DATA8015: Mathematical Foundation of Data Science **Fall 2025**

Part 2.3 Probability and Statistics: Selected Topics

Scriber: Yingyu Liang

1 Hypothesis Testing

1.1 Elements of Hypothesis Testing

First, review some basic concepts.

Definition 1 (Hypothesis)**.** A hypothesis is a statement about a population parameter.

The definition is rather general, but note that the hypothesis should be about the *population*, and it should be a statement (i.e., it should be true or false). Given samples from the distribution, the goal of a hypothesis test is to decide which of two complementary hypotheses is true.

Definition 2 (Null/Alternative Hypothesis)**.** The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

The typical null/alternative hypotheses for a population parameter θ are $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be the value of a sample $\boldsymbol{X} = (X_1, \cdots, X_n)$ from the population. A hypothesis test takes the sample and decides to accept or reject the null hypothesis.

Definition 3 (Hypothesis Test)**.** A hypothesis test is a rule that specifies:

i. For which sample values H_0 is accepted as true.

ii. For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

1.2 Likelihood Ratio Tests

For a population parameter θ of a distribution with pdf or pmf $p(x|\theta)$, consider the null hypothesis $H_0 : \theta \in \Theta_0$ and the alternative hypothesis $H_1 : \theta \in \Theta_0^c$. Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be the value of a sample $\boldsymbol{X} = (X_1, \cdots, X_n)$ from the population. Recall that the likelihood is $L(\theta|\boldsymbol{x}) = \prod_{i=1}^n p(x_i|\theta)$.

Definition 4. The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\theta \in \Theta} L(\theta|\boldsymbol{x})}.$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

For illustration, consider the case where $p(x|\theta)$ is the pmf of a discrete random variable. If the ratio is small, then there exists some parameter value outside Θ_0 for which the observed sample is much more likely than for any parameter value in Θ_0. Therefore, H_0 should be rejected.

Example: Normal LRT. Let $x_1, \cdots, x_n$ be iid from $\mathcal{N}(\theta, 1)$. Consider $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, where θ_0 is a fixed number.

The numerator of the LRT statistic $\lambda(\boldsymbol{x})$ is simply $L(\theta_0|\boldsymbol{x})$. The MLE was shown to be the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, so the denominator of $\lambda(\boldsymbol{x})$ is $L(\bar{x}|\boldsymbol{x})$.

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \frac{(2\pi)^{-n/2}\exp\left[-\sum_{i=1}^n (x_i-\theta_0)^2/2\right]}{(2\pi)^{-n/2}\exp\left[-\sum_{i=1}^n (x_i-\bar{x})^2/2\right]} \\ &= \exp\left[\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2 - \frac{1}{2}\sum_{i=1}^n (x_i-\theta_0)^2\right] \\ &= \exp\left[-\frac{n}{2}(\bar{x}-\theta_0)^2\right]\end{aligned}$$

where the last step follows from $\sum_i (x_i-\theta_0)^2 = \sum_i (x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2$. The rejection region of an LRT can then be written as:

$$\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\} = \{\boldsymbol{x} : |\bar{x}-\theta_0| \geq \sqrt{-2(\log c)/n}\}.$$

1.3 Bayesian Tests

Suppose, besides the distribution $p(x|\theta)$, we are also given a prior distribution $p(\theta)$ for the parameter. Then Bayes' rule gives the posterior

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}.$$

Then the posterior probabilities $P(\theta \in \Theta_0|\boldsymbol{x}) = P(H_0 \text{ is true}|\boldsymbol{x})$ and $P(\theta \in \Theta_0^c|\boldsymbol{x}) = P(H_1 \text{ is true}|\boldsymbol{x})$ may be computed.

With the posterior probabilities, one may choose to accept H_0 if $P(\theta \in \Theta_0|\boldsymbol{x}) \geq P(\theta \in \Theta_0^c|\boldsymbol{x})$ and to reject H_0 otherwise. In other words, the test statistic is $P(\theta \in \Theta_0^c|\boldsymbol{x})$ and the rejection region is $\{\boldsymbol{x} : P(\theta \in \Theta_0^c|\boldsymbol{x}) > 1/2\}$. More generally, one can set the rejection region to $\{\boldsymbol{x} : P(\theta \in \Theta_0^c|\boldsymbol{x}) > c\}$ for some other value $c \in (0, 1)$.

1.4 Error Probabilities and the Power Function

A hypothesis test can be one of two types of errors. If H_0 is true but the test rejects it, this is called a Type I Error. If H_0 is false but the test accepts it, this is called a Type II error. See the table below.

Let R denote the rejection region. For $\theta \in \Theta_0$, the probability of a Type I Error is $P_\theta(\boldsymbol{X} \in R)$. For $\theta \in \Theta_0^c$, the probability of a Type II Error is $P_\theta(\boldsymbol{X} \in R^c) = 1 - P_\theta(\boldsymbol{X} \in R)$.

Table 1: Types of Errors in Hypothesis Testing

Truth	**Reject** H_0	**Accept** H_0
H_0 is true	Type I Error	Correct Decision
H_0 is false	Correct Decision	Type II Error

So

$$P_\theta(\boldsymbol{X} \in R) = \begin{cases} \text{probability of a Type I Error} & \text{if } \theta \in \Theta_0, \\ 1 - \text{probability of a Type II Error} & \text{if } \theta \in \Theta_0^c. \end{cases}$$

This inspires the following concept.

Definition 5 (Power Function)**.** The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\boldsymbol{X} \in R)$.

Ideally, the power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$. So a good test should have a power function near 0 for most $\theta \in \Theta_0$ and near 1 for most $\theta \in \Theta_0^c$.

Usually, we consider tests that control the Type I Error probability at a specified level and then search for tests with as small Type II Error probability as possible. The following is useful for the discussion.

Definition 6. For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$. A test is a level α test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

1.5 p-Values

To report the result of a hypothesis test, one method is to report the decision to reject H_0 or not, together with the size α of the test. The size of the test is important for interpreting the decision: if α is small, then the decision to reject H_0 is fairly convincing, but otherwise the rejection decision is not very convincing. Another method is to report the value of a certain kind of test statistic called a p-value.

Definition 7 (p-Value)**.** A p-value $p(\boldsymbol{X})$ is a test statistic satisfying $0 \leq p(\boldsymbol{x}) \leq 1$ for every sample point $\boldsymbol{x}$. Small values of $p(\boldsymbol{X})$ give evidence that H_1 is true. A p-value is valide if, for every $\theta in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\boldsymbol{X}) \leq \alpha) \leq \alpha.$$

Given a valid p-value $p(\boldsymbol{X})$, it is easy to construct a level α test based on it: the test that rejects H_0 if and only if $p(\boldsymbol{X}) \leq \alpha$ is a level α test, by the valid property of the p-value. An advantage of reporting p-value is that each reader can choose the desired α for deciding to reject H_0 or not. Furthermore, the smaller the p-value, the stronger the evidence for rejecting H_0.

The following theorem is the most common way to construct p-values (proof omitted).

Theorem 8. Let $W(\boldsymbol{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\boldsymbol{x}$, define:

$$p(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\boldsymbol{X}) \geq W(\boldsymbol{x})).$$

Then $p(\boldsymbol{X})$ is a valid p-value.

Example: Two-sided Normal p-Value. Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be a sample from $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$. Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. It can be shown that the LRT rejects H_0 for large values of $W(\boldsymbol{X}) = |\bar{X} - \mu_0|/(S/\sqrt{n})$, where $\bar{X}$ is the sample mean and S is the sample variance. For $\theta \in \Theta_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}$ (i.e., $\mu = \mu_0$), regardless of the value of σ, $W(\boldsymbol{X})$ has a Student's t distribution with $n-1$ degrees of freedom. Thus, $P_\theta(W(\boldsymbol{X}) \geq W(\boldsymbol{x}))$ in Theorem 8 is the same for all $\theta \in \Theta_0$. Thus, the p-value from the theorem for this two-sided t test is $p(\boldsymbol{x}) = 2P(T_{n-1} \geq |\bar{x} - \mu_0|/(s/\sqrt{n}))$, where T_{n-1} has a Student's t distribution with $n-1$ degrees of freedom.

2 Interval Estimation

Point estimation for a parameter θ of a distribution infers a single value as the value of θ, given the realization $\boldsymbol{x} = (x_1, \cdots, x_n)$ for a sample $\boldsymbol{X} = (X_1, \cdots, X_n)$ from the distribution. Interval estimation and, more generally, set estimation, infer a set $C \subseteq \Theta$ and $C = C(\boldsymbol{x})$, which is determined by the observed data $\boldsymbol{x}$. (C is usually an interval for real-valued θ.)

Definition 9 (Interval Estimation)**.** An interval estimate of a real-valued parameter θ is any pair of functions $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ of a sample that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x}$. If $\boldsymbol{X} = \boldsymbol{x}$ is observed, the inference $L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$ is made. The random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called an interval estimator.

The interval estimator can provide some guarantee of capturing the parameter, as quantified below.

Definition 10 (Coverage Probability)**.** The coverage probability of an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{x})]$ of a parameter θ is the probability that the random interval covers the true parameter θ. In symbols, it is deonted by either $P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{x})])$ or $P(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{x})]|\theta)$.

Definition 11 (Confidence Coefficient)**.** The confidence coefficient of an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{x})]$ of a parameter θ is the infimum of the coverage probability $\inf_\theta P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{x})])$.

Interval estimators with the confidence coefficient are sometimes known as confidence intervals. Sometimes we will work with more general sets rather than intervals, i.e., confidence sets. A confidence set with confidence coefficient $1-\alpha$ is simply called a $1-\alpha$ confidence set.

There is a strong correspondence between hypothesis testing and interval estimation, so confidence intervals can be constructed by inverting the test statistic. In general, every confidence set corresponds to a test and vice versa. This is formalized in the following theorem (proof omitted).

Theorem 12. For each $\theta \in \Theta_0$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\boldsymbol{x}$ value, define a set $C(\boldsymbol{x})$ in the parameter space by

$$C(\boldsymbol{x}) = \{\theta_0 : \boldsymbol{x} \in A(\theta_0)\}.$$

Then the random set $C(\boldsymbol{X})$ is a $1 - \alpha$ confidence set.

Conversely, let $C(\boldsymbol{X})$ be a $1 - \alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{\boldsymbol{x} : \theta_0 \in C(\boldsymbol{x})\}.$$

Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Example: Inverting a Normal Test. Let $\boldsymbol{X}$ be a sample from $\mathcal{N}(\mu, \sigma^2)$ and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. For a fixed α interval, a reasonable test has rejection region $\{\boldsymbol{x} : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$. Note that H_0 is accepted for the sample points with

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Since the test has size α, $P(H_0 \text{ is accepted} | \mu = \mu_0) = 1 - \alpha$. Combining this with the characterization of the acceptance region, we have

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha.$$

This is true for every μ_0. Hence, we have

$$P_\mu\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

This means, the interval $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}]$, obtained by inverting the acceptance region of the level α test, is a $1 - \alpha$ confidence interval.

3 Elements of Information Theory

Information theory initially treats questions in the areas of data compression/transmission. These lead to the definitions of quantities measuring the intuitive notion of information, such as entropy and mutual information, which are functions of the probability distributions that underlie the process of compression/communication.

Entropy. The concept of information is usually vague and broad. But the concept of entropy has many properties that agree with the intuition about information.

Definition 13 (Entropy)**.** Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \mathbb{P}\{X = x\}, x \in \mathcal{X}$. The entropy $H(X)$ of the random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathbb{E} \log p(X).$$

It can also be written as $H(p)$. The log is to the base 2, and the entropy is expressed in bits. The entropy is a measure of the average uncertainty in the random variable. It is the number of bits on average required to describe the random variable.

For example, consider a random variable that has a uniform distribution over $2^5 = 32$ outcomes. To identify an outcome, we need a label that takes on 32 different values, for which 5-bit strings suffice. The entropy of this random variable is exactly $H(x) = 5$ bits.

We now extend the definition to a pair of random variables, viewing the two as a single vector-valued random variable.

Definition 14 (Joint Entropy)**.** The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) = -\mathbb{E} \log p(X, Y).$$

We also define the conditional entropy of a random variable given another.

Definition 15 (Conditional Entropy)**.** If $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$ is defined as

$$H(X, Y) = -\sum_{x \in \mathcal{X}} H(Y|X = x) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) = -\mathbb{E} \log p(Y|X).$$

The conditional entropy measures the uncertainty of a random variable conditional on the knowledge of another random variable. Indeed, the joint entropy is the entropy of one plus the conditional entropy of the other.

Theorem 16 (Chain Rule of Entropy)**.**

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

Proof. The chain rule of probability $p(X, Y) = p(X)p(Y|X)$ leads to

$$\log p(X, Y) = \log p(X) + \log p(Y|X).$$

Taking expectations on both sides leads to the theorem. □

Relative Entropy. The relative entropy $D(p||q)$ is a measure of the distance between two distributions p and q. It is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p: we could construct a code with average description length $H(p)$ if we knew the true distribution p, but would need $H(p) + D(p||q)$ bits if we used the code for a distribution q. In machine learning, it is often called the information gain if p would be used instead of q, which is currently used.

Definition 17 (Relative Entropy, or Kullback–Leibler (KL) divergence)**.** The relative entropy or Kullback–Leibler (KL) divergence between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p(X)}{q(X)}.$$

(Here we use the convention that $0 \log \frac{0}{0} = 0, 0 \log \frac{0}{q} = 0$, and $p \log \frac{p}{0} = \infty$.) The relative entropy is always nonnegative and is zero if and only if $p = q$. However, it is not symmetric and does not satisfy the triangle inequality, so it is not a true distance between distributions.

Mutual Information. The mutual information is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Definition 18 (Mutual Information)**.** Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)||p(x)p(y)) = \mathbb{E}_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)}. \end{aligned}$$

The mutual information $I(X;Y)$ is a measure of the dependence between the two random variables. It is symmetric in X and Y, and always nonnegative and is equal to zero if and only if X and Y are independent.

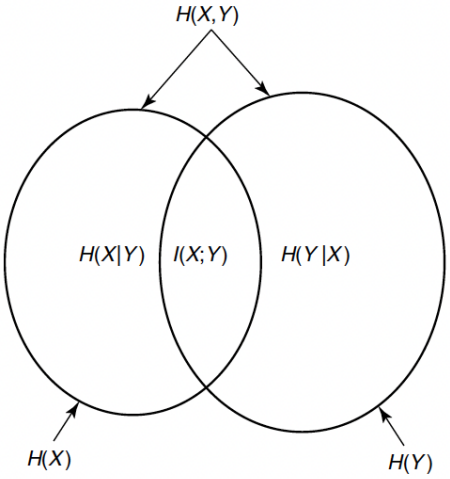

Figure 1: Relationship between entropy and mutual information. Credit: Figure 2.2 in *Elements of Information Theory*, second edition, by Thomas M. Cover and Joy A. Thomas.

Relationship between Entropy and Mutual Information. By their definitions, we can derive the following (proof omitted):

Theorem 19 (Mutual Information and Entropy)**.**

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) = H(Y) + H(X|Y) \\ I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y;X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) = I(Y;X) \\ I(X;X) &= H(X). \end{aligned}$$

The relationship is expressed in a Venn diagram in Figure 1.

Cross-Entropy. The cross-entropy is frequently used as a loss function in machine learning. It is essentially the uncentered version of relative entropy (KL-divergence).

Definition 20 (Cross-Entropy)**.** The cross-entropy between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$\mathrm{CE}(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{q(x)} = \mathbb{E}_p \log\frac{1}{q(X)} = -\mathbb{E}_p \log q(X).$$

By definition, the cross-entropy and the relative entropy have the following relationship:

$$D(p||q) = \mathrm{CE}(p||q) - H(p).$$

The relative entropy geometrically is a divergence (an asymmetric, generalized form of squared distance), measuring how far the distribution q is from the distribution p. The cross-entropy is also such a measure, but it is even further away from a distance, since $\mathrm{CE}(p||p) = H(p)$ is not zero. This can be fixed by subtracting $H(p)$ to obtain $D(p||q)$, which agrees more closely with the notion of distance. However, when learning a distribution q to approximate the true distribution p, minimizing $D(p||q)$ is equivalent to minimizing $\mathrm{CE}(p||q)$ since $H(p)$ is a constant, and moreover, $\mathrm{CE}(p||q)$ is more friendly for computation. Therefore, $\mathrm{CE}(p||q)$ is often used as the loss function.

For example, in classification, for a given input x the true label distribution is $p(y|x)$, while the current machine learning model predicts $q(y|x)$ (often by softmax on logits), the cross-entropy loss is then $\mathrm{CE}(p||q) = -\sum_{y\in\mathcal{Y}} p(y|x)\log q(y|x)$. For a training data point (x_i, y_i), the distribution $p(y|x_i)$ is a singleton, i.e., $p(Y = y_i|x_i) = 1$ and $p(Y = y|x_i) = 0, \forall y \neq y_i$. Then the cross-entropy loss on this data point is $\mathrm{CE}(p||q) = -\log q(y_i|x_i)$, which coincides with the negative log-likelihood loss. So this loss can be motivated either by MLE or by the KL-divergence.

4 Concentration Inequalities

The study of sums or averages of independent random variables is a key aspect of probability theory. For any independent random variables $X_1, X_2, \cdots, X_N$,

$$\mathrm{Var}(X_1 + \cdots + X_N) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_N).$$

If they have the same distribution with mean μ and variance σ^2, then

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{\sigma^2}{N}.$$

This means the variance of the sample mean shrinks to 0 as $N \to 0$, and we should expect that the sample mean concentrates around the expectation μ. This is precisely what the Law of Large Numbers is about.

Theorem 21 (Strong Law of Large Numbers)**.** Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with mean μ. Consider the sum $S_N = X_1 + \cdots + X_N$. Then, as $N \to \infty$,

$$\frac{S_N}{N} \to \mu \text{ almost surely.}$$

The Central Limit Theorem further identifies the limiting distribution of the properly scaled sum of X_i's as the normal distribution $\mathcal{N}(0, 1)$.

Theorem 22 (Lindeberg-Lévy Central Limit Theorem)**.** Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Consider the sum $S_N = X_1 + \cdots + X_N$ and normalize it to obtain a random variable with zero mean and unit variance as follows:

$$Z_N := \frac{S_N - \mathbb{E}S_N}{\sqrt{\mathrm{Var}(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu).$$

Then, as $N \to \infty$,

$$Z_N \to \mathcal{N}(0, 1) \text{ in distribution.}$$

The convergence in distribution means the CDF of the normalized sum converges pointwise to the CDF of the standard normal distribution.

Concentration inequalities aim to further quantify how a random variable X (usually the sum of independent random variables) deviates around its mean μ. It usually takes the form of two-sided bounds for the tails of $X - \mu$:

$$\mathbb{P}\{|X - \mu| > t\} \leq \text{ something small.}$$

The most standard concentration inequalities are Markov's and Chebyshev's Inequalities.

Theorem 23 (Markov's Inequality)**.** For any non-negative random variable X and $t > 0$, we have

$$\mathbb{P}\{X > t\} \leq \frac{\mathbb{E}X}{t}.$$

Proof. Fix $t > 0$.

$$\begin{aligned}
\mathbb{E}X &= \mathbb{E}\left[X\mathbf{1}\{X \geq t\} + X\mathbf{1}\{X < t\}\right] \\
&= \mathbb{E}X\mathbf{1}\{X \geq t\} + \mathbb{E}X\mathbf{1}\{X < t\} \\
&\geq \mathbb{E}t\mathbf{1}\{X \geq t\} + 0 \\
&= t \cdot \mathbb{P}\{X \geq t\}.
\end{aligned}$$

Dividing both sides by t completes the proof. □

Applying Markov's Inequality to $|X - \mu|^2$ for a random variable X with mean μ leads to Chebyshev's Inequality.

Corollary 24 (Chebyshev's Inequality)**.** Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have

$$\mathbb{P}\{|X - \mu| > t\} \leq \frac{\sigma^2}{t^2}.$$

Note that Markov's tail bound has a (inverse) linear dependence while Chebyshev's has a quadratic dependence on t. For the sum of independent random variables, we expect faster convergence to 0, since by Central Limit Theorem, the sum tends to a normal distribution whose tail decays to 0 exponentially fast. However, although the distribution of the sum can be approximated by the normal distribution, it turns out that the approximation error decays to 0 too slowly: $O(1/\sqrt{N})$ as in the sharp quantitative version, the Berry-Esseen Central Limit Theorem. Alternative, direct approaches to concentration are thus developed to bypass the Central Limit Theorem.

4.1 Hoeffding's Indequality

We first consider the concentration of the sum of i.i.d. symmetric Bernoulli random variables. A random variable X has a symmetric Bernoulli distribution (also called Rademacher distribution) if it takes values -1 and 1 with probabilities 1/2 each, i.e., $\mathbb{P}\{X = -1\} = \mathbb{P}\{X = 1\} = 1/2$.

Theorem 25 (Hoeffding's Inequality)**.** Let $X_1, \cdots, X_N$ be independent symmetric Bernoulli random variables, and $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$, we have

$$\mathbb{P}\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right),$$

and

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

Proof. Assume without loss of generality that $\|a\|_2 = 1$. Apply Markov's Inequality to a scaled and exponentiated version of the sum and get

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} &= \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right) \geq \exp(\lambda t)\right\} \\ &\leq \exp(-\lambda t) \cdot \mathbb{E}\exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right).\end{aligned}$$

The problem is reduced to bounding $\mathbb{E}\exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right)$, the moment generating function (MFG) of the sum. From independence, we have

$$\mathbb{E}\exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right) = \prod_{i=1}^{N} \mathbb{E}\exp\left(\lambda a_i X_i\right).$$

For each i, since X_i takes values -1 and 1 with probabilities 1/2,

$$\mathbb{E}\exp\left(\lambda a_i X_i\right) = \frac{\exp(\lambda a_i) + \exp(-\lambda a_i)}{2} = \cosh(\lambda a_i) \leq \exp(\lambda^2 a_i^2/2)$$

where the last step follows from the bound on the hyperbolic cosine function: $\cosh(x) \leq \exp(x^2/2), \forall x \in \mathbb{R}$, which can be verified by comparing the Taylor's expansions of both sides. Then we have

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} &\leq \exp(-\lambda t)\prod_{i=1}^N \exp(\lambda^2 a_i^2/2) \\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^N a_i^2\right) \\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2}\right)\end{aligned}$$

where the last step uses the assumption $\|a\|_2 = 1$.

This bound holds for any $\lambda > 0$. Minimizing the bound with $\lambda = t$ completes the proof of the first statement. The two-sidedd bound in the second statement follows from that for $S = \sum_{i=1}^N a_i X_i$,

$$\mathbb{P}\{|S| \geq t\} = \mathbb{P}\{S \geq t\} + \mathbb{P}\{-S \geq t\},$$

and applying the one-sided bound on both S and $-S$. □

The proof technique of bounding the MFG is generally applicable in many cases. It also proves the following extension of Hoeffding's Inequality for general bounded random variables.

Theorem 26 (Hoeffding's Inequality for General Bounded Random Variables)**.** Let $X_1, \cdots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have

$$\mathbb{P}\left\{\sum_{i=1}^N (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^N (M_i - m_i)^2}\right).$$

The following corollary for Bernoulli random variables is useful for generalization analysis in statistical learning.

Corollary 27 (Hoeffiding's Inequality for Bernoulli Random Variables)**.** Let $X_1, \cdots, X_N$ be i.i.d. Bernoulli random variables with $\mathbb{P}\{X_i = 1\} = p$ and $\mathbb{P}\{X_i = 0\} = 1 - p$. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^N X_i < p - \epsilon\right\} \leq \exp\left(-2N\epsilon^2\right),$$

and

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^N X_i \geq p + \epsilon\right\} \leq \exp\left(-2N\epsilon^2\right).$$

Example: Empirical Risk Concentrates around Population Risk. Recall that in the statistical learning framework, we have access to the sample from the data distribution, and would like to find a model with small risk over the whole distribution. Here we use concentration to analyze the generalization performance, i.e., how large is the population risk given a certain number of sample points.

Consider binary classification. The learning algorithm is given a set of labeled examples $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ where $(\boldsymbol{x}_i, y_i)$'s are drawn i.i.d. from some fixed but unknown distribution $\mathcal{D}$ over the instance space $\mathcal{X}$ and the label space $\mathcal{Y} = \{0, 1\}$. The goal is to find a hypothesis $h : \mathcal{X} \to \{0, 1\}$ from a hypothesis class $\mathcal{H}$ with small error over the whole distribution $\mathcal{D}$. The (population) risk is

$$R(h) = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}\{h(\boldsymbol{x}) \neq y\} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\mathbf{1}\{h(\boldsymbol{x}) \neq y\}$$

and the empirical risk is

$$R_{\mathcal{S}}(h) = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{S}}\{h(\boldsymbol{x}) \neq y\} = \frac{1}{N}\sum_{i=1}^{n}\mathbf{1}\{h(\boldsymbol{x}_i) \neq y_i\}.$$

Theorem 28 (Generalization Bound for Binary Classification, Finite Hypothesis Class)**.** Let $\mathcal{H}$ be a finite hypothesis class. Let $\mathcal{D}$ be an arbitrary, fixed unknown data distribution over $\mathcal{X} \times \{0, 1\}$. For any $\delta > 0$, if we draw a sample $\mathcal{S}$ from $\mathcal{D}$ of size N, then with probability at least $1 - \delta$, all hypotheses $h \in \mathcal{H}$ have

$$R(h) \leq R_{\mathcal{S}}(h) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{2N}}.$$

Proof. First fix a hypothesis h. Note that $\mathbf{1}\{h(\boldsymbol{x}_i) \neq y_i\}$'s are i.i.d. Bernoulli random variables with expectation $R(h)$. By Hoeffding's Inequality, we get that

$$\mathbb{P}\{R_{\mathcal{S}}(h) < R(h) - \epsilon\} \leq \exp\left(-2N\epsilon^2\right).$$

Choose $\epsilon = \sqrt{\frac{\ln(|\mathcal{H}|)+\ln(1/\delta)}{2N}}$ such that $\exp\left(-2N\epsilon^2\right) = \delta/|\mathcal{H}|$. By union bound over all $h \in \mathcal{H}$, we have with probability at least $1 - \delta$, all hypotheses $h \in \mathcal{H}$ satisfy $R_{\mathcal{S}}(h) \geq R(h) - \epsilon$. Rearranging and plugging in the chosen ϵ leads to the desired bound. □

4.2 Chernoff's Inequality

Hoeffiding's Inequality is not sharp when X_i's are Bernoulli random variables with small parameters p_i. Chernoff's Inequality gives sharper bounds in this case.

Theorem 29 (Chernoff's Inequality)**.** Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^N X_i$ and denote its mean by $\mu = \mathbb{E}S_N$. Then, for any $t > \mu$, we have

$$\mathbb{P}\{S_N \geq t\} \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t,$$

and for any $t < \mu$, we have

$$\mathbb{P}\{S_N \leq t\} \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

Proof. Here provides the proof for $t > \mu$; that for $t < \mu$ is similar.

Using the same method based on MGF as for Hoedffding's gives:

$$\mathbb{P}\{S_N \geq t\} \leq e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E} \exp(\lambda X_i).$$

Since X_i takes value 1 with proability p_i and value 0 with probability $1 - p_i$, we have

$$\mathbb{E} \exp(\lambda X_i) = e^{\lambda} p_i + (1 - p_i) = 1 + (e^{\lambda} - 1)p_i \leq \exp\left[(e^{\lambda} - 1)p_i\right]$$

where the last step follows from $1 + x \leq e^x, \forall x \in \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{P}\{S_N \geq t\} &\leq e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E} \exp(\lambda X_i) \\
&\leq e^{-\lambda t} \prod_{i=1}^{N} \exp\left[(e^{\lambda} - 1)p_i\right] \\
&= e^{-\lambda t} \exp\left[(e^{\lambda} - 1) \sum_{i=1}^{N} p_i\right] \\
&= e^{-\lambda t} \exp\left[(e^{\lambda} - 1)\mu\right].
\end{aligned}$$

This holds for any $\lambda > 0$. Setting $\lambda = \ln(t/\mu)$ (which is positive for $t > \mu$) and simplifying the expression, we complete the proof. □

Applying the theorem with $t = (1 \pm \delta)\mu$ and analyzing the bounds for small δ leads to the following useful corollary.

Corollary 30 (Chernoff's Inequality: Small Deviation)**.** Let X_i be independentnt Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^{N} X_i$ and denote its mean by $\mu = \mathbb{E}S_N$. Then, for any $\delta \in (0, 1]$, we have

$$\mathbb{P}\{|S_N - \mu| \geq \delta\mu\} \leq 2e^{-c\mu\delta^2}$$

where $c > 0$ is an absolute constant.

4.3 Sub-Gaussian Distributions

The above concentration inequalities apply to Bernoulli random variables. We would like to extend these results to a more general class of distributions. If we would like $\sum_{i=1}^{N} a_i X_i$ to have an exponential tail as in Hoeffding's, then by specializing a_i to let the sum consist only of one X_i, we know that each X_i should have a sub-gaussian tail. It turns out that the reverse is also true: for sub-gaussian distributions, concentration results like Hoeffding's can indeed be proved, as shown below.

We first begin with several equivalent definitions of sub-gaussianness (proof left as exercise).

Proposition 31 (Sub-Gaussian Properties)**.** Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.

(a) The tails of X satisify

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-t^2/K_1^2), \quad \forall t \geq 0.$$

(b) the moments of X satisfy

$$\|X\|_{L^p} := (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}, \quad \forall p \geq 1.$$

(c) The MGF of X^2 satisfies

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2), \quad \forall \lambda \in [-1/K_3, 1/K_3].$$

(d) The MGF of X^2 is bounded at some point, namely

$$\mathbb{E}\exp(X^2/K_4^2) \leq 2.$$

Moreover, if $\mathbb{E}X = 0$ then properties (a)-(d) are also equivalent to the following one.

(e) The MGF of X satisfies

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2), \quad \forall \lambda \in \mathbb{R}.$$

Definition 32 (Sub-Gaussian Random Variables)**.** A random variable X that satisfies one of the equivalent properties (a)-(d) in Proposition 31 is called a sub-gaussian random variable. The sub-gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_4 in property (d). In other words, we define

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}.$$

Then Proposition 31 can be restated in terms of the sub-gaussian norm. For example, property (a) states that every sub-gaussian random variable X has the tail bound

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-ct^2/\|X\|_{\psi_2}^2), \quad \forall t \geq 0. \tag{1}$$

Classical examples of sub-gaussians include Gaussian distribution $\mathcal{N}(0, \sigma^2)$ (with $\|X\|_{\psi_2} \leq C\sigma$ for some absolute constant C), Bernoulli (with $\|X\|_{\psi_2} = 1/\sqrt{\ln 2}$), and any bounded random variable X (with $\|X\|_{\psi_2} \leq \|X\|_{L^\infty}/\sqrt{\ln 2}$ where $\|X\|_{L^\infty}$ is the essential supremum of $|X|$). Furtheremore, the sum of indepedent sub-gaussians is also a sub-gaussian with the squared norm bounded by the sum of squared norms of the independent components. (This is analogous to the rotation invariance property of the Gaussian distribution.)

Proposition 33 (Sums of Independent Sub-Gaussians)**.** Let $X_1, \cdots, X_N$ be independent, mean zero, sub-gaussian random variables. Then $\sum_{i=1}^N X_i$ is also a sub-gaussian random variable, and

$$\left\| \sum_{i=1}^N X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2$$

where C is an absolute constant.

Proof. Simply check the MGF of the sum. □

The approximate rotation invariance above can be restated as a concentration inequality via the tail bound in (1).

Theorem 34 (General Hoeffding's Inequality)**.** Let $X_1, \cdots, X_N$ be independent, mean zero, sub-gaussian random variables. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left\{ \left| \sum_{i=1}^N X_i \right| \geq t \right\} \leq 2 \exp\left(- \frac{ct^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2} \right).$$

Let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$ and apply the theorem for $a_i X_i$ instead of X_i. Then we have the following corollary: for every $t \geq 0$,

$$\mathbb{P}\left\{ \left| \sum_{i=1}^N a_i X_i \right| \geq t \right\} \leq 2 \exp\left(- \frac{ct^2}{K^2 \|a\|_2^2} \right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

The above results assume that the random variables X_i's have zero means. Otherwise, we can always center X_i by subtracting the mean. The following lemma guarantees that centering does not harm the sub-gaussian property.

Lemma 35 (Centering Sub-Gaussians)**.** If X is a sub-gaussian random variable then $X - \mathbb{E}X$ is sub-gaussian, too, and

$$\|X - \mathbb{E}X\|_{\psi_2} \leq C \|X\|_{\psi_2}$$

where C is an absolute constant.

5 Further Reading

The sections on hypothesis testing and interval estimation are partially based on [1], which provide a thorough and in-depth treatment on statistical inference. The section on information theory is from [2], and only covers the introductory material due to time limitations. It is highly suggested to read more chapters in [2], especially Chapter 11 on Information Theory and Statistics which connects information theory to statistical inference, and Chapter 14 Kolmogorov Complexity which can relate learning to compression. The section on concentration is based on [3]. Many other topics in [3] are also useful for data science. For example, Chapters 3 and 4 cover random vectors/matrices which are basic objects in data science, and Chapter 8 covers chaining which is the basic tool for, e.g., uniform convergence in learning.

References

[1] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[3] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.