<u>Sample Space</u>

<u>*Problem of the Day*</u>: A family has two kids, one of them is a girl. What is the probability that both kids are girls?

- To solve the problem, we need to have a <u>statistical model</u>.

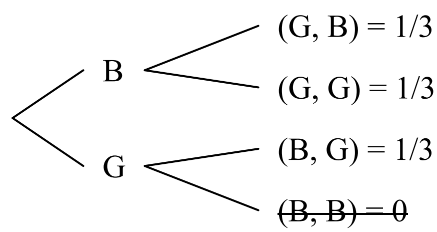
❖ ***Statistical model*** is a set of assumptions

- A1: For each kid, $\Pr(B) = \Pr(G) = 0.5$
- A2: The gender of the two kids are independent

❖ ***Sample Space*** is the list of elementary events with their probability of occurrence.

❖ Interpretation of the Problem

1) Suppose I have no additional information about the family. Then we have 4 elementary events, each with the same probability $1/4$.



Since my information is that at least one kid is a girl, this rules out the event (B, B). Thus, the sample space becomes {(G,G), (G,B), (B,G)}, in which each element has the same probability $1/3$.

Therefore, the answer under Interpretation 1 is $1/3$.

2) Suppose, in addition to the information given, I have also met *a girl* from this family. Now the experiment is simply about the other kid whom I have never met. In this case, the sample space is {B, G}, and each element has the same probability $1/2$.

Therefore, the answer under Interpretation 2 is $1/2$. The fact that I have met *a girl* in this family increases the probability of the family having two girls from $1/3$ to $1/2$.

- Conclusion: Always make explicit the following:
  - The statistical experiment
  - The sample space $\Omega = \{\omega_1, \dots, \omega_N\}$ where ω_n's are elementary events and their probabilities
  - The statistical model

❖ If possible, we should define the elementary events so that the statistical model leads to think that all elementary events have the same probability.

- If the sample space is <u>FINITE</u>,

$$P(A) = \frac{\text{number of elementary events in } A}{\text{number of elements in } \Omega} = \frac{\#A}{\#\Omega}$$

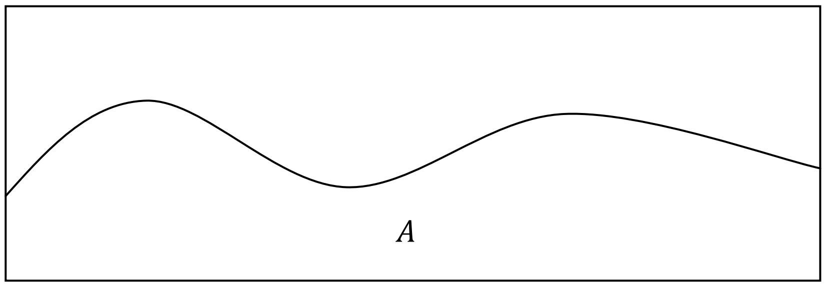
where A is an event not necessarily elementary, i.e. A is a list of elementary events.

- In this case, $P(A)$ is the uniform probability on Ω,

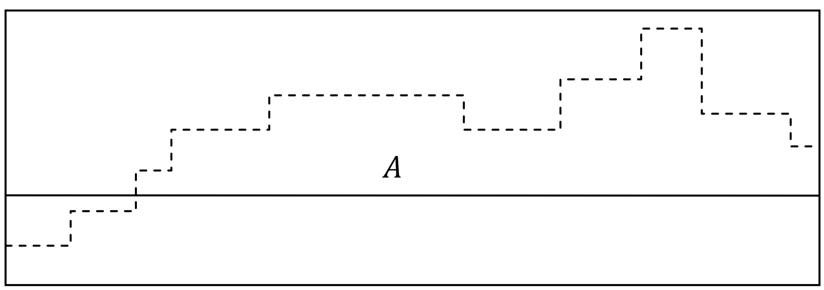
$$P : \mathcal{P}(\Omega) \to [0,1]$$
$$A \mapsto \frac{\#A}{\#\Omega}$$

Algebras and σ-Algebras of Events

Problem: I throw a piece of chalk on the blackboard. What is the probability for the chalk to hit below the given curve?



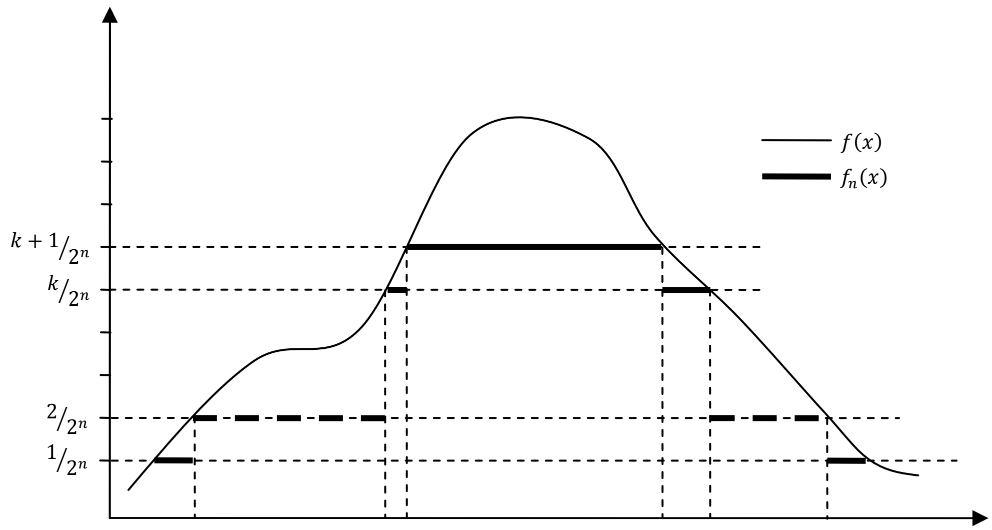
- Sample space: $\Omega = \{all\ the\ points\ on\ the\ blackboard\}$
- Statistical model
  - Assume that all the points on the blackboard have the same probability of being hit. Then, $\forall \omega \in \Omega : P(\{\omega\}) = \epsilon \;\; \Rightarrow \;\; \epsilon = 0$. That is, if we ask about the probability of a single point being hit, the answer is going to be zero. Therefore, we need to resort to a different measure of probability—expressed as the area under a curve.



A step-function

$f(x) = 1/3$

    In the cases of constant and step-functions, the probability of the chalk hitting inside A is $P(A) = \frac{area\ of\ A}{area\ of\ \Omega}$. We can extend this idea to any function $f(x)$.
  - To find the probability of an event in general, given a function $f(x)$.
    - Define the event A as
$$A \coloneqq \{\omega = (x, y) \in \Omega \ : y \le f(x)\}$$
    - Define the events A_n as
$$A_n \coloneqq \{\omega = (x, y) \in \Omega : y \le f_n(x)\}$$
      where
$$f_n(x) \coloneqq \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mathbf{1}_{\left\{\frac{k}{2^n} \le f(x) < \frac{k+1}{2^n}\right\}}, \qquad \forall n \in \mathbb{N}$$
$$\mathbf{1}_{\left\{\frac{k}{2^n} \le f(x) < \frac{k+1}{2^n}\right\}} = \begin{cases} 1 & \frac{k}{2^n} \le f(x) < \frac{k+1}{2^n} \\ 0 & otherwise \end{cases}$$
      - $f_n(x) \le f_{n+1}(x)$ for all x and all n
      - $\lim_{n \to \infty} f_n(x) = f(x)$


$f(x)$
$f_n(x)$
$k+1/_{2^n}$
$k/_{2^n}$
$2/_{2^n}$
$1/_{2^n}$

Algebras and σ-Algebras (Cont'd)

❖ Recap.
Let $A = \{\omega = (x, y) \in \Omega : y \leq f(x)\}$
Let $A_n = \{\omega = (x, y) \in \Omega : y \leq f_n(x)\}$ and $f_n(x) = f_n(x) \coloneqq \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\left\{\frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}\right\}}$
We can show
1) $\forall n \in \mathbb{N}, \forall x \in \mathbb{R} : f_n(x) \leq f_{n+1}(x)$
2) f_n is increasing implies that f_n has a limit. Thus, $\lim_{n\to\infty} f_n \to f$ for all x.

As a result of (1) and (2), $A_n = \{(x, y) : y \leq f_n(x)\}$ is such that $A_n \subset A_{n+1}$, i.e. A_n is increasing. Therefore, $\cup_n A_n = A$, and we can say

$$P(A) = \lim P(A_n)$$

We can define $P(A)$ for any set $A = \{(x, y) : y \leq f(x)\}$ such that

$$\lim P(A_n) = \lim \int_a^b f_n(x)\, dx \equiv \int_a^b f(x)\, dx$$

More generally, whenever, f is Riemann integrable, we can get $P(A)$ by using the step-wise approximation.

- Conclusion: For any sample Ω, the events A for which I can define $P(A)$ are the subsets of Ω including at least
  - Ω itself (by definition, $P(\Omega) = 1$)
  - If $P(A)$ exists, then $P(\bar{A}) = 1 - P(\bar{A})$
  - If A and B are included, then $A \cap B$ is included, and also $A \cup B$
  - If A_n is included in A_{n+1}, and all the A_n's are such that $P(A_n)$ exist, then $\cup_n A_n$ has to be included.

❖ *Definition*. Let Ω be the sample space. An ***algebra*** $\mathcal{A}$ of events of events of Ω is a family of subsets of Ω (i.e. $\mathcal{A} \subset \mathcal{P}(\Omega)$) such that
1) $\Omega \in \mathcal{A}$
2) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
3) $A, B \in \mathcal{A} \Rightarrow (A \cap B) \in \mathcal{A}$

$\mathcal{A}$ is a ***σ-algebra*** (or ***σ-field***) if in addition, we have
4) $A_n \in \mathcal{A}, \forall n \in \mathbb{N} : A_n \subset A_{n+1} \Rightarrow \bigcup_n A_n \in \mathcal{A}$

- Remark. If $\mathcal{A}$ is an algebra, $\forall A_1, \dots, A_N$ finite collection of sets such that $A_i \in \mathcal{A}$, $i = 1, \dots, N$, then $\bigcup_{i=1}^N A_i \in \mathcal{A}$, and $\bigcap_{i=1}^N A_i \in \mathcal{A}$
  - Note. By the De Morgan's Law, $A \cup B = \overline{(\bar{A} \cap \bar{B})}$.
- Remark. If Ω is finite and $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, then $\mathcal{A}$ algebra $\Leftrightarrow \sigma$-algebra
  - Note. If Ω is infinite, then σ-algebra $\Rightarrow \mathcal{A}$ algebra (but not reverse, see e.g. on P.11)

❖ **Theorem**. If Ω is countable and $\mathcal{A}$ is a σ-algebra of Ω such that $\forall \omega \in \mathcal{A} : \{\omega\} \in \mathcal{A}$, then $\mathcal{A} = \mathcal{P}(\Omega)$
- *Proof*. We will demonstrate the equality by showing $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ and $\mathcal{P}(\Omega) \subseteq \mathcal{A}$.
  First, by the definition of a σ-algebra, $\mathcal{A} \subseteq \mathcal{P}(\Omega)$.

Next, to show that $\mathcal{P}(\Omega) \subseteq \mathcal{A}$, we need to show that $\forall B : B \in \mathcal{P}(\Omega) \Rightarrow B \in \mathcal{A}$. Since Ω is countable, we can describe its elements as

$$\Omega = \{\omega_1, \dots, \omega_n, \dots\}$$

Consider the following sets A_n defined for $n \in \mathbb{N}$ as $A_n = \{\omega_1, \dots, \omega_n\}$. It is easy to show that $A_n \in \mathcal{A}$ and $A_n \subseteq A_{n+1}$ (e.g. by recurrence: $A_{n+1} = A_n \cup \{\omega_{n+1}\}$).

Now consider any set $B \in \mathcal{P}(\Omega)$, we can always rewrite B as

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap \bigcup_{n=1}^{\infty} A_n \\ &= B \cap \lim_{n\to\infty} \bigcup_{k=1}^{n} A_k \\ &= B \cap \lim_{n\to\infty} A_n \\ &= \lim_{n\to\infty} (B \cap A_n) \\ &\in \mathcal{A} \end{aligned}$$

Probability Measure

- *Definition.* Let $(\Omega, \mathcal{A})$ be a measurable space, where Ω is the sample space and $\mathcal{A}$ is a σ-algebra.
  - μ is a measure on $(\Omega, \mathcal{A})$ if and only if
$$\begin{aligned} \mu : \mathcal{A} &\to \overline{\mathbb{R}}_+ \\ A &\mapsto \mu(A) \end{aligned}$$
    is such that
    1) $\mu(\emptyset) = 0$
    2) $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

    where (A_i) are pair-wise disjoint sets with $A_i \in \mathcal{A}$ for all $i \in \mathbb{N}$.
  - P is a probability measure if and only if
    - P is a measure
    - $P(\Omega) = 1$

  Therefore,
$$\begin{aligned} P : \mathcal{A} &\to [0,1] \\ A &\mapsto P(A) \end{aligned}$$
  - Remark. If $A, B \in \mathcal{A} : A \cap B = \emptyset$, then $\mu(A \cup B) = \mu(A) + \mu(B)$.
    - Consider countable collection $A, B, \emptyset, \dots, \emptyset$, use (2) and (1)
  - Remark. If $A, B \in \mathcal{A} : A \subseteq B$, then $\mu(A) \leq \mu(B)$
  - Remark. Consider the measure space $(\Omega, \mathcal{A}, \mu)$, and $B \in \mathcal{A} : 0 < \mu(B) < \infty$. The associated probability measure on $(B, \mathcal{A} \cap B)$ is defined as
$$\forall S \in \mathcal{A} \cap B : P(S) = \frac{\mu(S)}{\mu(B)}$$
    - Note. $(\mathcal{A} \cap B) = \{S = (A \cap B) : \forall A \in \mathcal{A}\}$ $(\subseteq \mathcal{P}(B)?)$

- ***Uniform Probability Measure***
  - Ω is finite:
$$P(\{\omega\}) = \frac{1}{\#\Omega}, \qquad \forall \omega \in \Omega$$
  - Ω is infinite (and not countable), e.g. $\Omega = \mathbb{R}^2$. I can define the ***Lebesgue Measure*** on $\mathbb{R}^2$, which is
$$\mu(A) = Area\ of\ A, \qquad \forall A \in \mathcal{A} = \mathcal{P}(\mathbb{R}^2)$$
  - Consider $B \subset \Omega = \mathbb{R}^2$ (think of B as the blackboard), with $0 < \mu(B) < \infty$.
    From the Lebesgue measure (area of $\mathbb{R}^2$) on $\mathbb{R}^2$, I can define uniform probability measure P on $B \cap \mathcal{A} = \{B \cap \mathcal{A} : A \in \mathcal{A}\}$ as
$$P(S) = \frac{\mu(S)}{\mu(B)}, \qquad \forall S \in (B \cap \mathcal{A})$$

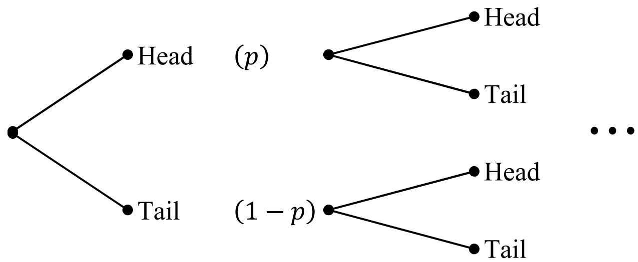
- ***Empirical Probability Measure***
  - Assume we have a statistical experiment with draws $\omega \in \Omega$.
  - After n repetitions of the experiment, we have $(\omega_1, \dots, \omega_n)$. The ***sampling frequency*** of A is given by

$$\forall A \in \mathcal{A} : f_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\omega_i \in A\}}(\omega_i)$$

- E.g. A could be "rolling a dice and get 3" or "rolling a dice and get 3, 4, 5"
- We can show that f_n is a probability measure.

➢ ***Law of Large Numbers*** (***LLN***)

- Remark. We want to understand the connection between $f_n(A)$ and $P(A)$, where the latter is the population ("genuine") probability of event A.
- E.g. Toss a coin



If I repeat my experiment n times

$$f_n(Head) = \frac{\#\ of\ Head\ observations}{n}$$

Sample space:

$$\Omega = \{(\omega_i)_{1\le i\le n} : \omega_i \in \{H, T\}\}$$

under the assumption that

1) Tosses are independent
2) $P(H) = 1/2$

$$P(\{(\omega_i)_{1\le i\le n}\}) = \frac{1}{2^n}, \qquad \forall (\omega_i)_{1\le i\le n} \in \Omega^{(n)}$$

- Strictly speaking, it is possible to only get heads with probability $1/2^n$. In this case, $f_n(H) = 1$, which does not converge to $P(H) = 1/2$. However,

$$P(\{f_n(H) = 1\}) = \frac{1}{2^n} \to 0$$

Here, we need to understand the meaning of $f_n(H)$ converges to $P(H)$ with probability approaching 1.

❖ *Definition.* ***Monotone sequence***

➢ A sequence $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$ is called an ***increasing sequence*** with limit:

$$\lim_n A_n = \bigcup_{n=1}^{\infty} A_n = \lim_n \uparrow A_n$$

➢ A sequence $A_1 \supseteq \cdots \supseteq A_n \supseteq \cdots$ is ***decreasing*** with limit:

$$\lim_n A_n = \bigcap_{n=1}^{\infty} A_n = \lim_n \downarrow A_n$$

➢ A ***monotone class*** is a class that contains the limits of all its increasing and decreasing sequences.

- A σ-algebra is a monotone class (this is true by the last theorem in homework 1).

- A ***class*** is a collection of sets.

- *Theorem*. ***Monotone Continuity of Probability Measure***
  - Consider a probability measure P on $(\Omega, \mathcal{A})$ where $\mathcal{A}$ is a σ-algebra
  - Suppose $(A_n)_n \subseteq \mathcal{A}$ is an increasing (and countable) sequence in $\mathcal{A}$, and $(B_n)_n \subseteq \mathcal{A}$ a decreasing (countable) sequence in $\mathcal{A}$. Then,
    1) $P(\lim \uparrow A_n) = \lim_{n\to\infty} P(A_n)$
    2) $P(\lim \downarrow B_n) = \lim_{n\to\infty} P(B_n)$

Monotone Continuity Theorem (cont'd)

- *Theorem.* Monotone Continuity of Probability Measure.
  - Consider a probability measure P on $(\Omega, \mathcal{A})$ where $\mathcal{A}$ is a σ-algebra
  - Suppose $(A_n)_n \in \mathcal{A}$ is an increasing (and countable) sequence in $\mathcal{A}$, and $(B_n)_n \in \mathcal{A}$ a decreasing (countable) sequence in $\mathcal{A}$. Then,
    1) $P(\lim \uparrow A_n) = \lim_{n\to\infty} P(A_n)$
    2) $P(\lim \downarrow B_n) = \lim_{n\to\infty} P(B_n)$

  Proof. We can define a sequence of disjoint sets (C_n) such that

$$C_{n+1} = A_{n+1} \backslash A_n, \qquad \forall n \in \mathbb{N}$$
$$C_1 = A_1$$

  - Note. $\bigcup_{k=1}^{n} C_k = \bigcup_{k=1}^{n} A_k = A_n$. (We can justify this claim by induction.)

$$\begin{aligned} P(\lim \uparrow A_n) &= P\left(\bigcup_{n=1}^{\infty} A_n\right) \\ &= P\left(\bigcup_{n=1}^{\infty} (A_{n+1}\backslash A_n)\right) \\ &= \sum_{n=1}^{\infty} P(A_{n+1}\backslash A_n) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} P(A_{k+1}\backslash A_k) \\ &= \lim_{n\to\infty} P\left(\bigcup_{k=1}^{n} (A_{k-1}\backslash A_k)\right) \\ &= \lim_{n\to\infty} P(A_n) \end{aligned}$$

■

- *Definition.* ***Limit Superior*** and ***Limit Inferior***:

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m\ge n} A_m$$

  - A_n occurs infinitely many times.

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m\ge n} A_m$$

  - A_n occurs eventually.
- $(\Omega, \mathcal{A})$ a σ-algebra, $(A_n) \in \mathcal{A}$, $\omega \in \Omega$. What does it mean to say $\omega \in (\limsup A_n)$??

$$\omega \in \left(\limsup_{n\to\infty} A_n\right) \iff \forall n \in \mathbb{N}, \exists m \ge n : \omega \in A_m$$

- Similarly,

$$\omega \in \left(\liminf_{n\to\infty} A_n\right) \iff \exists n \in \mathbb{N}, \forall m \ge n : \omega \in A_m$$

- Reference: http://en.wikipedia.org/wiki/Limit_superior_and_limit_inferior#Special_case:_discrete_metric

- $\{f_n(A) \to P(A)\}$ is an event because it is in the sample space $\Omega^{(\infty)}$ with

$$\Omega^{(\infty)} = \{(\omega_i)_{i\in\mathbb{N}} : \omega_i \in \{H,T\}\ \forall i \in \mathbb{N}\}$$

  - $\{f_n(A) \to P(A)\}$ is the set containing all the f_n's that converge to $P(A)$, i.e. $1/2$.

$$\begin{aligned}
\bar{\omega} \in \{f_n(A) \to P(A)\} &\Leftrightarrow f_n(A)_{(\bar{\omega})} \xrightarrow{n\to\infty} P(A) \\
&\Leftrightarrow \forall \epsilon > 0, \exists q \in \mathbb{N}, \forall n \geq q : \left|f_n(A)_{(\bar{\omega})} - P(A)\right| \leq \epsilon \\
&\Rightarrow \forall k \in \mathbb{N}, \exists q \in \mathbb{N}, \forall n \geq q : \left|f_n(A)_{(\bar{\omega})} - P(A)\right| \leq \frac{1}{k} \\
&\Leftrightarrow \bar{\omega} \in \bigcap_{k=1}^{\infty}\bigcup_{q=1}^{\infty}\bigcap_{n\geq q}\left\{\left|f_n(A)_{(\bar{\omega})} - P(A)\right| \leq \frac{1}{k}\right\}
\end{aligned}$$

  - Note. Union corresponds to existential quantifier, and intersection to universal quantifier:

$$\cup \leftrightarrow \exists, \quad \text{and,} \quad \cap \leftrightarrow \forall.$$

$$\begin{aligned}
P\big(f_n(A) \to P(A)\big) &= \lim_{k\to\infty}\downarrow \lim_{q\to\infty}\uparrow P\left(\bigcap_{n\geq q}\left\{|f_n(A) - P(A)| \leq \frac{1}{k}\right\}\right) \\
&= \lim_{k\to\infty}\downarrow \lim_{q\to\infty}\uparrow P\left(\sup_{n\geq q}|f_n(A) - P(A)| \leq \frac{1}{k}\right)
\end{aligned}$$

  - Recall:

$$f_n(A)_{(\omega)} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{\{\omega_k\in A\}}$$

with $\omega_i \in \{H,T\}$, $A =$ getting H

$$\Omega^{(n)} = \left\{(\omega_i)_{1\leq i\leq n} : \omega_i \in \{H < T\}\right\}$$

- *Definition.* ***Almost Surely Convergence*** and ***Probability Convergence***

$$f_n(A) \xrightarrow{a.s.} P(A) \Leftrightarrow \forall \epsilon > 0 : P\left(\sup_{q\geq n}\left|f_q(A) - P(A)\right| > \epsilon\right) \xrightarrow[n\to\infty]{} 0$$

$$f_n(A) \xrightarrow{p} P(A) \Leftrightarrow \forall \epsilon > 0 : P(|f_n(A) - P(A)| > \epsilon) \xrightarrow[n\to\infty]{} 0$$

<u>Convergence</u>

- Recall the two definitions
  - Convergence in probability

$$f_n(A) \xrightarrow{p} P(A) \;\Leftrightarrow\; \forall \epsilon > 0 : P(\{|f_n(A) - P(A)| > \epsilon\}) \xrightarrow[n\to\infty]{} 0$$
$$\Leftrightarrow\; \forall \epsilon > 0 : P(\{|f_n(A) - P(A)| \le \epsilon\}) \xrightarrow[n\to\infty]{} 1$$

  - Convergence almost sure

$$f_n(A) \xrightarrow{a.s.} P(A) \;\Leftrightarrow\; \forall \epsilon > 0 : P\left(\left\{\sup_{q\ge n}\left|f_q(A) - P(A)\right| > \epsilon\right\}\right) \xrightarrow[n\to\infty]{} 0$$
$$\Leftrightarrow\; \forall \epsilon > 0 : P\left(\left\{\sup_{q\ge n}\left|f_q(A) - P(A)\right| \le \epsilon\right\}\right) \xrightarrow[n\to\infty]{} 1$$

  - Note. $f_n(A) \xrightarrow{a.s.} P(A) \;\Rightarrow\; f_n(A) \xrightarrow{p} P(A)$
    - In general, however, the reverse does not hold. Consider

$$P\left(\left\{\sup_{q\ge n}(\left|f_q(A) - P(A)\right| > \epsilon\right\}\right) = P\left(\bigcup_{q\ge n}\{\left|f_q(A) - P(A)\right| > \epsilon\}\right)$$
$$\le \sum_{q\ge n} P\left(\{\left|f_q(A) - P(A)\right| > \epsilon\}\right)$$

    - Convergence almost sure is the <u>Strong</u> LLN
      - This is the stochastic analog of "pointwise convergence".
    - Convergence in probability is the <u>Weak</u> LLN
      - Continuous Mapping Theorem. For every continuous function g, if $x_n \xrightarrow{p} x$, then $g(x_n) \xrightarrow{p} g(x)$.

Quality Control and Sampling with / without Replacement

- Sampling with / without Replacement
  - Population of N individuals
  - Draw n individuals among N

- 1st experiment (with replacement):
  - Draw 1 individual from the population (this is #1). Put it back
  - Draw 1 individual from the population (this is #2). Put it back
  - …

  - Sample space

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in population\}$$

    where $\#\Omega = N^n$
    - Assuming independent draws, then

$$P(\{\omega\}) = \frac{1}{N^n}$$

- 2nd experiment (without replacement):
  - Draw 1 individual from the population
    - This is #1
  - Draw 1 individual from the remaining population
    - This is #2
  - …

  - Sample space:

$$\Omega^* = \left\{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in population \ \wedge \ \omega_i \neq \omega_j \ for\ i \neq j\right\}$$

    where $\Omega^* \subset \Omega$ with

$$\#\Omega^* = N(N-1)(N-2)\cdots$$

- The two experiments / models are compatible.

$$P \text{ defined on } \Omega \quad \Rightarrow \quad P \text{ defined on } \left(\Omega^*, \sigma(\Omega^*)\right)$$

  We can move from the 1st experiment to the 2nd experiment by precluding repetition
  - Probability of having no repetition
    $=$ probability of Ω^* within $(\Omega, \sigma(\Omega), P)$
    $$= \frac{\#\Omega^*}{\#\Omega} = \frac{(N)_n}{N^n}$$

  - In both experiments, order matters.

  - Intuition 1. When n is sufficiently smaller than N.
    - Then the probability of no repetition is really large ~ almost 1

$$(N)_n = N(N-1)\dots\underbrace{(N-n+1)}_{\sim N} \sim N^n$$

    - Application: for survey polls, N is usually quite large compared to n; so we can do calculations with repetitions.

- Intuition 2. When n is sufficiently close to N (extreme case: $n = N$)

N	1	2	3	4	5	6	7
Probability of no repetition ($(N)_N/N^N$)	1	0.5	0.222	0.094	0.038	0.015	0.006 (99.4% chance of having a repetition!)

  - Note. $(N)_N = N!$
    - The number of ways to choose an ordered sequence (without repetition) of all individuals.
      = number of permutations of the set of individuals
    - More generally

$$(N)_n = \frac{N!}{(N-n)!}$$

      - $(N)_n$ is the number of ordered (or arranged) samples of size n without repetitions in a population of size N.
      - Of course, several of these ordered samples share the exact same individuals but ordered in a different way (there are $n!$ ways of permuting n individuals)
      - The number of subsets of n individual in a population of size N without repetition is given by the ***Binomial coefficient***:

$$\frac{(N)_n}{n!} = \frac{N!}{(N-n)!\,n!}$$

      - Permutation (order matters)

$$(N)_n = \frac{N!}{(N-n)!}$$

      - Combination (order does not matter)

$$\binom{N}{n} = \frac{N!}{(N-n)!\,n!}$$

- Quality Control without Replacement in Sampling
  - N light bulbs with R deficient ones
    - Note. R is not random.
  - Minimum quality standard: No more than K among N are allowed to be deficient.
    - But it's too expensive to check the N light bulbs. So select randomly n light bulbs among N, observe k deficient ones.
  - Question: Given N, K, n, k, what values are likely for R? (want R to be smaller than K)
    - We want to assess: $P(\text{observed } k)$ and realize it depends on R (if R is large, then the probability of observing a large k is high, and vice versa).
    - Conversely: we have observed k. It makes more likely the value of R for which

$$P(\text{observed } k) = \text{large}$$

      - Given R, $f : k \to P_R(observed\ k)$
        - Note. ***Probability function*** indexed by R.

- Given k, $g : R \to P_R(observed\ k)$
  - Note. This is the ***likelihood function***.

- Sample Space
  - 1st choice: $\Omega = \{0,1,2,\dots,n\}$
    - However, this sample space is not convenient! Because the probabilities of the elementary events are not equal, i.e. probability distribution is not uniform.
  - 2nd choice: $\Omega =$ the $(N)_n$ ordered samples that can be drawn without replacement.
    - For this sample space, we can define a uniform probability distribution.

$$P_R\left(\underbrace{k\text{ deficient bulbs}}_{\text{elementary event }A}\right) = \frac{\#A}{\#\Omega}$$

$$= \underbrace{\frac{(R)_k(N-R)_{n-k}}{(N)_n}}_{\substack{\text{probability of getting } k \text{ deficient bulbs} \\ \text{from a sample of size } n \text{ in an ordered way}}} \times \underbrace{\binom{n}{k}}_{\substack{\text{number of ways to order} \\ \text{the } k \text{ defective bulbs}}}$$

$$= \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}$$

  - 3rd choice: $\Omega =$ the $\binom{N}{n}$ subsets
    - Uniform probability is induced from uniform probability with ordered samples.

$$P(k\text{ deficient bulbs}) = \frac{\#A}{\#\Omega} = c$$

    - This is the most appropriate sample space for the question.

- Remark. We end up with a probability that is not uniform on $\{1,2,\dots,n\}$

$$P(\{k\}) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}, \qquad if \begin{cases} n-(N-R) \le k \\ k \le R \end{cases}$$

  - This characterizes any event $A \subset \mathcal{P}(\Omega)$
  - This is the ***hypergeometric distribution***, $H(N,R,n)$

Quality Control (cont'd)

- Recap
  - Population: N
  - Defective light bulbs: R
  - Quality control: $R \leq K$
    - Sample n with defect k

  - Case 1. Draw without replacement
    - $\Omega = (N)_n$ arranged samples

$$P(\{k\}) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}} \quad \rightarrow \quad \Omega^* = \binom{N}{n}$$

  Each sample in Ω^* corresponds to exactly $n!$ arranged sample in Ω.
    - $P(\{k\})$ means the probability of getting k defective bulbs in a set of n bulbs.

  - Case 2. Draw with replacement
    - $\Omega = N^n$ samples that are arranged.

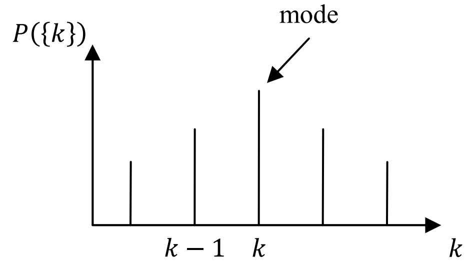
$$\begin{aligned} P(\{k\}) &= \frac{R^k (N-R)^{n-k} \binom{n}{k}}{N^n} \\ &= \frac{R^k (N-R)^{n-k}}{N^k N^{n-k}} \binom{n}{k} \\ &= \left(\frac{R}{N}\right)^k \left(1 - \frac{R}{N}\right)^{n-k} \binom{n}{k} \\ &= p^k (1-p)^{n-k} \binom{n}{k} \end{aligned}$$

      - $p = R/N$ is the population probability of picking a defective light bulb.
    - I have defined the ***Binomial probability distribution*** $B(n, p)$.
    - Remark. If I consider Ω^* sample where the arrangement does not matter (with replacement).
      There is no way we can define a uniform probability from Ω to Ω^*.
      - E.g. $n = 2$, and $\{\omega_1, \omega_2, \dots, \omega_N\}$

$$\begin{array}{ccc} \Omega & & \Omega^* \\ (\omega_1, \omega_1) & \leftrightarrow & (\omega_1, \omega_1) \\ (\omega_1, \omega_2) & \rightarrow & (\omega_1, \omega_2) \\ (\omega_2, \omega_1) & \nearrow & (\omega_2, \omega_2) \\ \vdots & & \vdots \end{array}$$

    - $P(\{k\}) = p^k (1-p)^{n-k} \binom{n}{k}$ gives the probability of success (picking a defective bulb) in one draw.
      - $p^k (1-p)^{n-k}$ is the probability of picking a sequence with k sucessses, exactly.
      - There are $\binom{n}{k}$ such sequences

- Suppose we know p and n. What is the most probable value for k? In other words, what is the mode?



- For $k \geq 1$

$$\frac{P(\{k\})}{P(\{k-1\})} = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k-1} p^{k-1} (1-p)^{n-k+1}} = \frac{n-k+1}{k} \cdot \frac{p}{1-p}$$

$$\frac{P(\{k\})}{P(\{k-1\})} > 1 \quad \Leftrightarrow \quad \frac{n-k+1}{k} \cdot \frac{p}{1-p} > 1 \quad \Leftrightarrow \quad (n+1)p > k$$

- 2 Cases:
  - If $(n+1)p$ is integer, then I have <u>2 modes</u>: $(n+1)p$ and $(n+1)p-1$
    - $(n+1)p \in \mathbb{Z} \ \Rightarrow \ \exists k : k = (n+1)p$
  - If $(n+1)p$ is not an integer, then I have <u>1 mode</u>: largest integer below $(n+1)p$

❖ Back to quality control problem (without replacement)
- If I know k (but not R), then the ***Maximum Likelihood Estimator*** of R is

$$MLE(R) = \arg\max_R P_R(\{k\})$$

- One way is

$$\frac{P_{R+1}(\{k\})}{P_R(\{k\})} \geq 1 \quad \Leftrightarrow \quad \frac{R+1}{R+1-k} \geq \frac{N-R}{N-R-(n-k)}$$

Quality Control (cont'd)

- We are not interested in estimating R, but testing whether $\underbrace{R}_{\text{\# deficient ones}} \leq \underbrace{K}_{\text{quality standard}}$
- Defining a test is equivalent to defining a critical region that tells me when I should reject

$$H_0 : R \leq K \quad \rightarrow \quad \boxed{MLE(R) \sim \frac{kN}{n}}$$

  - This is true because we assume that

$$\frac{R}{N} \sim \frac{k}{n} \Rightarrow R \sim \frac{kN}{n}$$

  - Critical region:

$$W = \{k \in \mathbb{N} : k \geq r\}$$

    - W is the "rejection zone," because we want $\frac{kN}{n} \leq K$. r is the critical value.
    - Have to pick r "much larger" than $\frac{Kn}{N}$; that is,

$$MLE \gg K \iff \frac{kN}{n} \gg K \iff k \gg \frac{Kn}{N}$$

- 2 situations and 2 errors associated with the decision I take after running the experiment:

Result / Truth	Reject H_0	Not Reject H_0
H_0 true $R \leq K$	**Type I Error**	
H_1 true $R > K$		**Type II Error**

- Neyman's approach:

$$\min[Type\ II\ Error], \qquad \text{subject to} \qquad Type\ I\ Error \leq \underbrace{\alpha}_{confidence\ level}$$

  - Pick α (e.g. 1% or 5%)
    - α is the probability of making Type I Error.
  - For each α, find r_α
  - Define W
  - Given k (result of your experiment), decide whether or not to reject H_0

- Quality Control when Sampling with Replacement
  - $H_0 : p \leq \frac{K}{N}$, where $p(= R/N)$ is the true probability of having deficient bulbs
  - MLE of p:

$$\arg\max_p \underbrace{\left[\binom{n}{k} p^k (1-p)^{n-k}\right]}_{\text{likelihood function of } p}$$

It is often useful to take the log-transformation of the likelihood function:

$$\arg\max_p(L(p)) = \arg\max_p\left(\ln\left\{\binom{n}{k} p^k (1-p)^{n-k}\right\}\right)$$

Then, we can differentiate w.r.t. p, set FOC equal zero, and solve for p.

$$\hat{p} = \frac{k}{n}$$

Define critical region:

$$W = \left\{ k : \hat{p} \geq \underbrace{p(\alpha)}_{\text{critical value associated with } \alpha} \right\}$$

- Extension: Multinomial distribution
  - We have K different colors, call them $a_1, \dots, a_K$, each with p_k probability of being picked.
  - Draw a sample of n with replacement and independence
    - $\Omega = \{(\omega_1, \dots, \omega_n) : \forall i, \omega_i \in \{a_1, \dots, a_K\}\}$
      - Here we care about the order of ω_i's
    - $P(\{\omega_1, \dots, \omega_n\}) = p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K}$ with

$$n_k = \sum_{i=1}^{n} \mathbf{1}_{\{\omega_i \in a_k\}}$$

  - The probability of observing (un-ordered)

$$P\begin{pmatrix} n_1 a_1 \\ n_2 a_2 \\ \vdots \\ n_K a_K \end{pmatrix} = \left(p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K}\right) T$$

with $\sum_{k=1}^{n} n_k = n$.

    - T is the number of configurations:
      - Choose $\binom{n}{n_1}$ pick n_1 spots among the n available
      - Choose $\binom{n - n_1}{n_2}$ pick n_2 spots among the $(n - n_1)$ left
      - $\vdots$

    Then,

$$T = \binom{n}{n_1}\binom{n - n_1}{n_2}\binom{n - n_1 - n_2}{n_3} \cdots \begin{pmatrix} n - \sum_{k=1}^{K-1} n_k \\ n_k \end{pmatrix} = \frac{n!}{\prod_{k=1}^{K} n_k!}$$

  - Result:

    For $p_1, \dots, p_K$ such that $p_k \in [0,1]$ and $\sum_{k=1}^{K} p_k = 1$,

    The ***multinomial distribution*** $M(n; p_1, \dots, p_K)$:

$$P\begin{pmatrix} n_1 a_1 \\ n_1 a_2 \\ \vdots \\ n_K a_K \end{pmatrix} = \begin{cases} \dfrac{n!}{\prod_{k=1}^{K} n_k!} \cdot \displaystyle\prod_{k=1}^{K} p_k^{n_k} & \text{if } \displaystyle\sum_{k=1}^{K} n_k = n \quad \text{with } n_k \in [0, n] \\ 0 & \text{otherwise} \end{cases}$$

    - This is a probability distribution on $\{0,1, \dots, n\}^K$
    - Multinomial when $K = 2$, $M(n; p_1, p_2)$ and $p_2 = 1 - p_1$. The distribution is about

$$(n_1, n_2) \in \{0,1, \dots, n\}^2 = \Omega$$

    - The binomial $B(n, p_1)$ is a distribution about

$$n_1 \in \{0,1, \dots, n\} = \Omega$$

Counting Process

- Events that occur over time (e.g. event could be a customer entering a store)
  - A ***counting process*** is a stochastic process $\{N(t) : t \geq 0\}$ such that
    - It is non-negative, i.e. $N(t) \geq 0$
    - $N(t)$ is an integer
    - Non-decreasing, i.e. $s \leq t \;\; \Rightarrow \;\; N(s) \leq N(t)$
- Independence of 2 events occurring in 2 different (disjoint) time intervals
- ***Poisson Process***. For an interval of size $\epsilon > 0$,

$$\frac{P(1 \text{ event})}{\epsilon} \xrightarrow{\epsilon \to 0} \lambda$$

  Here λ is the ***intensity of arrival***.

$$\frac{P(\text{more than 1 event})}{\epsilon} \xrightarrow{\epsilon \to 0} 0$$

  In the same sense, we are interested in events that do not happen too often.

- *Question*: What is the probability of observing k events in the time interval $[0, t]$?
  - We divide $[0, t]$ into n sub-intervals of length $\Delta t = t/n$.
    - Consider intervals $[0, \Delta t)$, $[\Delta t, 2\Delta t)$, …, and treat them as n consecutive experiments
    - For each interval,

$$\begin{aligned} p_n &\to \text{observe exactly 1 event} \\ p_n' &\to \text{observe more than 1 event} \\ 1 - p_n - p_n' &\to \text{observe 0 event} \end{aligned}$$

      I have

$$\frac{p_n}{t/n} \xrightarrow{n \to \infty} \lambda \quad \Leftrightarrow \quad np_n \xrightarrow{n \to \infty} \lambda t$$

      and

$$\frac{p_n'}{t/n} \xrightarrow{n \to \infty} 0 \quad \Leftrightarrow \quad np_n' \xrightarrow{n \to \infty} 0$$

    - Recall the multinomial formula:

$$P\begin{pmatrix} k \text{ interval with exactly 1 event} \\ k' \text{ interval with more than 1 event} \\ n - k - k' \text{ interval with 0 event} \end{pmatrix} = \frac{n!}{k!\,k'!\,(n-k-k')!} p_n^k (p_n')^{k'} (1 - p_n - p_n')^{n-k-k'}$$

    - What happens when $n \to \infty$ while k, k' remains fixed / finite?

$$\frac{n!}{(n-k-k')!} = \underbrace{n(n-1)(n-2)\cdots(n-k-k'+1)}_{k+k' \text{ terms}} \sim n^{k+k'}$$

$$\frac{n!}{k!\,k'!\,(n-k-k')!}p_n^k(p_n')^{k'}(1-p_n-p_n')^{n-k-k'}$$
$$\sim \frac{1}{k!\,k'!}\underbrace{(np_n)^k}_{\sim(\lambda t)^k}\underbrace{(np_n')^{k'}}_{\substack{\text{very small unless } k'=0 \\ \text{this is of order } o(1)}}(1-p_n-p_n')^{n-k-k'}$$

$$(1-p_n-p_n')^{n-k-k'} = \exp\left\{(n-k-k')\log\left(1-\underbrace{p_n}_{\frac{\lambda t}{n}}-\underbrace{p_n'}_{o\left(\frac{1}{n}\right)}\right)\right\}$$

For small x, $\log(1+x) \sim x$. So the probability is constant as $n \to \infty$.
Then,

$$P\binom{k \text{ interval with 1 event exactly}}{n-k \text{ interval with 0 event}} \sim \frac{1}{k!}(\lambda t)^k\underbrace{(1-p_n)^{n-k}}_{\sim\exp(-\lambda t)}$$
$$\sim \frac{1}{k!}(\lambda t)^k e^{-\lambda t}$$

This approximately works for k finite and n large enough.

- Therefore, we have defined the ***Poisson distribution*** with parameter over $\mathbb{N} \cup \{0\}$:

$$P(k \text{ events}) = \frac{1}{k!}(\lambda t)^k e^{-\lambda t}$$

  - Check that $\sum_{k=0}^{\infty} P(k \text{ events}) = 1$. this is true from the fact that the summation is a Taylor expansion for $e^{-\lambda t}$.
  - The only important parameter of the Poisson distribution is $\mu = \lambda t$.

Real Random Variable

- **The Lebesgue Measure**
  - Example: I draw randomly a number x between 0 and 1. What is the probability that $P(x < 0.47)$?
    - Choice for the first decimal number:
      $$\{0,1,2,3\} \rightarrow \text{choice for the second decimal is } \{0,\dots,9\}$$
      $$\{4\} \rightarrow \text{choice for the second decimal is } \{0,\dots,6\}$$
      Then, a total of 47 choices out of 100, namely
      $$P(x < 0.47) = \frac{47}{100} = 0.47$$
    - In the book, they calculate $P(x \leq 0.47) \rightarrow \neq^{+}$ proof $P(x = 0.47) = 0$.
      Define an infinite sequence of decimal digits and element must coincide with the decimal digits of x
      $$P(x = 0.47) = \lim \downarrow P(A_n) = \lim \downarrow \frac{1}{10^n} = 0$$

- *Definition*. The ***Borel sets*** $\mathcal{B}$ of $\mathbb{R}$ are the smallest σ-algebra of $\mathbb{R}$ containing all the open intervals in $\mathbb{R}$.
  - Any interval is a Borel set (but not every Borel set is an interval), and the set of all Borel sets is a σ-algebra.
  - (all possible) Borel sets = Borel ring = Borel field = Borel σ-algebra
  - **Theorem.**
    $$\mathcal{B} = \underbrace{\sigma\big((-\infty, x] : x \in \mathbb{R}\big)}_{\substack{\text{the smallest } \sigma\text{-algebra containing} \\ \text{all the semi-open intervals } (-\infty, x]}}$$

- Intervals $\rightarrow$ algebra $\mathcal{F}$ spanned by intervals
  $$\forall A \in \mathcal{F} : A = \bigcup_{i=1}^{n} F_i, \qquad \forall i \neq j : F_i \cap f_j = \emptyset$$
  $$P(A) = \sum_{i=1}^{n} P(F_i)$$

- *Definition*. ***Outer Measure***. Suppose
  - $\mathcal{F}$ is an algebra on Ω
  - P is σ-additive (i.e. countably additive) on $\mathcal{F}$ with $P(\Omega) = 1$

  Then, the ***outer measure*** of any $A \in \Omega$ is
  $$P^*(A) = \inf_{(A_i)_i \in \mathcal{F} \text{ such that } A \subseteq \left(\bigcup_{i=1}^{\infty} A_i\right)} \sum_{i=1}^{\infty} P(A_i)$$

- For any set $A \in \mathcal{F}$, we can show that $P^*(A) = P(A)$.
  - First, we show $P^*(A) \leq P(A)$
    Since $A \in \mathcal{F}$, we can define $A_1 = A$ and $A_i = \emptyset$ for all $i \geq 2$. Then,

$$(A_i)_{i\in\mathbb{N}} \in \mathcal{F} \;\Rightarrow\; A \subseteq \left(\bigcup_{i\in\mathbb{N}} A_i\right) = A \;\Rightarrow\; \sum_{i\in\mathbb{N}} P(A_i) = P(A)$$

Therefore, $P^*(A)$ is actually the inf over all possible sequences.

$$P^*(A) = \inf \sum P(F_i) \leq P(A)$$

- Second, we show that $P(A) \leq P^*(A)$.
  We know (by assumption) that $A \subseteq \left(\bigcup_{j\in\mathbb{N}} B_j\right)$. Define
  $$C_n = \bigcup_{j=1}^{n} B_j$$
  Clearly, C_n is increasing, and $(A \cap C_n)$ is also increasing to A.

Recap

- $\mathcal{F}$ is an algebra on Ω
- P is σ-additive on $\mathcal{F}$ with $P(\Omega) = 1$
- Outer measure of $A \subset \Omega$

$$P^*(A) = \inf_{(A_i)_i \in \mathcal{F} \text{ such that } A \subseteq \left(\bigcup_{i=1}^{\infty} A_i\right)} \left[\sum_{j=1}^{\infty} P(A_j)\right]$$

- We have shown that $P^*(A) \leq P(A)$

- Continue to prove that $P(A) \leq P^*(A)$
  For any $A_j \in \mathcal{F}$ such that $A \subseteq \left(\bigcup_{j=1}^{\infty} A_j\right)$, define

$$B_n = \bigcup_{j=1}^{n} A_j$$

  Note that $(B_n)_n$ is an increasing sequence, and $(A \cap B_n)_n$ is increasing towards A.
  We then have

$$P(A) = \lim \uparrow \underbrace{P(A \cap B_n)}_{\leq P(B_n) \leq \sum_{j=1}^{n} P(A_j)}$$

  At the limit,

$$P(A) \leq \sum_{j=1}^{\infty} P(A_j)$$

  This inequality is true for any sequence $(A_j) \in \mathcal{F}$ with $A \subseteq \left(\bigcup_{j=1}^{\infty} A_j\right)$. Therefore, we can conclude that the inequality remains over the infimum

$$P(A) \leq \inf_{(A_i)_i \in \mathcal{F} \text{ such that } A \subseteq \left(\bigcup_{i=1}^{\infty} A_i\right)} \sum_{j=1}^{\infty} P(A_j) \quad \Leftrightarrow \quad P(A) \leq P^*(A)$$

- **Theorem** (admitted). P^* is the unique probability measure on $(\Omega, \sigma(\mathcal{F}))$ such that

$$\forall A \in \mathcal{F} : P(A) = P^*(A)$$

  - **Remark.** P^* is defined for any $A \subset \Omega$, but we cannot say that P^* is a probability measure on $(\Omega, \mathcal{P}(\Omega))$.
    - This can be proved for the uniform probability measure on $[a, b]$

- The ***Lebesgue measure*** λ on $(\mathbb{R}, \mathcal{B})$ is defined such that

$$\forall A \in \mathcal{B} : \lambda(A) = \lim_{n \to \infty} \{2nP_n(A \cap [-n, n])\}$$

  where P_n is the uniform probability measure on $[-n, n]$

$$P_n(A \cap [-n, n]) = \frac{\text{length of } (A \cap [-n, n])}{\text{length of } [-n, n]} = \frac{\text{length of } (A \cap [-n, n])}{2n}$$

  - λ is a positive measure on $(\mathbb{R}, \mathcal{B})$ with convention

$$x + \infty = \infty, \qquad \forall x \in \mathbb{R}$$

  - Warning: $\lambda(A \setminus B) = \lambda(A) - \lambda(B)$ only if $\lambda(B) < \infty$

- Similarly, $\lambda(\lim \downarrow A_n) = \lim \downarrow \lambda(A_n)$ only if $\exists n^* : \lambda(A_n) < \infty$ for any $n \geq n^*$.
  - Counter-example: $A_k = [k, \infty)$ where $(A_k)_k \downarrow \emptyset$
  $$\lambda(A_k) = \lim_{n\to\infty}\{2nP_n(A_k \cap [-n, n])\} = +\infty$$
  However, $\lim_{k\to\infty} A_k = \emptyset$. This is not equal to $\lambda(\lim \downarrow A_k) = 0$. The disagreement results from the fact that we cannot find an n^* such that $\lambda(A_k) < \infty$ for $n \geq n^*$.

<u>Multivariate extension</u>

$$\mathcal{B}^d = \sigma\left(\prod_{j=1}^{d}(-\infty, x_j]\right)$$

is the smallest σ-field containing all $\prod_{j=1}^{d}(a_j, b_j)$

- ***Lebesgue measure on $(\mathbb{R}^d, \boldsymbol{\mathcal{B}}^d)$***

$$\lambda_d(A) = \lim_{n\to\infty}(2n)^d P_n(A \cap [-n, n]^d), \qquad \forall A \subset \mathbb{R}^d$$

Random Variable and Random Vectors (r.v.)

- (informally) A random variable is a function of the outcome of a statistical experiment.
  - Example.
    - Ω = sample space of sequences of Bernoulli trials $(\omega_1, \dots, \omega_n)$ with $\omega_i \in \{0,1\}$.
    - Ω is endowed with a probability measure:
$$\forall \omega \in \Omega : P(\{\omega\}) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}$$
So the probability space is $(\Omega, \mathcal{P}(\Omega), P)$.
      - We don't need the binomial coefficient here because we're only considering one observation.
    - The random variable X is defined as
$$X : \Omega \to \{0,1,\dots,n\}$$
$$\omega \mapsto X(\omega) = \sum_{i=1}^n \omega_i$$
The associated probability is
$$P(\{X=k\}) = P\big(X^{-1}(\{k\})\big) = P^X(k) = p^k(1-p)\binom{n}{k}$$
where $X^{-1}(A) = \{\omega \in \Omega : X(\omega) = k\}$ with $A \subset \Omega$ (i.e. $A \in \mathcal{P}(\Omega)$).
      - The probability measure P induces another probability measure P^X on $\{0,1,\dots,n\}$ defined by
$$\underbrace{P^X(k)}_{\substack{induced \\ probability \\ defined\ on\ X(\Omega)}} = \underbrace{P\big(X^{-1}(\{k\})\big)}_{\substack{initial\ probability \\ measure\ defined \\ on\ \Omega}}$$
      - Remark. We say that $X \sim \mathcal{B}(n,p)$. P^X is the probability distribution (or law) of r.v. X

- More general case. Consider a probability space $(\Omega, \mathcal{A}, P)$.
  - Define
$$X : \Omega \to \mathbb{R}^d$$
with $X(\Omega)$ is not only countable part of $\mathbb{R}^d$
    - $X(\Omega)$ is the range (i.e. the minimum codomain). If Ω is countable, then the range of $X(\cdot)$ should also be countable.
    - $P(\{X = x\})$ should not be sufficient to characterize P^X
      - This is true because singletons have probability zero if X is in a continuum.
      - Example. Suppose $X \sim U_{[a,b]}$. Then, $P^X(\{x\}) = 0$. So we cannot characterize P^X.
    - Hopefully, we can use intervals.
$$P^X\big((-\infty, x]\big) = \begin{cases} 1 & \text{if } x > b \\ \dfrac{x-a}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{if } x < a \end{cases} \qquad \forall x \in \mathbb{R}$$
    - We need to know that $P\left(X^{-1}\big((-\infty,x]\big)\right)$ makes sense, because
$$P\left(X^{-1}\big((-\infty,x]\big)\right) = P^X\big((-\infty,x]\big)$$

That is, I need to know that $X^{-1}\big((-\infty, x]\big) \in \mathcal{A}$ for all x.

- *Definition*. $(\Omega, \mathcal{A})$ measurable space
  - $X : \Omega \to \mathbb{R}$ is ***$\mathcal{A}$-measurable*** if
$$\forall x \in \mathbb{R} : X^{-1}\big((-\infty, x]\big) \in \mathcal{A}$$
  - $X : \Omega \to \mathbb{R}^d$ is ***$\mathcal{A}$-measurable*** if
$$\forall x \in \mathbb{R}^d : X^{-1}\left(\prod_{j=1}^{d}(-\infty, x_j]\right) \in \mathcal{A}$$
  - The pre-image of Borel sets should belong to the σ-algebra.
- *Definition*. If $(\Omega, \mathcal{A}, P)$ is a probability space, any function $X : \Omega \to \mathbb{R}$ which is $\mathcal{A}$-measurable is called ***random variable***.

- **Theorem.** Suppose
$$X : \Omega \to \mathbb{R}^d, \qquad with \ \ \Omega \in \mathcal{A} \ \ and \ \ \mathbb{R}^d \in \mathcal{B}^d$$
X is $\mathcal{A}$-measurable if and only if $\forall A \in \mathcal{B}^d : X^{-1}(A) \in \mathcal{A}$.
  - *Proof*. If $\forall A \in \mathcal{B}^d : X^{-1}(A) \in \mathcal{A}$ is true. Then, it must be true, in particular, that
$$\forall x \in \mathbb{R}^d : X^{-1}\left(\prod_{j=1}^{d}(-\infty, x_j]\right) \in \mathcal{A}.$$
Then, by definition X is $\mathcal{A}$-measurable.

    Suppose X is $\mathcal{A}$-measurable. That is,
$$\forall x \in \mathbb{R}^d : X^{-1}\left(\prod_{j=1}^{d}(-\infty, x_j]\right) \in \mathcal{A}.$$
Need to show that
$$\forall A \in \mathcal{B}^d : X^{-1}(A) \in \mathcal{A}.$$
Recall that $\mathcal{B}^d = \sigma\big(\prod_{j=1}^{d}(-\infty, x_j] : x \in \mathbb{R}^d\big) = \sigma(\mathcal{C})$. We have to show that
$$\forall A \in \mathcal{C} : X^{-1}(A) \in \mathcal{A} \ \ \Rightarrow \ \ \forall A \in \mathcal{B}^d = \sigma(\mathcal{C}) : X^{-1}(A) \in \mathcal{A}$$
Or we need to show that
$$X^{-1}(\mathcal{C}) \subset \mathcal{A} \ \ \Rightarrow \ \ X^{-1}\big(\sigma(\mathcal{C})\big) \subset \mathcal{A}.$$
    - <u>Comments</u>. We know that
$$X^{-1}(\mathcal{C}) \subset \mathcal{A} \ \ \Rightarrow \ \ \sigma\big(X^{-1}(\mathcal{C})\big) \subset \mathcal{A}.$$
But what is not clear is that
$$X^{-1}\big(\sigma(\mathcal{C})\big) \subset \sigma\big(X^{-1}(\mathcal{C})\big).$$
Note that the converse is clear, since
$$\sigma\big(X^{-1}(\mathcal{C})\big) \subset X^{-1}\big(\sigma(\mathcal{C})\big)$$
because
$$X^{-1}(\mathcal{C}) \subset \underbrace{X^{-1}\big(\sigma(\mathcal{C})\big)}_{\text{a } \sigma\text{-field}}$$
    - **Lemma 1.** Suppose
$$f : \Omega \to \Omega'$$
with $\mathcal{A}'$ being a σ-field on Ω'. Then, $f^{-1}(\mathcal{A}')$ is a σ-field on Ω.

- **Lemma 2.**

$$\sigma\left(X^{-1}(\mathcal{C})\right) = X^{-1}\left(\sigma(\mathcal{C})\right).$$

From the above discussion and Lemma 1, we have

$$\sigma\left(X^{-1}(\mathcal{C})\right) \subset X^{-1}\left(\sigma(\mathcal{C})\right)$$

It remains to be proved that

$$X^{-1}\left(\sigma(\mathcal{C})\right) \subset \sigma\left(X^{-1}(\mathcal{C})\right).$$

Define

$$\mathcal{F} = \left\{B \subset \mathbb{R}^d : X^{-1}(B) \in \sigma\left(X^{-1}(\mathcal{C})\right)\right\}.$$

It can be shown (verify!) that $\mathcal{F}$ is a σ-field.

$$\begin{aligned} \mathcal{C} \subset \mathcal{F} \;\; &\Rightarrow \;\; \sigma(\mathcal{C}) \subset \mathcal{F} \\ &\Rightarrow \;\; X^{-1}\left(\sigma(\mathcal{C})\right) \subset X^{-1}(\mathcal{F}) \subset \sigma\left(X^{-1}(\mathcal{C})\right) \\ &\Rightarrow \;\; X^{-1}\left(\sigma(\mathcal{C})\right) \subset \sigma\left(X^{-1}(\mathcal{C})\right). \end{aligned}$$

❖ Conclusion. When we have a function $X : \Omega \to \mathbb{R}^d$ with underlying probability space $(\Omega, \mathcal{A}, P)$, then we say that X is $\mathcal{A}$-measurable if and only if

$$\sigma(X) = X^{-1}(\mathcal{B}^d) = \{X^{-1}(B) : B \in \mathcal{B}^d\} \subseteq \mathcal{A}.$$

- The smallest σ-field that makes $X(\cdot)$ measurable is equal to the pre-image of the Borel σ-field.

➢ Note.

- $\sigma(X)$ is the smallest σ-field that makes X measurable.
- Then, the probability distribution P^X of X is a probability measure on $(\mathbb{R}^d, \mathcal{B}^d)$:

$$\forall B \in \mathcal{B}^d : P^X(B) = P\left(X^{-1}(B)\right) = P(X \in B)$$

Hence, P^X is induced by P.

- When we say that

$$X \sim U_{[a,b]}$$

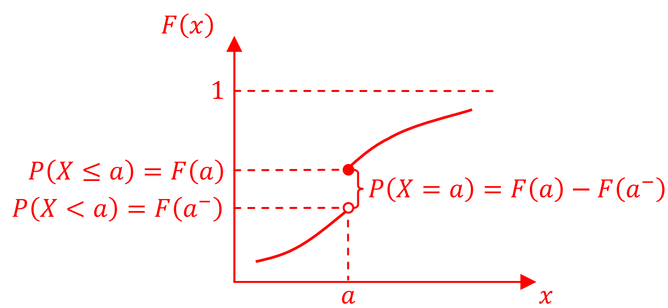
we mean

$$P\left(X \in (c,d)\right) = \frac{d-c}{b-a}$$

for any $(c,d) \subset [a,b]$. But we don't really care about the original $(\Omega, \mathcal{A}, P)$.

Distribution Function

- For any r.v. $X : \underset{(\mathcal{A},P)}{\Omega} \to \underset{X(\omega)}{\mathbb{R}}$
  - Probability distribution of X is P^X, which is a probability measure on $(\mathbb{R}, \mathcal{B})$, defined by
    $$P^X(A) = P\big(X^{-1}(A)\big)$$
    and characterized by
    $$\forall x \in \mathbb{R} : P^X\big((-\infty, x]\big) = P\left(X^{-1}\big((-\infty, x]\big)\right) = P(X \le x).$$
  - We can use a cumulative distribution function to characterize
    $$F_X : \mathbb{R} \to [0,1]$$
    $$x \mapsto F_X(x) = P(X \le x)$$
    - Remark. Can we characterize F_X?
      1) F_X must be non-decreasing
      2) $F_X(x) \xrightarrow{x \to -\infty} 0$ and $F_X(x) \xrightarrow{x \to +\infty} 1$
      3) F_X is right-continuous
      $$\lim \downarrow P\left(X \le x + \frac{1}{n}\right) = P(X \le x)$$



- Why F_X might not be left-continuous?
  $$\lim \uparrow P\left(X \le x - \frac{1}{n}\right) = P(X < x) = F_X(x^-)$$
  Thus, F_X is left-continuous at x if and only if $P(X = x) = 0$.

Cumulative Distribution Function

- $F_X : \mathbb{R} \to [0,1]$ such that F_X is
  - Non-decreasing
  - $F_X(x) \xrightarrow[x\to-\infty]{} 0$ and $F(x) \xrightarrow[x\to+\infty]{} 1$
  - Right-continuous

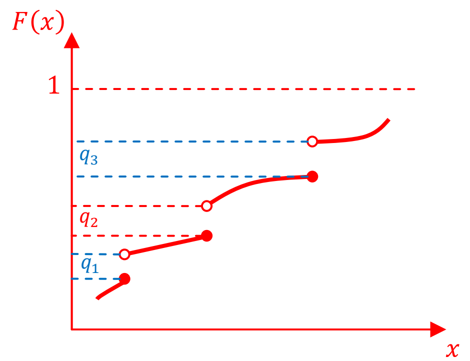
- Question: Is it sufficient to define F_X in order to characterize P^X?
  - Yes!
  - From F_X I can define a σ-additive function Q on all the intervals

$$Q\big((a,b]\big) = F_X(b) - F_X(a)$$
$$Q([a,b]) = F_X(b) - F_X(a)$$
$$Q\big((a,+\infty)\big) = 1 - F_X(a)$$
$$\vdots$$

  - Then, we can construct the outer measure Q^*
    - Unique
    - Coincides with Q on the set of the intervals
  - Q^* is a probability measure on $(\mathbb{R}, \mathcal{B})$

Density Function

- Any real r.v. X with probability distribution characterized by F_X
  - F_X is continuous $\Leftrightarrow$ $P(X = x) = 0,\ \forall x \in \mathbb{R}$
  - $F_X(x) > F_X(x^-)$ where both are real numbers
    - The interval $\big(F_X(x^-), F_X(x)\big)$ contains at least one rational number. We can therefore deduce that there are always at most a countable discontinuity points, i.e. points such that $P(X = x) > 0$.



      - There are only at most countable number of q_i's in the above diagram.

- 2 Extreme Cases
  - F_X only has discontinuity points.

$$\sum_{x\in\mathbb{R}} P(X = x) = 1$$

This is a discrete distribution. For example, Poisson distribution.

- F_X is continuous. If F_X is differentiable on $\mathbb{R}$ with continuous derivative f_X, then we need $F_X' \geq 0$. In addition,
$$\forall x \in \mathbb{R} : F_X(x) = \int_{-\infty}^{x} F_X'(u)du.$$
When $x \to +\infty$,
$$\int_{-\infty}^{\infty} F_X'(u)du = 1.$$

❖ (General Case) *Definition*. X is ***absolutely continuous*** if and only if
$$\begin{cases} \exists f_X \geq 0 \\ \forall x \in \mathbb{R} : F_X(x) = \int_{-\infty}^{\infty} f_X(u)du \end{cases}$$

- Remark. F_X may not be everywhere differentiable.
- Remark. f_X is not unique, (it is defined up to a set of measure zero).
- Absolutely continuous functions are those that can be differentiable almost everywhere.

❖ Example. Exponential Distribution.
$$F_X(x) = \mathbf{1}_{\{x\geq 0\}}\left(1 - e^{-x/\theta}\right) = \begin{cases} 1 - e^{-x/\theta} & if \;\; x \geq 0 \\ 0 & if \;\; x < 0 \end{cases}$$

- F_X is continuous
- F_X is not differentiable at $x = 0$
$$\lim_{h \to 0^+} \frac{F_X(x+h) - F_X(x)}{h} = \lim_{h \to 0^+} \frac{1 - e^{-h/\theta} - (1-1)}{h} = \lim_{h \to 0^+} - \frac{1 - e^{-h/\theta}}{h} = \frac{1}{\theta}$$
However, the derivative on the left is equal to zero.

Absolute Continuity

- *Definition.* X is absolutely continuous if and only if

$$\exists f_X \geq 0 : \forall x \in \mathbb{R} : F_X(x) = \int_{-\infty}^{x} F_X(u)du.$$

- Interpretation: When X is absolutely continuous, its probability distribution can be characterized in 2 ways:
  - The CDF F_X (with its 3 properties)
  - The PDF f_X with
    - $f_X(x) \geq 0$
    - $\int_{-\infty}^{+\infty} f_X(x)dx = 1$, where f_X is almost unique (cf Lebesgue measure zero)

- Connection between F_X and f_X:

$$F_X(x+\Delta x) - F_X(x) = P(X \in (x, x+\Delta x]) = \int_{x}^{x+\Delta x} f_X(u)du$$

$$f_X(x) = \lim_{\Delta x \to 0} \frac{F_X(x+\Delta x) - F_X(x)}{\Delta x}$$

Also, for Δx small enough, we can use the following approximation:

$$P(x < X \leq x + \Delta x) \approx \Delta x \cdot f_X(x)$$

- ***Gamma Distribution***, $\Gamma(p, \theta)$, $p, \theta > 0$.

$$f_X(x) = I_{\{x \geq 0\}} \frac{1}{\theta^p \Gamma(p)} e^{-x/\theta} x^{p-1}$$

  - Question: Is f_X a PDF?

$$\int_{-\infty}^{+\infty} f_X(x)dx = \int_0^{\infty} \frac{1}{\theta^p \Gamma(p)} e^{-x/\theta} x^{p-1} dx = \frac{1}{\theta^p \Gamma(p)} \int_0^{\infty} e^{-x/\theta} x^{p-1} dx$$

I want $\Gamma(p)$ to be such that

$$\Gamma(p) = \frac{1}{\theta^p} \int_0^{\infty} e^{-x/\theta} x^{p-1} dx$$

Change of variable: $y \coloneqq x/\theta$, so that $dy = (1/\theta)dx$

$$\Gamma(p) = \int_0^{\infty} e^{-y} y^{p-1} dy$$

This is the ***Gamma function***. There is no closed (or explicit) form for $\Gamma(\cdot)$. It is only defined through the integral.

    - The Gamma function is a continuous analog of factorials.

- Properties of the Gamma Function
  - If $p > 1$, then $\Gamma(p) = (p-1)\Gamma(p-1)$
  - If $p \in \mathbb{Z}$, then $\Gamma(p) = (p-1)!$
    - *Proof.* Use integration by parts:

$$\Gamma(p) = \int_0^{\infty} \underbrace{e^{-y}}_{u'} \underbrace{y^{p-1}}_{v} dy$$

Define

$$u'(y) = e^{-y} \;\Rightarrow\; u(y) = -e^{-y}$$
$$v(y) = y^{p-1} \;\Rightarrow\; v'(y) = (p-1)y^{p-2}$$

Apply integration by parts:

$$\Gamma(p) = \underbrace{\left[-e^{-y}y^{p-1}\right]_0^\infty}_{=0} + \underbrace{\int_0^\infty e^{-y}(p-1)y^{p-2}dy}_{(p-1)\Gamma(p-1)}$$

- $X \sim \Gamma(p,\theta)$, then

$$y = \frac{x}{\theta} \;\sim\; \Gamma(p,1) = \Gamma(p)$$

  - *Proof.*

$$P(Y \le y) = P\left(\frac{X}{\theta} \le y\right) = P(X \le \theta y) = \int_0^{y\theta} \frac{e^{-x/\theta}}{\theta^p \Gamma(p)} x^{p-1} dx = \int_0^y \underbrace{\frac{e^{-u}}{\Gamma(p)} u^{p-1}}_{PDF\ of\ \Gamma(p,1)} du$$

where $u = x/\theta$.

- Multivariate Extension:

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{x,y}(u,v)dvdu$$

where

$$\begin{aligned} f_{x,y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} &\approx \frac{1}{k}\left[\frac{\partial F}{\partial x}(x,y+k) - \frac{\partial F}{\partial x}(x,y)\right] \\ &\approx \frac{1}{hk}\left[F(x+h,y+k) - F(x,y+k) - F(x+h,y) + F(x,y)\right] \\ &\approx \frac{1}{hk} P(x < X \le x+h \;\wedge\; y < Y \le y+k) \end{aligned}$$

For small enough h and k:

$$P(x < X \le x+h \;\wedge\; y < Y \le y+k) \approx hk \cdot f_{X,Y}(x,y)$$

Lebesgue Integral and Mathematical Expectation

- **1st case**: X is discrete r.v.
  - $\mathcal{X}$ is finite or countable, and $P(x \in \mathcal{X}) = 1$.
    - $\mathcal{X}$ is like the Ω in previous lectures.
  - Assume we repeat n times the statistical experiment and we get: $X_1, X_2, \dots, X_n \sim p^x$
    - X_n's represent the nth experiment and they all follow the same distribution (iid)
  - For all $x \in \mathcal{X}$, the sampling distribution is

$$\frac{n_x}{n} = \frac{\#\text{ of times that value } x \text{ occurs}}{\#\text{ of experiments}} = \text{relative frequency of } x$$

  where n_x is the number of times I observe the value x.
  - Then, we can derive the ***mathematical*** (or ***population***) ***expectation of X***

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{x\in\mathcal{X}} x \cdot n_x = \sum_{x\in\mathcal{X}} x \cdot \frac{n_x}{n} \xrightarrow[\text{if } LLN \text{ applies}]{} \sum_{x\in\mathcal{X}} xP(X = x) = EX$$

    - Here we use x instead of X because we're talking about the *realizations*, not the random variables. We could have used X instead, in which case we'll be referring to the random variable before the experiments.

- Example 1. We draw (with replacement) N balls from a box that contains a proportion of p green balls.
  - x_i : number of green balls picked during experiment $\#i$
  - Here $x_i \sim \mathcal{B}(N, p)$. Then,

$$EX = \sum_{x=0}^{N} x \underbrace{P(X = x)}_{\mathcal{B}(N,p)} = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x}$$

$$= Np \sum_{x=0}^{N} x \frac{(N-1)!}{x!\,(N-x)!} p^{x-1}(1-p)^{N-x}$$

$$= Np \underbrace{\sum_{y=0}^{N-1} \frac{(N-1)!}{y!\,(N-y-1)!} p^y (1-p)^{N-(y+1)}}_{(p+(1-p))^{N-1}=1} = Np$$

  - Here $y = x - 1$

- Example 2. $X \sim Poisson(\lambda)$

$$EX = \sum_{x=0}^{\infty} xP(X = x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda \underbrace{\sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!}}_{=1} = \lambda$$

  - CDF of Poisson distribution:

$$F_X(x; k, \lambda) = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$$

- **2nd case** : X absolutely continuous

$$P(X \in (x, x + \Delta x]) \approx f_X(x)\Delta x$$

$$\sum xP(X \in (x, x+\Delta x]) \approx \sum x f_X(x)\Delta x$$

$$\Rightarrow \quad EX = \underbrace{\int_{-\infty}^{+\infty} x f_X(x) dx}_{\text{well-defined if } E|X|<\infty} \approx \sum x_i \underbrace{f_X(x)(x_{i+1} - x_i)}_{\approx P(x_i < X \le x_{i+1})}$$

- Example. $X \sim \Gamma(p, \theta)$

$$EX = \int_0^\infty x \frac{1}{\theta^p \Gamma(p)} x^{p-1} e^{-x/\theta} dx = p\theta \underbrace{\int_0^\infty x \underbrace{\frac{1}{\theta^{p-1}\Gamma(p+1)} x^{p-1} e^{-x/\theta}}_{=\Gamma(p+1,\theta)} dx}_{=1} = p\theta$$

  - This leads to the linearity of the expectation operation E:

$$E\left(\frac{X}{\theta}\right) = p$$

    - This property is not limited to the Gamma distribution.

Mathematical Expectation (cont'd)

- Want to define

$$EX = \int_{\mathbb{R}} x \underbrace{dP^X(x)}_{\substack{f_X(x)dx = P(x<X\leq x+dx) \\ \text{or } \ P(X=x)}}$$

  - We have shown for the cases $f_X(x)dx = P(x < X \leq x + dx)$ and $P(X = x)$.
  - We will see that

$$\int_{\Omega} X(\omega)dP(\omega) = \int_{\mathbb{R}} \underbrace{x}_{\text{Identity function}} dP^X(x)$$

  for $X : \Omega \to \mathbb{R}$.

- 1st case: X takes a finite number of values that are non-negative

$$X = \sum_{i=1}^{n} \alpha_i \, \mathbf{1}_{A_i}$$

  with $A_i = \{\omega : X(\omega) = \alpha_i\}$.
  - A_i is the pre-image of α_i.

  Integrate on both sides:

$$\int_{\Omega} XdP = \sum_{i=1}^{n} \alpha_i \underbrace{\int_{\Omega} \mathbf{1}_{A_i} dP}_{P(A_i)} = \sum_{i=1}^{n} \alpha_i P(A_i)$$

  where

$$P(A_i) = E(\mathbf{1}_{A_i}) = \int_{\Omega} \mathbf{1}_{\{A_i\}} dP = \int_{A_i} 1 \, dP$$

  - This extends to the case where X takes a countable number of non-negative values.

- 2nd case: X is measurable non-negative r.v. such that

$$X = \lim \uparrow \underbrace{\left\{ \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mathbf{1}_{\left\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\right\}} \right\}}_{X_n}$$

  We can use the monotone convergence theorem to conclude:

$$\int XdP = \lim \uparrow \int X_n dP$$

  In other words,

$$EX = E(\lim \uparrow X_n) = \lim \uparrow EX_n$$

- 3rd case: X is measurable (real) r.v.

$$X = X^+ - X^-$$

  where

$$X^+ = \max\{X, 0\} \quad \text{and} \quad X^- = \max\{-X, 0\}$$

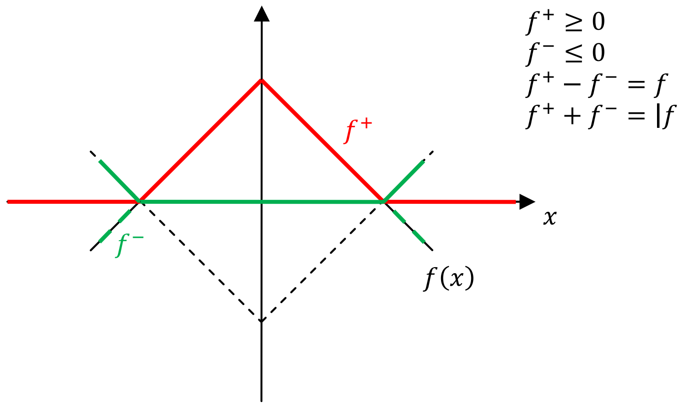
  Note that both X^+ and X^- are non-negative.

$$EX = EX^+ - EX^-$$

  are well-defined and finite if and only if

$$\left.\begin{array}{l} EX^+ < \infty \\ EX^- < \infty \end{array}\right\} \Leftrightarrow E|X| < \infty \Leftrightarrow X \text{ is integrable}$$

- Example.



❖ <u>Note 1</u>. $P = \alpha Q + (1-\alpha)\tilde{Q} \rightarrow$ measure

$$\int XdP = \alpha \int XdQ + (1-\alpha)\int Xd\tilde{Q}$$

❖ <u>Note 2</u>. **Transfer Theorem**: Suppose $Y = \phi(X)$ where Y is integrable, i.e. $E|Y| < \infty$

$$EY = E[\phi(X)] = \int_\Omega \phi\big(X(\omega)\big)dP(\omega) = \int_{\mathbb{R}} \phi(x)dP^X(x)$$

- X is a r.v., and Y is a r.v. generated by X. Then, to find expectation of Y, we can either evaluate it using the underlying probability space of X (i.e. Ω), or treating X as the probability that generates Y, and evaluate Y using the distribution of X.

Conditional Probability, Bayes' Rule, and Independence

- *Definition*. A and B are ***independent*** if and only if
$$P(A \cap B) = P(A)P(B).$$
  - Note. If $P(B) \neq 0$, then A and B are independent if and only if
$$Q^B(A) := \frac{P(A \cap B)}{P(B)} = P(A)$$
    - B is ***probable*** if $P(B) \neq 0$
    - We can call $Q^B(A)$ a probability measure with all the probability 1 put on B.
      - The probability space associated with Q^B is $(\Omega, \mathcal{A}, Q^B)$
$$Q^B : C \to Q^B(C) = \frac{P(B \cap C)}{P(B)}$$
      This formula describes the statistical model when
      - We draw from Ω
      - But we are sure that $\omega \in B$, because we have some additional information
    - Here $Q^B(\cdot)$ is a well-defined probability measure as long as $P(B) \neq 0$
      - $Q^B(\cdot)$ is called the ***conditional probability distribution***.
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  - A and B are independent if and only if
    - A has the same probability for $P(\cdot)$ and $P(\cdot\,|B)$
    - $P(A) = P(A|B)$

- Example 1. X represents duration (e.g. the Poisson process).
  - **No memory property**
$$P(X \geq t + h | X \geq t) = P(X \geq h) \;\Leftrightarrow\; P(X \geq t + h) = P(X \geq t)P(X \geq h)$$
    - For instance, $P(X \geq t)$ is modeled using the exponential distribution
$$P(X \geq t) = e^{-\theta t}$$
    - Given the Poisson,
$$F_X(t) = P(X \leq t) = \begin{cases} 1 - e^{-\theta t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$
      Then, the ***survival function*** is
$$S_X(t) = P(X > t) = 1 - F_X(t) = \begin{cases} e^{-\theta t} & \text{if } t \geq 0 \\ 1 & \text{if } t < 0 \end{cases}$$

- Example 2. A ***partition*** of Ω has the following properties:
  - $H_i \cap H_j = \emptyset$, for any $i \neq j$
  - $\bigcup_{i=1}^{n} H_i = \Omega$

  Decompose Ω into a partition.
$$\Omega = \bigcup_{i=1}^{n} H_i$$
  where $P(H_i) \neq 0$ for all i. Then,

$$P(A) = \sum_{i=1}^{n} P(A|H_i)P(H_i)$$

- Consider this:

$$P(A) = P(A \cap \Omega) = P\big(A \cap (H_1 \cup \cdots \cup H_n)\big) = P\big((A \cap H_1) \cup \cdots \cup (A \cap H_n)\big)$$
$$= P(A \cap H_1) + \cdots + P(A \cap H_n) = P(A|H_1)P(H_1) + \cdots + P(A|H_n)P(H_n)$$

- This is the key to define mixtures of distributions (cf. Wikipedia article)
- Example. $\Gamma(p)$

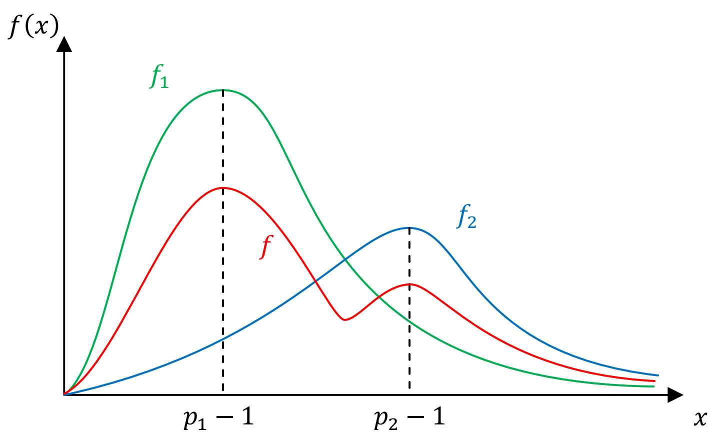
$$f(x) = \frac{1}{\Gamma(p)} e^{-x} x^{p-1}$$

  - This is unimodal.

$$f'(x) = \frac{1}{\Gamma(p)} (-e^{-x}x^{p-1} + (p-1)x^{p-2}e^{-x})$$
$$f'(x) = 0 \ \Rightarrow \ x = p - 1$$

- Suppose $X \sim [\alpha\Gamma(p_1) + (1-\alpha)\Gamma(p_2)]$.

$$F_X(x) = \alpha F_1(x) + (1-\alpha)F_2(x)$$
$$f_X(x) = \alpha f_1(x) + (1-\alpha)f_2(x)$$



  - If $F = \sum_i \alpha_i F_i$, with $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$, then $f = F' = \sum_i \alpha_i f_i$, where $f_i = F_i'$.
    - Here α_i's can be interpreted as PMF values (or probability of singletons).
    - This can extend to continuous cases, and the sum will be replaced by an integral.
    - This works for any distribution functions (CDF and PDF)

- Note 3. The statement A, B are independent $\Leftrightarrow$ $P(A \cap B) = P(A)P(B)$ is always true.
  - If $P(B) = 0$, then any set A is independent of B
    - Both sure and improbable sets are independent of anything, including themselves.
  - A, B are independent

$$\Leftrightarrow A^c \text{ and } B \text{ are independent}$$
$$\Leftrightarrow A \text{ and } B^c \text{ are independent}$$
$$\Leftrightarrow A^c \text{ and } B^c \text{ are independent}$$

    - This is the ***Independence Complement Theorem***.
    - For proof, use the following as initial step:

$$P(A) = P\big(A \cap (B \cup B^c)\big), \qquad P(B) = P\big((A \cup A^c) \cap B\big),$$
$$P(A^c \cap B^c) = P(A \cup B)^c$$

- Note 4. A, B, C are pairwise independent does NOT imply
$$P(A \cap B \cap C) = P(A)\underbrace{P(B)P(C)}_{=P(B\cap C)} = P(A)P(B \cap C)$$
where A is independent of $(B \cap C)$

❖ *Definition.* $(A_i)_{i\in I}$ are ***mutually independent*** if and only if for all $J \subseteq I$ with J finite,
$$P\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} P(A_i)$$

❖ **Theorem (0-1 Law of Borel-Cantelli).** Consider $(A_n)_n$ sequence of events.

1) If $\sum_{i=1}^{\infty} P(A_n) < \infty$, then
$$P(\limsup A_n) = 0$$
2) If $\sum_{n=1}^{\infty} P(A_n) = \infty$, and $(A_n)_n$ are mutually independent, then,
$$P(\limsup A_n) = 1$$

- *Proof.* Begin by recalling that
$$\limsup A_n = \bigcap_{n\in\mathbb{N}} \bigcup_{p\geq n} A_p$$
has the interpretation that A_n happens infinitely many times.

Proof of (1). Note that $\bigcup_{p\geq n} A_p$ is a decreasing sequence. Thus,
$$P(\limsup A_n) = \lim \downarrow P\left(\bigcup_{p\geq n} A_p\right) \leq \lim \downarrow \left\{\sum_{p=n}^{\infty} P(A_p)\right\} = 0$$
The last equality is justified by the fact that each $P(A_p)$ is finite, and the sequence of partial sums is decreasing.

Proof of (2).
$$P(\limsup A_n) = \lim \downarrow P\left(\bigcup_{p\geq n} A_p\right) = 1 - \lim \uparrow P\left(\bigcap_{p\geq n} (A_p)^c\right)$$
Note that
$$\begin{aligned} P\left(\bigcup_{p\geq n} A_p\right) &= \lim_N \uparrow \left\{P\left(\bigcup_{N\geq p\geq n} A_p\right)\right\} \\ &= 1 - \lim_N \downarrow P\left(\bigcap_{N\geq p\geq n} (A_p)^c\right) \\ &= 1 - \lim_N \downarrow \prod_{N\geq p\geq n} P\left((A_p)^c\right) \\ &= 1 - \underbrace{\lim_N \downarrow \prod_{N\geq p\geq n} \left(1 - P(A_p)\right)}_{=0} \end{aligned}$$

To show that the second term is indeed equal to zero,

$$\prod_{N\geq p\geq n}\left(1-P(A_p)\right)\leq\prod_{N\geq p\geq n}\exp\left(-P(A_p)\right)=\exp\underbrace{\left\{-\sum_{N\geq p\geq n}P(A_p)\right\}}_{\to-\infty}$$

The inequality is justified by $1-x\leq e^{-x}\approx 1-x+\frac{x^2}{2}\cdots$. Since

- Independence of r.v.
  - Suppose we have $(\Omega,\mathcal{A},P)$, and we have random variables $X_1\in\mathbb{R}^{d_1},X_2\in\mathbb{R}^{d_2},\ldots$ X_1,X_2 independent

$$\Leftrightarrow\ \forall A_1,A_2: P(X_1\in A_1,X_2\in A_2)=P(X_1\in A_1)P(X_2\in A_2)$$
$$\Leftrightarrow\ \forall A_1,A_2: X_1^{-1}(A_1)\text{ and }X_2^{-1}(A_2)\text{ are independent}$$
$$\Leftrightarrow\ X_1^{-1}(\mathbb{R}^{d_1})\text{ and }X_2^{-1}(\mathbb{R}^{d_2})\text{ are independent}$$

- *Definition.* $(\Omega,\mathcal{A},P)$ with $\mathcal{C}_i\subset\mathcal{A}$. Then, $(\mathcal{C}_i)_{i\in I}$ are independent if and only if
$$\forall A_i\in\mathcal{C}_i:(A_i)_{i\in I}\text{ are independent}$$

- *Definition.* $(\Omega,\mathcal{A},P)$ with X_i on $(\mathbb{R}^{d_i},\mathbb{R}^{d_i})$. $(X_i)_{i\in I}$ are independent if and only if
$$\left(\sigma(X_i)\right)_{i\in I}\text{ are independent}$$

- **Theorem.** $(\Omega,\mathcal{A},P)$ with $\mathcal{C}_i\subset\mathcal{A}$ for all $i\in I$. If
$$\forall i\in I: A,B\in\mathcal{C}_i\ \Rightarrow\ (A\cap B)\in\mathcal{C}_i$$
Then, $(\mathcal{C}_i)_{i\in I}$ are independent if and only if $\left(\sigma(\mathcal{C}_i)\right)_{i\in I}$ are independent.

  - One easy case is when $\mathcal{C}_i=\{A_i\}$. $(\{A_i\})_{i\in I}$ are independent if and only if $(\{\emptyset,\Omega,A_i,(A_i)^c\})_{i\in I}$ are independent.

  - <u>Case 1</u>: discrete r.v. $P(X_i\in\mathcal{X}_i)=1$ with $\mathcal{X}_i$ finite or countable. $(X_i)_{i\in I}$ are independent if and only if

$$\left(X_i^{-1}\left(\underbrace{\sigma(\mathcal{X}_i)}_{\substack{\sigma\text{ field defined by}\\ \text{the values }x_i\in\mathcal{X}_i}}\right)\right)_{i\in I}\quad\text{independent}$$

$$\Leftrightarrow\ \left(X_i^{-1}\left(\underbrace{\mathcal{P}\{x_i:x_i\in\mathcal{X}_i\}}_{\substack{\mathcal{P}(\mathcal{X}_i)=\sigma(\{\{x_i:x_i\in\mathcal{X}_i\}\})\\ \text{stable by intersection}}}\right)\right)_{i\in I}\quad\text{independent}$$

$$\Leftrightarrow\ \forall J\subseteq I:|J|<\infty,\forall x_i\in\mathcal{X}_i: P(X_i=x_i: i\in J)=\prod_{i\in J}P(X_i=x_i)$$

    - $\mathcal{X}_i$ is the support of X_i.

Independence or r.v.

- **Theorem.** $(\Omega, \mathcal{A}, P)$ with $\mathcal{C}_i \subset \mathcal{A}$ for all $i \in I$. If $\forall i \in I : A, B \in \mathcal{C}_i \;\Rightarrow\; A \cap B \in \mathcal{C}_i$, then $(\mathcal{C}_i)_{i\in I}$ are independent if and only if $\left(\sigma(\mathcal{C}_i)\right)_{i\in I}$ are independent.
  - 2 discrete r.v. X, Y are independent if and only if
$$P(X = x, Y = y) = P(X = x)P(Y = y), \qquad \forall x, y$$
$$\Leftrightarrow \; P^{(X,Y)}(\{x, y\}) = P^X(\{x\})P^Y(\{y\})$$

- *Definition*. Given $(\Omega_i, \mathcal{A}_i, P_i)_{i\in I}$,
$$P \equiv \bigotimes_{i\in I} P_i$$
is the probability measure on $(\prod_{i\in I} \Omega_i, \otimes_{i\in I} \mathcal{A}_i)$, where
$$\bigotimes_{i\in I} \mathcal{A}_i = \sigma\left(\prod_{i\in I} A_i : A_i \in \mathcal{A}_i \middle| \forall i, \text{and } \exists J_{\text{finite}} \subset I : \forall i \notin J : A_i = \Omega_i\right)$$
such that
$$P\left(\prod_{i\in I} A_i\right) = \prod_{i\in I} \underbrace{P_i(A_i)}_{\substack{\text{only a finite}\\ \text{number of}\\ \text{them are not 1}}}$$
  - Note. $\otimes$ means the cross-product of collection of sets.
  - $\Omega_1 = \{a, b\}, \Omega_2 = \{c, d\}, \mathcal{A}_1 = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}, \; \mathcal{A}_2 = \{\emptyset, \{c\}, \{d\}, \{c, d\}\}$. Then,
$$\Omega_1 \times \Omega_2 = \{(a, c), (a, d), (b, c), (b, d)\}$$
$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \{\emptyset, \{a\} \times \{c\}, \dots\}$$
  - Example.
$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$$
$$(x, y) \in \mathbb{R}^2 \;\Leftrightarrow\; x \in \mathbb{R}, y \in \mathbb{R}$$
$A_1 \times A_2$ with $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.

- **Theorem.** $(X_i)_{i\in I}$ are independent if and only if
$$\underbrace{P^{(X_i)_{i\in I}}}_{\substack{\text{Global CDF}\\ \text{or Joint CDF}}} = \underbrace{\prod_{i\in I} P^{X_i}}_{\text{induced probability}}$$
  - For us, the index set I is most of time finite, and sometimes countable (if we are dealing with sequences)
  - When X's are independent, then the global CDF is equal to the product of individual CDF.

- Case 2 (continuation from last class): X_i is a real-valued r.v. (extension to $\mathbb{R}^d$ is "easy")
$$\sigma(X_i) = X_i^{-1}(\mathcal{B}) = X_i^{-1}\left(\sigma\left((-\infty, x] : x \in \mathbb{R}\right)\right) \underset{\text{by } \textit{Lemma } 2}{=} \sigma\left(\underbrace{X_i^{-1}\left((-\infty, x] : x \in \mathbb{R}\right)}_{\text{stable by intersection}}\right)$$
$(X_i)_{i\in I}$ are independent if and only if
$$\left(X_i^{-1}\left((-\infty, x] : x \in \mathbb{R}\right)\right)_{i\in I} \text{ are independent}$$

$$\Leftrightarrow \quad \forall J_{\text{finite}} \subset I, \forall x_i \in \mathbb{R} : P(X_i \leq x_i : i \in J) = \prod_{i \in J} P(X_i \leq x_i)$$

$$\Leftrightarrow \quad F_{(X_i)_{i \in J}}\big((x_i)_{i \in J}\big) = \prod_{i \in J} F_{X_i}(x_i)$$

Therefore, if I is finite, I only have to check this last equality on I.

- Random variables are independent if and only if their joint CDF is a product of their respective CDF's.

Expectation and Independence

❖ *Definition.* For a real r.v. X that is integrable,

$$\underbrace{VarX}_{\text{variance of } X} = \underbrace{E\left[\left(X - \underbrace{EX}_{\text{not } r.v.}\right)^2\right]}_{\text{not } r.v.}$$

❖ **Proposition.** Suppose

$$VarX < \infty \iff EX^2 < \infty \iff X \text{ square integrable.}$$

Then, $VarX = E(X^2) - (EX)^2$.

- Square integrable means $\int X^2 dF_X$ exists.
- *Proof.* By definition,

$$\begin{aligned} Var\,X &= E[(X - EX)^2] \\ &= E[X^2 + (EX)^2 - 2XEX] \\ &= EX^2 + (EX)^2 - 2EX \cdot EX \\ &= EX^2 - (EX)^2 \end{aligned}$$

  - Note. $Var\,X \geq 0 \Rightarrow EX^2 \geq (EX)^2$.
    - The inequality in this case is due to the convexity of the square function
    - Note that $VarX$ is a number, not a random variable, so $VarX \geq 0$.
  - Note. Since X is r.v., we have to say $X \geq 0$ almost surely

❖ ***Jensen Inequality***:

- If ϕ is a concave function, then $E[\phi(X)] \leq \phi[EX]$.
- If ϕ is a convex function, then $E[\phi(X)] \geq \phi[EX]$.

❖ **Proposition.** If X is square integrable, and a is a parameter, then

$$\underbrace{E[(X-a)^2]}_{\substack{\text{mean squared error} \\ \text{with respect to } a}} = MSE = \underbrace{Var\,X}_{\substack{\text{measure of} \\ \text{variability}}} + \underbrace{(EX-a)^2}_{\text{bias squared}}$$

- Note. a does not have to be EX.
- Note. $Var\,X = Var(X - a)$.

$$\begin{aligned} Var(X-a) &= E(X-a)^2 - \big(E(X-a)\big)^2 \\ &= E(X^2 - 2aX + a^2) - (EX - a)^2 \\ &= EX^2 - 2aEX + a^2 - ((EX)^2 - 2aEX + a^2) \\ &= EX^2 - (EX)^2 = Var(X) \end{aligned}$$

- Note. $Var(aX + b) = a^2 Var\,X$

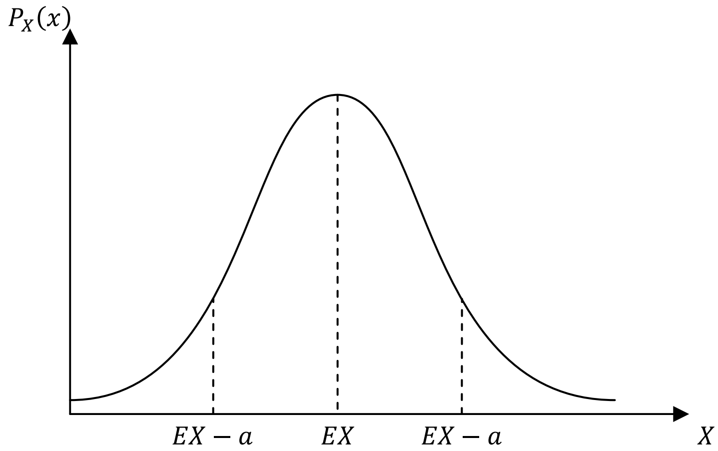
$$\begin{aligned} &Var(aX+b) = Var(aX) \\ &= E(aX)^2 - \big(E(aX)\big)^2 \\ &= a^2EX^2 - a^2(EX)^2 \\ &= a^2(EX^2 - (EX)^2) = a^2Var(X) \end{aligned}$$

❖ Property of variance.

- ***Markov inequality***

$$P(|X - EX| \geq a) \leq \frac{1}{a} E|X - EX|, \qquad \forall a > 0$$

$$P(|X - EX| \leq a) \geq 1 - \frac{1}{a} E|X - EX|, \qquad \forall a > 0$$



- ***Bienayme-Chebyshev***

$$P(|X - EX| \geq a) \leq \frac{1}{a^2} Var\ X$$

- *Proof.*
  - Proof of Markov inequality.

$$a\mathbf{1}_{\{|X-EX|\geq a\}} \leq |X - EX|, \qquad almost\ surely$$
$$\Leftrightarrow \quad \mathbf{1}_{\{|X-EX|\geq a\}} \leq \frac{1}{a}|X - EX|, \qquad almost\ surely$$

  Take expectation of this inequality:

$$E\left[\mathbf{1}_{\{|X-EX|\geq a\}}\right] \leq \frac{1}{a}E|X - EX| \quad \Leftrightarrow \quad P(|X - EX| \geq a) \leq \frac{1}{a}E|X - EX|$$

    - The key is to use the fact that the expectation of the indicator is the probability of the events.
  - Proof of Bienayme-Chebyshev. Same as in the Markov case. Just to square everything.

$$a\mathbf{1}_{\{|X-EX|\geq a\}} \leq |X - EX| \quad \Rightarrow \quad \left(a\mathbf{1}_{\{|X-EX|\geq a\}}\right)^2 \leq (|X - EX|)^2$$
$$\Rightarrow \quad E\left(a\mathbf{1}_{\{|X-EX|\geq a\}}\right)^2 \leq E(|X - EX|)^2$$
$$\Rightarrow \quad P(|X - EX| \geq a) \leq \frac{1}{a^2} VarX$$

❖ Special cases of the above two inequalities.

- Pick $a = kE|X - EX|$. Then the Markov inequality is

$$P\left(\frac{|X - EX|}{E|X - EX|} \geq k\right) \leq \frac{1}{k}$$

- Pick $a = k\sqrt{Var\ X} = ks(X)$, where $s(X)$ is the standard deviation. The B-C inequality is

$$P\left(\frac{|X - EX|}{s(X)} \geq k\right) \leq \frac{1}{k^2}$$

  - Example. $k = 2$

$$P\left(\frac{|X - EX|}{s(X)} \geq 2\right) \leq \frac{1}{4} \quad \Leftrightarrow \quad P(X \in [EX - 2s(X),\ \ EX + 2s(X)]) \geq \frac{3}{4}$$

- Averaging reduces variability?
  - If $X_1, \ldots, X_n$ are n r.v.'s that are identically and distributed and independent (iid) $\sim X$, then

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\underbrace{\left[E\left(\sum_{i=1}^{n} X_i\right)^2 - \left(E\sum_{i=1}^{n} X_i\right)^2\right]}_{n\,VarX} = \frac{1}{n}Var\,X$$

Variance (cont'd)

- $X_1, \ldots, X_n$ is iid P^X
$$Var\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\bar{X}_n}\Bigg) = \frac{1}{n}Var(X), \qquad X \text{ is a representative } r.v. \text{ of } X_i \text{ due to } iid$$
Consequently,
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
is a <u>consistent</u> estimator of EX if
   - $E\bar{X}_n = EX$
   $$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\big(E(X_1) + \cdots + E(X_n)\big) = \frac{nE(X)}{n} = E(X)$$
   - $Var\bar{X}_n \underset{n}{\to} 0$

- Example. $X_i = \mathbf{1}_A(Y_i)$, $\bar{X}_n = f_n(A)$ is a consistent estimator of $P(A)$.

- *Definition.* $X_n \xrightarrow[MSE]{L^2} X \iff E(X - X_n)^2 \overset{n}{\to} 0$
   - $\overset{MSE}{\longrightarrow} \iff \overset{L^2}{\to} \Rightarrow \overset{Prob}{\longrightarrow} \Rightarrow \overset{Distribution}{\longrightarrow}$
   - <u>Note</u>. $\|X\| = \sqrt{EX^2}$ is the norm in the space of square-integrable r.v.
      - L^2 is a normed space (a Hilbert space with $\langle X, Y\rangle = E(XY)$)

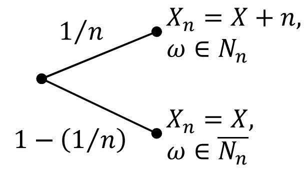
- Property 1:
$$X_n \overset{L^2}{\to} X \iff \begin{cases} EX_n \to EX \\ Var(X_n - X) \to 0 \end{cases}$$
$$X_n \overset{L^2}{\to} a \iff \begin{cases} EX_n \to EX \\ VarX_n \to 0 \end{cases}$$
Proof.
$$\underbrace{E[(X - X_n)^2]}_{\to 0} = Var(X - X_n) + (EX - EX_n)^2.$$

- Property 2:
$$X_n \overset{L^2}{\to} X \Rightarrow X_n \overset{P}{\to} X$$
Proof. For any $\epsilon > 0$
$$P(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon^2}\underbrace{E(X_n - X)^2}_{\substack{\to 0 \text{ by } L^2 \\ \text{convergence}}}$$
   - <u>Note</u>. Convergence in probability does not imply convergence in L^2
      - Counter example:

Suppose $X_n = X$ for all $\omega \notin N_n$ with $P(N_n) = 1/n$. If $\omega \in N_n$, then $X_n = X + n$.

$$P(|X_n - X| > \epsilon) = P(N_n) = \frac{1}{n} \to 0$$

$$E(X - X_n)^2 = \frac{1}{n} n^2 = n \to \infty$$



- In this example, N_n is a sequence of sets in the sample space.
- X_n is almost equal to X except when $\omega \in N_n$.

▪ Let $D_n = X - X_n$.

$$ED_n^2 = P(N_n)E(D_n^2|N_n) + P(\overline{N_n})E\left(\overline{D_n}^2 \middle| \overline{N_n}\right)$$

❖ *Definition.* Suppose X, Y are square integrable.
- $Cov(X,Y) = E(XY) - EX \cdot EY = E\{(X - EX)(Y - EY)\}$
- X, Y are ***uncorrelated*** if and only if $Cov(X,Y) = 0$.

❖ **Theorem (Law of Large Numbers for uncorrelated r.v.).** Consider X_n such that

$$EX_n = m, \qquad VarX_n = s^2 < \infty, \qquad \forall i \neq j : Cov(X_i, X_j) = 0$$

Then, $\bar{X}_n \xrightarrow{L^2} m$.
- This a strong LLN because it implies the weak LLN.

❖ ***Characteristic Function***.
- Covariance of X, Y does not characterize independence. The reason is that, knowing $EX, EY, VarX, VarY$ only characterize the marginal distribution of X and Y, but not the joint distribution of X and Y. But we need the joint distribution to determine independence.
- We need to know $E[g(X)]$ for any g, which is equivalent to knowing P^X
- Similarly, $E[g(X)h(Y)]$ for any g, h is equivalent to knowing $P^{X,Y}$.

- *Definition*. Given a r.v. $X : \Omega \to \mathbb{R}^d$, the characteristic equation is

$$\phi_X(u) = E[\exp(\mathrm{i}\, u^T X)], \qquad \forall u \in \mathbb{R}^d$$
$$= \cos(u^T X) + \mathrm{i} \sin(u^T X)$$

This is bounded within the unit circle.

Characteristic Function (cont'd)

- *Definition.* The ***characteristic function*** of a r.v. $X : \Omega \to \mathbb{R}^d$ is
$$\phi_X(u) = E\left(e^{\mathrm{i}u^T X}\right), \qquad \forall u \in \mathbb{R}^d$$
  - Note. $u^T X \in \mathbb{R}$ is a scalar.
  - Note. Knowing ϕ_X on $\mathbb{R}^d$ is equivalent to
    - Knowing $E[h_u(X)]$, where $h_u(X) = \exp(\mathrm{i}u^T X)$ are a basis of function
    - Knowing $E[h(X)]$
    - Knowing P^X

- **Theorem.** Consider 2 r.v. X, Y.
$$P^X = P^Y \iff \phi_X = \phi_Y$$

- **Theorem.** Given a one-dimensional real random variable, if $E|X|^n < \infty$, then ϕ_X is n-times differentiable and
$$\phi_X^{(k)}(0) = \mathrm{i}^k E X^k, \qquad \forall k = 0, \ldots, n$$
When n is finite, we can switch the differentiation and expectation operation.

- Example. Calculation of moments of a r.v. (one-dimensional)
$$\phi_X(u) = E\left(e^{\mathrm{i}uX}\right) \Rightarrow \phi_X'(u) = \mathrm{i}E\left(Xe^{\mathrm{i}uX}\right) \Rightarrow \phi_X'(0) = \mathrm{i}E(X)$$
$$\Rightarrow \phi_X^{(2)}(u) = \mathrm{i}^2 E\left(X^2 e^{\mathrm{i}uX}\right) \Rightarrow \phi_X^{(2)}(0) = -E(X^2)$$
  - This is an "efficient" way to get higher order moments

  - Another way is to use the ***moment generating function*** (***MGF***):
$$L_X(u) = E\left(e^{u^T X}\right)$$
This is the LaPlace transformation.
$$\underbrace{\frac{\partial L_X(u)}{\partial u}}_{\text{column vector}} = E\left(Xe^{u^T X}\right) \Rightarrow \left.\frac{\partial L_X(u)}{\partial u}\right|_{u=0} = EX$$
$$\frac{\partial}{\partial u}\left(\underbrace{\frac{\partial L_X(u)}{\partial u^T}}_{\text{row vector}}\right) = E\left(XX^T e^{u^T X}\right) \Rightarrow \left.\frac{\partial^2 L_X(u)}{\partial u \partial u^T}\right|_{u=0} = E(XX^T)$$
    - Note. $\partial L_X / \partial u$ will give a column vector, $\partial L_X / \partial u^T$ will give a row vector.
    - Note. Since $L_X(u)$ is a scalar, the order of differentiation does not matter, i.e. can differentiate w.r.t u and then u^T. However, if $L_X(u)$ is a column vector, must differentiate w.r.t. a row vector u^T.
    - For random vector X of dimension d
$$\underbrace{Var\ X}_{d\times d} = E(XX') - (EX)(EX') = E[(X - EX)(X - EX)']$$
where $\left[E\left(X_i X_j\right) - E(X_i)E\left(X_j\right)\right]_{1 \le i,j \le d}$ is a typical element of $Var\ X$.
      - On the main diagonal ($i = j$), we have $Var\ X_i$
      - Off the main diagonal ($i \neq j$), we have $Cov\left(X_i, X_j\right)$

- Covariance of random vectors: X of dimension d_X and Y of dimension d_Y
$$Cov(X,Y) = \underbrace{E(XY')}_{d_X \times d_Y} - E(X)E(Y')$$
  - With linear combination of X and Y, where A is $(n \times d_X)$ and B is (m, d_Y)
$$Cov(AX,BY) = \underbrace{A}_{n\times d_X} \cdot \underbrace{Cov(X,Y)}_{d_X\times d_Y} \cdot \underbrace{B'}_{d_Y\times m}$$

- Example of using MGF on Poisson distribution. Let $X \sim \mathcal{P}(\lambda)$
$$E(e^{uX}) = \sum_{k=0}^{\infty} e^{uk} P(X=k) = \sum_{k=0}^{\infty} e^{uk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^u\lambda)^k}{k!} = e^{-\lambda} e^{\lambda e^u}$$
$$= \exp[\lambda(e^u - 1)]$$
Let $L_X(u) = \exp[\lambda(e^u - 1)]$. Then,
$$L_X'(u) = \lambda e^u \cdot \exp[-\lambda(1-e^u)] \;\Rightarrow\; L_X'(0) = \lambda \;\Rightarrow\; EX = \lambda$$
$$L_X''(u) = [\lambda e^u + (\lambda e^u)^2]\exp[-\lambda(1-e^u)] \;\Rightarrow\; L_X''(0) = \lambda + \lambda^2 \;\Rightarrow\; EX^2 = \lambda + \lambda^2$$
Therefore, $Var\, X = EX^2 - E^2X = \lambda + \lambda^2 - \lambda^2 = \lambda$.

- **Theorem.** 2 r.v. X, Y are independent if and only if
$$\begin{aligned}\forall u, v : \phi_{X,Y}(u,v) &= \phi_X(u)\phi_Y(v)\\ &= E[\exp(\mathrm{i}u^T X + \mathrm{i}v^T Y)]\\ &= \phi_{\binom{X}{Y}}\binom{u}{v}\\ &= E\left[\exp\left(\mathrm{i}\binom{u}{v}^T \binom{X}{Y}\right)\right]\end{aligned}$$

- **Theorem.** Let X, Y be independent random vectors of size d. Then
$$\forall u \in \mathbb{R}^d : \phi_{X+Y}(u) = \phi_X(u)\phi_Y(u)$$

  - Example (with Poisson distribution). $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$, and X, Y are independent.
$$\begin{aligned}\phi_{X+Y}(u) &= \phi_X(u)\phi_Y(u)\\ &= E\left(e^{\mathrm{i}uX}\right)E\left(e^{\mathrm{i}uY}\right)\\ &= \exp\left(-\lambda\left(1-e^{\mathrm{i}u}\right)\right)\exp\left(-\mu\left(1-e^{\mathrm{i}u}\right)\right)\\ &= \exp\left(-(\lambda+\mu)\left(1-e^{\mathrm{i}u}\right)\right)\end{aligned}$$
    Therefore, $(X+Y) \sim \mathcal{P}(\lambda+\mu)$.

  - Other ways to show this result. Let $k \in \mathbb{Z}$.
$$\begin{aligned}P(X+Y=k) &= \sum_{i=0}^{k} P(X=i, Y=k-i)\\ &= \sum_{i=0}^{k} P(X=i)P(Y=k-i)\\ &= \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^i}{i!} \cdot \frac{e^{-\mu}\mu^{k-i}}{(k-i)!}\end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{-(\lambda+\mu)}}{k!} \underbrace{\sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!} \lambda^i \mu^{k-i}}_{=(\lambda+\mu)^k} \\
&= \frac{e^{-(\lambda+\mu)}}{k!} (\lambda+\mu)^k \\
&= \mathcal{P}(\lambda+\mu)
\end{aligned}
$$

❖ Let $X_1, \dots, X_n$ be n iid r.v. with $EX_j = \frac{1}{n}\sum_{j=1}^{n} X_j = m$ and $Var\ X_j = \Sigma$

$$
\begin{aligned}
\phi_{\bar{X}_n}(u) &= E[\exp(\mathrm{i}u^T \bar{X}_n)] \\
&= E\left[\exp\left(\mathrm{i}u^T \frac{1}{n} \sum_{j=1} X_j\right)\right] \\
&= E\left[\prod_{j=1}^{n} \exp\left(\frac{\mathrm{i}u^T}{n} X_j\right)\right] \\
&= \prod_{j=1}^{n} \underbrace{E\left[\exp\left(\frac{\mathrm{i}u^T}{n} X_j\right)\right]}_{\phi_{X_j}(u/n)}, \qquad \text{by independence} \\
&= \left[\phi_{X_j}\left(\frac{u}{n}\right)\right]^n, \qquad \text{by identically distributed}
\end{aligned}
$$

➢ This result is useful to understand the asymptotic behavior of $\bar{X}_n$:

$$
\begin{aligned}
\phi_{\sqrt{n}(\bar{X}_n - m)}(u) &= E\left[\exp\left(\mathrm{i}u^T \sqrt{n}(\bar{X}_n - m)\right)\right] \\
&= E\left[\exp\left(\mathrm{i}\frac{u}{\sqrt{n}} \sum_{j=1}^{n} (X_j - m)\right)\right] \\
&= \left[\phi_{X_j - m}\left(\frac{u}{\sqrt{n}}\right)\right]^n
\end{aligned}
$$

Deriving the Normal Distribution (cont'd)

- We have $X_1, \dots, X_n$ iid with $EX_j = m$ and $Var\ X_j = \Sigma$

$$\phi_{\bar{X}_n}(u) = \left[\phi_{X_j}\left(\frac{u}{n}\right)\right]^n$$

$$\begin{aligned}\phi_{\sqrt{n}(\bar{X}_n - m)}(u) &= E\left[\exp\left(\mathrm{i}u^T\sqrt{n}(\bar{X}_n - m)\ \right)\right] \\ &= E\left[\exp \mathrm{i}\frac{1}{\sqrt{n}}\sum_{j=1}^{n} u^T\left(X_j - m\right)\right] \\ &= \left[\phi_{u^T(X_j - m)}\left(\frac{1}{\sqrt{n}}\right)\right]^n\end{aligned}$$

where the second equality is justified by:

$$\begin{aligned}u^T\sqrt{n}(\bar{X}_n - m) &= \sqrt{n}u^T\left[\frac{1}{n}\sum_{j=1}^{n} X_j - m\right] \\ &= \sqrt{n}\cdot\frac{u^T}{n}\sum_{j=1}^{n}\left(X_j - m\right) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n} u^T\left(X_j - m\right)\end{aligned}$$

and the third equality is justified by:

$$\begin{aligned}E\left[\exp\left(\frac{\mathrm{i}}{\sqrt{n}}\sum_{j=1}^{n} u^T\left(X_j - m\right)\right)\right] &= E\left[\prod_{j-1}^{n}\exp\left(\frac{\mathrm{i}}{\sqrt{n}}\sum_{j=1}^{n} u^T\left(X_j - m\right)\right)\right] \\ &= \prod_{j=1}^{n}\underbrace{E\left[\exp\left(\frac{\mathrm{i}}{\sqrt{n}}\sum_{j=1}^{n} u^T\left(X_j - m\right)\right)\right]}_{\phi_{u^T\left(X_j - m\right)}\left(\frac{1}{\sqrt{n}}\right)}\end{aligned}$$

- So we have transformed the a function of a d-dimensional vector u into a function of a real number $1/\sqrt{n}$. Let

$$f(x) = \phi_{u^T(X_j - m)}(x)$$

Taking the Taylor expansion of $f(\cdot)$

$$f(x) \sim \left[f(0) + \underbrace{f'(0)(x - 0)}_{=0} + \frac{f''(0)(x - 0)}{2!} + \underbrace{o(x^2)}_{1/n}\right]$$

- Substitute back the original function

$$\phi_{\sqrt{n}(\bar{X}_n - m)}(u) = \left[\phi_{u^T(X_j - m)}\left(\frac{1}{\sqrt{n}}\right)\right]^n$$

$$= \left[1 + \underbrace{\phi'_{u^T(X_j-m)}(0)}_{=\mathrm{i}E\left(u^T(X_j-m)\right)=0} \frac{1}{\sqrt{n}} + \underbrace{\phi''_{u^T(X_j-m)}(0)}_{\substack{\mathrm{i}^2 E\left[\left(u^T(X_j-m)\right)^2\right] \\ =-Var\left(u^T(X_j-m)\right) \\ =-u^T\Sigma u}} \frac{1}{2n} + o\left(\frac{1}{n}\right) \right]^n$$

$$= \left[1 - \frac{u^T\Sigma u}{2n} + o\left(\frac{1}{n}\right)\right]^n$$

$$= \exp\left\{\ln\left\{\left[1 - \frac{u^T\Sigma u}{2n} + o\left(\frac{1}{n}\right)\right]^n\right\}\right\}$$

$$= \exp\left\{ n \cdot \underbrace{\ln\left\{\left[1 - \frac{u^T\Sigma u}{2n} + o\left(\frac{1}{n}\right)\right]\right\}}_{\approx -\frac{u^T\Sigma u}{2n}} \right\}$$

$$\approx \exp\left(-\frac{u^T\Sigma u}{2}\right)$$

- Conclusion. For any X_j iid with $EX_j = m$ and $Var\, X_j = \Sigma$,

$$\phi_{\sqrt{n}(\bar{X}_n - m)}(u) \underset{n\to\infty}{\longrightarrow} \exp\left(-\frac{u^T\Sigma u}{2}\right)$$

  - Question 1: What does it mean to have $\phi_{Y_n}(u) \to \phi_Y(u)$?
    - Convergence in distribution
  - Question 2: What is Y when $\phi_Y(u) = \exp\left(-\frac{u^T\Sigma u}{2}\right)$?
    - Normal r.v.

- *Definition.* For r.v. X_n in $\mathbb{R}^d$ and X

$$X_n \xrightarrow{d} X \iff \forall u \in \mathbb{R}^d : \phi_{X_n}(u) \xrightarrow{n\to\infty} \phi_X(u)$$

- **Theorem.** $X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$.

- Recall that

$$L^2 \;\Rightarrow\; a.s. \;\Rightarrow\; P \;\Rightarrow\; d$$

- Example. $X_i \xrightarrow{d} X_1$, but

$$P(|X_n - X_1| > \epsilon) = P(|X_2 - X_1| > \epsilon) \nrightarrow 0$$

  where the equality is justified by

$$P^{(X_1,X_n)} = P^{(X_1 \otimes X_n)} = P^{X_1} \otimes P^{X_n} = P^{X_1} \otimes P^{X_2} = P^{(X_1,X_2)}$$

- Convergence in distribution is about $E[h(X_n)] \to E[h(X)]$ as long as h is continuous. But $P(X \in A) = E[\mathbf{1}_A(X)]$ is not continuous.

- ***Central Limit Theorem.*** Suppose X_i's are iid with $EX_i = \mu < \infty$ and $Var(X_i) = \sigma < \infty$.

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1).$$

Let

$$S_n \coloneqq X_1 + \cdots + X_n$$

$$Z_n \coloneqq \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}, \qquad where\ \bar{X}_n \coloneqq \frac{S_n}{n}$$

$$Y_i \coloneqq \frac{\bar{X}_i - \mu}{\sigma}.$$

Then,

$$Z_n = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$$

http://en.wikipedia.org/wiki/Central_limit_theorem#Proof

Convergence in Distribution

- *Definition.* X_n, X are r.v. in $\mathbb{R}^d$

$$X_n \xrightarrow{d} X \iff \forall u \in \mathbb{R}^d : \phi_{X_n}(u) \underset{n\to\infty}{\longrightarrow} \phi_X(u)$$

- **Theorem.** $X \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$.
  - **Lemma.** $X_n \xrightarrow{d} X \iff \forall u \in \mathbb{R}^d : u^T X_n \xrightarrow{d} u^T X$

  Recall that

$$X_n \xrightarrow{p} X \iff \forall \epsilon > 0, \exists n_0 \in \mathbb{N}, \forall n \geq n_0, \forall \eta > 0 : P(|X_n - X| > \epsilon) < \eta$$

  Want to show that the distance of two characteristic functions goes to zero:

$$\begin{aligned}
&\left|\phi_{X_n}(u) - \phi_X(u)\right| = \left|E\left(e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right)\right| \\
&\leq E\left[\left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right|\right] \\
&= \int \left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| dP^X \\
&= \int_{|X_n - X| \leq \epsilon} \left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| dP^X + \int_{|X_n - X| > \epsilon} \left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| dP^X
\end{aligned}$$

  Note that in the second term,

$$\begin{aligned}
&\left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| \leq \left|e^{\mathrm{i}uX_n}\right| + \left|e^{\mathrm{i}uX}\right| \leq 2 \\
\Rightarrow &\int_{A_n} \left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| dP^X \leq \int_{A_n} 2 dP^X \\
\Leftrightarrow &\int_{A_n} \left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| dP^X \leq 2 \int_{A_n} \mathbf{1}_{A_n} dP = 2P(A_n)
\end{aligned}$$

  where

$$A_n = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}$$

  For the first term, since $e^{\mathrm{i}uX}$ is continuous

$$\left|e^{\mathrm{i}uX_n} - e^{\mathrm{i}uX}\right| \leq \alpha(\epsilon) P(|X_n - X| \leq \epsilon)$$

- In the end, what we have "shown" is

$$\left|\phi_{X_n}(u) - \phi_X(u)\right| < \eta$$

  - To prove convergence, separate the set into two: one that has probability zero, the other that doesn't have probability zero, but get a bound for the thing that's inside the intergral.

- **Theorem.** Characterization of convergence in distribution:

$$X_n \xrightarrow{d} X \iff E[h(X_n)] \to E[h(X)]$$

  for any *continuous* and *bounded* function h.

- **Theorem.**

$$X_n \xrightarrow{d} X \iff F_{X_n}(x) \to F_X(x)$$

  for any x where $F_X(x)$ is continuous.

  - Example. Let $X_n \sim \delta_{\frac{1}{n}}$ and $X \sim \delta_0$. Consider $\delta_{\frac{1}{n}} \to \delta_0$.

$$X_n \sim \delta_{\frac{1}{n}} \;\; \Rightarrow \;\; F_{X_n}(x) = \delta_{\frac{1}{n}}\big((-\infty, x]\big) = \begin{cases} 1 & \text{if } \dfrac{1}{n} \le x \\ 0 & \text{if } \dfrac{1}{n} > x \end{cases}$$

Thus,

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 1 & \text{if } 0 < x \\ 0 & \text{if } 0 \ge x \end{cases}$$

But

$$F_X(x) = \delta_0\big((-\infty, x]\big) = \begin{cases} 1 & \text{if } 0 \le x \\ 0 & \text{if } 0 > x \end{cases}$$

Therefore, F_{X_n} converges to F_X everywhere except at point 0.

- Note. If $f_{X_n} \xrightarrow{a.s.} f_X$, then $X_n \xrightarrow{d} X$.

Normal Distribution

- What we have done so far is to consider

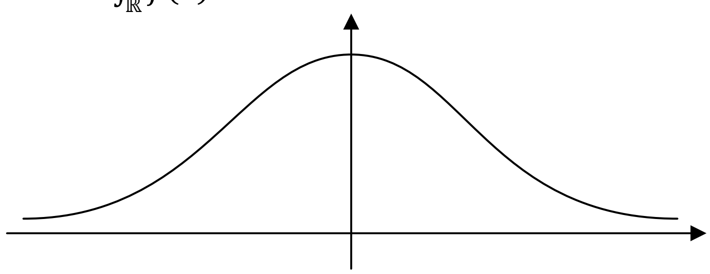
$$X_i \sim iid : EX_i = m \ \& \ VarX_i = \Sigma$$

$$\sqrt{n}(\bar{X}_n - m) \xrightarrow{d} Y \quad such\ that \quad \phi_Y(u) = \exp\left(-\frac{u^T \Sigma u}{2}\right)$$

- *Definition.* The ***standard normal distribution*** $\mathcal{N}(0,1)$ with density function

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

  - It is not easy to define in close form $F(x)$
  - Difficult to show that $\int_{\mathbb{R}} f(x)dx = 1$



- *Definition.* Normal distribution $\mathcal{N}(m, \sigma^2)$ where $m \in \mathbb{R}$ and $\sigma^2 \in \overline{\mathbb{R}}_+ = (0, \infty)$

$$\begin{aligned} Y \sim \mathcal{N}(m, \sigma^2) \ &\Leftrightarrow \ Y = m + \sigma X \\ &\Leftrightarrow \ F_Y(y) = \Phi\left(\frac{y-m}{\sigma}\right) \end{aligned}$$

where $X \sim \mathcal{N}(0,1)$, and

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

- Moment generating function

$$\begin{aligned} L_Y(u) &= E[\exp(uY)] \\ &= E[\exp(um + u\sigma X)] \\ &= \exp(um) \cdot E(\exp(u\sigma X)) \\ &= \exp(um) \cdot L_{\sigma X}(u) \\ &= \exp(um) \cdot L_X(\sigma u) \end{aligned}$$

Here,

$$\begin{aligned} L_X(u) &= E[\exp(uX)] \\ &= \int_{-\infty}^{+\infty} e^{ux} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2} + \frac{u^2}{2}} dx \\ &= e^{\frac{u^2}{2}} \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2}} dx}_{=1} \\ &= e^{\frac{u^2}{2}} \end{aligned}$$

- Moments.

$$\frac{\partial L_Y(u)}{\partial u} = (m + u\sigma^2)\exp\left(um + \frac{u^2\sigma^2}{2}\right) \;\Rightarrow\; \left.\frac{\partial L_Y(u)}{\partial u}\right|_{u=0} = m = EY$$

$$\frac{\partial^2 L_Y(u)}{\partial u^2} = [(m + u\sigma^2)^2 + \sigma^2]\exp\left(um + \frac{u^2\sigma^2}{2}\right) \;\Rightarrow\; \left.\frac{\partial^2 L_Y(u)}{\partial u^2}\right|_{u=0} = m^2 + \sigma^2 = EY^2$$

Therefore,

$$VarY = E(Y^2) - (EY)^2 = \sigma^2$$

$\boldsymbol{d}$-Dimensional Normal Distribution

- *Definition.* Let X be a random vector in $\mathbb{R}^d$.
$$X \text{ is a normal vector } \Leftrightarrow \forall u \in \mathbb{R}^d : u^T X \sim \mathcal{N}.$$

- If $EX = m$ and $VarX = \Sigma$. We have
$$E(u^T X) = u^T m, \qquad Var(u^T X) = u^T \Sigma u.$$
$$\underbrace{L_X(u)}_{E(\exp(u^T X))} = L_{u^T X}(1) = \exp\left(1 \cdot E(u^T X) + 1 \cdot \frac{Var(u^T X)}{2}\right) = \exp\left(u^T m + \frac{u^T \Sigma u}{2}\right)$$

- Thus,
$$X \sim \mathcal{N}(m, \Sigma) \Leftrightarrow L_X(u) = \exp\left(u^T m + \frac{u^T \Sigma u}{2}\right) \Leftrightarrow \varphi_X(u) = \exp\left(\mathrm{i} u^T m - \frac{u^T \Sigma u}{2}\right)$$

- We can also show that
$$f_X(x) = \frac{\exp\left[-\frac{1}{2}(x-m)^T \Sigma (x-m)\right]}{(2\pi)^{d/2} (\det \Sigma)^{1/2}}$$
as long as $\det \Sigma \neq 0$.
  - In dimension 1, we have:
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \qquad \forall x \in \mathbb{R}$$

- Recall that: If X_i iid with $EX_i = m$ and $VarX_i = \Sigma$, then
$$\sqrt{n}(\bar{X}_n - m) \xrightarrow{d} Y \text{ with } \varphi_Y(u) = \exp\left(-\frac{u^T \Sigma u}{2}\right)$$
Thus, $Y \sim \mathcal{N}(0, \Sigma)$.

- **Central Limit Theorem.**
  - Let X_i be iid with $EX_i = m$ and $VarX_i = \Sigma$. Then,
$$\sqrt{n}(\bar{X}_n - m) \xrightarrow{d} \mathcal{N}(0, \Sigma) \text{ and } \bar{X}_n \to \mathcal{N}\left(m, \frac{\Sigma}{n}\right)$$

- Consider 2 r.v. X, Y
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(m, \Sigma) \Leftrightarrow \varphi_{\begin{pmatrix} X \\ Y \end{pmatrix}} \begin{pmatrix} u \\ v \end{pmatrix} = \exp\left(\mathrm{i} \begin{pmatrix} u \\ v \end{pmatrix}^T m - \frac{1}{2} \begin{pmatrix} u \\ v \end{pmatrix}^T \Sigma \begin{pmatrix} u \\ v \end{pmatrix}\right)$$

  Suppose X, Y are not correlated. This is true if and only if
$$Cov(X, Y) = 0 \Leftrightarrow \Sigma = \begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{bmatrix}$$
  where $\Sigma_X = VarX$ and $\Sigma_Y = VarY$. Then,
$$\begin{pmatrix} u \\ v \end{pmatrix}^T \Sigma \begin{pmatrix} u \\ v \end{pmatrix} = u^T \Sigma_X v + v^T \Sigma_Y v$$
  Then,

$$\varphi_{\binom{u}{v}}\binom{u}{v} = \exp\left(\mathrm{i}u^T m_X - \frac{u^T \Sigma_X v}{2}\right) \cdot \exp\left(\mathrm{i}v^T m_Y - \frac{v^T \Sigma_Y v}{2}\right) = \varphi_X(u)\varphi_Y(v)$$

where

$$m = \binom{m_x}{m_Y} = \binom{EX}{EY}$$

This implies that X, Y are independent.

In general,

$$\binom{X}{Y} \sim \mathcal{N} \;\; \Rightarrow \;\; Cov(X,Y) = 0 \;\; \Leftrightarrow \;\; X, Y \text{ are independent}$$

- Transformation of r.v.
  - Let X be r.v. in $\mathbb{R}^d$.

    $$g : X \to Y$$

    where g is bijective.

    $$\begin{aligned} E(Y) = E[g(X)] &= \int g(x) f_X(x) dx \\ &= \int g\, \underbrace{[g^{-1}(y)]}_{=x} f_Y(y) dy \\ &= \int g(x) f_Y\big(g(x)\big) |J_{g(x)}| dx \end{aligned}$$

    Here we have

    $$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|J_{g^{-1}(y)}\right|$$