Introduction to Statistical Inference

Lecture 2: Probability review

Mohammad-Reza A. Dehaqani

dehaqani@ut.a

Sample Spaces, Realizations, Events

Probability Theory is the mathematical language for uncertainty quantification.

The starting point in developing the probability theory is to specify 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 want to assign a real number $\mathbb{P}(A)$, called the probability of A. We call function $\mathbb{P} : \{\text{subsets of } \Omega\} \to \mathbb{R}$ a probability distribution.

We don't want $\mathbb{P}$ to be arbitrary, we want it to satisfy some 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. $\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$

Probability on Finite Sample Spaces

Suppose that the sample space is finite $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$.

Example:

If we toss a die twice, then Ω has $n = 36$ elements:

$$\Omega\{\ (i,j)\ :\ i,j = 1,2,3,4,5,6\}$$

If each outcome is equally likely, then $\mathbb{P}(A) = |A|/36$, where $|A|$ denotes the number of elements in A.

In general, if Ω is finite and if each outcome is equally likely, then

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

To compute the probability $\mathbb{P}(A)$, we need to count the number of points in an event A. Methods for counting points are called combinatorial methods.

Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2} \times \frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. We can formalize this useful notion of independence as follows:

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. For example, in tossing a coin twice, we usually assume that the tosses are independent which reflects the fact that the coin has no memory of the first toss.

2. We derive independence of A and B by verifying that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. For example, in tossing a fair die, let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Are A and B independent?
Yes! Since $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 2/3$, $AB = \{2, 4\}$,
$\mathbb{P}(AB) = 1/3 = (1/2) \times (2/3)$

Examples

- Suppose that A and B are disjoint events, each with positive probability. Can they be independent?

Answer: No! $\mathbb{P}(AB) = \mathbb{P}(\emptyset) = 0$, but $\mathbb{P}(A)\mathbb{P}(B) > 0$

Conditional Probability

the sample space is the set of all possible outcomes of an experiment. Suppose we are interested only in part of the sample space, the part where we know some event – call it A – has happened, and we want to know how likely it is that various other events ($B, C, D \ldots$) have also happened.

What we want is the **conditional probability** of B given A.

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

Here are some facts about 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:
   - $0 \leq \mathbb{P}(B|A) \leq 1$
   - $\mathbb{P}(\Omega|A) = 1$
   - $\mathbb{P}(\emptyset|A) = 0$
   - $\mathbb{P}(B|A) + \mathbb{P}(\bar{B}|A) = 1$
   - $\mathbb{P}(B + C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A) - \mathbb{P}(BC|A)$
2. Important: The rules of probability apply to events on the left of the bar.
3. In general
$$\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$$
4. What 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

A useful tool for computing probabilities is the following law.

Law of Total Probability

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

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

Then for any event B

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

Bayes' Theorem

Conditional probabilities can be inverted. That is,

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

This relationship is called **Bayes' Rule** after Thomas Bayes (1702-1761) who did not discover it (in this form, Bayes' Rule was proved by Laplace).

Discrete Random Variables

Statistics is concerned with data.

Question: How do we link sample spaces and events to data?

Answer: The link is provided by the concept of a **random variable**.

Definition

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

Example: Flip a coin 10 times. Let $X(\omega)$ be the number of heads in the sequence. For example, if $\omega = HHTHTTTHTH$, then $X(\omega) = 5$.

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

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) \\ \mathbb{P}(X = x) &= \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}) \end{aligned}$$

The Cumulative Distribution Function

Definition

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

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

Question: Why do we bother to define CDF?

Answer: CDF effectively contains all the information about the random variable

Theorem

Let X have CDF F and Y have CDF G. If $F(x) = G(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$. In words, the CDF completely determines the distribution of a random variable.

Properties of CDFs

Question: Given a function $F(x)$, can we find a random variable X such that $F(x)$ is the CDF of X, $F_X(x) = F(x)$?

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

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

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

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

Important Examples

- **The Point Mass Distribution**
  X has a point mass distribution at a, denoted $X \sim \delta_a$, if $\mathbb{P}(X = a) = 1$.
  In this case

$$F(x) = \begin{cases} 0, & x < a; \\ 1, & x \geq a. \end{cases}$$

  and

$$f(x) = \begin{cases} 1, & x = a; \\ 0, & x \neq a. \end{cases}$$

- **The Discrete Uniform Distribution**
  Let $n > 1$ be a given integer. Suppose that X has probability mass function given by

$$f(x) = \begin{cases} 1/n, & \text{for } x = 1, \ldots, n; \\ 0, & \text{otherwise.} \end{cases}$$

  We say that X has a uniform distribution on $1, \ldots, n$.

Important Examples

- **The Bernoulli Distribution**
  Let X represents a coin flip. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has a Bernoulli distribution, denoted $X \sim \text{Bernoulli}(p)$. The probability mass function is

$$f(x|p) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

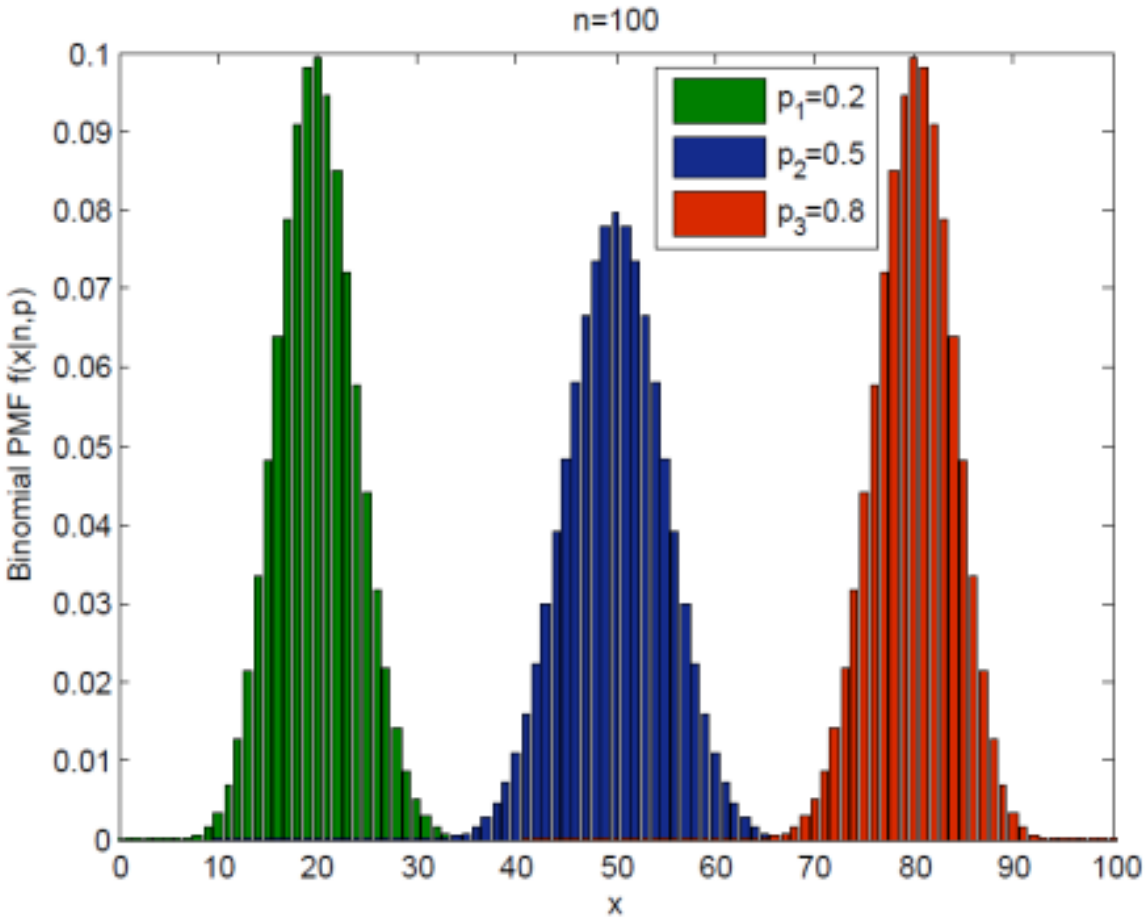
- **The Binomial Distribution**
  Suppose we have a coin which falls heads with probability p for some $p \in [0, 1]$. Flip the coin n times and let X be the number of heads. Assume that the tosses are independent. The probability mass function of X is then

$$f(x|n, p) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x}, & \text{if } x = 0, 1, \ldots, n; \\ 0, & \text{otherwise.} \end{cases}$$

  A random variable with this mass function is called a Binomial random variable and we write $X \sim \text{Bin}(n, p)$.

  Remark: X is a random variable, x denotes a particular value of the random variable, n and p are parameters, that is, fixed real numbers. The parameter p is usually unknown and must be estimated from data.

Binomial Distribution $\mathrm{Bin}(n, p)$



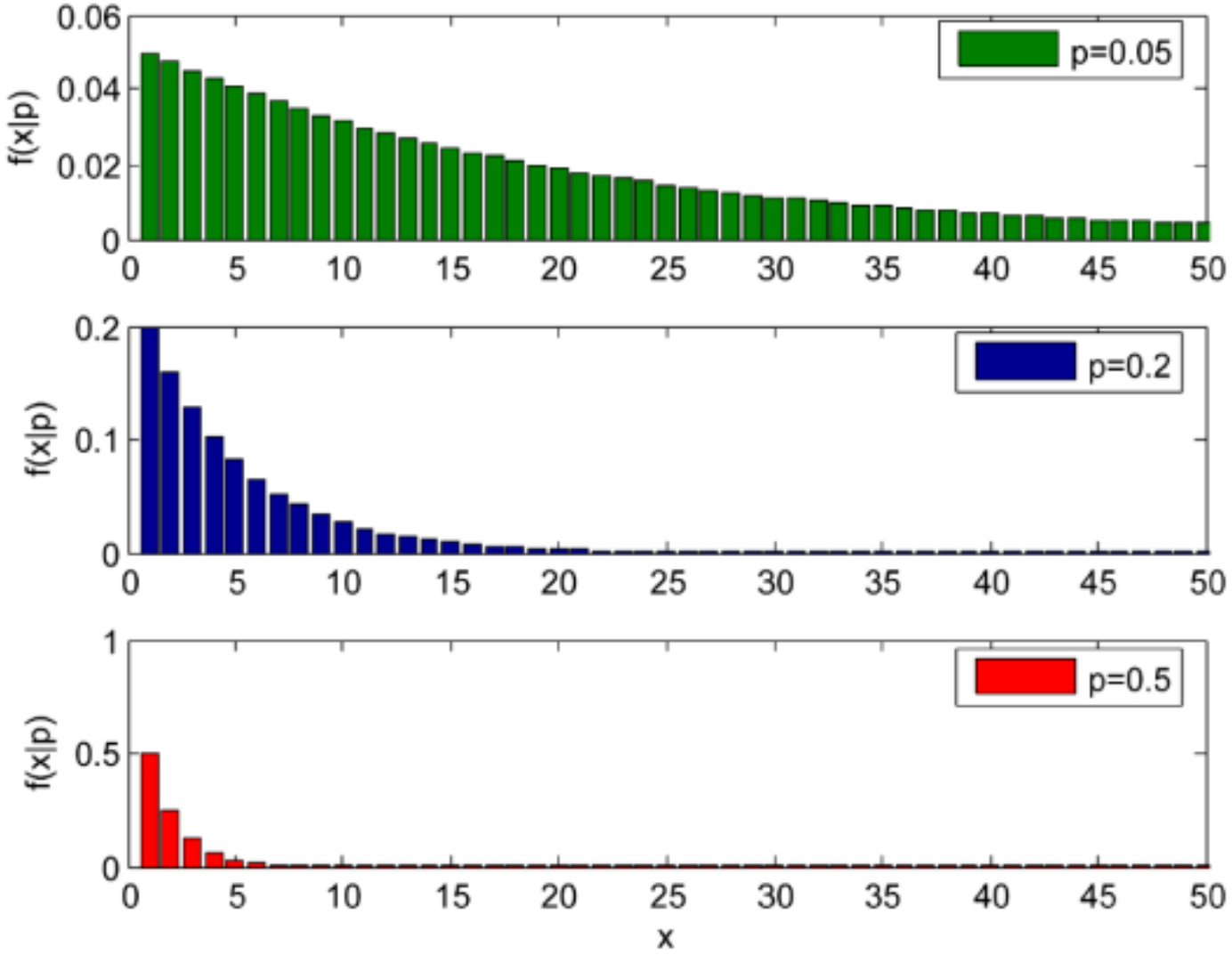
Important Examples

- **The Geometric Distribution**
  X has a geometric distribution with parameter $p \in (0, 1)$, denoted $X \sim \text{Geom}(p)$, if

$$f(x|p) = p(1-p)^{x-1}, \quad x = 1, 2, 3 \ldots$$

Think of X as the number of flips needed until the first heads when flipping a coin. Geometric distribution is used for modeling the number of trials until the first success.

Geometric Distribution $\mathrm{Geom}(p)$



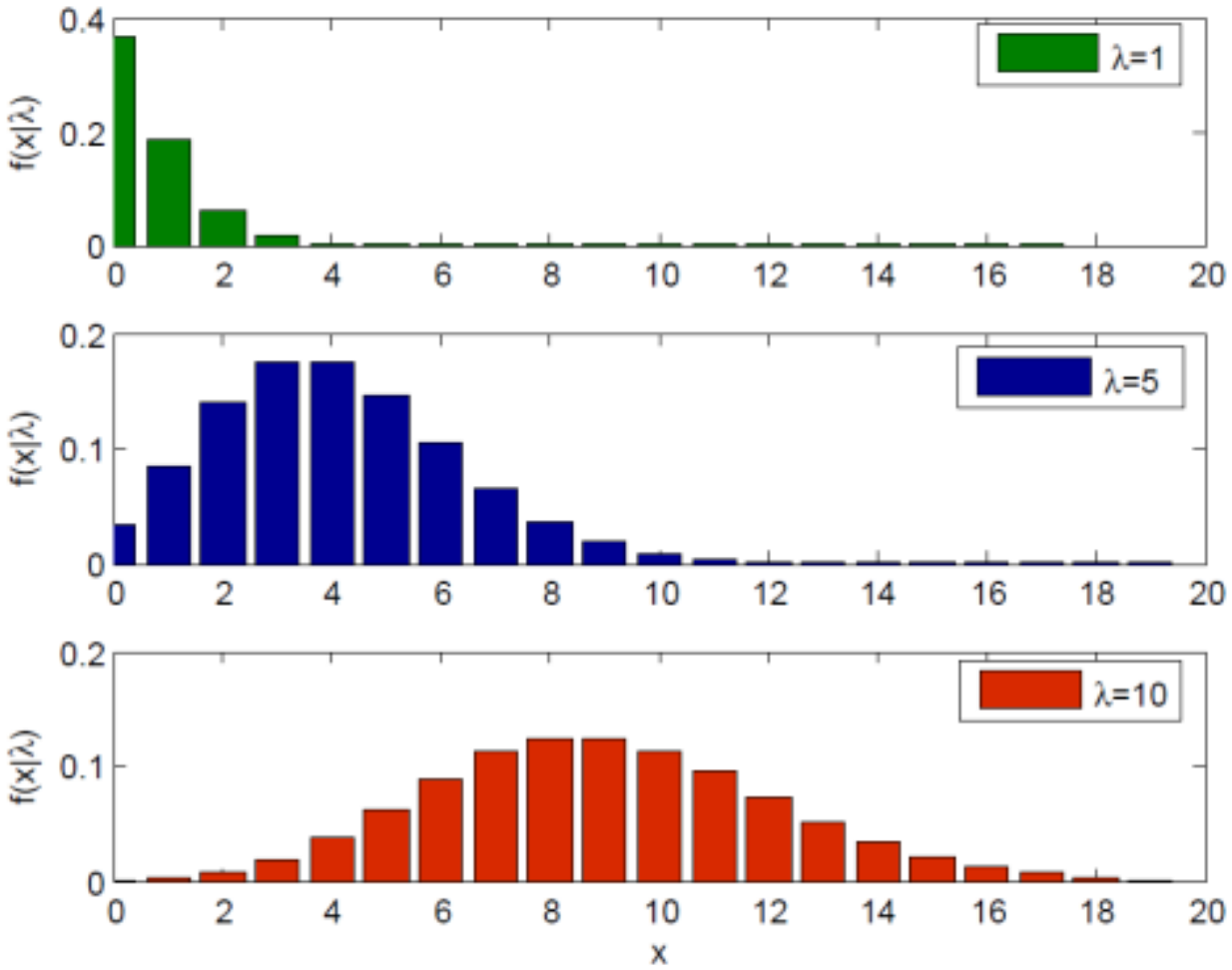
Important Examples

- **The Poisson Distribution**
  X has a Poisson distribution with parameter λ, denoted $X \sim \text{Poisson}(\lambda)$ if

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The Poisson distribution is often used as a model for counts of rare events like traffic accidents. $f(x|\lambda)$ expresses the probability of a given number of events x occurring in a fixed interval of time if these events occur with a known average rate λ and independently of the time since the last event.

Poisson Distribution Poisson(λ)



Continuous Random Variables

Definition

Recall that a random variable is a (deterministic) map $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each (random) realization $\omega \in \Omega$.

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 $a \leq b$

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

- The function $f_X(x)$ is called the probability density function (PDF)

- Relationship between the CDF $F_X(x)$ and PDF $f_X(x)$:

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

Important Remarks

- If X is continuous then $\mathbb{P}(X = x) = 0$ for every x.
- Don't think of $f_X(x)$ as $\mathbb{P}(X = x)$. This is only true for discrete random variables.
- For continuous random variables, we get probabilities by integrating.
- A PDF can be bigger than 1 (unlike PMF!). For example:

$$f_X(x) = \begin{cases} 10, & x \in [0, 0.1] \\ 0, & x \notin [0, 0.1] \end{cases}$$

Important Examples

- **The Uniform Distribution**
  X has a uniform distribution on $[a, b]$, denoted $X \sim U[a, b]$, if

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

- **Normal (Gaussian) Distribution**
  X has a Normal (or Gaussian) distribution with parameters μ and σ, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

  - Many phenomena in nature have approximately Normal distribution.
  - Distribution of a sum of random variables can be approximated by a Normal distribution (central limit theorem)

PROPOSITION A

If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$.

Suppose that $X \sim N(\mu, \sigma^2)$ and we wish to find $P(x_0 < X < x_1)$

Consider the random variable

$$Z = \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma}$$

$a = 1/\sigma$ and $b = -\mu/\sigma$, we see that $Z \sim N(0, 1)$,

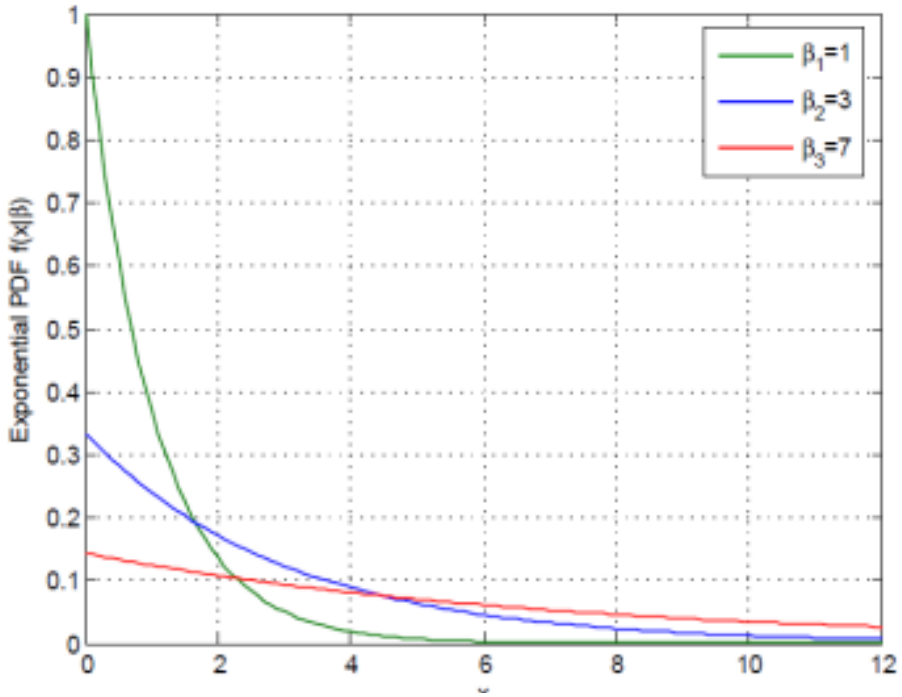
$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P\left(\frac{X - \mu}{\sigma} \le \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z \le \frac{x - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right) \end{aligned}$$

$$\begin{aligned} P(x_0 < X < x_1) &= F_X(x_1) - F_X(x_0) \\ &= \Phi\left(\frac{x_1 - \mu}{\sigma}\right) - \Phi\left(\frac{x_0 - \mu}{\sigma}\right) \end{aligned}$$

Important Examples

- **Exponential Distribution**
  X has an Exponential distribution with parameter $\beta > 0$, $X \sim \text{Exp}(\beta)$, if

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0$$



The exponential distribution is used to model the life times of electronic components and the waiting times between rare events. β is a survival parameter: the expected duration of survival of the system is β units of time.

Important Examples

- **Gamma Distribution**
  X has a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$,
  $X \sim \text{Gamma}(\alpha, \beta)$, if

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

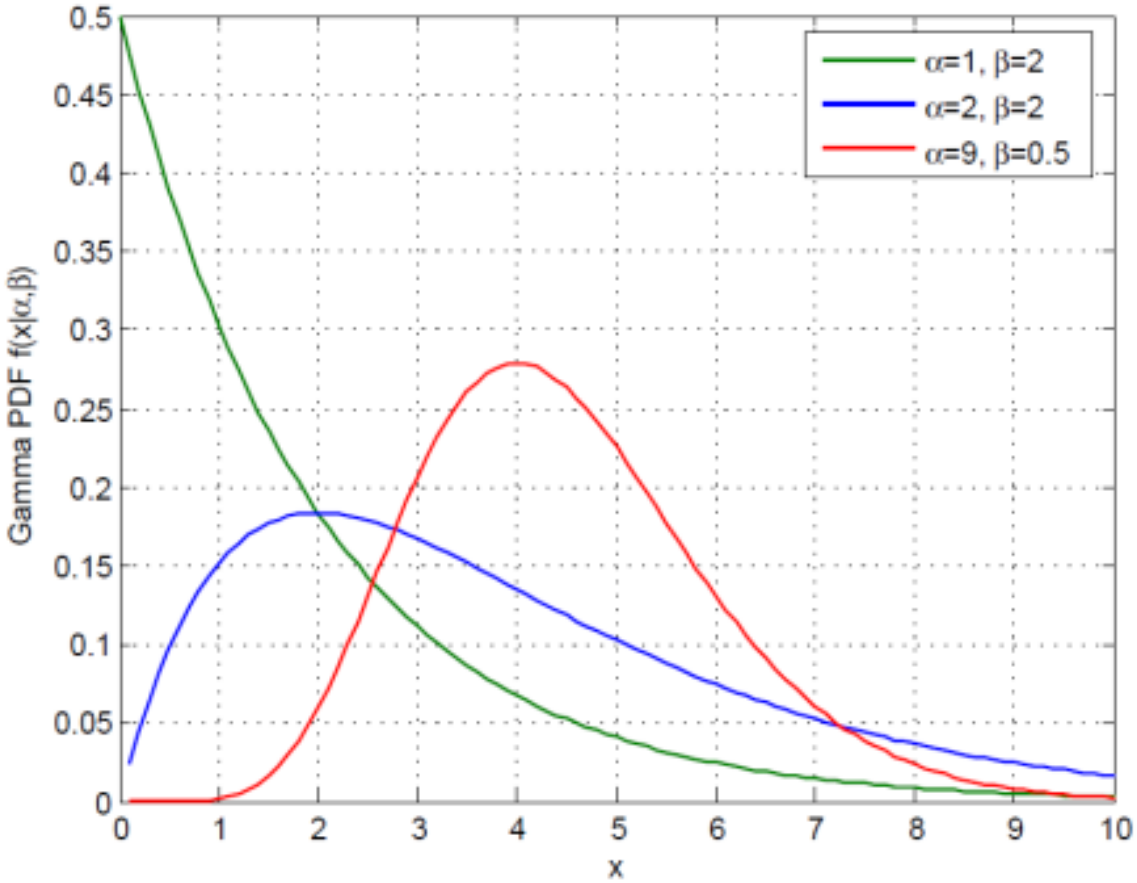
  - $\Gamma(\alpha)$ is the Gamma function

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

  - The Gamma distribution is frequently used to model waiting times.
  - Exponential distribution is a special case of the Gamma distribution:

$$\text{Gamma}(1, \beta) = \text{Exp}(\beta)$$

Gamma Distribution



Important Examples

- **Beta Distribution**
  X has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, $X \sim \text{Beta}(\alpha, \beta)$, if

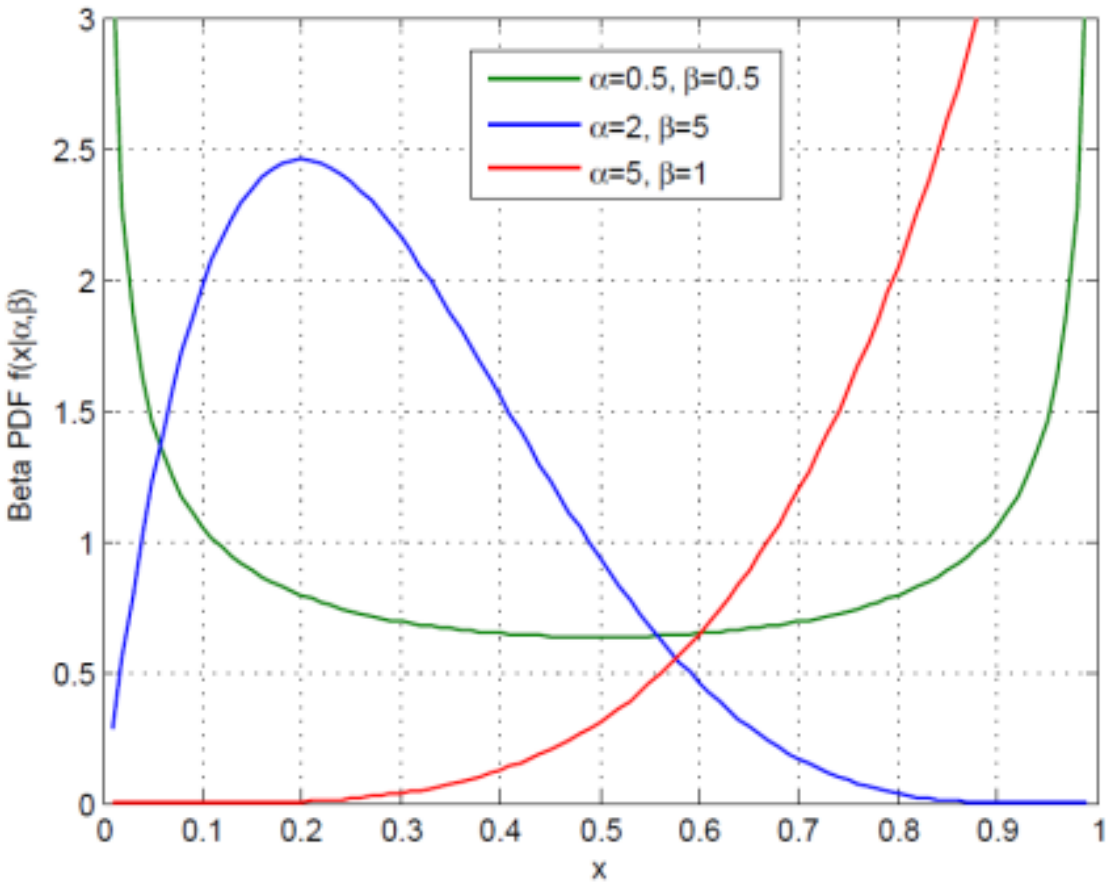
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

  - The beta distribution is often used for modeling of proportions.

  - The beta distribution has an important application in the theory of order statistics. A basic result is that the distribution of the k^{th} largest $X_{(k)}$ of a sample of size n from a uniform distribution $X_1, \ldots, X_n \sim U(0,1)$ has a beta distribution:

$$X_{(k)} \sim \text{Beta}(k, n-k+1)$$

Beta Distribution



Joint Distributions

Bivariate 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

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

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

Let X be a continuous random variable with density $f(x)$ and let $Y = g(X)$ where g is a differentiable, strictly monotonic function on some interval I. Suppose that $f(x) = 0$ if x is not in I. Then Y has the density function

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

for y such that $y = g(x)$ for some x, and $f_Y(y) = 0$ if $y \neq g(x)$ for any x in I. Here g^{-1} is the inverse function of g; that is, $g^{-1}(y) = x$ if $y = g(x)$. ■

Let $Z = F(X)$; then Z has a uniform distribution on $[0, 1]$.

Proof

$$P(Z \leq z) = P(F(X) \leq z) = P(X \leq F^{-1}(z)) = F(F^{-1}(z)) = z$$

This is the uniform cdf. ■

Let U be uniform on $[0, 1]$, and let $X = F^{-1}(U)$. Then the cdf of X is F.

Proof

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

apply F^{-1} to uniform random numbers. This is quite practical as long as F^{-1} can be calculated easily.

Generate Exponential Random Numbers

$F(t) = 1 - e^{-\lambda t}$. F^{-1} can be found by solving $x = 1 - e^{-\lambda t}$ for t:

$$e^{-\lambda t} = 1 - x$$
$$-\lambda t = \log(1 - x)$$
$$t = -\log(1 - x)/\lambda$$

Thus, if U is uniform on [0, 1], then $T = -\log(1 - U)/\lambda$ is an exponential random

$$V = 1 - U$$
$$P(V \leq v) = P(1 - U \leq v) = P(U \geq 1 - v) = 1 - (1 - v) = v$$

We may thus take $T = -\log(V)/\lambda$, where V is uniform on [0, 1].

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

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

In principle, to check whether X and Y are independent, we need to check the above equation for all subsets A and B. Fortunately, we have the following result:

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
  If X and Y are discrete, then we can compute the conditional probability of the event $\{X = x\}$ given that we have observed $\{Y = y\}$:

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

This leads to the following definition of 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
  For continuous random variables, 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 of a Random Variable

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

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.

Remarks:

- The expectation is a one-number summary of the distribution.
- Think of $\mathbb{E}[X]$ as the average value you would obtain if you computed the numerical average $\frac{1}{n}\sum_{i=1}^{n} X_i$ of a large number of i.i.d. draws $X_1, \ldots, X_n$.

  The fact that

$$\mathbb{E}[X] \approx \frac{1}{n}\sum_{i=1}^{n} X_i$$

  is a theorem called the law of large numbers.

Let $Y = r(X)$. How do we compute $\mathbb{E}[Y]$?

$$\boxed{\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) f_X(x) dx}$$

if $Z = r(X, Y)$, then

$$\boxed{\mathbb{E}[Z] = \mathbb{E}[r(X, Y)] = \int \int r(x, y) f_{X,Y}(x, y) dx dy}$$

Properties of Expectations

- If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

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

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

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

Remark: Note the the summation rule does not require independence but the multiplication rule does.

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

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

A Model for Measurement Error

A sequence of repeated independent measurements made with no deliberate change in the apparatus or experimental procedure may not yield identical values, and the **uncontrollable fluctuations are often modeled as random**

If the true value of the quantity being measured is denoted by x_0, the measurement, X, is modeled as

$$X = x_0 + \beta + \varepsilon$$

where β is the constant, or systematic, error and ε is the random component of the error; ε is a random variable with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$. We then have

$$E(X) = x_0 + \beta \qquad \text{Var}(X) = \sigma^2$$

β is often called the **bias**

The two factors affecting the size of the error are the bias and the size of the variance,

Mean Squared Error

$$\text{MSE} = E[(X - x_0)^2]$$

$\text{MSE} = \beta^2 + \sigma^2$.

Proof

From Theorem B of Section 4.2,

$$\begin{aligned} E[(X - x_0)^2] &= \text{Var}(X - x_0) + [E(X - x_0)]^2 \\ &= \text{Var}(X) + \beta^2 \\ &= \sigma^2 + \beta^2 \end{aligned}$$

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 $X_1, \ldots, X_n$

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

Expectation and Variance of Important Random Variables

Distribution	**Mean**	**Variance**
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Bin(n,p)	p	$np(1-p)$
Geom(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
Exp(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$

Conditional Expectation
Conditional Variance

Suppose that X and Y are random variables.

Q: What is the mean of X among those times when $Y = y$?

A: It is the mean of X as before, but instead of $f_X(x)$ we use $f_{X|Y}(x|y)$.

Definition

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

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

If $Z = r(X, Y)$ is a new random variable, then

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

Important Remark:

- $\mathbb{E}[X]$ is a number
- $\mathbb{E}[X|Y=y]$ is a function of y

Conditional Expectation

Question: What is $\mathbb{E}[X|Y=y]$ before we observe the value y of Y?

Answer: Before we observe Y, we don't know the value of $\mathbb{E}[X|Y=y]$, it is uncertain, so it is a random variable which we denote $\mathbb{E}[X|Y]$.

$\mathbb{E}[X|Y]$ is the random variable whose value is $\mathbb{E}[X|Y=y]$ when $Y=y$.

The Rule of Iterated Expectations

Theorem

For random variables X and Y, assuming the expectations exist, we have

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

More generally, for any function $r(x, y)$ we have

$$\mathbb{E}[\mathbb{E}[r(X, Y)|X]] = \mathbb{E}[r(X, Y)] \quad \textit{and} \quad \mathbb{E}[\mathbb{E}[r(X, Y)|Y]] = \mathbb{E}[r(X, Y)]$$

Example 1:
Suppose we draw

$$X \sim U(0,1)$$

After we observe $X = x$, we draw

$$Y|X = x \sim U(x,1)$$

Find $\mathbb{E}[Y|X = x]$.
Answer:

$$\mathbb{E}[Y|X = x] = \frac{x+1}{2}, \quad \text{as intuitively expected}$$

Note that $\mathbb{E}[Y|X] = \frac{X+1}{2}$ is a random variable whose value is the number $\mathbb{E}[Y|X = x] = \frac{x+1}{2}$ once $X = x$ is observed.

Example 2: Compute $\mathbb{E}[Y]$ in Example 1.

Answer:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}\left[\frac{X+1}{2}\right] = \frac{1/2+1}{2} = \frac{3}{4}$$

Prediction

To attempt to predict Y from X

Let us first consider a relatively trivial situation: the problem of predicting Y by means of a constant value, c.

$$\text{MSE} = E[(Y - c)^2]$$

$$\begin{aligned} E[(Y - c)^2] &= \text{Var}(Y - c) + [E(Y - c)]^2 \\ &= \text{Var}(Y) + (\mu - c)^2 \end{aligned}$$

The first term of the last expression does not depend on c, and the second term is minimized for $c = \mu$, which is the optimal choice of c.

Now let us consider predicting Y by some function $h(X)$ in order to minimize

$$\text{MSE} = E\{[Y - h(X)]^2\}$$

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

For every x, the inner expectation is minimized by setting $h(x)$ equal to the constant

$$E(Y|X = x)$$

We thus have that the minimizing function $h(X)$ is

$$h(X) = E(Y|X)$$

For the bivariate normal distribution, we found that

$$E(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$$

Conditional Variance

Recall, that "unconditional" variance of random variable Y is

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

Therefore, it is natural to define **conditional variance** of Y given that $X = x$ as follows (replace all expectations by conditional expectations):

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

Denote $\mathbb{E}[Y|X = x]$ by $\mu_Y(x)$. Then

$$\mathbb{V}[Y|X = x] = \int (y - \mu_Y(x))^2 f_{Y|X}(y|x)dy$$

- $\mathbb{V}[Y]$ is a number, $\mathbb{V}[Y|X = x]$ is a function of x

Theorem

For random variables X and Y

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

Inequalities

Markov Inequality

Inequalities are useful for bounding quantities that might otherwise be hard to compute. They will be used in the large sample theory

Markov Inequality

Let X be a non-negative random variable and suppose that $\mathbb{E}[X]$ exists. Then for any $a > 0$

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

Remark:

- This result says that the probability that X is much bigger than $\mathbb{E}[X]$ is small: Let

$$a = k\mathbb{E}[X]$$

Then

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

Chebyshev Inequality

Chebyshev Inequality

Let X be a random variable with mean μ and variance σ^2. Then for any $a > 0$

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

Remarks:

- This result says that if σ^2 is small, then there is a high probability that X will not deviate much from μ.
- If $a = k\sigma$, then

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

- If $Z = \frac{X-\mu}{\sigma}$, then

$$\mathbb{P}(|Z| \geq a) \leq \frac{1}{a^2}$$

Cauchy-Schwarz and Jensen Inequalities

These are two inequalities on expected values that are often useful.

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 (x^2, e^x, etc), then*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

- *If g is concave ($-x^2$, $\log x$, etc), then*

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

Examples: $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, $\mathbb{E}(1/X) \geq 1/\mathbb{E}[X]$, $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$.

Reference

The slides contents come from USC mathematical statistics course and John A. Rice's book