Chapter 1

Background

In this chapter we review some basic mathematics that will be useful for this course.

1.1 Infinite Series

Geometric Series

Geometric series concerns about the sum of a finite or an infinite sequence of numbers with progressive increase in the power. Geometric series appear often in discrete random variables (Chapter 3) when we discuss expectation and moments.

**Definition 1.** Let $0 < r < 1$, a finite geometric series of power $n$ is a sequence of numbers \{1, r, r^2, \ldots, r^n\}. An infinite geometric series is a sequence of numbers \{1, r, r^2, \ldots\}.

**Theorem 1.** The sum of a finite geometric series of power $n$ is

$$
\sum_{k=0}^{n} r^k = 1 + r + r^2 + \ldots + r^n = \frac{1 - r^{n+1}}{1 - r}.
$$

**Proof.** We multiply both sides by $1 - r$. The left hand side becomes

$$
\left(\sum_{k=0}^{n} r^k\right) (1 - r) = (1 + r + r^2 + \ldots + r^n) (1 - r)
$$

$$
= (1 + r + r^2 + \ldots + r^n) - (r + r^2 + r^3 + \ldots + r^{n+1})
$$

$$
\overset{(a)}{=} 1 - r^{n+1},
$$

where (a) holds because terms are canceled due to subtractions. \qed
An immediate corollary of Equation (??) is the sum of an infinite geometric series.

**Corollary 1.** Let \( 0 < r < 1 \). The sum of an infinite geometric series is

\[
\sum_{k=0}^{\infty} r^k = 1 + r + r^2 + \ldots = \frac{1}{1 - r}.
\]

(1.2)

**Proof.** We take limit in Equation (??). This yields

\[
\sum_{k=0}^{\infty} r^k = \lim_{n \to \infty} \sum_{k=0}^{n} r^k = \lim_{n \to \infty} \frac{1 - r^{n+1}}{1 - r} = \frac{1}{1 - r}.
\]

Remark: Note that the condition \( 0 < r < 1 \) is important. If \( r > 1 \), then the limit \( \lim_{n \to \infty} r^{n+1} \) in Corollary 1 will diverge. The constant \( r \) cannot equal to 1, for otherwise the fraction \( \frac{1 - r^{n+1}}{1 - r} \) is undefined. We are not interested in the case when \( r = 0 \), because the sum is trivially 1: \( \sum_{k=0}^{\infty} 0^k = 1 + 0^1 + 0^2 + \ldots = 1 \).

**Corollary 2.** Let \( 0 < r < 1 \). It holds that

\[
\sum_{k=1}^{\infty} kr^{k-1} = 1 + 2r + 3r^2 + \ldots = \frac{1}{(1 - r)^2}.
\]

(1.3)

**Proof.** Take derivative on both sides of Equation (??). The left hand side becomes

\[
\frac{d}{dr} \sum_{k=0}^{\infty} r^k = \frac{d}{dr} (1 + r + r^2 + \ldots)
\]

\[
= 1 + 2r + 3r^3 + \ldots = \sum_{k=1}^{\infty} kr^{k-1}
\]

The right hand side becomes

\[
\frac{d}{dr} \left( \frac{1}{1 - r} \right) = \frac{1}{(1 - r)^2}.
\]

\( \square \)

© 2018 Stanley Chan. All Rights Reserved.
Binomial Series

When handling polynomials, it is common to encounter problems involving powers of a sum of two numbers, e.g., \((a + b)^2\) or \((1 + x)^3\). The Binomial Theorem provides a useful formula when computing these powers.

**Theorem 2** (Binomial Theorem). For any real numbers \(a\) and \(b\), the binomial series of power \(n\) is

\[
(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k,
\]

where \(\binom{n}{k} = \frac{n!}{k!(n-k)!}\).

The Binomial theorem is valid for any real number \(a\) and \(b\). The quantity \(\binom{n}{k}\) reads as “\(n\) choose \(k\)”. Its definition is

\[
\binom{n}{k} \overset{\text{def}}{=} \frac{n!}{k!(n-k)!},
\]

where \(n! = n(n - 1)(n - 2)\ldots 3 \cdot 2 \cdot 1\). We shall discuss the physical meaning of \(\binom{n}{k}\) in Section ??

**Theorem 3** (Pascal Identity). Let \(n\) and \(k\) be positive integers such that \(k \leq n\). Then,

\[
\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.
\]

**Proof.** We start by recalling the definition of \(\binom{n}{k}\). This gives us

\[
\binom{n}{k} + \binom{n}{k-1} = \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-(k-1))!}
\]

\[
= n! \left( \frac{1}{k!(n-k)!} + \frac{1}{(k-1)!(n-(k-1))!} \right),
\]

where we factor out \(n!\) to obtain the second equation. Next, we observe that

\[
\frac{1}{k!(n-k)!} = \frac{n-k+1}{k!(n-k+1)!}, \quad \text{and} \quad \frac{1}{(k-1)!(n-k+1)!} = \frac{k}{k!(n-k+1)!}.
\]

Substituting into the previous equation we obtain

\[
\binom{n}{k} + \binom{n}{k-1} = n! \left( \frac{n-k+1}{k!(n-k+1)!} + \frac{k}{k!(n-k+1)!} \right)
\]

\[
= n! \left( \frac{n+1}{k!(n-k+1)!} \right) = \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}.
\]

\(\square\)

© 2018 Stanley Chan. All Rights Reserved.
With Pascal Identity we can derive the so-called Pascal triangle, illustrated in Figure 1.1.

\[
\begin{array}{cccccc}
 & 1 & & & & \\
1 & 1 & & & & \\
 & 1 & 2 & 1 & & \\
1 & 4 & 6 & 4 & 1 & \\
 & 1 & 5 & 10 & 10 & 5 & 1 \\
1 & 7 & 21 & 35 & 35 & 21 & 7 & 1 \\
\end{array}
\]

Figure 1.1: Pascal triangle for \( n = 1, \ldots, 7 \). Courtesy: Wikipedia.org

Pascal triangle provides a simple visualization of computing the coefficient of \((a + b)^n\). For example, when \( n = 5 \), the number 10 corresponds to \( \binom{5}{3} \). By Pascal Identity, we know that \( \binom{5}{3} = \binom{4}{2} + \binom{4}{3} \), which is the sum of the numbers 4 and 6 of the previous row.

**Example.** Find \((1 + x)^3\). Using Binomial Theorem, we can show that

\[
(1 + x)^3 = \sum_{k=0}^{n} \binom{3}{k} x^{3-k} = 1 + 3x + 3x^2 + x^3.
\]

**Example.** Let \( 0 < p < 1 \). Find \( \sum_{k=0}^{n} \binom{n}{k} p^{n-k}(1 - p)^k \). Again by using Binomial Theorem, we have

\[
\sum_{k=0}^{n} \binom{n}{k} p^{n-k}(1 - p)^k = (p + (1 - p))^n = 1.
\]

This result will be useful when evaluating Binomial random variables in Chapter 3.

**Proof of Binomial Theorem** (Optional). We now prove the Binomial Theorem.

**Proof.** We prove by induction. Consider the base case \( n = 1 \). When \( n = 1 \),

\[
(a + b)^1 = a + b = \sum_{k=0}^{1} a^{1-k}b^k.
\]

Therefore, the base case is verified. Assume up to case \( n \). We need to verify case \( n + 1 \).

\[
(a + b)^{n+1} = (a + b)(a + b)^n = (a + b)\sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k
\]

\[
= \sum_{k=0}^{n} \binom{n}{k} a^{n-k+1}b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^{k+1}.
\]

© 2018 Stanley Chan. All Rights Reserved.
We want to apply Pascal Identity to combine the two terms. In order to do so, we note that the second term in this sum can be rewritten as
\[
\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^{k+1} = \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k-1} b^{k+1} \\
= \sum_{\ell=1}^{n+1} \binom{n}{\ell-1} a^{n+1-\ell} b^\ell, \quad \text{where} \quad \ell = k + 1 \\
= \sum_{\ell=1}^{n} \binom{n}{\ell-1} a^{n+1-\ell} b^\ell + b^{n+1}.
\]
The first term in the sum can be written as
\[
\sum_{k=0}^{n} \binom{n}{k} a^{n-k+1} b^k = \sum_{\ell=1}^{n} \binom{n}{\ell} a^{n+1-\ell} b^\ell + a^{n+1}, \quad \text{where} \quad \ell = k.
\]
Therefore, the two terms can be combined using Pascal Identity to yield
\[
(a + b)^{n+1} = \sum_{\ell=1}^{n} \left[ \binom{n}{\ell} + \binom{n}{\ell-1} \right] a^{n+1-\ell} b^\ell + a^{n+1} + b^{n+1} \\
= \sum_{\ell=1}^{n} \binom{n+1}{\ell} a^{n+1-\ell} b^\ell + a^{n+1} + b^{n+1} = \sum_{\ell=0}^{n+1} \binom{n+1}{\ell} a^{n+1-\ell} b^\ell.
\]
Hence, the \(n+1\)-th case is also verified. This completes the proof.

1.2 Taylor Approximation

Given a function \(f : \mathbb{R} \to \mathbb{R}\), it is often useful to analyze its behavior by approximating \(f\) using its local information. Taylor approximation (or Taylor series) is one of the most powerful tools for such tasks. We will use Taylor approximation in many occasions.

**Definition 2 (Taylor Approximation).** Let \(f : \mathbb{R} \to \mathbb{R}\) be a continuous function with infinite derivatives. Let \(a \in \mathbb{R}\) be a fixed constant. The Taylor approximation of \(f\) at \(x = a\) is
\[
f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots \\
= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n,
\]
where \(f^{(n)}\) denotes the \(n\)-th order derivative of \(f\).
The Taylor approximation defined in Definition ?? has an infinite number of terms. If we use a finite number of terms, we obtain an \( n \)-th order Taylor approximation:

First Order: \( f(x) = f(a) + f'(a)(x - a) + \mathcal{O}((x - a)^2) \)

Second Order: \( f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \mathcal{O}((x - a)^3) \).

Here, the big-O notation \( \mathcal{O}(\varepsilon^k) \) means any term that has an order at least power \( k \). For small changes, i.e., \( \varepsilon \ll 1 \), a high-order term \( \mathcal{O}(\varepsilon^k) \approx 0 \) for large \( k \).

**Exponential Series**

An immediate application of Taylor approximation is to derive the exponential series.

**Theorem 4.** Let \( x \) be any real number. Then,

\[
e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.
\]

**(1.6)**

**Proof.** Let \( f(x) = e^x \) for any \( x \). Then, Taylor approximation around \( x = 0 \) is

\[
f(x) = f(0) + f'(0)(x - 0) + \frac{f''(0)}{2!}(x - 0)^2 + \ldots
\]

\[
= e^0 + e^0(x - 0) + \frac{e^0}{2!}(x - 0)^2 + \ldots
\]

\[
= 1 + x + \frac{x^2}{2} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.
\]

\[\square\]

**Example.** Evaluate the sum

\[
\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^\lambda = 1.
\]

This result will be useful to **Poisson random variables** in Chapter 3.

**Logarithmic Approximations**

Taylor approximation also allows us to find approximations to logarithm functions.

**Lemma 1.** Let \( 0 < x < 1 \) be a constant. Then,

\[
\log(1 + x) = x - x^2 + \mathcal{O}(x^3).
\]

**(1.7)**
Proof. Let \( f(x) = \log(1 + x) \). Then, the derivatives of \( f \) are

\[
f'(x) = \frac{1}{1 + x}, \quad \text{and} \quad f''(x) = -\frac{1}{(1 + x)^2}.
\]

Taylor approximation at \( x = 0 \) gives

\[
f(x) = f(0) + f'(0)(x - 0) + \frac{f''(0)}{2}(x - 0)^2 + O(x^3)
\]

\[
= \log 1 + \left(\frac{1}{1 + 0}\right)x - \left(\frac{1}{(1 + 0)^2}\right)x^2 + O(x^3)
\]

\[
= x - x^2 + O(x^3).
\]

\[
\square
\]

Example. Show that

\[
\lim_{N \to \infty} \left(1 + \frac{s^2}{2N}\right)^N = e^{s^2/2}.
\] (1.8)

Solution: To prove the above equation, we consider \( N \log \left(1 + \frac{s^2}{N}\right) \). By the logarithmic lemma, we can obtain the second order approximation:

\[
\log \left(1 + \frac{s^2}{2N}\right) = \frac{s^2}{2N} - \frac{s^4}{4N^2}.
\]

Therefore, multiplying both sides by \( N \) yields

\[
N \log \left(1 + \frac{s^2}{2N}\right) = \frac{s^2}{2} - \frac{s^4}{4N}.
\]

Putting the limit \( N \to \infty \) we can show that

\[
\lim_{N \to \infty} \left\{N \log \left(1 + \frac{s^2}{2N}\right)\right\} = \frac{s^2}{2}.
\]

Take exponential on both sides yields the results.

Remark: This result will be useful to prove the Central Limit Theorem.

1.3 Integration

There are a few key results in calculus we need for this course, especially when we discuss cumulative distribution functions in Chapter 4, and transformation of random variables in Chapter 5.
Theorem 5 (Fundamental Theorem of Calculus). Let \( f : [a, b] \to \mathbb{R} \) be a continuous function defined on a closed interval \([a, b]\). Then,

\[
f(x) = \frac{d}{dx} \int_a^x f(t) dt,
\]

for any \( x \in (a, b) \).

Proof. (Optional).
Define the integral as a function \( F \):

\[
F(x) = \int_a^x f(t) dt.
\]

The derivative of \( F \) with respect to \( x \) is

\[
\frac{d}{dx} F(x) = \lim_{h \to 0} \frac{F(x + h) - F(x)}{h} = \lim_{h \to 0} \frac{1}{h} \left( \int_a^{x+h} f(t) dt - \int_a^x f(t) dt \right)
\]

\[
= \lim_{h \to 0} \frac{1}{h} \int_x^{x+h} f(t) dt
\]

\[
(a) \leq \lim_{h \to 0} \frac{1}{h} \int_x^{x+h} \left\{ \max_{x \leq \tau \leq x+h} f(\tau) \right\} dt,
\]

\[
= \lim_{h \to 0} \left\{ \max_{x \leq \tau \leq x+h} f(\tau) \right\} = f(x).
\]

Here, the inequality in \((a)\) holds because \( f(t) \leq \max_{x \leq \tau \leq x+h} f(\tau) \) for all \( x \leq t \leq x+h \). The maximum exists because \( f \) is continuous in a closed interval.

Remark: An alternative proof is to use Mean Value Theorem in terms of Riemann-Stieltjes integrals (See, e.g., T. Apostol, “Mathematical Analysis”, Pearson 2004, Theorem 7.34). To handle more general functions such as delta functions, one can use techniques in Lebesgue’s integration. However, this is beyond the scope of our course.

Corollary 3. Let \( f : [a, b] \to \mathbb{R} \) be a continuous function defined on a closed interval \([a, b]\). Let \( g : \mathbb{R} \to [a, b] \) be a continuously differentiable function. Then,

\[
\frac{d}{dx} \int_a^{g(x)} f(t) dt = g'(x)f(g(x)),
\]

for any \( x \in (a, b) \).
Proof. We can prove this by chain rule: Let $y = g(x)$, then we have

$$
\frac{d}{dx} \int_{a}^{g(x)} f(t) dt = \frac{dy}{dx} \cdot \frac{d}{dy} \int_{a}^{y} f(t) dt = g'(x)f(y),
$$

which completes the proof. \hfill \Box

Example. Evaluate the integral

$$
\frac{d}{dx} \int_{0}^{x-\mu} \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{t^2}{2\sigma^2} \right\} dt.
$$

Solution. Let $y = x - \mu$. Then by using the fundamental theorem of calculus, we can show that

$$
\frac{d}{dx} \int_{0}^{x-\mu} \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{t^2}{2\sigma^2} \right\} dt = \frac{dy}{dx} \cdot \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{y^2}{2\sigma^2} \right\}
$$

$$
= \frac{d(x - \mu)}{dx} \cdot \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}
$$

$$
= \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}.
$$

1.4 Linear Algebra

We use matrices and vectors to study high dimensional problems, e.g., multi-variate Gaussian in Chapter 6 and regression in Chapter 7.

When we discuss matrices and vectors, we use bold letter $x \in \mathbb{R}^N$ to denote a (column) vector, and a bold upper case letter to denote a matrix $A \in \mathbb{R}^{M \times N}$.

Definition 3 (Inner Product). Let $x = [x_1, x_2, \ldots, x_N]^T$, and $y = [y_1, y_2, \ldots, y_N]^T$. The inner product $x^T y$ is

$$
x^T y = \sum_{i=1}^{N} x_i y_i.
$$

Geometrically, the inner product of two vectors $x$ and $y$ can be considered as the projection of one onto the other.

Example. Let $x = [1, 0, -1]^T$, and $y = [3, 2, 0]^T$. Find $x^T y$.

Solution. The inner product is $x^T y = (1)(3) + (0)(2) + (-1)(0) = 3$. 

© 2018 Stanley Chan. All Rights Reserved.
Definition 4 (Norm). Let \( \mathbf{x} = [x_1, \ldots, x_N]^T \) be a vector. The \( \ell_p \)-norm of \( \mathbf{x} \) is

\[
\|\mathbf{x}\|_p = \left( \sum_{i=1}^{N} x_i^p \right)^{1/p},
\]

for any \( p \geq 1 \).

There are a few important norms we will encounter in Chapter 7 regression. The first one is the \( \ell_2 \)-norm

\[
\|\mathbf{x}\|_2 = \left( \sum_{i=1}^{N} x_i^2 \right)^{1/2}.
\]

By taking the square on both sides, one can show that \( \|\mathbf{x}\|_2^2 = \mathbf{x}^T \mathbf{x} \). This is called the squared \( \ell_2 \)-norm, and is the sum of the squares.

The \( \ell_1 \)-norm will be used occasionally in this course, but it has significant implication in sparse models. The \( \ell_1 \)-norm of \( \mathbf{x} \) is

\[
\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|,
\]

which is the sum of absolute values.

The \( \ell_\infty \)-norm picks the maximum of \( \{x_1, \ldots, x_N\} \):

\[
\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \left( \sum_{i=1}^{N} x_i^p \right)^{1/p} = \max \{x_1, \ldots, x_N\},
\]

because as \( p \to \infty \), only the largest element will be amplified.

Definition 5 (Weighted \( \ell_2 \)-norm Square). Let \( \mathbf{x} = [x_1, \ldots, x_N]^T \) and let \( \mathbf{W} = \text{diag}(w_1, \ldots, w_N) \) be a non-negative diagonal matrix. The weighted \( \ell_2 \)-norm square of \( \mathbf{x} \) is

\[
\|\mathbf{x}\|_W^2 = \mathbf{x}^T \mathbf{W} \mathbf{x} = \left[ x_1 \ldots x_N \right] \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_N \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \sum_{i=1}^{N} w_i x_i^2.
\]

The geometry of the weighted \( \ell_2 \)-norm is determined by the matrix \( \mathbf{W} \). For example, if \( \mathbf{W} = \mathbf{I} \) (the identity operator), then \( \|\mathbf{x}\|_W^2 = \|\mathbf{x}\|_2^2 \). This defines a circle. If \( \mathbf{W} \) is any non-negative matrix, then \( \|\mathbf{x}\|_W^2 \) defines an ellipse.

© 2018 Stanley Chan. All Rights Reserved.
While in this course we do not need excessive knowledge about matrix calculus, some basic results would become useful for discussing least squares regression in Chapter 7.

**Definition 6.** Scalar Differentiation of Vectors Let $f : \mathbb{R}^N \to \mathbb{R}$ be a differentiable scalar function, and let $y = f(x)$ for some input $x \in \mathbb{R}^N$. Then,

$$\frac{dy}{dx} = \begin{bmatrix} \frac{dy}{dx_1} \\ \vdots \\ \frac{dy}{dx_N} \end{bmatrix}. \quad (1.13)$$

**Example.** Let $y = x^T Ax$ for any matrix $A \in \mathbb{R}^{N \times N}$. Find $\frac{dy}{dx}$.

**Solution.**

$$\frac{d}{dx} (x^T Ax) = Ax + A^T x. \quad (1.14)$$

Now, if $A$ is symmetric, i.e., $A = A^T$, then

$$\frac{d}{dx} (x^T Ax) = 2Ax. \quad (1.15)$$

**Example.** Let $\epsilon = \|Ax - y\|^2_2$, where $A \in \mathbb{R}^{N \times N}$ is symmetric. Find $\frac{d\epsilon}{dx}$.

**Solution.** First, we note that

$$\epsilon = \|Ax - y\|^2_2 = x^T A^T Ax - 2y^T Ax + y^T y.$$ 

Taking the derivative with respect to $x$ yields

$$\frac{d\epsilon}{dx} = 2A^T Ax - 2A^T y = 2A^T (Ax - y). \quad (1.16)$$

**Definition 7 (Inverse).** For an $N \times N$ matrix $A$, the inverse $A^{-1}$ is a matrix such that

$$AA^{-1} = I, \quad (1.17)$$

where $I$ is the identity matrix.

**Remark:** The inverse of a matrix does not always exist.

**Example.** Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, find the inverse $A^{-1}$. When will $A^{-1}$ exist?

**Solution.** $A^{-1}$ exists as long as $ad \neq bc$.

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \quad (1.18)$$
The denominator of the previous equation is known as the determinant of a matrix. In general we use
\[ \det A \quad \text{or} \quad |A| \]
to denote the determinant of a matrix \( A \).

### 1.5 Basic Combinatorics

Permutations and combinations are useful to study the number of possible configurations in an experiment. We use them often when we discuss discrete random variables in Chapter 3.

**Permutation**

Permutation concern about the following questions:

Consider a set of \( n \) distinct balls. Suppose we want to pick \( k \) balls from the set without replacement. How many ordered configurations can we obtain?

Note that in the above question, the word “ordered” is crucial. For example, the set \( A = \{a, b, c\} \) can lead to 6 different ordered configurations

\[ (a, b, c), \ (a, c, b), \ (b, a, c), \ (b, c, a), \ (c, a, b), \ (c, b, a). \]

**Theorem 6.** The number of permutations of choosing \( k \) out of \( n \) is

\[
\frac{n!}{(n-k)!}
\]

where \( n! = n(n-1)(n-2) \ldots 3 \cdot 2 \cdot 1. \)

**Proof.** We can prove the above theorem by listing out all possible ways of getting the result.

<table>
<thead>
<tr>
<th>Which ball to pick</th>
<th>Number of choices</th>
<th>Why?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The 1st ball</td>
<td>( n )</td>
<td>No has been picked, so we have ( n ) choices</td>
</tr>
<tr>
<td>The 2nd ball</td>
<td>( n - 1 )</td>
<td>The first ball has been picked</td>
</tr>
<tr>
<td>The 3rd ball</td>
<td>( n - 2 )</td>
<td>The first two balls have been picked</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>The ( k )th ball</td>
<td>( n - k + 1 )</td>
<td>The first ( k - 1 ) balls have been picked</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td>( n(n-1)(n-2) \ldots (n-k+1) )</td>
<td></td>
</tr>
</tbody>
</table>

© 2018 Stanley Chan. All Rights Reserved. 12
As shown above, the total number of ordered configurations is $n(n-1)(n-2)\ldots(n-k+1)$. This can be simplified as

$$n(n-1)(n-2)\ldots(n-k+1) = n(n-1)(n-2)\ldots(n-k+1) \cdot \frac{(n-k)(n-k-1)\ldots3\cdot2\cdot1}{(n-k)(n-k-1)\ldots3\cdot2\cdot1} = \frac{n!}{(n-k)!}.$$

\[\square\]

**Example.** Consider a set of 4 balls \{1, 2, 3, 4\}. We want to pick 2 balls at random without replacement. The ordering matters. How many permutations can we obtain?

**Solution.** The possible configurations are (1,2), (2,1), (1,3), (3,1), (1,4), (4,1), (2,3), (3,2), (2,4), (4,2), (3,4), (4,3). So totally there are 12 configurations. We can also verify this number by noting that there are 4 balls altogether and so the number of choices for picking the first ball is 4 and the number of choices for picking the second ball is (4-1)=3. Thus, the total is $4 \cdot 3 = 12$. Referring to the formula, this result coincides with the theorem, which states that the number of permutations is $\frac{4!}{(4-2)!} = \frac{4\cdot3\cdot2\cdot1}{2\cdot1} = 12$.

**Combination**

Combination concern about the following questions:

*Consider a set of \(n\) distinct balls. Suppose we want to pick \(k\) balls from the set without replacement. How many un-ordered configurations can we obtain?*

Unlike permutation, combination treats a subset of balls with whatever ordering as one single configuration. For example, the subset \((a, b, c)\) is considered no difference with \((a, c, b)\), or \((b, c, a)\), etc.

**Theorem 7.** The number of combinations of choosing \(k\) out of \(n\) is

$$\frac{n!}{k!(n-k)!}$$

where \(n! = n(n-1)(n-2)\ldots3\cdot2\cdot1\).

**Proof.** We start with the permutation result, which gives us $\frac{n!}{(n-k)!}$ permutations. Note that every permutation has exactly \(k\) balls. However, while these \(k\) balls can be arranged in any
order, in combination we treat them as one single configuration. Therefore, the task is to
count the number of possible orderings for these \( k \) balls.

To this end, we note that for a set of \( k \) balls, there are totally \( k! \) possible ways of ordering
them. The number \( k! \) comes from the following table.

<table>
<thead>
<tr>
<th>Which ball to pick</th>
<th>Number of choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>The 1st ball</td>
<td>( k )</td>
</tr>
<tr>
<td>The 2nd ball</td>
<td>( k - 1 )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>The ( k )th ball</td>
<td>1</td>
</tr>
<tr>
<td>Total:</td>
<td>( k(k-1) \ldots 3 \cdot 2 \cdot 1 )</td>
</tr>
</tbody>
</table>

Therefore, the total number of orderings for a set of \( k \) balls is \( k! \).

Since permutation gives us \( \frac{n!}{(n-k)!} \) and every permutation has \( k! \) repetitions due to ordering, we divide the number by

\[ \frac{n!}{k!(n-k)!} \]

This completes the proof.

\[ \square \]

**Example.** Consider a set of 4 balls \( \{1,2,3,4\} \). We want to pick 2 balls at random without
replacement. The ordering does not matter. How many combinations can we obtain?

**Solution.** The permutation result gives us 12 permutations. However, among all these 12
permutations, there are only 6 distinct pair of numbers. For example, the pair (1,2) and
(2,1) are the same if we do not consider ordering. We can confirm this by noting that since
we picked 2 balls, there are exactly 2 possible orderings for these 2 balls. Therefore, we
have \( \frac{12}{2} = 6 \) number of combinations. Using the formula of the theorem, we check that the
number of combinations is \( \frac{4!}{2!(4-2)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6 \).

### 1.6 Set Theory

To study probability, we need some basic concepts in set theory. These involves operations
of sets that allow us to combine, separate and decompose sets.

**Set, Subset, Empty Set, Universal Set**

| **Definition 8 (Set).** A set is a collection of objects. We denote |
|--------------------|-------------------|
| \( A = \{\xi_1, \xi_2, \ldots, \xi_n\} \) | as a set, where \( \xi_i \) is the \( i \)-th element in the set. |

© 2018 Stanley Chan. All Rights Reserved.
A set can be finite, infinite and countable, or infinite and uncountable. See the following three examples.

**Example.** $A = \{1, 2, 3, 4, 5, 6\}$ is a finite set.

**Example.** $A = \{2, 4, 6, 8, \ldots\}$ is a countable but infinite set.

**Example.** $A = \{x \mid 0 < x < 1\}$ is a un-countable infinite set.

**Remark:** The elements of a set can also be sets. For example, $A = \{\{1, 2\}, \{3\}\}$ is a set containing two sets $\{1, 2\}$ and $\{3\}$. The cardinality (i.e. the size) of $A$ is 2.

**Definition 9** (Subset). A set $B$ is a **subset** of $A$ if for any $\xi \in B$, $\xi$ is also in $A$. We write

$$B \subseteq A$$

**Definition 10** (Empty Set). A set is **empty** is it contains no element. We denote an empty set as

$$A = \emptyset.$$  

An empty set is a subset of any set, i.e., $\emptyset \subseteq A$ for any $A$.

**Definition 11** (Universal Set). The **universal set** is the set containing all elements. We denote a universal set as

$$A = \Omega.$$  

The universal set $\Omega$ contains itself, i.e., $\Omega \subseteq \Omega$.

**Example.** Let $A = \{1, 2, 3\}$. List out all the subsets of $A$.

**Solution.** The subsets of $A$ include:

$$A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$
Union and Intersection

Definition 12 (Finite Union). The union of two sets $A$ and $B$ contains all elements in $A$ or in $B$. That is,

$$A \cup B = \{\xi \mid \xi \in A \text{ or } \xi \in B\}. \tag{1.19}$$

If $A \subseteq B$, then $A \cup B = B$ because $B$ is a bigger set than $A$.

Example. If $A = \{1, 2, 3, 4\}$, $B = \{1, 5, 6\}$, then $A \cup B = \{1, 2, 3, 4, 5, 6\}$.

Example. If $A = \{t \mid 3 < t \leq 4\}$, $B = \{t \mid t \geq 3.5\}$, then $A \cup B = \{t \mid t > 3\}$.

Definition 13 (Infinite Union). For an infinite sequence of sets $A_1, A_2, \ldots$, the infinite union is defined as

$$\bigcup_{n=1}^{\infty} A_n = \{\xi \mid \xi \in A_n \text{ for at least one } n = 1, 2, \ldots\}. \tag{1.20}$$

Here, in this definition, the index $n$ can be arbitrarily large but it should still be finite.

Example. Find the infinite union of $\bigcup_{n=1}^{\infty} A_n$ if

$$A_n = \left[0, 1 - \frac{1}{n}\right].$$

Solution. First, we note that $[0, 1) \subset A_n$ for any $n$. However, $1 \not\in \bigcup_{n=1}^{\infty} A_n$ because 1 is not an element in any $A_n$. Therefore,

$$\bigcup_{n=1}^{\infty} A_n = [0, 1).$$

Definition 14 (Finite Intersection). The intersection of two sets $A$ and $B$ contains all elements in $A$ and in $B$. That is,

$$A \cap B = \{\xi \mid \xi \in A \text{ and } \xi \in B\}. \tag{1.21}$$

Example. If $A = \{1, 2, 3, 4\}$, $B = \{1, 5, 6\}$, then $A \cap B = \{1\}$.

Example. If $A = \{t \mid 3 < t \leq 4\}$, $B = \{t \mid t \geq 3.5\}$, then $A \cap B = \{t \mid 3.5 \leq t \leq 4\}$.

Definition 15 (Infinite Intersection). For an infinite sequence of sets $A_1, A_2, \ldots$, the infinite intersection is defined as

$$\bigcap_{n=1}^{\infty} A_n = \{\xi \mid \xi \in A_n \text{ for every } n = 1, 2, \ldots\} \tag{1.22}$$
Figure 1.2: [Left] Union and [Right] Intersection.

**Example.** Find the infinite intersection of $\bigcap_{n=1}^{\infty} A_n$ if

$$A_n = \left[0, 1 + \frac{1}{n}\right].$$

**Solution.** The set $A = [0, 1]$ is contained in all $A_n$. Therefore,

$$\bigcap_{n=1}^{\infty} A_n = [0, 1].$$

**Complement, Difference**

**Definition 16 (Complement).** The **complement** of a set $A$ is the set containing all elements in $\Omega$ but not in $A$. That is

$$A^c = \{\xi \mid \xi \in \Omega \text{ and } \xi \notin A\}.$$  

**Example.** Let $A = \{\text{even integers}\}$, $\Omega = \{\text{integers}\}$, then $A^c = \{\text{odd integers}\}$.

Note that when defining complement, the universal set $\Omega$ is critical. For example, if in the previous example we set $\Omega = \mathbb{R}$, then $A^c$ will become the set of all non-even integer real numbers. That would include 1.3, $\pi$, $\sqrt{2}$, etc.

**Definition 17 (Difference).** The **difference** $A \setminus B$ is the set containing all elements in $A$ but not in $B$.

$$A \setminus B = \{\xi \mid \xi \in A \text{ and } \xi \notin B\}.$$  

**Example.** Let $A = \{1, 3, 5, 6\}$, $B = \{2, 3, 4\}$, then $A \setminus B = \{1, 5, 6\}$. 

© 2018 Stanley Chan. All Rights Reserved.
It is possible to define difference in terms of intersection and complement.

**Theorem 8.** Let $A$ and $B$ be two sets. Then

$$A \setminus B = A \cap B^c$$

(1.23)

**Proof.** Let $x \in A \setminus B$. Then $x \in A$ and $x \notin B$. Since $x \notin B$, we have $x \in B^c$. Therefore, $x \in A$ and $x \in B^c$. By definition of intersection, we have $x \in A \cap B^c$. This shows $A \setminus B \subseteq A \cap B^c$.

Conversely, let $x \in A \cap B^c$. Then, $x \in A$ and $x \in B^c$, which implies that $x \in A$ and $x \notin B$. By definition of $A \setminus B$, we have that $x \in A \setminus B$. This shows $A \cap B^c \subseteq A \setminus B$. \qed

**Disjoint and Partition**

**Definition 18 (Disjoint).** Two sets $A$ and $B$ are **disjoint** if

$$A \cap B = \emptyset.$$ 

(1.24)

For a collection of sets $\{A_1, A_2, \ldots, A_n\}$, we say that the collection is disjoint if

$$A_i \cap A_j = \emptyset,$$

(1.25)

for any pair $i \neq j$.

**Example.** Let $A = \{x > 1\}$ and $B = \{x < 0\}$. Then $A$ and $B$ are disjoint.

**Example.** Let $A = \{1, 2, 3\}$ and $B = \emptyset$. Then $A$ and $B$ are disjoint.
Definition 19 (Partition). A collection of sets \( \{A_1, \ldots, A_n\} \) is a **partition** to the universal set \( \Omega \) if it satisfies the following conditions:

- (non-overlap) \( \{A_1, \ldots, A_n\} \) is disjoint:

\[
A_i \cap A_j = \emptyset. \tag{1.26}
\]

- (decompose) Union of \( \{A_1, \ldots, A_n\} \) gives the universal set:

\[
\bigcup_{i=1}^{n} A_i = \Omega. \tag{1.27}
\]

As its name suggests, a partition is a collection of non-overlapping subsets such that when taking the union they will form the universal set. Partition is important in studying probability because it allows us to tackle complex events by considering smaller / simple sub-events.

![Figure 1.4: A partition of \( \Omega \)](image)

**Example.** Let \( \Omega = \{1, 2, 3, 4, 5, 6\} \). The collection

\[
A_1 = \{1, 2, 3\}, \quad A_2 = \{4, 5\}, \quad A_3 = \{6\}
\]

forms a partition.

**Example.** Let \( \Omega = \{1, 2, 3, 4, 5, 6\} \). The collection

\[
A_1 = \{1, 2, 3\}, \quad A_2 = \{4, 5\}, \quad A_3 = \{5, 6\}
\]

does not form a partition because \( A_2 \cap A_3 = \{5\} \).
Set Operations

When handling multiple sets, it would be useful to have some basic set operations. The followings are the four very basic set operations we will use in this course:

<table>
<thead>
<tr>
<th>Theorem 9 (Set Operations)</th>
<th>Let $A$, $B$ and $C$ be three sets. The following identities hold.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Commutative (Order does not matter).</strong></td>
<td>$A \cap B = B \cap A,$ and $A \cup B = B \cup A.$ (1.28)</td>
</tr>
<tr>
<td><strong>Associative (How to do multiple union and intersection)</strong></td>
<td>$A \cup (B \cup C) = (A \cup B) \cup C$ $A \cap (B \cap C) = (A \cap B) \cap C.$ (1.29)</td>
</tr>
<tr>
<td><strong>Distributive (How to mix union and intersection)</strong></td>
<td>$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$ (1.30)</td>
</tr>
<tr>
<td><strong>De Morgan’s Law (How to complement over intersection and union)</strong></td>
<td>$(A \cap B)^c = A^c \cup B^c$ $(A \cup B)^c = A^c \cap B^c.$</td>
</tr>
</tbody>
</table>