Probability and Statistics

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2014

Outline

Outline

Probability Spaces (1/3)

- An experiment (e.g., tossing a coin) is called ***random experiment*** iff its outcome is uncertain in advance

Definition (Probability Space)

A ***probability space*** is a triple $(\Omega, \mathcal{F}, P)$ where:
a) The ***sample space*** Ω is a non-empty set containing all possible outcomes of a random experiment;
b) The σ***-algebra*** $\mathcal{F} \subseteq 2^{\Omega}$ is a set of subsets (i.e., events) of Ω such that: b-1) $\Omega \in \mathcal{F}$; b-2) If $A \in \mathcal{F}$, then $A^c = \Omega \backslash A \in \mathcal{F}$; b-3) If $A_i \in \mathcal{F}$ for $i = 1, 2, \cdots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$;
c) The ***probability*** $P : \mathcal{F} \to [0,1]$ is a function satisfying: c-1) $P(\Omega) = 1$; c-2) For mutually exclusive events A_i, $i = 1, 2, \cdots$, where $A_i \cap A_j \neq \emptyset$, $i \neq j$, we have $P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

- Based on the De Morgan's law, properties b-2) and b-3) also imply that if $A_i \in \mathcal{F}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$

Probability Spaces (2/3)

- Consider a random experiment tossing two coins:
  - $\Omega = \{HH, HT, TH, TT\}$[1]
  - If we define the events "the first coin lands head" $A_1 = \{HH, HT\}$ and "the first coin lands tail" $A_2 = \{TH, TT\}$, then $\mathcal{F} = \{\Omega, \emptyset, A_1, A_2\}$
  - If we define the events "at least one head" $B_1 = \{HH, HT, TH\}$ and "two heads" $B_2 = \{HH\}$, then $\mathcal{F} = \{\Omega, \emptyset, B_1, B_1^c, B_2, B_2^c, B_1^c \cup B_2, (B_1^c \cup B_2)^c\}$
  - A nature way to define probability is by frequency, i.e., $P(A) = \lim_{n\to\infty} times_n(A)/times_n(\Omega) = \lim_{n\to\infty} times_n(A)/n$, where $times_n(\cdot)$ denotes how many times an event occurs when repeating the experiment n times
- What if the experiment is not repeatable?
  - P can also be defined to represent the degree of believe
- Note Ω may be infinite (e.g., consider an experiment throwing a dart and the outcome is "at x meters from the center of the target")

[1] HT means the first coin lands head and the second lands tail

Probability Spaces (3/3)

- By definition, we have [Proof]:
  - If $P(A) = p$, then $P(A^c) = 1 - p$
  - $P(\emptyset) = 0$
  - $0 \leqslant P(A) \leqslant 1$
  - If $A \subseteq B$, then $P(A) \leqslant P(B)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leqslant P(A) + P(B)$ (equality holds when A and B are mutually exclusive)
- We call $P(A)$ the ***marginal probability*** of A and $P(A \cap B)$ the ***joint probability*** of A and B

Theorem (Law of Total Probability)

Let $\{B_i\}_{i=1}^{\infty}$ be a partition of Ω (i.e., $\bigcup_{i=1}^{\infty} B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$), then for any A we have $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$.

Sure and Almost Sure Events

- An event A happens ***surely*** if no outcome not in this event can occur
- An event A happens ***almost surely*** if $P(A) = 1$
- What's the difference?

Sure and Almost Sure Events

- An event A happens ***surely*** if no outcome not in this event can occur
- An event A happens ***almost surely*** if $P(A) = 1$
- What's the difference?
- The event "zero, one, or two heads" $A = \Omega$ is a sure event in the coin-tossing experiment
- The event "not at 4.3 meters from the center" is an almost sure even in the dart-throwing experiment
  - Define probability of an event as the proportion of the event's corresponding area to the area of the target
  - Since the event "at 4.3 meters from the center" is a circle without area, its probability is 0
  - That is, the event "not at 4.3 meters from the center" has probability 1
- An almost sure event can still not happen

Conditional Probability and Independence

- Define the ***conditional probability*** $P(A|B) = P(A \cap B)/P(B)$ as the probability of the occurrence of A given that B occurred
  - The basic idea is to reduce the sample space to B: $P(A|B) = \lim_{n\to\infty} \frac{times_n(A\cap B)}{times_n(B)} = \lim_{n\to\infty} \frac{times_n(A\cap B)/times_n(\Omega)}{times_n(B)/times_n(\Omega)} = P(A \cap B)/P(B)$
- Events A and B are ***independent*** iff their occurrence has nothing to do with each other, i.e., $P(A|B) = P(A)$
  - Or equivalently, $P(A \cap B) = P(A)P(B)$
  - Don't mix this up with the mutual exclusiveness: $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Bayes' Rule

- Given $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$, we can easily see that:

Theorem (Bayes' Rule)

$P(A|B) = P(B|A)P(A)/P(B)$.

- Bayes' Rule is so important to ML such that each term is given a name: ***posterior*** (of A given B) $=$ ***likelihood*** $\times$ ***prior*** $/$ ***evidence***

Example (From Predicting the Cause to Historical Statistics)

Given an event B "Having a suntan." We want to infer whether the event A_1 "Mountain climbing" or A_2 "Sleeping" is the cause. In other words, we want to find an event A_i such that the posterior $P(A_i|B)$ is higher. From Bayes' rule, we can instead seeking for the event maximizing the product of likelihood and prior, which, in this case, can be obtained from historical statistics.

Bayes' Rule

- If A or B is continuous, we can instead formulate Bayes' Rule in terms of the probability density $p(A)$ or $p(B)$.
  - If A is continuous and B is discrete,

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

  - If A is discrete and B is continuous,

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

  - If both A and B are continuous,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

Outline

1. **Probability, The Basics**
   - Events and Probability
   - Random Variables
   - Expectations and Variances
   - Common Distributions
2. **Statistics**
   - Point Estimation
   - The Central Limit Theorem
   - Interval Estimation**
   - Hypothesis Testing**
3. **Multivariate Probability****
   - Multivariate Random Variables
   - Multivariate Normal Distribution

Random Variables (1/2)

Definition (Random Variable)

A ***random variable*** $X : \Omega \to \mathcal{W}$, $\mathcal{W} \subseteq \mathbb{R}$, defined on a probability space $(\Omega, \mathcal{F}, P)$ is a function that assigns a number to each outcome $\omega \in \Omega$ such that for every $x \in \mathbb{R}$, $-\infty < x < \infty$, the set $\{\omega | X(\omega) \leqslant x\}$ is an event in $\mathcal{F}$.

- In the coin-tossing experiment, we can define X that sums up the total number of heads such that $X(TT) = 0$, $X(HT) = 1$, and so on
  - Denote $P(X \leqslant 1)$ the probability of the event "less than or equal to one head"
- In the dart-throwing experiment, we define Y as the distance from the center
  - Denote $P(Y \leqslant 4.3)$ the probability of the event "within 4.3 meters"
- A random variable is ***discrete*** if $\mathcal{W}$ is countable; otherwise ***continuous***

Random Variables (2/2)

- We can perform arithmetic (e.g., $X+Y$, X^2, $2X$) or conditioning (e.g., $X|Y=y$, $X|Y \leqslant y$) on random variables to get a new one
- X and Y are said to be ***equal in distribution*** (or ***stochastically equal***), denote by $X =_{s.t.} Y$, iff $P(X \leqslant a) = P(Y \leqslant a)$ for all $a \in \mathbb{R}$
  - If $X =_{s.t.} Y$, does $X+Y =_{s.t.} 2X$ hold?

Random Variables (2/2)

- We can perform arithmetic (e.g., $X+Y$, X^2, $2X$) or conditioning (e.g., $X|Y=y$, $X|Y \leqslant y$) on random variables to get a new one
- X and Y are said to be ***equal in distribution*** (or ***stochastically equal***), denote by $X =_{s.t.} Y$, iff $P(X \leqslant a) = P(Y \leqslant a)$ for all $a \in \mathbb{R}$
  - If $X =_{s.t.} Y$, does $X+Y =_{s.t.} 2X$ hold? No, as the domains of X and Y may be different
- X and Y are said to be ***equal***, denote by $X = Y$, iff $X(\omega) = Y(\omega)$ for all $\omega \in \Omega$
- X and Y are independent iff $P(X \leqslant x | Y \leqslant y) = P(X \leqslant x)$ (or equivalently, $P(X \leqslant x, Y \leqslant y) = P(X \leqslant x)P(Y \leqslant y)$)

Distributions and Densities (1/2)

Definition (Probability Distribution Function)

Given a random variable X. A function $F_X : \mathbb{R} \to [0,1]$, defined by $F_X(x) = P(X \leqslant x)$, is called the ***probability distribution function*** of X.

Definition (Probability Mass Function)

If X is discrete, we have $F_X(x) = \sum_{s \leqslant x} P_X(s)$, where $P_X(s) = P(X = s)$ is called the ***probability mass function*** of X.

Definition (Probability Density Function)

If X is continuous and F_X is differentiable such that $F_X(x) = \int_{-\infty}^{x} p_X(s) ds$, we call p_X the ***probability density function*** of X.

- Is $p_X(s)$ a probability?

Distributions and Densities (1/2)

Definition (Probability Distribution Function)

Given a random variable X. A function $F_X : \mathbb{R} \to [0,1]$, defined by $F_X(x) = P(X \leqslant x)$, is called the ***probability distribution function*** of X.

Definition (Probability Mass Function)

If X is discrete, we have $F_X(x) = \sum_{s \leqslant x} P_X(s)$, where $P_X(s) = P(X = s)$ is called the ***probability mass function*** of X.

Definition (Probability Density Function)

If X is continuous and F_X is differentiable such that $F_X(x) = \int_{-\infty}^{x} p_X(s)ds$, we call p_X the ***probability density function*** of X.

- Is $p_X(s)$ a probability? ***No***, it is the "rate of increase" of F_X at s
  - $P(X = x)$ always equals to 0 when X is continuous

Distributions and Densities (2/2)

- From now on, we focus on the continuous random variables
- The ***joint distribution*** of X and Y is defined by $F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} p_{X,Y}(s,t)dsdt$
  - $p_{X,Y}$ is the ***joint density***
- We may obtain the ***marginal distribution*** of X by $F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} p_{X,Y}(s,t)dsdt = \int_{-\infty}^{x} p_X(s)ds$
  - $p_X(s) = \int_{-\infty}^{\infty} p_{X,Y}(s,t)dt$ is the ***marginal density*** of X (by law of total probability)
- The ***conditional distribution*** of X on Y is $F_{X|Y=y}(x|y) = \frac{\int_{-\infty}^{x} p_{X,Y}(s,y)ds}{\int_{-\infty}^{\infty} p_{X,Y}(s,y)ds} = \frac{\int_{-\infty}^{x} p_{X,Y}(s,y)ds}{p_Y(y)} = \int_{-\infty}^{x} p_{X|Y=y}(s|y)ds$
  - $p_{X|Y=y}(s|y) = p_{X,Y}(s,y)/p_Y(y)$ is the ***conditional density***
  - X and Y are independent iff $F_{X|Y=y}(x) = F_X(x)$ (or $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ or $p_{X,Y}(s,y) = p_X(s)p_Y(y)$)

Bayes' Rule for Random Variables

- Generally, $P(X \leqslant x | Y \leqslant y) = \frac{P(Y \leqslant y | X \leqslant x) P(X \leqslant x)}{P(Y \leqslant y)}$
- Can be written as as different forms in terms of mass/density functions:
  - $P_{X|Y=y}(x|y) = \frac{P_{Y|X=x}(y|x) P_X(x)}{P_Y(y)}$ for discrete X and Y
  - $p_{X|Y=y}(x|y) = \frac{p_{Y|X=x}(y|x) p_X(x)}{p_Y(y)}$ for continuous X and Y
  - $P_{X|Y=y}(x|y) = \frac{p_{Y|X=x}(y|x) P_X(x)}{p_Y(y)}$ for discrete X and continuous Y
  - $p_{X|Y=y}(x|y) = \frac{P_{Y|X=x}(y|x) p_X(x)}{P_Y(y)}$ for continuous X and discrete Y

Outline

Expectations

Definition (Expectation)

The ***expectation*** (or ***expected value*** or ***mean***) of a real-valued function f whose domain is the values of a continuous random variable X is defined by $E[f(X)] = \int_{-\infty}^{\infty} f(x) p_X(x) dx$.

- E is a functional of f
- For convenience, in $E[f(X)]$ we may expand f directly:
  - E.g., if $f(x) = x$, then $E[f(X)] = E[X] = \int_{-\infty}^{\infty} x p_X(x) dx = \mu_X$ is called the expectation of X
  - $E[X^n] = \int_{-\infty}^{\infty} x^n p_X(x) dx$ is called the ***nth moment*** of X
- $E[X|Y = y] = \int_{-\infty}^{\infty} x p_{X|Y=y}(x|y) dx$ is called the ***conditional expectation***
- We may consider expectation of functions defined over multiple variables:
  - $E[X + Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) p_{X,Y}(x, y) dxdy$
- We can subscript E to average f with respect to some particular variables
  - E.g., $E_X[X + Y] = \int_{-\infty}^{\infty} (x + y) p_{X,Y}(x, y) dx$
  - Note that $E_X[X + Y]$ is a function of y

Properties

- $E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)p_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}x\left(\int_{-\infty}^{\infty}p_{X,Y}(x,y)dy\right)dx + \int_{-\infty}^{\infty}y\left(\int_{-\infty}^{\infty}p_{X,Y}(x,y)dx\right)dy = \int_{-\infty}^{\infty}xp_X(x)dx + \int_{-\infty}^{\infty}yp_Y(y)dy = E[X]+E[Y]$
- Also, $E[aX+b] = aE[X]+b$ where a and b are a constants [Proof]
- $E[E[X]] = E[X]$ ($E[X]$ is a constant)
- $E[XY] = E[X]E[Y]$ if X and Y are independent [Proof]

Jensen's Inequality

Theorem (Jensen's Inequality)

Given a convex, differentiable function f defined on the values of a random variable X, we have $E[f(X)] \geqslant f(E[X])$.

Proof.

Define a linear function $g(x,a) = f(a) + f'(a) \cdot (x - a)$ that is tangent to f at some point a. Since f is convex, we have $g(x, E[X]) \leqslant f(x)$ for all x. This implies that $E[f(X)] = \int f(x)p(x)dx \geqslant \int g(x, E[X])p(x)dx = E[g(x, E[X])] = E[f(E[X]) + f'(E[X]) \cdot (X - E[X])] = f(E[X])$. □

Variance

Definition (Variance)

The ***variance*** of a real-valued function f whose domain is the values of a continuous random variable X is defined as
$Var[f(X)] = E[(f(X) - E[f(X)])^2]$.

- Variance measures how much a function f varies from its expected value in average
- In particular, $Var(X) = E\left[(X - E[X])^2\right] = \sigma_X^2$ is called the variance of X
- We have $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2 - 2E[X]X + E[X]\right] = E[X^2] - E[X]^2$
- $\sigma_X = \sqrt{Var(X)}$ is called the ***standard deviation*** of X

Covariance

Definition (Covariance)

The ***covariance*** between two random variable X and Y, denoted by $Cov[X, Y]$, is defined as $Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$.

- If X and Y are related in a linear way (e.g., $Y = aX + b$), covariance measures how much these two variables change together
  - Positive (resp. negative) covariance implies that Y grows (resp. shrinks) as X increases
- $Cov[X, Y] = 0$ if X and Y are independent [Proof]
  - The converse is ***not*** true as X and Y may be related in a nonlinear way (e.g., $Y = \sin(X)$)

Properties

- $Var[aX+b]=a^2Var[X]$ where a and b are constants [Proof]
- $Var[aX+bY]=a^2Var[X]+b^2Var[Y]+2abCov[X,Y]$ [Proof]
  - $Var[X+Y]=Var[X]+Var[Y]$ if X and Y are independent
- $Cov[aX+b,cY+d]=acCov[X,Y]$ [Proof]
- $Cov[aX+bY,cW+dV]=$ $acCov[X,W]+adCov[X,V]+bcCov[Y,W]+bdCov[Y,V]$ [Proof]

Correlation

Definition (Correlation)

The ***correlation*** between two random variable X and Y, denoted by $Corr[X,Y]$, is defined as $Corr[X,Y] = Cov[X,Y]/\sqrt{Var[X]Var[Y]}$.

- Correlation is the normalized covariance with respect to X's and Y's variances
  - The value always lies between $[-1,1]$
- Remember how a search engine calculates the similarity between two documents?
  - In addition to the cosine function, the correlation is another similarity measure (if we think the attributes of a document version as the values of a random variable)
  - What's the difference?

Correlation

Definition (Correlation)

The ***correlation*** between two random variable X and Y, denoted by $Corr[X,Y]$, is defined as $Corr[X,Y] = Cov[X,Y]/\sqrt{Var[X]Var[Y]}$.

- Correlation is the normalized covariance with respect to X's and Y's variances
  - The value always lies between $[-1, 1]$
- Remember how a search engine calculates the similarity between two documents?
  - In addition to the cosine function, the correlation is another similarity measure (if we think the attributes of a document version as the values of a random variable)
  - What's the difference?Correlation measures the similarity between the trends of the change across attributes; while the cosine function measures the similarity between corresponding attributes directly

Markov's Inequality

Theorem (Markov's Inequality)

Let h be a real-valued, nonnegative, and nondecreasing function defined over the values of a random variable X, we have $P(X \geqslant t) \leqslant \frac{E[h]}{h(t)}$ for any $t \in \mathbb{R}$.

Proof.

By definition, $E[h] = \int_{-\infty}^{\infty} h(z) p_X(z) dz$. Since h is nonnegative, we have $\int_{-\infty}^{\infty} h(z) p_X(z) dz \geqslant \int_{t}^{\infty} h(z) p_X(z) dz$. Furthermore, $\int_{t}^{\infty} h(z) p_X(z) dz \geqslant h(t) \int_{t}^{\infty} p_X(z) dz = h(t) P(X \geqslant t)$ as h is nondecreasing. We obtain the proof. □

- By letting $h(x) = x^+$ we have $P(X \geqslant t) \leqslant \frac{\mu_X}{t}$ for $t > 0$ [Proof]
- Provides a quick check for some statement about the tail of a distribution
  - E.g., If we know that the average response time of a web site is 1 second. How many users will experience delay longer than 10 seconds?

Markov's Inequality

Theorem (Markov's Inequality)

Let h be a real-valued, nonnegative, and nondecreasing function defined over the values of a random variable X, we have $P(X \geqslant t) \leqslant \frac{E[h]}{h(t)}$ for any $t \in \mathbb{R}$.

Proof.

By definition, $E[h] = \int_{-\infty}^{\infty} h(z) p_X(z) dz$. Since h is nonnegative, we have $\int_{-\infty}^{\infty} h(z) p_X(z) dz \geqslant \int_{t}^{\infty} h(z) p_X(z) dz$. Furthermore, $\int_{t}^{\infty} h(z) p_X(z) dz \geqslant h(t) \int_{t}^{\infty} p_X(z) dz = h(t) P(X \geqslant t)$ as h is nondecreasing. We obtain the proof. □

- By letting $h(x) = x^{+}$ we have $P(X \geqslant t) \leqslant \frac{\mu_X}{t}$ for $t > 0$ [Proof]
- Provides a quick check for some statement about the tail of a distribution
    - E.g., If we know that the average response time of a web site is 1 second. How many users will experience delay longer than 10 seconds? Markov's Inequality tells us that there will be no more than $1/10 = 10\%$ of total users that will experience this

Chebyshev's Inequality

- If we know σ_X, we can have a more specific bound:

Theorem (Chebyshev's Inequality)

$P(|X-\mu_X| \geqslant t) \leqslant \frac{\sigma_X^2}{t^2}$ *for any* $t > 0$.

Proof.

Let $Y =_{s.t.} (X-\mu_X)^2$ and $h(x) = x$. By Markov's Inequality we have $P(Y \geqslant t^2) \leqslant \frac{\mu_Y}{t^2}$. Note that

$P(Y \geqslant t^2) = P\left((X-\mu_X)^2 \geqslant t^2\right) = P(|X-\mu_X| \geqslant t)$ and

$\mu_Y = E\left[(X-\mu_X)^2\right] = \sigma_X^2$. So $P(|X-\mu_X| \geqslant t) \leqslant \frac{\sigma_X^2}{t^2}$. $\square$

- Setting $t = c\sigma_X$ for some $c > 0$, we have $P(|X-\mu_X| \geqslant c\sigma_X) \leqslant \frac{1}{c^2}$

Outline

Describing the Distribution of a Random Variable

- Given a random variable X and a function *Dist* parametrized by θ, we say$X$ has distribution $Dist(\theta)$, denoted by $X \sim Dist(\theta)$, iff
  - $P_X(x) = Dist(x|\theta)$ when X is discrete, or
  - $p_X(x) = Dist(x|\theta)$ when X is continuous
- Next, we study common *Dist* functions

Bernoulli Distribution (Discrete)

- The distribution of a random variable X depends on how the experiment is defined
- The simplest experiment is to perform a trial whose outcome can be either 0 (failure) or 1 (success)
- Let p be the probability of success, we have $P_X(1) = P(X = 1) = p$ and $P_X(0) = P(X = 0) = (1 - p)$
- $X \sim Ber(p)$, where $Ber(x|p) = p^x(1-p)^{1-x}$ for $x = 0, 1$
- $F_X(x) = \sum_{k \leqslant x} Ber(k|p)$ for $x = 0, 1$
- $E[X] = p$, $Var[X] = p(1-p)$ [Proof]

Binomial Distribution (Discrete)

- How about the experiment that performs the Bernoulli trial independently for n times and counts the times of success?
- We have $P_X(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$
- $X \sim Bin(n, p)$, where $Bin(x|n, p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$ for $0 \leqslant x \leqslant n$
- $F_X(x) = \sum_{k \leqslant x} Bin(k|n, p)$
- $E[X] = np$, $Var[X] = np(1-p)$ [Proof]
- Let $X^{(i)} \sim Ber(p)$, we can see that $X^{(1)} + \cdots + X^{(n)} \sim Bin(n, p)$

Multinomial Distribution (Discrete)

- Now, what if each trial in the Binomial distribution can have K possible outcomes (e.g., rolling a die) instead of 2?
- Let p_i be the possibility the ith possible outcome occurs in a trial, where $\sum_{i=1}^{K} p_i = 1$, we have $P_X(x_1, \cdots, x_K | \mathbf{p}) = \frac{n!}{x_1 \cdots x_K} \prod_{i=1}^{K} p_i^{x_i}$ for $\sum_{i=1}^{K} x_i = n$
- $X \sim Mul(n, K, \mathbf{p})$, where $Mul(x_1, \cdots, x_K | n, K, \mathbf{p}) = \frac{n!}{x_1 \cdots x_k} \prod_{i=1}^{K} p_i^{x_i}$ for $\sum_{i=1}^{K} x_i = n$
- Distributions are discussed separately in terms of each x_i, i.e., $F_X(x_i | x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_K) = \sum_{s \leqslant x_i} Mul(x_1, \cdots, s, \cdots, x_K | n, K, \mathbf{p})$, where $\sum_{j=1}^{i-1} x_j + s + \sum_{j=i+1}^{k} x_j = n$

Dirichlet Distribution (Continuous) (1/3)

- If $\mathbf{p} = (p_1, \cdots, p_K) \sim \text{Dirichlet}(\boldsymbol{\alpha})$, then

$$P(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

  for all $p_1, \cdots, p_K > 0$ satisfying $p_1 + \cdots p_K = 1$.
  - $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^\top$ and $\alpha_0 = \sum_i \alpha_i$.
  - $\Gamma(\alpha)$ is the Gamma function defined as $\Gamma(\alpha) \equiv \int_0^\infty u^{\alpha-1} e^{-u} du$.
    - Note that $\Gamma(\alpha) = (\alpha - 1)!$ if α is a positive integer.

Dirichlet Distribution (Continuous) (2/3)

- If we use the Dirichlet distribution as the prior for the multinomial (i.e., $\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha})$), we have

$$
\begin{aligned}
P(\mathbf{p}|x_1,\cdots,x_K) &= \frac{P(x_1,\cdots,x_K|\mathbf{p})\,P(\mathbf{p}|\boldsymbol{\alpha})}{\int P(x_1,\cdots,x_K|\mathbf{p})\,P(\mathbf{p}|\boldsymbol{\alpha})\,d\mathbf{p}} \\
&= \frac{\left(\frac{n!}{x_1\cdots x_K}\prod_{i=1}^K p_i^{x_i}\right)\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{i=1}^K p_i^{\alpha_i-1}\right)}{\int_{\mathbf{p}}\left(\frac{n!}{x_1\cdots x_K}\prod_{i=1}^K p_i^{x_i}\right)\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{i=1}^K p_i^{\alpha_i-1}\right)d\mathbf{p}} \\
&= \frac{\prod_{i=1}^K p_i^{\alpha_i+x_i-1}}{\frac{\Gamma(\alpha_1+x_1)\cdots\Gamma(\alpha_K+x_K)}{\Gamma(\alpha_0+n)}\times\int_{\mathbf{p}}\frac{\Gamma(\alpha_0+n)}{\Gamma(\alpha_1+x_1)\cdots\Gamma(\alpha_K+x_K)}\prod_{i=1}^K p_i^{\alpha_i+x_i-1}\,d\mathbf{p}} \\
&= \frac{\Gamma(\alpha_0+n)}{\Gamma(\alpha_1+x_1)\cdots\Gamma(\alpha_K+x_K)}\prod_{i=1}^K p_i^{\alpha_i+x_i-1} \\
&\sim \text{Dirichlet}(\boldsymbol{\alpha}+\mathbf{x})
\end{aligned}
$$

  where $\mathbf{x} = [x_1,\cdots,x_K]^\top$.

- We see that the posterior has the same form as the prior and we call such a prior a *conjugate prior*.

Dirichlet Distribution (Continuous) (3/3)

- As x_i are counts of occurrences of state i in a sample of x, we can view α_i as counts of occurrences of state i in some imaginary sample of α_0 instances. In defining the prior, we are subjectively saying that in a sample of α_0, we expect α_i of them to belong to state i.
  - Note that larger α_0 implies that we have a higher confidence in our subjective proportions.
- In a sequential setting where we receive a sequence of instances, because the posterior and the prior have the same form, the current posterior accumulates information from all past instances and becomes the prior for the next instance.

Dirichlet-Multinomial Distribution (Continuous)

- In the case the Dirichlet distribution is used as the prior for the multinomial, by integrating out $\mathbf{p}$, we get the marginal joint distribution

$$\begin{aligned} P(x_1, \cdots, x_K | \boldsymbol{\alpha}) &= \int_{\mathbf{p}} P(x_1, \cdots, x_K | \mathbf{p}) P(\mathbf{p}|\alpha) \, d\mathbf{p} \\ &= \int_{\mathbf{p}} \left(\frac{n!}{x_1 \cdots x_K} \prod_{i=1}^{K} p_i^{x_i} \right) \left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \right) d\mathbf{p} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \left(\prod_{k=1}^{K} \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\alpha_k)} \right) \end{aligned}$$

which is called the *Dirichlet-multinomial distribution*.

Uniform Distribution (Continuous)

- We say that X is uniformly distributed within $[a,b]$ if $X \sim Uni(a,b)$, where $Uni(x|a,b) = 1/(b-a)$ for $a \leqslant x \leqslant b$
- $F_X(x) = \int_a^x Uni(x|a,b)dx = (x-a)/(b-a)$
- $E[X] = (a+b)/2$, $Var[X] = (b-a)^2/12$ [Proof]
- [Homework] Plot the above density/mass and distribution functions using Matlab

Convergence of Random Variables (1/2)

Theorem (Convergence in Distribution)

A sequence of random variables $\{X^{(1)}, X^{(2)}, \cdots\}$ ***converges in distribution*** *to* X *iff* $\lim_{n\to\infty} F_{X^{(n)}}(x) = F(x)$.

Theorem (Convergence in Probability)

A sequence of random variables $\{X^{(1)}, X^{(2)}, \cdots\}$ ***converges in probability*** *to* X *iff for any* $\varepsilon > 0$, $\lim_{n\to\infty} P[|X^{(n)} - X| < \varepsilon] = 1$.

Theorem (Convergence Almost Surely)

A sequence of random variables $\{X^{(1)}, X^{(2)}, \cdots\}$ ***converges almost surely*** *to* X *iff* $P\left[\lim_{n\to\infty} X^{(n)} = X\right] = 1$.

Convergence of Random Variables (2/2)

- What's the difference between the convergence in probability and almost surely?

Convergence of Random Variables (2/2)

- What's the difference between the convergence in probability and almost surely?
  - The former leaves open the possibility that $|X^{(n)} - X| > \varepsilon$ happens an infinite number of times; while the latter guarantees that this almost surely will not occur
  - Convergence almost surely implies convergence in probability

Outline

The Sample Mean and Variance

- ***Statistics*** refer to numeric quantities derived from sample data of a population
- Common statistics?

The Sample Mean and Variance

- ***Statistics*** refer to numeric quantities derived from sample data of a population
- Common statistics?Let $\mathcal{X} = \{X^{(1)}, \cdots, X^{(n)}\}$ be a set of n independent and identically distributed (i.i.d.) random variables drawn (or sampled) from a population X of unknown mean μ_X and variance σ_X^2
  - Sample mean: $m_X = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$
  - Sample variance: $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X^{(i)} - m_X)^2$ (Why $\frac{1}{n-1}$ instead of $\frac{1}{n}$?)
- The process of estimating the values (resp. intervals) of parameters of a population using statistics is known as the ***point (resp. interval) estimation***

Bias and Variance (1/2)

- Let θ be an unknown parameter and $d_{\mathcal{X}}$ be its statistic (a random variable) obtained from $\mathcal{X}$, we want to measure how "good" $d_{\mathcal{X}}$ is
  - ***Bias***: $E[d_{\mathcal{X}}]-\theta$ (here the expectation is averaged over all possible $\mathcal{X}$ of the same size, i.e., $E[d_{\mathcal{X}}]=\int d_{\mathcal{X}}p(\mathcal{X})d\mathcal{X}$)
  - ***Variance***: $E\left[(d_{\mathcal{X}}-E[d_{\mathcal{X}}])^2\right]$
  - ***Mean square error***:

$$\begin{aligned}
E_{\mathcal{X}}\left[(d_{\mathcal{X}}-\theta)^2\right] &= E\left[(d_{\mathcal{X}}-E[d_{\mathcal{X}}]+E[d_{\mathcal{X}}]-\theta)^2\right] \\
&= E\left[(d_{\mathcal{X}}-E[d_{\mathcal{X}}])^2+(E[d_{\mathcal{X}}]-\theta)^2+2\,(d_{\mathcal{X}}-E[d_{\mathcal{X}}])\,(E[d_{\mathcal{X}}]-\theta)\right] \\
&= E\left[(d_{\mathcal{X}}-E[d_{\mathcal{X}}])^2\right]+E\left[(E[d_{\mathcal{X}}]-\theta)^2\right]+2E\left[(d_{\mathcal{X}}-E[d_{\mathcal{X}}])\,(E[d_{\mathcal{X}}]-\theta)\right] \\
&= E\left[(d_{\mathcal{X}}-E[d_{\mathcal{X}}])^2\right]+(E[d_{\mathcal{X}}]-\theta)^2 = variance + bias^2
\end{aligned}$$

- We call a statistic ***unbiased estimator*** iff it has zero bias
  - m_X is an unbiased estimator of μ_X, as $E[m_X]=E\left[\frac{1}{n}\sum_{i=1}^{n}X^{(i)}\right]=\frac{1}{n}\sum_{i=1}^{n}E[X^{(i)}]=\mu_X$
  - But $\tilde{s}_X^2=\frac{1}{n}\sum_{i=1}^{n}(X^{(i)}-m_X)^2$ is ***not*** an unbiased estimator of σ_X^2

Bias and Variance (2/2)

$$
\begin{aligned}
Var(m_X) &= E\left[(m_X - E[m_X])^2\right] = E\left[m_X^2 - 2\mu_X m_X + \mu_X^2\right] = E\left[m_X^2\right] - \mu_X^2 \\
&= \frac{1}{n^2}\sum_{ij} E[X^{(i)}X^{(j)}] - \mu_X^2 = \frac{1}{n^2}\left(\sum_{i=j} E[X^{(i)}X^{(j)}] + \sum_{i\neq j} E[X^{(i)}X^{(j)}]\right) - \mu_X^2 \\
&= \frac{1}{n^2}\left(\sum_{i} E[X^{(i)2}] + n(n-1)E[X^{(i)}]E[X^{(j)}]\right) - \mu_X^2 \\
&= \frac{1}{n}E[X^2] + \frac{(n-1)}{n}\mu_X^2 - \mu_X^2 = \frac{1}{n}\left(E[X^2] - \mu_X^2\right) = \sigma_X^2/n
\end{aligned}
$$

$$
\begin{aligned}
E[\widetilde{s}_X^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - m_X)^2\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n} X^{(i)2} - 2\sum_{i=1}^{n} X^{(i)} m_X + \sum_{i=1}^{n} m_X^2\right)\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n} X^{(i)2} - n m_X^2\right)\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} E\left[X^{(i)2}\right] - nE[m_X^2]\right) = E\left[X^2\right] - E[m_X^2] = \left(\sigma_X^2 + \mu_X^2\right) - \left(Var(m_X) + E[m_X]^2\right) \\
&= \sigma_X^2 + \mu_X^2 - \frac{1}{n}\sigma_X^2 - \mu_X^2 = \frac{n-1}{n}\sigma_X^2 \neq \sigma_X^2
\end{aligned}
$$

- We can see from above that $s_X^2 = \frac{n}{n-1}\widetilde{s}_X^2$ is an unbiased estimator

Law of Large Numbers (1/2)

- Let $\{X^{(i)}\}_{i=1}^{n}$ be a set of n i.i.d. random variables drawn from a population X of unknown mean μ_X and variance σ_X, and $m_X^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean

Theorem (Weak Law of Large Numbers)

For any $\varepsilon > 0$, $\lim_{n\to\infty} P\left(\left|m_X^{(n)} - \mu_X\right| < \varepsilon\right) = 1$.

Proof.

By Chebyshev's inequality we have
$P\left(\left|m_X^{(n)} - \mu_X\right| \geqslant \varepsilon\right) = P\left(\left|m_X^{(n)} - E[\overline{x}]\right| \geqslant \varepsilon\right) \leqslant \frac{Var(\overline{x}_n)}{\varepsilon^2} = \frac{\sigma_X}{n\varepsilon^2}$, implying
$\lim_{n\to\infty} P\left(\left|m_X^{(n)} - \mu_X\right| \geqslant \varepsilon\right) \leqslant \lim_{n\to\infty} \frac{\sigma_X}{n\varepsilon^2} = 0$ and therefore
$\lim_{n\to\infty} P\left(\left|m_X^{(n)} - \mu_X\right| < \varepsilon\right) = 1$. □

Law of Large Numbers (2/2)

- More complex arithmetic shows that $m_X^{(n)}$ converges almost surely to μ_X:

Theorem (Strong Law of Large Numbers)

$$P\left(\lim_{n\to\infty} m_X^{(n)} = \mu_X\right) = 1.$$

Outline

Central Limit Theorem

- Now, let's study "how" $m_X^{(n)}$ deviates from μ_X
- Let $Y^{(n)} =_{s.t.} m_X^{(n)} - \mu_X$, we want to know the distribution of $Y^{(n)}$ as $n \to \infty$
  - But the law of large numbers tells us that $P\left(\lim_{n\to\infty} Y^{(n)} = 0\right) = 1$ so the distribution is trivial
- We study the enlarged[2] deviation instead: $Y^{(n)} =_{s.t.} \sqrt{n}(m_X^{(n)} - \mu_X)$

Theorem (Central Limit Theorem)

$\{Y^{(n)}\}_n$ converges in distribution to a random variable of distribution $\mathcal{N}(0, \sigma_X^2)$; that is, $\lim_{n\to\infty} Y^{(n)} \sim \mathcal{N}(0, \sigma_X^2)$, where $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-(x-\mu)^2}{2\sigma^2})$.

- $\lim_{n\to\infty} F_{Y^{(n)}}(x) = \lim_{n\to\infty} P(\sqrt{n}(m_X^{(n)} - \mu_X) \leqslant x) = \int_{-\infty}^{x} \mathcal{N}(x|0, \sigma_X^2)dx$.

[2] It can be shown that $\sqrt{n}$ is the only enlarge coefficient such that $Y^{(n)}$ converges and has nontrivial distribution

The Normal Distribution

- $\mathcal{N}(\mu, \sigma^2)$ is called the ***normal*** (or ***Gaussian***) distribution
- Central limit theorem tells us that ***no matter what the original distribution of X was***, if n is very large, the (enlarged) deviation of the sample mean from μ_X has probability looks like below:

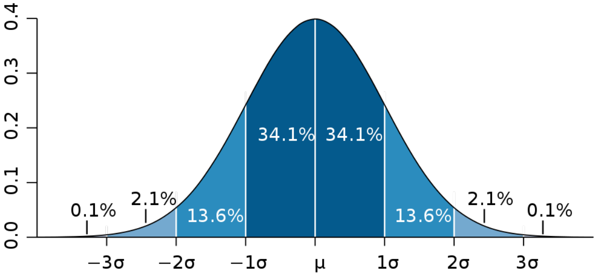


Figure : Density of a normal random variable. The probability that the deviation falls within $[-2\sigma, 2\sigma]$ is about 95%.

Properties

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, (a\sigma)^2)$ for any $a, b \in \mathbb{R}$ [Proof]
  - We call $Z =_{s.t.} \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ the ***z-normalization*** fo X
- Given two functions $f(x) = \mathcal{N}(x|\mu_1, \sigma_1^2)$ and $g(x) = \mathcal{N}(x|\mu_2, \sigma_2^2)$, we have $(f \circ g)(x) = \int f(x-t)g(t)dt = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ [Proof]
  - The convolution of two normal distributions is still a normal distribution
- If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ is independent with $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
  - Not true if X_1 and X_2 are dependent
  - E.g., let $X_1 \sim \mathcal{N}(\mu_{X_1}, \sigma_{X_1}^2)$ and $X_2 =_{s.t.} \begin{cases} X_1, & |X_1| \leqslant c \\ -X_1, & otherwise \end{cases}$ for some $c \in \mathbb{R}$, then both X_1 and X_2 are univariate normal but $X_1 + X_2$ is not

When Should We Assume Normal?

- When should we assume that a random variable is normal?

When Should We Assume Normal?

- When should we assume that a random variable is normal?
  - Given n i.i.d. random variables $X^{(i)}$, $1 \leqslant i \leqslant n$, of mean μ_X and variance σ_X^2, the distribution of random variable $\sqrt{n}\left(\frac{\sum_{i=1}^{n} X^{(i)}}{n} - \mu_X\right)$ approaches $\mathcal{N}(0, \sigma_X^2)$ when n is large
  - That is, the distribution of $\sum_{i=1}^{n} X^{(i)}$ is close to $\mathcal{N}(n\mu_X, n\sigma_X^2)$ when n is large
  - We can assume a random variable to be normal if 1) its values can be regarded as deviations from some prototype (i.e., mean); 2) it can be regarded as the sum of many random variables
- The binomial distribution (sum of outcomes of n Bernoulli experiments) can be approximated by the normal distribution when n is large

Outline

Interval Estimation

- In interval estimation, we specify an interval within which θ lies with a certain degree of confidence.
- To obtain such an interval estimator, we make use of the probability distribution of the point estimator.

Two-sided Confidence Interval

- Suppose $\mathcal{X} = \{X^{(i)}\}_{i=1}^{n}$ is a sample from a normal density with the mean μ_X and variance σ^2.
- Can we find a interval $[u(\mathcal{X}), v(\mathcal{X})]$ such that $P\left(u(\mathcal{X}) < \mu_X < v(\mathcal{X})\right) = \gamma$?
- Let's start from analyzing the property of the sample mean $m_X = \sum_{i=1}^{n} X^{(i)}/n$.

Two-sided Confidence Interval

- m_X is the sum of normals and therefore is also normal, $m_X \sim \mathcal{N}\left(\mu_X, \sigma^2/n\right)$. We can also define the statistic with a *unit normal distribution* $\mathcal{Z} \sim \mathcal{N}(0,1)$:

$$\frac{(m_X - \mu_X)}{\sigma/\sqrt{n}} \sim \mathcal{Z}$$

- We know that $P(-1.96 < \mathcal{Z} < 1.96) = 0.95$, and we can write

$$P(-1.96 < \sqrt{n}\frac{(m_X - \mu_X)}{\sigma} < 1.96) = 0.95$$

or

$$P(m_X - 1.96\frac{\sigma}{\sqrt{n}} < \mu_X < m_X + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

  - That is "with 95 percent confidence," μ_X will lie within $1.96\sigma/\sqrt{n}$ units of the sample mean.

Generalized for Any Required Confidence

- Let us denote z_α such that $P(\mathcal{Z} > z_\alpha) = \alpha$, $0 < \alpha < 1$.
- Because $\mathcal{Z}$ is symmetric around the mean, $z_{1-\alpha/2} = -z_{\alpha/2}$, and $P(X < -z_{\alpha/2}) = P(X > z_{\alpha/2}) = \alpha/2$. Hence,

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} < \mathcal{Z} < z_{\alpha/2}) \\
&= P(-z_{\alpha/2} < \sqrt{n}\frac{(m_X - \mu_X)}{\sigma} < z_{\alpha/2}) \\
&= P(m_X - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_X < m_X + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})
\end{aligned}$$

- Hence, a $100(1-\alpha)$ percent two-sided confidence interval for μ_X can be computed for any α.

One-sided Confidence Interval

- Similarly, knowing that $P(\mathcal{Z} < 1.64) = 0.95$, we have

$$
\begin{aligned}
0.95 &= P(\sqrt{n}\frac{(m_X - \mu_X)}{\sigma} < 1.64) \\
&= P(m_X - 1.64\frac{\sigma}{\sqrt{n}} < \mu_X)
\end{aligned}
$$

  - $(m - 1.64\sigma/\sqrt{n}, \infty)$ is a 95 percent one-sided upper confidence interval for μ_X.

- Generalizing, a $100(1-\alpha)$ percent one-sided confidence interval for μ_X can be computed from

$$
P(m_X - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu_X) = 1 - \alpha
$$

Sample Variance?

- In the previous intervals, we assume the variance σ^2 is known. However, we only have sample variance $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X^{(i)} - m_X)^2$ in usual.
- Then, $\sqrt{N}(m_X - \mu_X)/s_X$ is t-distributed with $N-1$ degrees of freedom, denoted as

$$\frac{\sqrt{N}(m_X - \mu_X)}{s_X} \sim t_{N-1}$$

- Hence for any $\alpha \in (0, 1/2)$, we can define an interval, using the values specified by the *t*-distribution, instead of unit normal $\mathcal{Z}$:

$$P(t_{1-\alpha/2,N-1} < \sqrt{N}\frac{(m_X - \mu_X)}{s_X} < t_{\alpha/2,N-1}) = 1 - \alpha$$

or using $t_{1-\alpha/2,N-1} = -t_{1\alpha/2,N-1}$, we can write

$$P(m_X - t_{\alpha/2,N-1}\frac{s_X}{\sqrt{N}} < \mu_X < m_X + t_{\alpha/2,N-1}\frac{s_X}{\sqrt{N}}) = 1 - \alpha$$

Properties of Student t-distribution

- We say $\sqrt{N}(m_X - \mu_X)/s_X$ is t-distributed with $\nu = N - 1$ degrees of freedom.
- As N becomes larger, t density becomes more and more like the unit normal, the difference being that t has thicker tails, indicating greater variability than does normal.

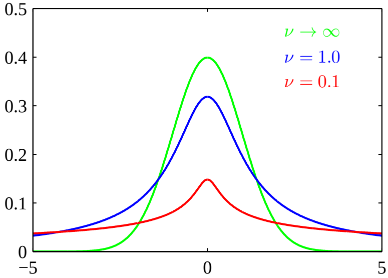


Figure : The limit $\nu \to \infty$ corresponds to a Gaussian distribution

Outline

Hypothesis Testing

- We will come back to this later if we have time to talk about the ML experiments

Outline

Multivariate Random Variables (1/2)

- Now, let's extend the notion of random variable to the multivariate cases: $\boldsymbol{X} = [X_1, \cdots, X_d]^\top$
  - We discuss the distribution of $\boldsymbol{X}$, which is a joint distribution of $X_1, \cdots, X_d$
- Typically, the ***attributes*** X_i (or ***variables*** or ***features***) of $\boldsymbol{X}$ are correlated (otherwise, they can be discussed individually)
  - The ***mean vector*** of $\boldsymbol{X}$ can be defined as $\boldsymbol{\mu}_{\boldsymbol{X}} = E[\boldsymbol{X}] = [\mu_{X_1}, \cdots, \mu_{X_d}]^\top$
  - Denoting $\sigma_{X_i,X_j} = Cov[X_i, X_j] = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})] = E[X_iX_j] - \mu_{X_i}\mu_{X_j}$, we define the ***covariance matrix*** of $\boldsymbol{X}$ as

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = Cov[\boldsymbol{X}] = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_d} \\ \sigma_{X_2,X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2,X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_d,X_1} & \sigma_{X_d,X_2} & \cdots & \sigma^2_{X_d} \end{bmatrix} =$$

$$E[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top] = E[\boldsymbol{X}\boldsymbol{X}^\top] - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top.$$

Multivariate Random Variables (2/2)

- $\Sigma_{\boldsymbol{X}}$ is always symmetric and positive semidefinite
  - $\boldsymbol{v}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{v} = \boldsymbol{v}^\top \left(\int_{\boldsymbol{X}} (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top d\boldsymbol{X} \right) \boldsymbol{v} = \int_{\boldsymbol{X}} \left(\boldsymbol{v}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top \boldsymbol{v} \right) d\boldsymbol{X} = E[\boldsymbol{v}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top \boldsymbol{v}] = E\left[\left(\boldsymbol{v}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}}) \right)^2 \right] \geqslant 0$
- $\Sigma_{\boldsymbol{X}}$ is positive definite iff it is nonsingular
  - We write $Var[\boldsymbol{X}] > 0$ when $\Sigma_{\boldsymbol{X}}$ is positive definite
- $\Sigma_{\boldsymbol{X}}$ is singular (i.e., $det(\Sigma_{\boldsymbol{X}}) = 0$) implies that $\boldsymbol{X}$ has either
  - Deterministic attributes causing zero rows, or
  - Redundant attributes causing linear dependence between rows
- How to measure the variance of $\boldsymbol{X}$?

Multivariate Random Variables (2/2)

- $\Sigma_{\boldsymbol{X}}$ is always symmetric and positive semidefinite
  - $\boldsymbol{v}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{v} = \boldsymbol{v}^\top \left(\int_{\boldsymbol{X}} (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top d\boldsymbol{X} \right) \boldsymbol{v} = \int_{\boldsymbol{X}} \left(\boldsymbol{v}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top \boldsymbol{v} \right) d\boldsymbol{X} = E[\boldsymbol{v}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top \boldsymbol{v}] = E\left[\left(\boldsymbol{v}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}}) \right)^2 \right] \geqslant 0$
- $\Sigma_{\boldsymbol{X}}$ is positive definite iff it is nonsingular
  - We write $Var[\boldsymbol{X}] > 0$ when $\Sigma_{\boldsymbol{X}}$ is positive definite
- $\Sigma_{\boldsymbol{X}}$ is singular (i.e., $det(\Sigma_{\boldsymbol{X}}) = 0$) implies that $\boldsymbol{X}$ has either
  - Deterministic attributes causing zero rows, or
  - Redundant attributes causing linear dependence between rows
- How to measure the variance of $\boldsymbol{X}$? By $det(\Sigma_{\boldsymbol{X}})$
- Suppose $d = 2$, we can see that a small

  $$det(\Sigma_{\boldsymbol{X}}) = det\left(\begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1,X_2} \\ \sigma_{X_2,X_1} & \sigma^2_{X_2} \end{bmatrix} \right) = \sigma^2_{X_1}\sigma^2_{X_2} - \sigma_{X_1,X_2}\sigma_{X_2,X_1}$$

  implies either
  - $\boldsymbol{X}$ does not vary much from $\mu_{\boldsymbol{X}}$, or
  - The attributes of $\boldsymbol{X}$ are highly correlated

Properties and Point Estimation

- Consider $\boldsymbol{w} \in \mathbb{R}^d$ and a random variable $\boldsymbol{w}^\top \boldsymbol{X}$
  - $\mu_{\boldsymbol{w}^\top \boldsymbol{X}} = E[\boldsymbol{w}^\top \boldsymbol{X}] = \boldsymbol{w}^\top E[\boldsymbol{X}] = \boldsymbol{w}^\top \mu_{\boldsymbol{X}}$
  - $\sigma^2_{\boldsymbol{w}^\top \boldsymbol{X}} = Var(\boldsymbol{w}^\top \boldsymbol{X}) = E[(\boldsymbol{w}^\top \boldsymbol{X} - \boldsymbol{w}^\top \mu_{\boldsymbol{X}})^2] =$ $E[(\boldsymbol{w}^\top \boldsymbol{X} - \boldsymbol{w}^\top \mu_{\boldsymbol{X}})(\boldsymbol{X}^\top \boldsymbol{w} - \mu_{\boldsymbol{X}}^\top \boldsymbol{w})] = E[\boldsymbol{w}^\top (\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top \boldsymbol{w}] =$ $\boldsymbol{w}^\top E[(\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})^\top] \boldsymbol{w} = \boldsymbol{w}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{w}$
- Given $\mathcal{X} = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(n)}\}$ a set of n i.i.d. random variables drawn from a population $\boldsymbol{X}$
  - Sample mean: $\boldsymbol{m}_{\boldsymbol{X}} = \frac{\sum_{t=1}^{n} \boldsymbol{X}^{(t)}}{n}$
  - Sample covariance matrix: $\boldsymbol{S}_{\boldsymbol{X}} = \frac{1}{n-1} \sum_{t=1}^{n} (\boldsymbol{X}^{(t)} - \boldsymbol{m}_{\boldsymbol{X}})(\boldsymbol{X}^{(t)} - \boldsymbol{m}_{\boldsymbol{X}})^\top$
    - $s^2_{X_i} = \frac{\sum_{t=1}^{n} (X_i^{(t)} - m_{X_i})^2}{n-1}$
    - $s^2_{X_i, X_j} = \frac{\sum_{t=1}^{n} (X_i^{(t)} - m_{X_i})(X_j^{(t)} - m_{X_j})}{n-1}$

Outline

Multivariate Normal Distribution

Definition (Multivariate Normal Distribution)

A multivariate random variable $\boldsymbol{X} = [X_1, \cdots, X_d]^\top$ is said to have the ***multivariate normal distribution***, denote as $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, iff for any $\boldsymbol{w} \in \mathbb{R}^d$, the random variable $\boldsymbol{w}^\top \boldsymbol{X}$ (that is, the projection of $\boldsymbol{X}$ on $\boldsymbol{w}$) is univariate normal.

- $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} det(\boldsymbol{\Sigma})^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$ provided $\boldsymbol{\Sigma}$ is nonsingular
  - If $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is singular (i.e., $det(\boldsymbol{\Sigma}_{\boldsymbol{X}}) = 0$), we can remove the deterministic/redundant attributes of $\boldsymbol{X}$ to make $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ nonsingular

Distributions of Components

- If $\boldsymbol{X} \sim \mathcal{N}(\mu_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$, then each attribute of $\boldsymbol{X}$ is univariate normal
- Is converse true?

Distributions of Components

- If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, then each attribute of $\boldsymbol{X}$ is univariate normal
- Is converse true? No
  - Again, let $X_1 \sim \mathcal{N}(\mu_{X_1}, \sigma^2_{X_1})$ $X_2 =_{s.t.} \begin{cases} X_1, & |X_1| \leqslant c \\ -X_1, & otherwise \end{cases}$ for some $c \in \mathbb{R}$, and $\boldsymbol{w} = [1,1]^\top$, then both X_1 and X_2 are univariate normal but $\boldsymbol{w}^\top \boldsymbol{X} = X_1 + X_2$ is not
- However, if $X_1, \cdots, X_d$ are i.i.d. and $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, where $\boldsymbol{\mu}_{\boldsymbol{X}} = [\mu_1, \cdots, \mu_d]^\top$ and
$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{bmatrix} \sigma_i^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_i^2 \end{bmatrix}$$ [Proof]

The Mahalanobis Distance

Definition (Mahalanobis Distance)

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two specific values (vectors) of a random variable $\boldsymbol{X}$ with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, the ***Mahalanobis distance*** between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as $(\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}-\boldsymbol{y})$.

- The larger the distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_{\boldsymbol{X}}$, the smaller the multivariate normal density $p_{\boldsymbol{X}}(\boldsymbol{x})$
- Mahalanobis distance degenerates into the Euclidean distance when $\boldsymbol{\Sigma}_{\boldsymbol{X}} = c\boldsymbol{I}$, as $(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})^\top (c\boldsymbol{I})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}}) = \frac{1}{c}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})^\top(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}}) = \frac{1}{c}\|\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}}\|$
- How does $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ affect the distance?
    - The level set $\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}}) = c^2, c \in \mathbb{R}\}$ is an ellipsoid (a surface) centered at $\boldsymbol{\mu}_{\boldsymbol{X}}$ and its shape/orientation are determined by $\boldsymbol{\Sigma}_{\boldsymbol{X}}$

Bivariate Examples (1/3)

- Let's consider an example where $d = 2$,
$$\mathbf{\Sigma_X} = \begin{bmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix}, \text{ and } \rho = \frac{\sigma_{X_1,X_2}}{\sigma_{X_1}\sigma_{X_2}}$$
- If $|\rho| < 1$, then $\mathbf{\Sigma_X}$ is positive definite and nonsingular
  - As $det(\mathbf{\Sigma_X})$ and all the leading principle minors are greater than 0
  - In particular when $|\rho| = 0$, the attributes of $\boldsymbol{X}$ are independent and $p_{\mathbf{X}}(\boldsymbol{x}) = \prod_{i=1}^{d} p_{X_i}(x_i)$
- If $|\rho| = 1$, the two attribute of $\boldsymbol{X}$ are linearly related and one of them can be eliminated

Bivariate Examples (2/3)

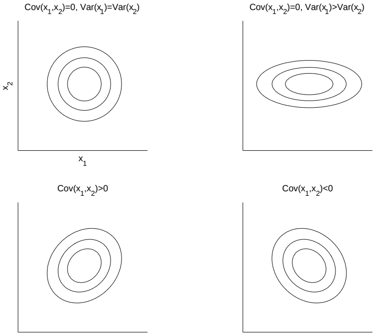
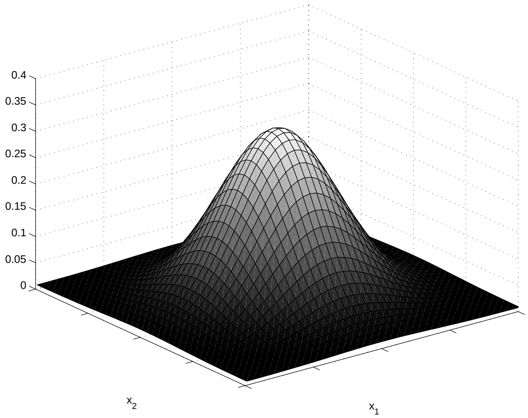


Figure : The level sets closer to the center μ_X are defined with lower c. (a) When $Cov[X_1,X_2]=0$ and $Var[X_1]=Var[X_2]\neq 0$, the level sets are spheres and the Mahalanobis distance degenerates into the Euclidean distance. (b) By increasing $Var[X_1]$, we stretch the level sets (and squeeze the distance) horizontally along the X_1 axis. (c) By increasing $Cov[X_1,X_2]$ (or ρ), we stretch the level sets along the 45° axis. The closer the ρ to 1, the thinner the sets. (d) By decreasing $Cov[X_1,X_2]$ (or ρ), we stretch the level sets along the -45° axis.

Bivariate Examples (3/3)

- The shape of $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}) = \frac{1}{(2\pi)^{d/2}det(\boldsymbol{\Sigma}_{\boldsymbol{X}})^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})\right]$ in a graph is also determined by $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, as it is proportional to the inverse of Mahalanobis distance



Properties (1/2)

- Given $\boldsymbol{X} \sim \mathcal{N}(\mu_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$ and $\boldsymbol{w} \in \mathbb{R}^d$, we have $\boldsymbol{w}^\top \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{w}^\top \mu_{\boldsymbol{X}}, \boldsymbol{w}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{w})$
  - By definition $\boldsymbol{w}^\top \boldsymbol{X}$ is normal and we have $\mu_{\boldsymbol{w}^\top \boldsymbol{X}} = \boldsymbol{w}^\top \mu_{\boldsymbol{X}}$ and $\sigma^2_{\boldsymbol{w}^\top \boldsymbol{X}} = \boldsymbol{w}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{w}$
- More generally, given any $\boldsymbol{W} \in \mathbb{R}^{d \times k}$, $k \leqslant d$, we have $\boldsymbol{W}^\top \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{W}^\top \mu_{\boldsymbol{X}}, \boldsymbol{W}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{W})$ which is k-variate normal
  - The projection of $\boldsymbol{X}$ onto a k-dimensional space is still normal

Properties (2/2)

- Applying Bayes' rule to normal variables we get [Proof]:

Theorem

Given two dependent random variables $\boldsymbol{X} = [X_1, \cdots, X_d]^\top$ *and* $\boldsymbol{Y} = [Y_1, \cdots, Y_k]^\top$ *such that*

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \text{ and } (\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L})$$

for some $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$, $\boldsymbol{W} \in \mathbb{R}^{d \times k}$, $\boldsymbol{b} \in \mathbb{R}^k$ *and* $\boldsymbol{L} \in \mathbb{R}^{k \times k}$, *then we have*

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{W}^\top \boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L} + \boldsymbol{W}^\top \boldsymbol{\Lambda} \boldsymbol{W}) \text{ and}$$
$$(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\Sigma}(\boldsymbol{W}\boldsymbol{L}^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}), \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{W}^\top)^{-1}$.

- The mean of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ is a linear combination of the conditioned values $\boldsymbol{x}$
- $p(\boldsymbol{Y})$ marginalized from $p(\boldsymbol{X}, \boldsymbol{Y})$ is a normal distribution if $p(\boldsymbol{Y}|\boldsymbol{X})$ and $p(\boldsymbol{X})$ are normal distributions satisfying the above relation
  - Note that when $\boldsymbol{W} = \boldsymbol{I}$, $p(\boldsymbol{Y})$ is just the convolution of two normal distributions $\mathcal{N}(\boldsymbol{b}, \boldsymbol{L})$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$