quad B = \frac{A}{A+1}$$

with posterior mean:

$$\hat{\mu}_{\text{Bayes}} = (1 - B)\mu_0 + BX = \mu_0 + B(X - \mu_0)$$

Using $\mathrm{Var}[x] = \mathbb{E}\{\mathrm{Var}[X \mid Y]\} + \mathrm{Var}\{\mathbb{E}[X \mid Y]\}$, the MSE of this Bayes estimator is:

$$\begin{aligned}
\mathrm{MSE}(\hat{\mu}_{\text{Bayes}}) &= \left(\mathbb{E}[(\hat{\mu}_{\text{Bayes}} - \mu)]\right)^2 + \mathrm{Var}(\hat{\mu}_{\text{Bayes}} - \mu) \\
&= \left(\mathbb{E}(\mathbb{E}[(\hat{\mu}_{\text{Bayes}} - \mu) \mid \mu])\right)^2 + \mathrm{Var}(\hat{\mu}_{\text{Bayes}} - \mu) \\
&= (\mu_0 - \mu_0)^2 + \mathbb{E}\{\mathrm{Var}[(\hat{\mu}_{\text{Bayes}} - \mu) \mid \mu]\} + \mathrm{Var}\{\mathbb{E}[(\hat{\mu}_{\text{Bayes}} - \mu) \mid \mu]\} \\
&= B^2 + \mathrm{Var}((B-1)\mu) \\
&= \frac{A^2 + A}{(A+1)^2} = \frac{A}{A+1} = B.
\end{aligned}$$

while

$$\mathrm{MSE}(\hat{\mu}_{\text{MLE}}) = 1$$

James–Stein Estimator

In the multidimensional case:

$$\mu_i \sim \mathcal{N}(\mu_0, A), \quad X_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1) \text{ , independent}, \quad i = 1, \ldots, n$$

Then:

$$\hat{\mu}_i^{\text{Bayes}} = \mu_0 + B(X_i - \mu_0), \quad \mathrm{MSE} = nB$$
$$\hat{\mu}^{\text{MLE}} = \bar{X}, \quad \mathrm{MSE} = n$$

If $X\&A$ (or B) is unknown, assume:

$$X_i \sim \mathcal{N}(\mu_i, A+1), \quad \mathrm{Var}(X_i) = \mathbb{E}[\mathrm{Var}(X_i \mid \mu_i)] + \mathrm{Var}(\mathbb{E}(X_i \mid \mu_i)) = 1 + A$$

Let:

$$\hat{\mu} = \bar{X}, \quad \hat{A} = \frac{S}{\alpha(n) - 1}, \quad \alpha(n) : \text{ sample size}$$

$$S = \sum_{i=1}^{n} (X_i - \bar{X})^2, \quad \hat{B} = \frac{\hat{A}}{\hat{A} + 1} = 1 - \frac{\alpha(n)}{S}$$

To make $\hat{B}$ unbiased, use: $\mathbb{E}\left[\frac{1}{\chi^2_{n-1}}\right] = \frac{1}{n-3}$, we have:

$$\mathbb{E}(\frac{1}{S}) = \frac{1}{A+1}, \mathbb{E}(\frac{1}{Y}) = \frac{1}{(A+1)(n-3)} \Rightarrow \mathbb{E}(\hat{B}) = B \Rightarrow \alpha(n) = n - 3$$

Finally, we have the **James–Stein Estimator:**

$$\hat{\mu}_i^{\text{JS}} = \hat{\mu} + \hat{B}(X_i - \hat{\mu}), \quad i = 1, \ldots, n$$

$$\text{MSE}(\hat{\mu}^{\text{JS}}) = nB + 3(1 - B)$$

Theorem

James–Stein Theorem
Suppose $X_i \sim \mathcal{N}(\mu_i, 1)$ independently for $i = 1, \ldots, N$, with $N \geq 4$. Then

$$\mathbb{E}\left[\|\hat{\mu}^{\text{JS}} - \mu\|^2\right] < \mathbb{E}\left[\|\hat{\mu}^{\text{MLE}} - \mu\|^2\right] = N, \quad \text{for all } \mu \in \mathbb{R}^N.$$

Proof. The risk of the Maximum Likelihood Estimator (MLE) $\hat{\mu}^{\text{MLE}} = X$ is:

$$\mathbb{E}\left[\|\hat{\mu}^{\text{MLE}} - \mu\|^2\right] = \mathbb{E}\left[\|X - \mu\|^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[(X_i - \mu_i)^2\right] = \sum_{i=1}^{N} \text{Var}(X_i) = N.$$

To evaluate the risk of the James-Stein estimator $\hat{\mu}^{\text{JS}}$, we use Stein's Lemma. For $X \sim \mathcal{N}(\mu, I_N)$ and a differentiable function $h : \mathbb{R}^N \to \mathbb{R}^N$, the lemma states:

$$\mathbb{E}\left[(X - \mu) \cdot h(X)\right] = \mathbb{E}\left[\nabla \cdot h(X)\right], \quad \text{where } \nabla \cdot h = \sum_{i=1}^{N} \frac{\partial h_i}{\partial X_i}.$$

The risk of $\hat{\mu}^{\text{JS}} = \left(1 - \frac{N-2}{\|X\|^2}\right) X$ is $\mathbb{E}\left[\|\hat{\mu}^{\text{JS}} - \mu\|^2\right]$. We expand the squared norm:

$$\begin{aligned}
\|\hat{\mu}^{\text{JS}} - \mu\|^2 &= \left\| (X - \mu) - \frac{N-2}{\|X\|^2} X \right\|^2 \\
&= \|X - \mu\|^2 - 2(N-2)\frac{X \cdot (X - \mu)}{\|X\|^2} + \frac{(N-2)^2}{\|X\|^2}.
\end{aligned}$$

Taking the expectation:

$$\mathbb{E}\left[\|\hat{\mu}^{\text{JS}} - \mu\|^2\right] = \mathbb{E}\left[\|X - \mu\|^2\right] - 2(N-2)\,\mathbb{E}\left[\frac{X \cdot (X - \mu)}{\|X\|^2}\right] + (N-2)^2\,\mathbb{E}\left[\frac{1}{\|X\|^2}\right].$$

We apply Stein's Lemma to the middle term with $h(X) = \frac{X}{\|X\|^2}$. The divergence of $h(X)$ is:

$$\nabla \cdot h(X) = \sum_{i=1}^{N} \frac{\partial}{\partial X_i} \left(\frac{X_i}{\sum_j X_j^2} \right) = \sum_{i=1}^{N} \frac{\|X\|^2 - 2X_i^2}{\|X\|^4} = \frac{N\|X\|^2 - 2\|X\|^2}{\|X\|^4} = \frac{N-2}{\|X\|^2}.$$

By the lemma, $\mathbb{E}\left[\frac{X \cdot (X-\mu)}{\|X\|^2}\right] = \mathbb{E}\left[\frac{N-2}{\|X\|^2}\right]$. Substituting this back:

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\mu}^{\text{JS}} - \mu\|^2\right] &= N - 2(N-2)\mathbb{E}\left[\frac{N-2}{\|X\|^2}\right] + (N-2)^2\mathbb{E}\left[\frac{1}{\|X\|^2}\right] \\
&= N - 2(N-2)^2\mathbb{E}\left[\frac{1}{\|X\|^2}\right] + (N-2)^2\mathbb{E}\left[\frac{1}{\|X\|^2}\right] \\
&= N - (N-2)^2\mathbb{E}\left[\frac{1}{\|X\|^2}\right].
\end{aligned}$$

For $N \geq 3$, the term $(N-2)^2$ is strictly positive. Since $\|X\|^2 > 0$ with probability 1, its expectation $\mathbb{E}\left[\frac{1}{\|X\|^2}\right]$ is also positive. Therefore,

$$\mathbb{E}\left[\|\hat{\mu}^{\text{JS}} - \mu\|^2\right] = N - (\text{a positive quantity}) < N.$$

This holds for all $\mu \in \mathbb{R}^N$. □

Shrinkage Effect

Assume the mean vector is zero: $\mu_i = 0$, and the observations are:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, A+1), \quad i = 1, \ldots, n.$$

We compute the expected sum of squares:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] = n(A+1).$$

This is larger than $\mathbb{E}\left[\sum_{i=1}^{n} \mu_i^2\right] = nA$, indicating that the raw observations are overdispersed.

Now let us shrink each X_i using $\tilde{\mu}_i = \sqrt{B}X_i$, where $B = \frac{A}{A+1}$ is the Bayes shrinkage factor.

Then the expected squared norm becomes:

$$\mathbb{E}\left[\sum_{i=1}^{n}(\tilde{\mu}_i)^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(\sqrt{B}X_i)^2\right] = B \cdot \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] = B \cdot n(A+1) = nA.$$

Using the shrinkage estimate $\tilde{\mu}_i = \sqrt{B}X_i$ recovers the correct expected sum of squares nA, matching the true dispersion of the latent μ_i's. This motivates shrinkage: although raw data $X_i \sim \mathcal{N}(0, A+1)$ are noisy (overdispersed), shrinking them appropriately yields a better estimate of the underlying signal μ_i.

Ridge Regression

For high-dimensional linear regression, we consider the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}[\boldsymbol{\varepsilon}] = 0, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n.$$

In such settings, the ordinary least squares (OLS) estimator:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \ \textbf{i.e., } \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}$$

may not exist if $\mathbf{X}^\top\mathbf{X}$ is not invertible, e.g., when $p > n$.

It pushes us to make some regularizations, like **Ridge Estimator:**

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}.$$

Bayesian Interpretation

There is a Bayesian rationale for ridge regression.
Assume a prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda}\mathbf{I}_p)$, and observation model $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. Then the posterior distribution is:

$$\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}} \sim \mathcal{N}\left((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}},\ \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\right),$$

and the posterior mean is:

$$\mathbb{E}[\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}.$$

Penalized Optimization Formulation

Regularization methods provide a principled way to control model complexity and improve generalization, especially in high-dimensional settings. The key idea is to add a penalty term to the loss function that discourages overly large coefficients $\boldsymbol{\beta}$. This penalization reduces variance at the cost of introducing some bias.

- **Ridge Regression:** The ridge estimator solves the following penalized least squares problem:

  $$\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\right\}.$$

  When $\lambda > 0$, it shrinks the estimated coefficients toward the origin, introducing bias but reducing variance. This formulation corresponds to what is called penalized likelihood, MAP estimation, or more generally, regularized regression.

- **Lasso Regression:** An alternative and highly influential penalty is the ℓ_1-norm:

  $$\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\right\},$$

  where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$. This encourages sparsity—many coefficients in $\hat{\boldsymbol{\beta}}$ may be exactly zero. The resulting estimator is known as the **lasso**, and is widely used for variable selection in addition to regularization.

Lecture 17 Causal Inference

This lecture is based on [Was20].

Causal Questions

For years, causal inference was studied by statisticians, epidemiologists and economists. We first have the difference between prediction and causation. Prediction passively observes $X = x$ but causation actively intervening and setting $X = x$ is significant and requires different techniques and, typically, much stronger assumptions.

Prediction vs. Causation

- **Prediction:** Predict y after *observing* $X = x$. This is passive observation.
- **Causation:** Predict y after *setting* $X = x$. This reflects active intervention.

Definition

Elements in Causal Inference:

- X: Treatment or exposure (the cause)
- Y: Outcome or response (the effect)
- Causal questions typically ask:
  - Is there a causal effect? (Yes/No)
  - What is the size of the effect?
- More generally, we may ask: What are the causal relationships among a set of variables? This is known as causal discovery.

Types of Data

In causal inference, we commonly encounter three types of data sources:

- **Controlled, Randomized Experiments:** Treatment is randomly assigned to subjects, ensuring independence between treatment and potential outcomes.
- **Observational Studies:** No randomization is involved. Treatment assignment may be related to confounding variables, making causal inference more challenging.
- **Quasi-Experiments:** Although there is no direct intervention by the researcher, external events or policies create variation in treatment exposure that may approximate random assignment.

Framework

There are several mathematical frameworks to formalize and study causality:

- **Counterfactual / Potential Outcomes:** Think in terms of what would happen under different hypothetical interventions.
- **Causal Graphs:** Use directed acyclic graphs (DAGs) to represent and reason about causal structures.
- **Structural Equation Models (SEMs):** Represent each variable as a function of its causes plus randomness.

Remark

Association does not imply causation (may have confounders).
In general, even if X and Y are statistically dependent (associated), it does not mean that changing X would cause a change in Y. This distinction is captured by:

$$\mathbb{P}(Y \mid X = x) \neq \mathbb{P}(Y \mid \mathrm{do}(X = x))$$

The left-hand side refers to passive observation, while the right-hand side refers to an active intervention.

Counterfactual Framework

In the counterfactual framework, we think of potential outcomes for each individual under different treatment conditions.

Suppose we observe data $(X_i, Y_i) \sim (X, Y)$. For each unit, we define:

$$Y(x) := \text{the value of } Y \text{ if we intervene and set } X = x, \text{ i.e., } Y(x) = Y \mid \text{do}(X = x)$$

Only one of these potential outcomes is ever observed — the one corresponding to the actual treatment received. The rest are counterfactuals, and causal inference is essentially about reasoning from the observed data to learn about these unobserved outcomes.

Binary Treatment

We consider the case of a binary treatment: $X \in \{0, 1\}$. For each individual, define:

$$Y(0),\ Y(1) \quad \text{as the potential outcomes under control and treatment}$$

$$Y = X \cdot Y(1) + (1 - X) \cdot Y(0)$$

Only one of $Y(0)$ or $Y(1)$ is observed, depending on whether the individual received treatment.

Complete data: $(X, Y(0), Y(1), Y)$

X	Y	$Y(0)$	$Y(1)$
1	1	*	1
1	0	*	0
0	1	1	*
0	0	0	*

* indicates the counterfactual (unobserved) outcome.

Definition

Average Treatment Effect (ATE) :
The parameter of interest is:

$$\theta := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Remark

The parameter θ has the following interpretation: θ is the mean response if we exposed everyone minus the mean response if we exposed no-one.

Theorem

Lemma: In general,

$$\mathbb{E}[Y(1)] \neq \mathbb{E}[Y \mid X = 1], \quad \mathbb{E}[Y(0)] \neq \mathbb{E}[Y \mid X = 0].$$

Proof. Note that:

$$\mathbb{E}[Y \mid X = 1] = \mathbb{E}[Y(1) \mid X = 1],$$

because if $X = 1$, then $Y = Y(1)$. However, unless $X \perp Y(1)$, we cannot conclude that

$$\mathbb{E}[Y(1) \mid X = 1] = \mathbb{E}[Y(1)].$$

Indeed, selection bias may cause treated individuals to differ systematically from the general population. So:

$$\mathbb{E}[Y(1)] = \mathbb{E}_X[\mathbb{E}[Y(1) \mid X]] \neq \mathbb{E}[Y(1) \mid X = 1],$$

in general.

The same argument holds for $Y(0)$. Hence:

$$\mathbb{E}[Y(1)] \neq \mathbb{E}[Y \mid X = 1], \quad \mathbb{E}[Y(0)] \neq \mathbb{E}[Y \mid X = 0].$$

□

Theorem

Theorem: In general, there does not exist a uniformly consistent estimator of

$$\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

based solely on observable data (X, Y).

Proof. The proof proceeds by construction. We will define two joint distributions, $p(x, y_0, y_1)$ and $q(x, y_0, y_1)$, over the full potential outcome space such that their corresponding causal effects are different, i.e., $\theta(p) \neq \theta(q)$, but they induce the same distribution over the observable data (X, Y). If such distributions exist, no estimator based on observables can distinguish between them.

Let treatment assignment $X \sim \text{Bernoulli}(1/2)$.

Distribution p: Let the potential outcomes $(Y(0), Y(1))$ be independent of X and fixed for all individuals:

$$P_p\big((Y(0), Y(1)) = (0, 1)\big) = 1.$$

The observable outcome is $Y = X \cdot Y(1) + (1 - X) \cdot Y(0) = X \cdot 1 + (1 - X) \cdot 0 = X$. Thus, the observable data distribution is $P_p(Y = X) = 1$, which means $P_p(X = 1, Y = 1) = 1/2$ and $P_p(X = 0, Y = 0) = 1/2$. The causal effect under p is:

$$\theta(p) = \mathbb{E}_p[Y(1)] - \mathbb{E}_p[Y(0)] = 1 - 0 = 1.$$

Distribution q: Let the joint distribution of $(X, Y(0), Y(1))$ be defined as follows:

$$P_q(X = 1, Y(0) = 0, Y(1) = 1) = 1/2$$
$$P_q(X = 0, Y(0) = 0, Y(1) = 0) = 1/2$$

and all other combinations have probability zero.

Let's check the observable data distribution under q.

- The probability of observing $(X = 1, Y = 1)$ is $P_q(X = 1, Y(1) = 1) = P_q(X = 1, Y(0) = 0, Y(1) = 1) = 1/2$.
- The probability of observing $(X = 0, Y = 0)$ is $P_q(X = 0, Y(0) = 0) = P_q(X = 0, Y(0) = 0, Y(1) = 0) = 1/2$.

This is identical to the observable distribution under p.

Now, let's compute the causal effect under q. We need the marginal expectations of the potential outcomes.

- For $Y(1)$:

$$P_q(Y(1) = 1) = P_q(X = 1, Y(0) = 0, Y(1) = 1) = 1/2.$$
$$P_q(Y(1) = 0) = P_q(X = 0, Y(0) = 0, Y(1) = 0) = 1/2.$$
$$\mathbb{E}_q[Y(1)] = 1 \cdot P_q(Y(1) = 1) + 0 \cdot P_q(Y(1) = 0) = 1/2.$$

- For $Y(0)$:

$$P_q(Y(0) = 0) = P_q(X = 1, Y(0) = 0, Y(1) = 1) + P_q(X = 0, Y(0) = 0, Y(1) = 0) = 1.$$
$$\mathbb{E}_q[Y(0)] = 0 \cdot P_q(Y(0) = 0) = 0.$$

The causal effect under q is:

$$\theta(q) = \mathbb{E}_q[Y(1)] - \mathbb{E}_q[Y(0)] = \frac{1}{2} - 0 = \frac{1}{2}.$$

Conclusion: We have constructed two worlds, p and q, where the distribution of observable data (X, Y) is identical, but the true causal effects are different $(\theta(p) = 1 \neq \theta(q) = 1/2)$. Any estimator $\hat{\theta}_n$ that is a function of the observed data $(X_i, Y_i)_{i=1}^n$ will have the same sampling distribution under both p and q. Therefore, it cannot converge to both 1 and 1/2.

This formally means that for any such estimator $\hat{\theta}_n$, its worst-case error over a set of possible distributions $\mathcal{P}$ (containing distributions like p and q) does not go to zero. For any $\epsilon < 1/4$,

$$\sup_{P \in \{p,q\}} \mathbb{P}_P\left(|\hat{\theta}_n - \theta(P)| > \epsilon\right) \not\to 0 \quad \text{as } n \to \infty.$$

This proves that no uniformly consistent estimator of θ exists in general without further assumptions. □

Remark

Continuous Treatment: Let $X \in \mathbb{R}$ be continuous. Then the causal quantity becomes:

$$\theta(x) := \mathbb{E}[Y(x)] \quad \text{vs.} \quad m(x) := \mathbb{E}[Y \mid X = x].$$

Again, unless we assume ignorability or randomization, $\theta(x) \neq m(x)$.

Fortunately, there are two ways1 to make θ estimable.

Method 1: Randomization

One way to make the causal effect θ identifiable is to conduct a randomized experiment. If the treatment X is randomly assigned, then $X \perp (Y(0), Y(1))$. This assumption eliminates confounding.

Theorem

Theorem (Causal Identifiability under Randomization):
If $X \perp (Y(0), Y(1))$, then:

$$\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0] =: \alpha.$$

A consistent estimator is the plug-in (difference-in-means) estimator:

$$\hat{\alpha} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i} - \frac{\sum_{i=1}^n (1 - X_i) Y_i}{\sum_{i=1}^n (1 - X_i)}.$$

This estimator is uniformly consistent under the condition $\mathbb{P}(X = 1), \mathbb{P}(X = 0) > \delta > 0$.

Proof. Since $X \perp (Y(0), Y(1))$, we have:

$$\mathbb{E}[Y \mid X = 1] = \mathbb{E}[Y(1)], \quad \mathbb{E}[Y \mid X = 0] = \mathbb{E}[Y(0)],$$

which implies:

$$\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0] = \alpha.$$

To prove that $\hat{\alpha} \to \theta$ in probability uniformly over a class of distributions, we proceed by expressing $\hat{\alpha}$ explicitly in terms of empirical averages.

Define the components:

$$A_n = \sum_{i=1}^n X_i Y_i, \quad B_n = \sum_{i=1}^n X_i, \quad C_n = \sum_{i=1}^n (1 - X_i) Y_i, \quad D_n = \sum_{i=1}^n (1 - X_i).$$

Then the plug-in estimator becomes:

$$\hat{\alpha} = \frac{A_n}{B_n} - \frac{C_n}{D_n} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i} - \frac{\frac{1}{n}\sum_{i=1}^n (1 - X_i) Y_i}{\frac{1}{n}\sum_{i=1}^n (1 - X_i)}.$$

We define the population quantities:

$$A := \mathbb{E}[XY] = \mathbb{E}[Y \mid X = 1] \cdot \mathbb{E}[X], \quad B := \mathbb{E}[X], \quad C := \mathbb{E}[(1-X)Y] = \mathbb{E}[Y \mid X = 0] \cdot \mathbb{E}[1-X], \quad D := \mathbb{E}[1-X].$$

Then the population version of $\hat{\alpha}$ is:

$$\alpha = \frac{A}{B} - \frac{C}{D} = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0] = \theta,$$

under the randomization assumption $X \perp (Y(0), Y(1))$.

We now analyze the convergence of $\hat{\alpha}$ to θ. Observe that:

By the Law of Large Numbers, for each term (e.g., A_n/n), we have:

$$\frac{A_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i Y_i \xrightarrow{P} \mathbb{E}[XY] = A,$$

and similarly for $B_n/n \to B$, $C_n/n \to C$, and $D_n/n \to D$.

Using Hoeffding's inequality (for bounded $Y_i \in [0,1]$ or sub-Gaussian bounded support), we can get high-probability deviation bounds of the form:

$$\mathbb{P}\left(\left|\frac{A_n}{n} - A\right| > \epsilon\right) \leq 2\exp(-2n\epsilon^2),$$

and similarly for the other terms.

Hence, for any small $\epsilon > 0$, with high probability we have:

$$\left|\frac{A_n}{B_n} - \frac{A}{B}\right| \leq \epsilon_1, \quad \left|\frac{C_n}{D_n} - \frac{C}{D}\right| \leq \epsilon_2,$$

for some constants $\epsilon_1, \epsilon_2 \to 0$ as $n \to \infty$.

By a union bound over all four empirical sums, we get:

$$\mathbb{P}\left(|\hat{\alpha} - \theta| > \epsilon\right) \to 0.$$

In fact, since the convergence is uniform over the class $\mathcal{P}$ of distributions satisfying $\mathbb{P}(X=1), \mathbb{P}(X=0) > \delta > 0$, we conclude:

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P\left(|\hat{\alpha} - \theta(P)| > \epsilon\right) \to 0.$$

This establishes the uniform consistency of $\hat{\alpha}$ under randomization. □

Method 2: Adjusting for Confounding

In observational studies, treatment X is not randomized, and may be related to $(Y(0), Y(1))$. To identify θ, we must condition on observed covariates Z that remove this dependence.

Assumption (Conditional Ignorability):

$$X \perp (Y(0), Y(1)) \mid Z.$$

This means that within each level of Z, treatment assignment behaves like randomization and no unmeasured confounders except Z.

Theorem

Theorem (Identification under Conditional Ignorability)
Under the assumption above, we have:

$$\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \int \mu(1,z)p(z)dz - \int \mu(0,z)p(z)dz,$$

where $\mu(x,z) = \mathbb{E}[Y \mid X = x, Z = z]$.
An estimator for θ is:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1, Z_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(0, Z_i),$$

where $\hat{\mu}(x,z)$ is a consistent estimator of the regression function.

Proof. From the conditional ignorability assumption and the definition of potential outcomes, we have:

$$\mathbb{E}[Y(1)] = \int \mathbb{E}[Y(1) \mid Z = z]p(z)dz = \int \mathbb{E}[Y \mid X = 1, Z = z]p(z)dz,$$

and similarly for $Y(0)$. Subtracting gives the result. □

Remark

- The procedure is known as adjusting for confounders, and the resulting estimator is called the adjusted treatment effect.
- Unlike prediction, the bias-variance tradeoff does not apply symmetrically: bias must be minimized more aggressively, often requiring semi-parametric methods.
- In linear regression $\mu(x,z) = \beta_0 + \beta_1 x + \beta_2^T z$, then $\theta = \beta_1$ if the model is correctly specified and all confounders are included.

Causal Graph

One powerful framework for representing causal relationships is the use of directed acyclic graphs (DAGs). A DAG is a graph over a set of variables $\{Y_1, \ldots, Y_k\}$ with no directed cycles. It encodes the factorization of the joint distribution as:

$$p(Y_1, \ldots, Y_k) = \prod_{j=1}^{k} p(Y_j \mid \text{parents}(Y_j)),$$

where $\text{parents}(Y_j)$ denotes the set of variables with arrows pointing into Y_j.

A DAG becomes a **causal graph** if the arrows not only encode statistical dependence, but also represent causal mechanisms — that is, if the graph encodes how the system responds to interventions (e.g., setting a variable $X := x$).

Example

Example: Causal Effect via Intervention in a DAG
Consider the DAG:

$$Z \to X \to Y,$$

where:

- Z: a confounder (e.g., health status),
- X: treatment (e.g., taking vitamins),
- Y: outcome (e.g., health status).

The causal effect of X on Y is **not** identified from $p(Y \mid X)$, because Z affects both X and Y, introducing confounding.
To compute the causal effect $p(y \mid \text{do}(X = x))$, we apply the truncated factorization rule: remove all arrows into X, fix $X = x$, and compute the interventional distribution as:

$$p(y \mid \text{do}(X = x)) = \int p(y \mid x, z) p(z) dz.$$

This formula matches the result from the counterfactual framework under the assumption of conditional ignorability: $X \perp (Y(0), Y(1)) \mid Z$.

Remark

The distinction between $p(Y \mid X = x)$ and $p(Y \mid \text{do}(X = x))$ can be viewed as the difference between statistical association and causal effect — and the DAG helps us keep track of that distinction formally.

Structural Equation Modeling (SEM)

A Structural Equation Model (SEM) represents each variable as a function of its causal inputs and noise:

$$Z = g_1(U), \quad X = g_2(Z, V), \quad Y = g_3(Z, X, W),$$

where U, V, W are independent noise terms, and g_1, g_2, g_3 describe the data-generating process.

Intervening on a variable (e.g., $\text{do}(X = x)$) means replacing the equation for X with $X := x$, and propagating its effect through the system to compute outcomes like Y.

Example

Example: Intervention in a SEM
Given:

$$Z = g_1(\epsilon_1), \quad X = g_2(Z, \epsilon_2), \quad Y = g_3(Z, X, \epsilon_3),$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are independent noise variables.
Under $\text{do}(X = x)$, we fix $X := x$, so the system becomes:

$$Z = g_1(\epsilon_1), \quad Y = g_3(Z, x, \epsilon_3).$$

This modified system describes the causal effect of setting X to x.

References

[Bre01] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[CB02] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition, 2002.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

[EH16] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2016.

[Hil95] Theodore P. Hill. The significant-digit phenomenon. *The American Mathematical Monthly*, 102(4):322–327, 1995.

[HRD18] N. A. Heard and P. Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, 2018.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

[Leh90] Erich L. Lehmann. What is a statistical model? *Annals of Statistics*, 18(4):1221–1237, 1990.

[LSE00] Lawrence M. Leemis, Bruce W. Schmeiser, and Diane L. Evans. Survival distributions satisfying benford's law. *The American Statistician*, 54(4):236–241, 2000.

[Pol23] Russell A. Poldrack. *Statistical Thinking: Analyzing Data in an Uncertain World*. Princeton University Press, 2023.

[SL03] George A. F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2003.

[TdCP17] Sitsofe Tsagbey, Miguel de Carvalho, and Garritt L. Page. All data are wrong, but some are useful? advocating the need for data auditing. *The American Statistician*, 71(3):231–235, 2017.

[Was20] Larry Wasserman. Causal inference, 2020. Lecture notes, Carnegie Mellon University.