Statistics, Bayes Rule, Limit Theorems

CS 5960/6960: Nonparametric Methods
Tom Fletcher

January 26, 2009
Independent, Identically Distributed RVs

**Definition**

The random variables $X_1, X_2, \ldots, X_n$ are said to be **independent, identically distributed (iid)** if they share the same probability distribution and are independent of each other.

Recall that independence means

$$F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$
Random Samples

Definition

A **random sample** from the distribution $F$ of length $n$ is a set $(X_1, \ldots, X_n)$ of iid random variables with distribution $F$. The length $n$ is called the **sample size**.

- A random sample represents $n$ experiments in which the same quantity is measured.
- A **realization** of a random sample, denoted $(x_1, \ldots, x_n)$ are the values we get when we take the measurements.
A **statistic** on a random sample \((X_1, \ldots, X_n)\) is a function \(T(X_1, \ldots, X_n)\).

**Examples:**

- **Sample Mean**
  \[
  \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
  \]

- **Sample Variance**
  \[
  S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
  \]
Estimation

Given data $x_1, \ldots, x_n$, we want to define a probability model that describes the data as a realization of a random sample, $X_i$.

This involves defining a probability distribution for the $X_i$ with parameters $\theta = (\theta_1, \ldots, \theta_m)$ and then estimating these parameters from the data.

Example: If we decide to model our population with a Gaussian distribution, we would need to estimate the parameters $(\mu, \sigma)$, the mean and standard deviation.
The **likelihood function**, \( L(\theta, \{x_i\}) \), is the probability of observing the data \( x_i \) given that they came from the density \( f(X|\theta) \), where \( \theta \) is a parameter. That is,

\[
L(\theta; \{x_i\}) = \prod_{i=1}^{n} f(x_i|\theta).
\]
Maximum Likelihood Estimation (MLE)

Definition

The **maximum likelihood estimate** of a parameter $\theta$ given data $x_i$ is the parameter that maximizes the likelihood function:

$$\hat{\theta} = \arg \max_{\theta} L(\theta; \{x_i\}).$$
Log-Likelihood

Typically, it is easier to maximize the *log-likelihood* function, \( l(\theta; \{x_i\}) = \log L(\theta; \{x_i\}) \), because it converts products into sums:

\[
\hat{\theta} = \arg \max_\theta l(\theta; \{x_i\}) = \arg \max_\theta \log \left( \prod_{i=1}^{n} f(x_i | \theta) \right) = \arg \max_\theta \sum_{i=1}^{n} \log f(x_i | \theta)
\]
Example: MLE of Gaussian Mean
Exercise: MLE of Exponential Distribution

The exponential distribution with rate parameter $\lambda$ is defined by the pdf

$$f(x) = \lambda e^{-\lambda x}.$$

What is the MLE for $\lambda$ given data $x_1, \ldots, x_n$?
Error of an Estimator

**Definition**

The **mean-squared error (MSE)** of an estimator \( \hat{\theta} \) for a parameter \( \theta \) is

\[
\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].
\]

Note that the estimator \( \hat{\theta} \) is considered as a function of the random sample \( \{X_i\} \), and these random variables are what the expectation is taken over.
Bias and Variance of an Estimator

**Definition**

The **bias** of an estimator $\hat{\theta}$ is given by

$$\text{Bias}(\hat{\theta}) = E[\theta - \hat{\theta}].$$

**Definition**

The **variance** of an estimator $\hat{\theta}$ is given by

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2].$$
Decomposition of MSE

The MSE of an estimator $\hat{\theta}$ decomposes into the bias and variance as follows:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2.$$
Example: Bias of Sample Mean
Exercise: Variance of Sample Mean

Let $X_i$ be a continuous iid random variables with mean $\mu$ and variance $\sigma^2 < \infty$. What is the variance of the sample mean $\bar{X}$?
Bayes’ Rule

Bayes’ rule is the following formula for the probability of an event $A$ conditioned on the event $B$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This formula is easily derived from the definition of conditional probability:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$
Bayesian Statistics

The posterior distribution combines the likelihood with a prior on the parameters, which represents our existing beliefs or knowledge about the parameters.

Likelihood: \( f(x_1, \ldots, x_n | \theta) \)

Prior: \( f_\theta(\theta) \)

Posterior: \( f_{\theta|x}(\theta) \)

According to Bayes’ Rule:

\[
f_{\theta|x}(\theta) = \frac{f(x_1, \ldots, x_n | \theta) f_\theta(\theta)}{\int f(x_1, \ldots, x_n | \theta) f_\theta(\theta) \, d\theta}.
\]
Frequentist vs. Bayesian

A **Frequentist** thinks of data coming from a fixed “true” distribution of the population. The parameters of this distribution are unchanging.

A **Bayesian** thinks of the parameters of the population distribution as *random variables*. The prior specifies a belief about the nature of the parameters.
Limit Theorems

Limit theorems state some property of a statistic $T(X_1, \ldots, X_n)$ in the limit as the sample size goes to infinity, i.e., as $n \to \infty$.

Typically, a limit theorem will state that a statistic converges to the corresponding quantity for the population. For example, the sample mean converging to the population mean.
Weak vs. Strong Convergence

Given iid random variables $X_1, X_2, \ldots$, a statistic $T_n(X_1, \ldots, X_n)$, we can define two types of convergence for the statistic: weak or strong.
We say that the statistic $T_n(X_1, \ldots, X_n)$ converges weakly or converges in probability to $\tau \in \mathbb{R}$ if for every $\epsilon > 0$

$$\lim_{n \to \infty} P(|T_n - \tau| < \epsilon) = 1$$
Strong Convergence

We say that the statistic $T_n(X_1, \ldots, X_n)$ converges strongly or converges almost surely to $\tau \in \mathbb{R}$ if for every $\epsilon > 0$

$$P \left( \lim_{n \to \infty} |T_n - \tau| < \epsilon \right) = 1$$
Theorem

Let $X_1, X_2, \ldots$ be iid random variables, with $E|X_i| < \infty$ and population mean $\mu = E[X_i]$. Then the sample mean $\bar{X}_n$ converges weakly to $\mu$ as $n \to \infty$. 
Chebyshev’s Inequality

**Theorem**

Let $X$ be a random variable with mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Then for any real number $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$ 

This tells us that “most” values of a random variable are “close” to the mean.

**Example:** $(k = 2)$

No more than $1/4$ of the values are farther than 2 standard deviations from the mean.
Proof of Chebyshev’s Inequality
Proof of Weak Law of Large Numbers
Strong Law of Large Numbers

**Theorem**

Let $X_1, X_2, \ldots$ be iid random variables, with $E|X_i| < \infty$ and population mean $\mu = E[X_i]$. Then the sample mean $\bar{X}_n$ converges strongly to $\mu$ as $n \to \infty$.

Proof is much more difficult than the weak case.
Central Limit Theorem

Let $X_1, X_2, \ldots$ be iid random variables from a distribution with mean $\mu$ and variance $\sigma^2 < \infty$. Then in the limit as $n \to \infty$, the statistic

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

has a standard normal distribution.
Importance of the Central Limit Theorem

- Applies to real-world data when the measured quantity comes from the average of many small effects.
- Examples include electronic noise, interaction of molecules, exam grades, etc.
- This is why a Gaussian distribution model is often used for real-world data.