Math 408 - Mathematical Statistics

Lecture 11. Probability Theory: an Overveiw

February 11, 2013

The starting point in developing the probability theory is the notion of a sample space = the set of possible outcomes.

Definition

- The **sample space** Ω is the set of possible outcomes of an "experiment"
- Points $\omega \in \Omega$ are called **realizations**
- **Events** are subsets of Ω

Next, to every event $A \subset \Omega$, we assign a real number $\mathbb{P}(A)$, called the probability of A. We call function $\mathbb{P}$: {subsets of Ω} $\to \mathbb{R}$ a probability distribution.

Function $\mathbb{P}$ is not arbitrary, it satisfies several natural properties (called axioms of probability):

1. $0 \leq \mathbb{P}(A) \leq 1$ (Events range from never happening to always happening)
2. $\mathbb{P}(\Omega) = 1$ (Something must happen)
3. $\mathbb{P}(\emptyset) = 0$ (Nothing never happens)
4. $\mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1$ (A must either happen or not-happen)
5. $\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$

Statistical Independence

Definition

Two events A and B are **independent** if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$$

Independence can arise in two distinct ways:

1. We explicitly assume that two events are independent.
2. We derive independence of A and B by verifying that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.

Conditional Probability

Definition

If $\mathbb{P}(A) > 0$, then the conditional probability of B given A is

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

Useful Interpretation:

Think of $\mathbb{P}(B|A)$ as the

fraction of times B occurs among those in which A occurs

Properties of Conditional Probabilities:

1. For any fixed A such that $\mathbb{P}(A) > 0$, $\mathbb{P}(\cdot|A)$ is a probability, i.e. it satisfies the rules of probability.
2. In general $\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$
3. If A and B are independent then $\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$
   Thus, another interpretation of independence is that knowing A does not change the probability of B.

Law of Total Probability and Bayes' Theorem

Law of Total Probability

Let $A_1, \ldots, A_n$ be a partition of Ω, i.e.

- $\bigcup_{i=1}^{n} A_i = \Omega$ *($A_1, \ldots, A_n$ are jointly exhaustive events)*
- $A_i \bigcap A_j = \emptyset$ *for $i \neq j$ ($A_1, \ldots, A_n$ are mutually exclusive events)*
- $\mathbb{P}(A_i) > 0$

Then for any event B

$$\boxed{\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

Bayes' Theorem

Conditional probabilities can be inverted. That is,

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

Random Variables

We need the random variables to link sample spaces and events to data.

Definition

A random variable is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

This mapping induces probability on $\mathbb{R}$ from Ω as follows:
Given a random variable X and a set $A \subset \mathbb{R}$, define

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and let

$$\boxed{\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})}$$

Definition

The cumulative distribution function (CDF) $F_X : \mathbb{R} \to [0, 1]$ is defined by

$$\boxed{F_X(x) = \mathbb{P}(X \leq x)}$$

CDF contains all the information about the random variable

Properties of CDFs

Theorem

A function $F : \mathbb{R} \to [0, 1]$ is a CDF for some random variable if and only if it satisfies the following three conditions:

1. *F is non-decreasing:*

$$x_1 < x_2 \quad \Rightarrow \quad F(x_1) \leq F(x_2)$$

2. *F is normalized:*

$$\lim_{x \to -\infty} F(x) = 0 \quad \textit{and} \quad \lim_{x \to +\infty} F(x) = 1$$

3. *F is right-continuous:*

$$\lim_{y \to x+0} F(y) = F(x)$$

Discrete Random Variables

Definition

X is **discrete** if it takes countable many values $\{x_1, x_2, \ldots\}$.
We define the **probability mass function** (PMF) for X by

$$\boxed{f_X(x) = \mathbb{P}(X = x)}$$

Relationships between CDF and PMF:

- The CDF of X is related to the PMF f_X by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

- The PMF f_X is related to the CDF F_X by

$$f_X(x) = F_X(x) - F_X(x^-) = F_X(x) - \lim_{y \to x-0} F(y)$$

Continuous Random Variables

Definition

A random variable is **continuous** if there exists a function f_X such that

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1$, and
- For every $a \leq b$

$$\boxed{P(a < X \leq b) = \int_a^b f_X(x)dx}$$

- The function $f_X(x)$ is called the probability density function (PDF)
- Relationship between the CDF $F_X(x)$ and PDF $f_X(x)$:

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t)dt} \qquad \boxed{f_X(x) = F_X'(x)}$$

Transformation of Random Variables

Suppose that X is a random variable with PDF f_X and CDF F_X.
Let $Y = r(X)$ be a function of X.

Q: How to compute the PDFand CDF of Y?

1. For each y, find the set $A_y = \{x : r(x) \leq y\}$
2. Find the CDF $F_Y(y)$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \in A_y) = \int_{A_y} f_X(x)dx$$

3. The PDF is then $f_Y(y) = F_Y'(y)$

Important Fact: When r is strictly monotonic, then r has an inverse $s = r^{-1}$ and

$$\boxed{f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|}$$

Joint Distributions

- Discrete Case

Definition

Given a pair of discrete random variables X and Y, their **joint PMF** is defined by

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$$

- Continuous Case

Definition

A function $f_{X,Y}(x,y)$ is called the **joint PDF** of continuous random variables X and Y if

- $f_{X,Y}(x,y) \geq 0, \qquad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$
- For any set $A \subset \mathbb{R} \times \mathbb{R}$

$$\mathbb{P}((X,Y) \in A) = \int\int_A f_{X,Y}(x,y)dxdy$$

The **joint CDF** of X and Y is defined as $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$

Marginal Distributions

- Discrete Case
  If X and Y have joint PMF $f_{X,Y}$, then the **marginal PMF** of X is

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

  Similarly, the **marginal PMF** of Y is

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f_{X,Y}(x, y)$$

- Continuous Case
  If X and Y have joint PDF $f_{X,Y}$, then the **marginal PDFs** of X and Y are

$$f_X(x) = \int f_{X,Y}(x, y)dy \quad \text{and} \quad f_Y(y) = \int f_{X,Y}(x, y)dx$$

Independent Random Variables

Definition

Two random variables X and Y are **independent** if, for every A and B

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

Criterion of independence:

Theorem

Let X and Y have joint PDF/PMF $f_{X,Y}$. Then X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

Conditional Distributions

- Discrete Case
  The **conditional PMF**:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- Continuous Case
  The **conditional PDF** is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

  Then,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$$

Expectation and its Properties

The expectation (or mean) of a random variable X is the average value of X.

Definition

The **expected value**, or **mean**, or **first moment** of X is

$$\mu_X \equiv \mathbb{E}[X] = \begin{cases} \sum_x x f_X(x), & \text{if } X \text{ is discrete} \\ \int x f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

assuming that the sum (or integral) is well-defined.

- Let $Y = r(X)$, then $\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) f_X(x) dx$
- If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

- Let $X_1, \ldots, X_n$ be independent random variables. Then,

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

Variance and its Properties

The variance measures the "spread" of a distribution.

Definition

Let X be a random variance with mean μ_X.
The **variance** of X, denoted $\mathbb{V}[X]$ or σ_X^2, is defined by

$$\sigma_X^2 \equiv \mathbb{V}[X] = \mathbb{E}[(X-\mu_X)^2] = \begin{cases} \sum_x (x-\mu_X)^2 f_X(x), & \text{if } X \text{ is discrete} \\ \int (x-\mu_X)^2 f_X(x)dx, & \text{if } X \text{ is continuous} \end{cases}$$

The **standard deviation** is $\sigma_X = \sqrt{\mathbb{V}[X]}$

Important Properties of $\mathbb{V}[X]$:

- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mu_X^2$
- If a and b are constants, then $\mathbb{V}[aX+b] = a^2\mathbb{V}[X]$
- If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{V}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 \mathbb{V}[X_i]$$

Covariance and Correlation

If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the **covariance** between X and Y by

$$\boxed{\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}$$

and the **correlation** by

$$\boxed{\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}}$$

Properties of Covariance and Correlation

- The covariance satisfies (useful in computations):

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- The correlation satisfies:

$$-1 \leq \rho(X, Y) \leq 1$$

- If $Y = aX + b$ for some constants a and b, then

$$\rho(X, Y) = \begin{cases} 1, & \text{if } a > 0 \\ -1, & \text{if } a < 0 \end{cases}$$

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = \rho(X, Y) = 0$. The converse is not true.
- For random variables $X_1, \ldots, X_n$

$$\mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

Conditional Expectation and Conditional Variance

- The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_x x f_{X|Y}(x|y), & \text{discrete case;} \\ \int x f_{X|Y}(x|y)dx, & \text{continuous case.} \end{cases}$$

  - $\mathbb{E}[X]$ is a number
  - $\mathbb{E}[X|Y=y]$ is a function of y
  - $\mathbb{E}[X|Y]$ is the random variable whose value is $\mathbb{E}[X|Y=y]$ when $Y=y$
- The Rule of Iterated Expectations

$$\mathbb{E}\mathbb{E}[Y|X] = \mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}\mathbb{E}[X|Y] = \mathbb{E}[X]$$

- The conditional variance of X given $Y = y$ is

$$\mathbb{V}[X|Y=y] = \mathbb{E}[(X - \mathbb{E}[X|Y=y])^2|Y=y]$$

  - $\mathbb{V}[X]$ is a number
  - $\mathbb{V}[X|Y=y]$ is a function of y
  - $\mathbb{V}[X|Y]$ is the random variable whose value is $\mathbb{V}[X|Y=y]$ when $Y=y$
- For random variables X and Y

$$\mathbb{V}[X] = \mathbb{E}\mathbb{V}[X|Y] + \mathbb{V}\mathbb{E}[X|Y]$$

Inequalities

- Markov inequality: If X is a non-negative random variable, then for any $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then for any $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

- Hoeffding inequality: Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, then for any $\varepsilon > 0$

$$\mathbb{P}(|\overline{X}_n - p| \geq a) \leq 2e^{-2na^2}$$

- Cauchy-Schwarz inequality: If X and Y have finite variances, then

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

- Jensen Inequality:
  - If g is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$
  - If g is concave, then $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$

Convergence of Random Variables

There are two main types of convergence: convergence in probability and convergence in distribution.

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

- X_n **converges** to X **in probability**, written $X_n \xrightarrow{\mathbb{P}} X$, if for every $\epsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

- X_n **converges** to X **in distribution**, written $X_n \xrightarrow{\mathcal{D}} X$, if
$$\lim_{n\to\infty} F_n(x) = F(x)$$
for all x for which F is continuous.

$$\boxed{X_n \xrightarrow{\mathbb{P}} X \quad \text{implies that} \quad X_n \xrightarrow{\mathcal{D}} X}$$

Law of Large Numbers and Central Limit Theorem

The LLN says that the mean of a large sample is close to the mean of the distribution.

The Law of Large Numbers

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu} \quad \text{as } n \to \infty$$

The CLT says that $\overline{X}_n$ has a distribution which is approximately Normal with mean μ and variance σ^2/n. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Then

$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)} \quad \text{as } n \to \infty$$

The Central Limit Theorem

The central limit theorem tells us that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \;\dot{\sim}\; \mathcal{N}(0,1)$$

However, in applications, we rarely know σ. We can estimate σ^2 from $X_1, \ldots, X_n$ by sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Question: If we replace σ with S_n is the central limit theorem still true?

Answer: Yes!

Theorem

Assume the same conditions as in the CLT. Then,

$$\boxed{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)} \quad \text{as } n \to \infty$$

Multivariate Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. random vectors with mean μ and covariance matrix Σ:

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_{1i}] \\ \mathbb{E}[X_{2i}] \\ \vdots \\ \mathbb{E}[X_{ki}] \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \mathbb{V}[X_{1i}] & \mathrm{Cov}(X_{1i}, X_{2i}) & \ldots & \mathrm{Cov}(X_{1i}, X_{ki}) \\ \mathrm{Cov}(X_{2i}, X_{1i}) & \mathbb{V}[X_{2i}] & \ldots & \mathrm{Cov}(X_{2i}, X_{ki}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{ki}, X_{1i}) & \ldots & \mathrm{Cov}(X_{ki}, X_{k-1i}) & \mathbb{V}[X_{ki}] \end{pmatrix}$$

Let $\overline{X}_n = (\overline{X}_{1n}, \ldots, \overline{X}_{kn})^T$. Then

$$\boxed{\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)} \quad \text{as } n \to \infty$$