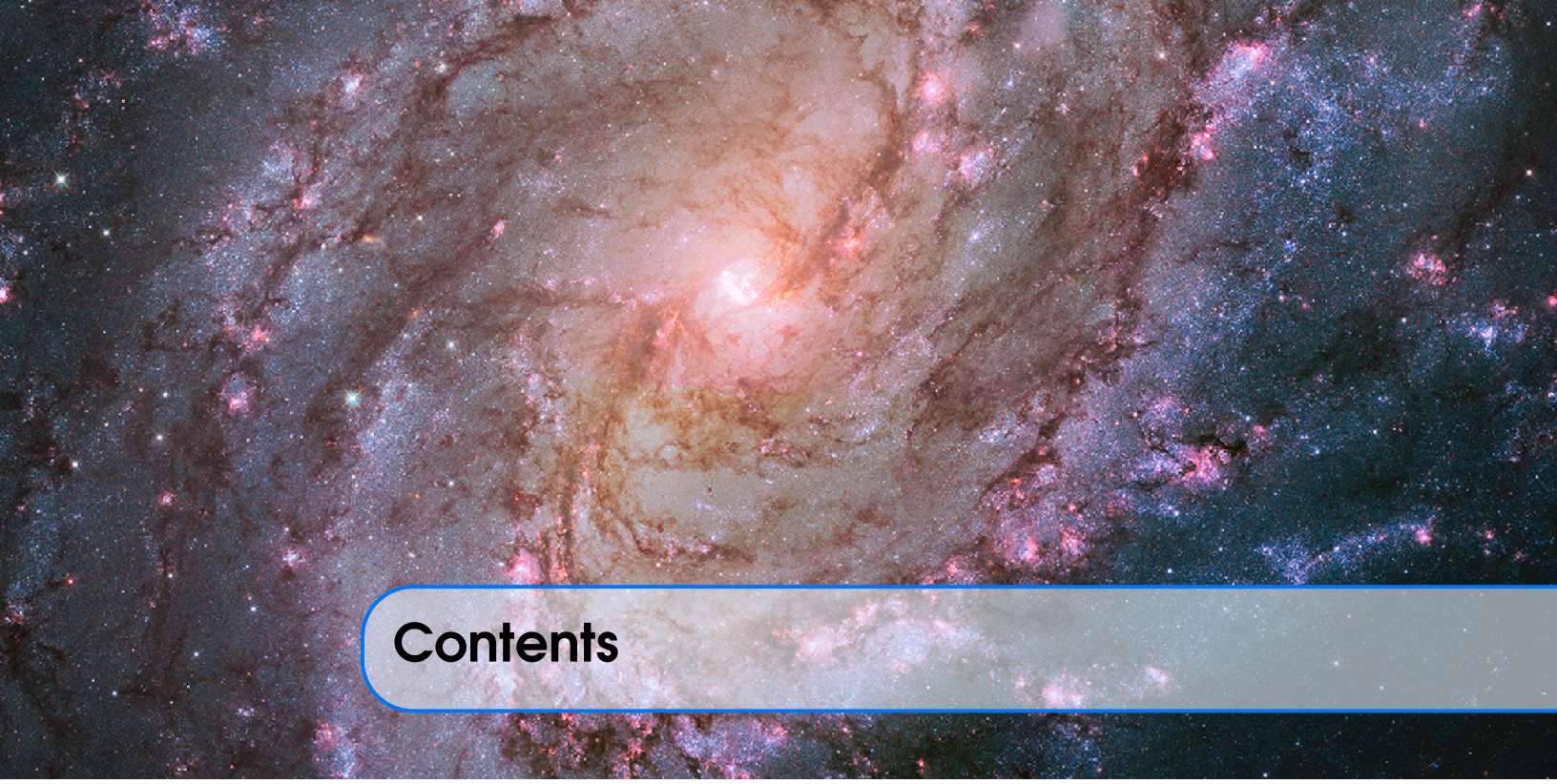
STAT 330 SPRING 2021

Mathematical Statistics

Instructor: Lucy Gao

Lecture Notes

Latex by Justin Li



Contents

1. (Univariate) Random Variables

1.1 Probability Models and random experiments

Random Experiment:

1. Outcome is random
2. (Theoretically) repeatable

■ **Example 1.1**
1. Connecting raw data (eg. heights of 30 randomly selected students)
2. Summary of raw data (eg. mean of heights of randomly selected students) ■

Definition 1.1.1 — Probability Model (describes a random experiment).
1. Sample Space S - Set of all possible experiments
2. Events A - Subset of the sample space
3. Probability function $P(A)$ satisfying:
(a) $P(A) \geq 0 \ \ \forall A$
(b) $P(S) = 1$
(c) If we have $A_1, A_2, A_3, ...$, that are mutually exclusive ($A_i \cap A_J = \emptyset, \ \forall i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Proposition 1.1.1 — Properties of Probability Function P.
Let A, B be events in sample space S, then

1. $P(\emptyset) = 0$

2. A and B mutually exclusive $\implies P(A \cup B) = P(A) + P(B)$. Generally, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. $P(A \cap \overline{B}) = P(A) - P(A \cap B)$
4. $P(\overline{A} = 1 - P(A)$
5. $A \subseteq B \implies P(A) \leq P(B)$
6. $0 \leq P(A) \leq 1$

Definition 1.1.2 — Conditional Probability.
Let A, B be events with $P(B) > 0$, then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Definition 1.1.3 — Independence.
Let A, B be events, we say $A \perp\!\!\!\perp B$ (independent) if

$$P(A \cap B) = P(A)P(B)$$

and we have $P(A \mid B) = P(A)$, $P(B \mid A) = P(A)$

■ **Example 1.2** Flip Coin Twice - Random Experiment
Sample Space $S = \{(H,H),(H,T),(T,H),(T,T)\}$. Event: "First coin up heads" = A, "Second coin up Tail" = A

$$A = \{(H,T),(H,H)\} \subseteq S$$

Probability function

$$P(A) = \frac{|A|}{|S|} = \frac{|A|}{4} \qquad P(S) = \frac{|S|}{|S|} = 1$$

It's easy to see $P(A) = P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{4}$. Then $P(A \cap B) = P(A)P(B)$, so $A \perp\!\!\!\perp B$ ■

Definition 1.1.4 — Random Variable.
$X : S \to \mathbb{R}$ satisfying

$$\{X \leq x\} = \{A \in S : X(A) \leq x\}$$

is a valid event for all $x \in \mathbb{R}$

Definition 1.1.5 — Cumulative Distribution Function (CDF).
Let X be a random variable, then the CDF of X is defined as

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$$

Proposition 1.1.2 — Properties of CDFs.
1. F is a non-decreasing function: $x_1 \leq x_2 \implies F(x_1) \leq F(x_2)$
2. $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$
3. $F(x)$ is right continuous: $\forall a \in \mathbb{R}$, $\lim_{x \to a^+} F(x) = F(a)$
4. $\forall a < b$, $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

5. $\forall a \in \mathbb{R}, P(X=a) = \lim_{x\to a^+} F(x) - \lim_{x\to a^-} F(x) = F(a) - \lim_{x\to a^-} F(x)$

Definition 1.1.6 — Discrete Variables.
Finite or countable number of values with positive probability. If there exists $A \subseteq \mathbb{R}$ that is finite or countable and $P(X \in A) = 1$, then X is discrete random variable.

probability mass function: $f(x) = P(X = x)$, support $A = \{x : f(x) > 0\}$.

Properties of pmf:

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\sum_{x\in A} f(x) = 1$

pmf to cdf:

$$F(x) = \sum_{y\in A, y\leq x} f(y) = \sum_{y\in A, y\leq x} P(X=y) = P\left(\bigcup_{y\in A, y\leq x} \{X=y\}\right) = P(\{X\leq x\}\cap\{X\in A\}) = P(X\leq x)$$

cdf to pmf:

$$f(x) = P(X=a) = F(a) - \lim_{t\to a^-} F(t)$$

Definition 1.1.7 — Bernoulli Distribution.
$X \sim Bernoulli(p)$, $X \in \{0,1\}$, $P(X=1) = p$ and $P(X=0) = 1-p$, then

$$f(x) = \begin{cases} P(X=x) & x \in \{0,1\} \\ 0 & \textbf{otherwise} \end{cases} = \begin{cases} p^x(1-p)^{1-x} & x \in \{0,1\} \\ 0 & \textbf{otherwise} \end{cases}$$

Definition 1.1.8 — Binomial Distribution.
1. n experiments
2. each experiment is independent
3. each experiment has 2 outcomes: 0 (prob=$1-p$) or 1 (prob=p)

Let X be the number of experiments with outcome 1, so $X \sim Bin(n,p)$, then for $X_i \sim Bernoulli(p)$, $X = \sum_{i=1}^{\infty} X_i$ and

$$f(x)\cdot P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Definition 1.1.9 — Geometric Distribution.
Let X be the number of outcomes before the first 1 outcome in repeated Bernoulli random trials
For example, let X be number of trials before first head, so $x = 0,1,2...$, then

$$f(x)\cdot P(X=x) = (1-p)^x p \quad x = 0,1,2.....$$

$f(x) = 0$ otherwise.

Definition 1.1.10 — Negative Binomial Distribution.
Let X be the 0 outcomes before the r^{th} outcome of 1, so $X \sim NegBin(r,p)$ in repeated Bernoulli(p) experiments. For $x = 0,1,2....$,

$$f(x) = P(X = x) = \binom{x+r-1}{x}(1-p)^x p^r$$

and $X = \sum_{i=1}^{r} X_i$ with $X_i \sim Geo(p)$

Definition 1.1.11 — Poisson Distribution.
Let $X \sim Poi(\mu)$, then for $x = 0,1,2.....$ we have

$$f(x) = P(X = x) = \frac{\mu^x}{x!}e^{-\mu}$$

otherwise we have $f(x) = 0$

Definition 1.1.12 — Continuous Variables.
If X is a random variable with CDF $F(x)$ s.t.

1. $F(x)$ is continuous at x for all $x \in \mathbb{R}$
2. $F(x)$ is differentiable everywhere on $\mathbb{R}$ **except** at countably many points

Then X is a continuous random variable

probability density function: $f(x) = F'(x)$ where F is differentiable, support set $A = \{x : f(x) > 0\}$

Properties of pdf:

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)dx = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1 - 0 = 1$
3. $f(x) = F'(x)$ if the derivative exists
4. $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
5. $P(a < x \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt = \int_{a}^{b} f(t)dt$
6. $P(X = b) = 0$

■ **Example 1.3**

$$F(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases} \quad \Longrightarrow \quad f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

■

■ **Example 1.4**
Let the pdf be

$$f(x) = \begin{cases} \frac{1}{x^2} & x \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. check if this pdf valid: 1. $f(x) \geq 0$ for all x and

$$\int_{\mathbb{R}} f(t)dt = \int_1^{\infty} \frac{1}{x^2}dx = \left[-\frac{1}{x}\right]_1^{\infty} = 1$$

2. Find the cdf:

$$P(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = 0 \quad x < 1$$

and

$$P(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = \int_1^{x} f(x)dx = 1 - \frac{1}{x} \quad x \geq 1$$

Then we have

$$F(x) = \begin{cases} 1 - \frac{1}{x} & x \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

3.

$$P(-2 < X < 3) = F(3) - F(-2) = 1 - \frac{1}{3} = \frac{2}{3} = \int_{-2}^{3} f(x)dx = \left[-\frac{1}{x}\right]_{-2}^{3}$$

■

Definition 1.1.13 — Gamma Function.

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy\,, \quad \alpha > 0$$

Proposition 1.1.3 — Properties of Gamma Function.

1. $\Gamma(\alpha) = (\alpha-1)\cdot\Gamma(\alpha-1)$ for $\alpha > 1$
2. $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

■ **Example 1.5** Let $X \sim N(0,1)$ and define

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}}$$

so let $Y = -\frac{X^2}{2}$, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dX = 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dX &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-Y} \frac{\sqrt{2}}{2} Y^{-\frac{1}{2}} dY \\
&= \frac{1}{\sqrt{\pi}} \underbrace{\int_0^{\infty} Y^{\frac{1}{2}-1} e^{-Y} dY}_{\Gamma(\frac{1}{2})} \\
&= \frac{1}{\sqrt{\pi}} \cdot \sqrt{\pi} \\
&= 1
\end{aligned}$$

■

■ **Example 1.6** Let $X \sim N(\mu, \sigma^2)$ and define

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

so let $Z = \frac{X-\mu}{\sigma}$, we have

$$\int_{-\infty}^{\infty} f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} dX = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{Z^2}{2}} \sqrt{\sigma^2} dZ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} dZ = 1$$

■

1.2 Expectation

Definition 1.2.1 — Expectation.
Let X be a discrete random variable with support A and pmf $f(x)$, then if

$$\sum_{x \in A} |x| \cdot f(x) = \infty$$

we say that $E[X]$ is **DNE**, otherwise we have

$$E[X] = \sum_{x \in A} x \cdot f(x)$$

Similarly, let X be a continuous random variable with pdf $f(x)$, then if

$$\int_{-\infty}^{\infty} |x| \cdot f(x) dx = \infty$$

we say that $E[X]$ is **DNE**, otherwise we have

$$E[x] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

■ **Example 1.7** Let X be a random variable with pmf

$$f(x) = \begin{cases} \frac{1}{x(x+1)} & \textbf{if } x \in \mathbb{N} \\ 0 & \textbf{otherwise} \end{cases}$$

Then

$$\sum_{x \in A} |x| f(x) = \sum_{x=1}^{\infty} x \cdot \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = \infty$$

so $E[X]$ **DNE** ■

■ **Example 1.8** Let X be a random variable with pmf

$$f(x) = \begin{cases} \frac{\theta}{x^{\theta+1}} & \textbf{if } x \geq 1 \\ 0 & \textbf{otherwise} \end{cases}$$

for $\theta > 0$. Then we have

$$\int_{-\infty}^{\infty} |x| f(x)dx = \int_{1}^{\infty} x \cdot \frac{\theta}{x^{\theta+1}} = \theta \underbrace{\int_{1}^{\infty} \frac{1}{x^{\theta}}}_{<\infty \text{ iff } \theta > 0} dx$$

Then $E[X]$ exists iff $\theta > 1$.
For $\theta > 1$,

$$E[X] = \int_{-\infty}^{\infty} |x| f(x)dx = \int_{1}^{\infty} x \cdot \frac{\theta}{x^{\theta+1}} = \theta \cdot \int_{1}^{\infty} \frac{1}{x^{\theta}} dx = \theta \cdot \left(\frac{1}{\theta - 1} \right) = \frac{\theta}{\theta - 1}$$

■

Proposition 1.2.1
Let X be a discrete random variable, then

$$E[g(X)] = \sum_{x \in A} g(x) f(x)$$

provided it exists (i.e. $\sum_{x \in A} |g(x)| f(x) < \infty$)

Similarly, if X is a continuous random variable, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)dx$$

provided it exists (i.e. $\int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty$)

Proposition 1.2.2 — Linearity of Expectation.
1. For all $a,b \in \mathbb{R}$, $E[aX+b] = aE[X]+b$
2. For all $a,b \in \mathbb{R}$, $E[ag(X)+bh(X)] = aE[g(X)]+bE[h(X)]$

Definition 1.2.2 — Variance.
Let X be a random variable, then the variance of X is defined by

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

Definition 1.2.3 — k^{th} moment of X.
Let X be a random variable, then the k^{th} moment of X is defined by

$$E[(X_\mu)^k]$$

for $k \in \mathbb{N}$ where $\mu = E[X]$

Proposition 1.2.3

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

Proof: Let $\mu = E[X]$, then

$$\text{Var}[X] = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - (E[X])^2$$

as desired.

■ **Example 1.9** Let $X \sim Unif(0,1)$, so pdf

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

then we have $E[X] = \frac{1}{2}$ so that $E[2X+1] = 2E[X]+1 = 2$. Note that

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^2 dx = \frac{1}{3}$$

this gives us that $\text{Var}[X] = E[X^2] - (E[X])^2 = \frac{1}{12}$ ■

1.3 Moment Generating Functions

Definition 1.3.1
Let X be a random variable, if $\exists h > 0$ s.t. $E[e^{tx}]$ exists for all $t \in (-h,h)$, then the **moment generating function** of X exists and

$$M(t) = E[e^{tx}] \quad \forall t \text{ s.t. } E[e^{tx}] \text{ exists}$$

■ **Example 1.10** Let $X \sim Exp(\theta)$, so pdf

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \textbf{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

for $\theta > 0$. Then we have

$$\int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_0^{\infty} e^{tx} \frac{1}{\theta} e^{-\frac{x}{\theta}} dx = \underbrace{\frac{1}{\theta} \int_0^{\infty} e^{-x(\frac{1}{\theta}-t)} dx}_{\frac{1}{\theta}-t>0 \iff t<\frac{1}{\theta}}$$

and we see that

$$t < \frac{1}{\theta} \iff \textbf{integral converge} \iff E[e^{tx}] \textbf{ exists}$$

so h can be $h = \frac{1}{k\theta}$ for $k \in \mathbb{N}$. This gives us that

$$\begin{aligned} M(t) &= \frac{1}{\theta} \int_0^{\infty} e^{-x(\frac{1}{\theta}-t)} dx \qquad \forall\ t < \frac{1}{\theta} \\ &= \frac{1}{\theta} \left[-\frac{1}{\frac{1}{\theta}-t} e^{-x(\frac{1}{\theta}-t)} \right]_0^{\infty} \qquad \forall\ t < \frac{1}{\theta} \\ &= \frac{1}{1-t\theta} \qquad \forall\ t < \frac{1}{\theta} \end{aligned}$$

■

Proposition 1.3.1
If the **MGF** of X exists and its domain is τ, then the **MGF** of $Y = aX + b$ exists for all $a, b \in \mathbb{R}$ with $a \neq 0$ and

$$M_Y(t) = e^{bt} M_X(at) \quad \forall\ t \in \{t \in \mathbb{R} : at \in \tau\}$$

Proposition 1.3.2
If the **MGF** of X exists, then

$$M_X(0) = 1 \qquad \textbf{and} \qquad M_X^{(k)}(0) = E[X^k]$$

for $k \in \mathbb{N}$ where

$$M_X^{(k)}(t) = \frac{d^k}{dt^k} M_X(t)$$

■ **Example 1.11** Let $X \sim Exp(\theta)$, so we have

$$M_X(t) = \frac{1}{1-t\theta} \qquad \forall\ t < \frac{1}{\theta}$$

so that

$$E[X] = M_X'(0) = \frac{-1}{(1-t\theta)^2} \cdot (-\theta)\Big|_{t=0} = \frac{1}{\theta}$$

■

Proposition 1.3.3 Let X, Y be random variables with $M_X = M_Y \iff X$ and Y have the same distribution (i.e. their CDFs are equal)

2. Multivariate Random Variables - I

2.1 Joint and Marginal CDFs

Definition 2.1.1 — Joint CDF.
Suppose X and Y are random variables defined on sample space S. Then the **joint CDF of X and Y** is defined by

$$F(x,y) = P(X \leq x, Y \leq y) \qquad \forall \ (x,y) \in \mathbb{R}^2$$

Similarly, let $X_1, X_2, ..., X_n$ be random variables on sample space S, then the **joint CDF of these variables is**

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n) \qquad \forall \ x \in \mathbb{R}^n$$

Proposition 2.1.1 — Properties of Joint CDF F(x,y).
1. F is non-decreasing in x: Fix y, $\forall\ x_1 < x_2$, we have $F(x_1,y) \leq F(x_2,y)$
2. F is non-decreasing in y: Fix x, $\forall\ y_1 < y_2$, we have $F(x,y_1) \leq F(x,y_2)$
3.

$$\lim_{x \to -\infty} F(x,y) = 0 \qquad \textbf{and} \qquad \lim_{y \to -\infty} F(x,y) = 0$$

4.

$$\lim_{(x,y) \to (-\infty,-\infty)} F(x,y) = 0 \qquad \textbf{and} \qquad \lim_{(x,y) \to (\infty,\infty)} F(x,y) = 1$$

Definition 2.1.2 — Marginal CDF of X,Y in $F(x,y)$.
The **marginal CDF of X**

$$\underbrace{F_1(x)}_{F_X(x)} = \lim_{y \to \infty} F(x,y) = P(X \leq x) \qquad \forall \ x \in \mathbb{R}$$

and the **marginal CDF of Y**

$$\underbrace{F_2(y)}_{F_Y(y)} = \lim_{x\to\infty} F(x,y) = P(Y \le y) \quad \forall\ y \in \mathbb{R}$$

Note Given joint CDFs, we can find marginal CDFs **but** given marginal CDFs we **cannot** find joint CDFs.

It's possible to have (X_1, Y_1) and (X_2, Y_2) s.t.

$$F_{X_1}(x) = F_{X_2}(x) \quad \textbf{and} \quad F_{Y_1}(y) = F_{Y_2}(y) \qquad \textbf{but} \qquad F_{X_1,Y_1}(x,y) \neq F_{X_2,Y_2}(x,y)$$

2.2 Bivariate Discrete Distributions

Definition 2.2.1 — Bivariate discrete random variables.
Let X, Y be random variables on sample space S. IF $\exists A \subseteq \mathbb{R}^2$ s.t. A is countable and $P((x,y) \in A) = 1$, then X, Y are a pair of bivariate discrete random variables.

The joint probability (mass) function (joint pf/pmf)

$$f(x,y) = P(X = x, Y = y) \quad \forall\ (x,y) \in \mathbb{R}^2$$

and the joint support of (x,y) is

$$A = \{(x,y) \in \mathbb{R}^2 : f(x,y) > 0\}$$

Proposition 2.2.1 — Properties of f(x,y).
1. $f(x,y) \ge 0\ \forall (x,y) \in \mathbb{R}^2$
2. $\sum_{(x,y)\in A} f(x,y) = 1$
3. For $R \subseteq \mathbb{R}^2$, $P((x,y) \in R) = \sum_{(x,y)\in R} f(x,y)$

■ **Example 2.1**

$$P(X \le Y) = P((x,y) \in R) \qquad \textbf{for} \quad R = \{(x,y) \in \mathbb{R}^2 : x \le y\}$$

and

$$P(X + Y \le 1) = P((x,y) \in R) \qquad \textbf{for} \quad R = \{(x,y) \in \mathbb{R}^2 : x + y \le 1\}$$

■

Definition 2.2.2 — Joint pmf $f(x,y)$ for (X,Y) random variables.
The marginal probability function of X:

$$\underbrace{f_1(x)}_{f_X(x)} = P(X = x) = \sum_{\textbf{all y}} f(x,y) \quad \forall\ x \in \mathbb{R}$$

and marginal probability function of Y

$$\underbrace{f_2(y)}_{f_Y(y)} = P(Y = y) = \sum_{\textbf{all x}} f(x,y) \quad \forall\ y \in \mathbb{R}$$

*** want to get the marginal from joint, sum out variable you don't want**

■ **Example 2.2** Let X, Y be discrete random variables with joint pmf

$$f(x,y) = \begin{cases} k(1-p)^2 p^{x+y} & \textbf{if } x,y \in \mathbb{N} \cup \{0\} \\ 0 & \text{otherwise} \end{cases}$$

with $0 < p < 1$.

1. Find k
2. Find marginal pmfs
3. Find $P(X \leq Y)$

Solution:
1. First, we note that $f(x,y) \geq 0 \implies k \geq 0$ and

$$\begin{aligned} \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} f(x,y) = 1 \quad &\implies \quad k\sum_{x=0}^{\infty}\sum_{y=0}^{\infty}(1-p)^2 p^x p^y = 1 \\ &\implies \quad k(1-p)^2\left(\sum_{x=0}^{\infty} p^x\right)\left(\sum_{y=0}^{\infty} p^y\right) = 1 \\ &\implies \quad k(1-p)^2 \cdot \frac{1}{1-p} \cdot \frac{1}{1-p} = 1 \\ &\implies \quad k = 1 \end{aligned}$$

2.

$$f_1(x) = \sum_{y=0}^{\infty}(1-p)^2 p^{x+y} = (1-p)^2 p^x \cdot \sum_{y=0}^{\infty} p^y = (1-p)p^x \quad \textbf{for } x \in \mathbb{N} \cup \{0\}$$

Similarly,

$$f_2(y) = \sum_{x=0}^{\infty}(1-p)^2 p^{x+y} = (1-p)^2 p^y \cdot \sum_{x=0}^{\infty} p^x = (1-p)p^y \quad \textbf{for } y \in \mathbb{N} \cup \{0\}$$

3. Note that

$$P(X \leq Y) = \sum_{x=0}^{\infty}\sum_{y=x}^{\infty}(1-p)^2 p^{x+y} = (1-p)^2 \sum_{x=0}^{\infty} p^x \sum_{y=x}^{\infty} p^y$$

and

$$\sum_{y=x}^{\infty} p^y = p^x \sum_{y=0}^{\infty} p^y = \frac{p^x}{1-p}$$

Then we have

$$P(X \leq Y) = \frac{(1-p)^2}{1-p}\sum_{x=0}^{\infty} p^{2x} = (1-p) \cdot \frac{1}{1-p^2} = \frac{1}{1+p}$$

■

2.3 Bivariate continuous random variables

Definition 2.3.1 — Bivariate continuous random variables.
If $F(x,y)$ is continuous and $\frac{\partial^2}{\partial x \partial y} F(x,y)$ exists and is continuous except perhaps only finite number of points. Then we say that X, Y are **bivariate continuous random variables** and we define **joint pdf** (prob density function) to be

$$f(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} & \textbf{when it exists} \\ 0 & \text{otherwise} \end{cases}$$

and the joint support is

$$A = \{(x,y) \in \mathbb{R}^2 : f(x,y) > 0\}$$

Proposition 2.3.1 — Properties of joint pdf.
1. $f(x,y) \geq 0 \; \forall \; (x,y) \in \mathbb{R}^2$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) = 1$
3. $P((x,y) \in R) = \int_R f(x,y) dx dy$ for $R \subseteq \mathbb{R}^2$

■ **Example 2.3**

$$P(X \leq Y) = \int_{x \leq y} f(x,y) dx dy \quad \textbf{for} \; R \subseteq \mathbb{R}^2$$

■

2.4 Appendix 1 (provided by Lucy Gao)

When Can You Take Functions Outside Expectation?

When Can You Take Functions Outside Expectation?

Introduction

Let X be a random variable with $E[X] = \mu$. A useful property of expectation is *linearity*, in the sense that if g is a linear function defined by $g(x) = ax + b$ for some constant real numbers a and b, then

$$E[g(X)] = E[ax + b] = aE[x] + b = g(E[X]).$$

We will now discuss an important *non-property* of expectation. A non-property is a statement that looks true that you may be tempted to use, but actually is *not* true. The reason that it looks true is usually due to misapplication of analogy and/or guessing. The non-property is as follows:

$$E[g(X)] = g(E[X]), \text{ for any non-linear function } g. \tag{1}$$

This non-property is *not always true.* It is *not always false.* It is *sometimes true and sometimes false.*

Examples

We will demonstrate this through a series of examples.

Example 1 *Let $X \in \{1, 2\}$ with $P(X = 1) = P(X = 2) = \frac{1}{2}$. Let $g(x) = \frac{1}{x}$.*

We can find $E[X]$ in at least two ways. The first is by definition:

$$E[X] = \sum_{x \in \{1,2\}} xP(X = x) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2}.$$

The second way is to notice that if $Y \sim Bernoulli(1/2)$, then $Y+1$ and X have the same distribution. Therefore,

$$E[X] = E[Y + 1] = E[Y] + 1 = \frac{1}{2} + 1 = \frac{3}{2}.$$

It follows that

$$g(E[X]) = \frac{1}{\frac{3}{2}} = \frac{2}{3}.$$

We will now find $E[g(X)]$. By definition,

$$E[g(X)] = E[1/X] = \sum_{x \in \{1,2\}} \left(\frac{1}{x}\right) P(X = x) = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}.$$

Thus, we have shown that

$$\frac{3}{4} = E[g(X)] \neq g(E[X]) = \frac{2}{3}.$$

Example 2 *Let $X \sim Uniform(0,1)$. Let $g(x) = -\log(x)$. (Unless otherwise stated, when I write* log*, I always mean the natural logarithm.)*

We will find $E[X]$ by definition:

$$E[X] = \int_0^1 x dx = \left[\frac{1}{2}x^2\right]_0^1 = \frac{1}{2}.$$

It follows that

$$g(E[X]) = -\log\left(\frac{1}{2}\right) = \log(2) \approx 0.693.$$

We will now find $E[g(X)]$. By definition,

$$E[g(X)] = E[-log(X)] = \int_0^1 -\log(x)dx = x(1-\log(x))\big|_{x=1} - \lim_{x\to 0} x(1-\log(x)) = 1.$$

Thus, we have shown that

$$1 = E[g(X)] \neq g(E[X]) = 0.693.$$

Example 3 *Let $X \in \{1,2\}$ with $P(X=1) = P(X=2) = 1/2$ and $g(x) = \begin{cases} \frac{1}{x}, & x \neq \frac{3}{2}, \\ \frac{3}{4}, & x = \frac{3}{2} \end{cases}$.*

We already showed in Example 1 that $E[X] = \frac{3}{2}$ in two different ways. Let's try a third. Recall that X and $Y+1$ have the same distribution, where $Y \sim Bernoulli(1/2)$. The table of commonly used distributions in the textbook tells you that Y has MGF $M_Y(t) = \frac{1}{2}e^t + \frac{1}{2}$. It follows that the MGF of $Y+1$, which is given by:

$$M_{Y+1}(t) = E[e^{(Y+1)t}] = e^t E[e^{Yt}] = e^t\left(\frac{1}{2}e^t + \frac{1}{2}\right) = \frac{1}{2}e^{2t} + \frac{1}{2}e^t.$$

Since X and $Y+1$ have the same distribution, and MGFs uniquely determine the distribution of a random variable, we can say that the MGF of X is given by

$$M_X(t) = \frac{1}{2}e^{2t} + \frac{1}{2}e^t.$$

Finally,

$$E[X] = M_X'(0) = \left[e^{2t} + \frac{1}{2}e^t\right]_{t=0} = 1 + \frac{1}{2} = \frac{3}{2}.$$

So it follows that $g(E[X]) = \frac{3}{4}$. We will now calculate $E[g(X)]$ by definition.

$$E[g(X)] = \sum_{x\in\{1,2\}} g(x)f(x) = g(1)f(1) + g(2)f(2) = 1\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

This means that we have shown that

$$E[g(X)] = g(E[X]) = \frac{3}{4}.$$

Conclusion

What have we shown? We have shown that non-property (1) does not hold for g and X in Examples 1 and 2, but *does* hold for g and X in Example 3. So we have shown that when g is a non-linear function, all bets are off – depending on what choice of g and X we take, we can sometimes take g outside the expectation, and we can sometimes *not* take g outside the expectation.

However, note that Example 3 is *extremely* contrived. So if you really must guess whether $E[g(X)] = g(E[X])$ or $E[g(X)] \neq g(E[X])$ for a random variable X and a non-linear function g, it is safer to guess the latter than the former.

■ **Example 2.4** Let X, Y be random variables with joint pdf

$$f(x,y) = \begin{cases} x+y & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Question 1: Is this a valid joint pdf?
Note that $f(x,y) \ge 0$, let's check the following:

$$\begin{aligned} \int_0^1 \int_0^1 f(x,y)dxdy = \int_0^1 \int_0^1 f(x,y)dydx &= \int_0^1 \int_0^1 (x+y)dydx \\ &= \int_0^1 \left[xy + \frac{1}{2}y^2\right]_0^1 dx \\ &= \int_0^1 x + \frac{1}{2}dx \\ &= \left[\frac{1}{2}x^2 + \frac{1}{2}x\right]_0^1 \\ &= 1 \end{aligned}$$

Therefore, $f(x,y)$ is a valid pdf

Question 2: Find $P\left(X \le \frac{1}{3}, Y \le \frac{1}{2}\right)$

$$P\left(X \le \frac{1}{3}, Y \le \frac{1}{2}\right) = \int_0^{\frac{1}{3}} \int_0^{\frac{1}{2}} (x+y)dydx = \int_0^{\frac{1}{3}} \left[xy + \frac{1}{2}y^2\right]_0^{\frac{1}{2}} dx = \int_0^{\frac{1}{3}} \left[\frac{1}{2}x + \frac{1}{8}\right] dx = \left[\frac{1}{4}x^2 + \frac{1}{8}x\right]_0^{\frac{1}{3}} = \frac{5}{72}$$

Question 3: Find $P\left(X+Y \le \frac{1}{2}\right)$

$$\begin{aligned} P\left(X+Y \le \frac{1}{2}\right) = \int_{x+y \le \frac{1}{2}} f(x,y)dxdy &= \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-x} (x+y)dydx \\ &= \int_0^{\frac{1}{2}} \left[xy + \frac{1}{2}y^2\right]_0^{\frac{1}{2}-x} dx \\ &= \int_0^{\frac{1}{2}} -\frac{1}{2}x^2 + \frac{1}{8}dx \\ &= \left[-\frac{1}{6}x^3 + \frac{1}{8}x\right]_0^{\frac{1}{2}} \\ &= \frac{1}{24} \end{aligned}$$

Question 4: Find $P\left(XY \leq \frac{1}{2}\right)$

Solution 1:

$$P\left(XY \leq \frac{1}{2}\right) = \int_{xy \leq \frac{1}{2}} f(x,y)dxdy = \int_0^{\frac{1}{2}} \int_0^1 (x+y)dxdy + \int_{\frac{1}{2}}^1 \int_0^{\frac{1}{2x}} (x+y)dxdy$$

Note that

$$\int_0^{\frac{1}{2}} \int_0^1 (x+y)dxdy = \frac{3}{8}$$

and

$$\int_{\frac{1}{2}}^1 \int_0^{\frac{1}{2x}} (x+y)dxdy = \int_{\frac{1}{2}}^1 \left[xy + \frac{1}{2}y^2\right]_0^{\frac{1}{2x}} dx = \int_{\frac{1}{2}}^1 \frac{1}{2} + \frac{1}{8x^2}dx = \left[\frac{1}{2}x - \frac{1}{8}x\right]_{\frac{1}{2}}^1 = \frac{3}{8}$$

Then we have

$$P\left(XY \leq \frac{1}{2}\right) = \int_0^{\frac{1}{2}} \int_0^1 (x+y)dxdy + \int_{\frac{1}{2}}^1 \int_0^{\frac{1}{2x}} (x+y)dxdy = \frac{3}{8} + \frac{3}{8} = \frac{3}{4}$$

Solution 2:

$$P\left(XY \leq \frac{1}{2}\right) = 1 - P\left(XY > \frac{1}{2}\right) = 1 - \int_{\frac{1}{2}}^1 \int_{\frac{1}{2y}}^1 (x+y)dxdy$$

Note that

$$\int_{\frac{1}{2}}^1 \left[\frac{1}{2}x^2 + xy\right]_{\frac{1}{2y}}^1 dy = \int_{\frac{1}{2}}^1 y - \frac{1}{8y^2}dy = \left[\frac{1}{2}y^2 + \frac{1}{8y}\right]_{\frac{1}{2}}^1 = \frac{1}{4}$$

Then we have

$$P\left(XY \leq \frac{1}{2}\right) = 1 - P\left(XY > \frac{1}{2}\right) = 1 - \frac{1}{4} = \frac{3}{4}$$

Question 5: Find the marginal pdf of X and Y.

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 (x+y)dy = \left[xy + \frac{1}{2}y^2\right]_0^1 = x + \frac{1}{2}$$

Then the marginal pdf of X is :

$$f_x(x) = \begin{cases} x + \frac{1}{2} & 0 < x < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Similarly, the marginal pdf of Y is :

$$f_y(y) = \begin{cases} y + \frac{1}{2} & 0 < y < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

■

■ **Example 2.5** Let X, Y be continuous random variables with joint pdf:

$$f(x,y) = \begin{cases} ke^{-x-y} & 0 < x < y < \infty \\ 0 & \textbf{otherwise} \end{cases}$$

Question 1: What is k?

$$1 = \int_0^\infty \int_0^y f(x,y)dxdy = \int_0^\infty \int_0^y ke^{-x-y}dxdy = k\int_0^\infty e^{-y}(1-e^{-y})dy = k\left[-e^{-y} + \frac{1}{2}e^{-2y}\right]_0^\infty = \frac{k}{2}$$

Then we have

$$1 = \frac{k}{2} \qquad \Longleftrightarrow \qquad k = 2$$

as desired.

Question 2: Find $P\left(X \leq \frac{1}{3}, Y \leq \frac{1}{2}\right)$

$$\begin{aligned} P\left(X \leq \frac{1}{3}, Y \leq \frac{1}{2}\right) = \int_0^{\frac{1}{3}} \int_x^{\frac{1}{2}} 2e^{-x-y}dydx &= \int_0^{\frac{1}{3}} \left[-2e^{-x-y}\right]_x^{\frac{1}{2}} dx \\ &= \int_0^{\frac{1}{3}} \left[-2e^{-x-\frac{1}{2}} + 2e^{-2x}\right] dx \\ &= \left[e^{-x-\frac{1}{2}} - e^{-2x}\right]_0^{\frac{1}{3}} \\ &= 1 - e^{-\frac{2}{3}} + 2\left(e^{-\frac{5}{6}} - e^{-\frac{1}{2}}\right) \\ &\approx 0.14 \end{aligned}$$

Question 3: Find $P(X \leq Y)$

$$P(X \leq Y) = \int_{x \leq y} f(x,y)dxdy = 1$$

Question 4: Find $P(X+Y \geq 1)$

$$\begin{aligned} P(X+Y \geq 1) = 1 - P(X+Y < 1) &= 1 - \int_0^{\frac{1}{2}} \int_x^{1-x} 2e^{-x-y}dydx \\ &= 1 - \int_0^{\frac{1}{2}} \left[-2e^{-x-y}\right]_x^{1-x} dx \\ &= 1 - \int_0^{\frac{1}{2}} \left[-2e^{-1} + 2e^{-2x}\right] dx \\ &= 1 - \left[-2e^{-1}x - e^{-2x}\right]_0^{\frac{1}{2}} \\ &= \frac{2}{e} \end{aligned}$$

Question 5: Find the marginal pdfs.

$$f_x(x) = \int_x^\infty 2e^{-x-y}dy = 2e^{-2x}$$

for $0 < x < \infty$. Then we have

$$f_x(x) = \begin{cases} 2e^{-2x} & 0 < x < \infty \\ 0 & \textbf{otherwise} \end{cases}$$

Similarly,

$$f_y(y) = \int_0^y 2e^{-x-y}dx = 2e^{-y} - 2e^{-2y}$$

for $0 < y < \infty$, then we have

$$f_y(y) = \begin{cases} 2e^{-y} - 2e^{-2y} & 0 < x < \infty \\ 0 & \textbf{otherwise} \end{cases}$$

■

2.5 Independent Random Variables

Definition 2.5.1 — Independent.
Let X, Y be two random variables, X is independent of Y if and only if for all $A, B \in \mathbb{R}$

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

Theorem 2.5.1
Let X, Y be random variables, if joint CDF $F(x,y)$, marginal CDFs $F_1(x)$ and $F_2(y)$, then

$$\textbf{X and Y are independent} \iff F(x,y) = F_1(x) \cdot F_2(y) \quad \forall\, (x,y) \in \mathbb{R}^2$$

Theorem 2.5.2
Let X, Y be random variables, if joint pdf/pmf $f(x,y)$, marginal pdfs/pmfs $f_1(x)$ and $f_2(y)$ with

$$A_1 = \{x \in \mathbb{R} : f_1(x) > 0\} \qquad \textbf{and} \qquad A_2 = \{y \in \mathbb{R} : f_2(y) > 0\}$$

Then

$$\textbf{X and Y are independent} \iff f(x,y) = f_1(x) \cdot f_2(y) \quad \forall\, (x,y) \in A_1 \times A_2$$

■ **Example 2.6** Let X,Y be discrete random variables with joint pmf

$$f(x,y) = \begin{cases} (1-p)^2 p^{x+y} & x,y \in \mathbb{N} \\ 0 & \textbf{otherwise} \end{cases}$$

for $0 < p < 1$. It's easy to get that

$$f_1(x) = \begin{cases} (1-p)p^x & x \in \mathbb{N} \\ 0 & \textbf{otherwise} \end{cases} \quad \textbf{and} \quad f_2(x) = \begin{cases} (1-p)p^y & y \in \mathbb{N} \\ 0 & \textbf{otherwise} \end{cases}$$

Are X,Y independent?
Note that

$$f(x,y) = f_1(x)f_2(y)$$

Then by **Theorem 2.5.2** we have X,Y are independent. ■

■ **Example 2.7** Let X,Y be continuous random variables with joint pdf

$$f(x,y) = \begin{cases} x+y & 0 < x < 1 \textbf{ and } 0 < y < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

It's easy to see that

$$f_1(x) = \begin{cases} x+\frac{1}{2} & 0 < x < 1 \\ 0 & \textbf{otherwise} \end{cases} \quad \textbf{and} \quad f_2(y) = \begin{cases} y+\frac{1}{2} & 0 < t < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Are X,Y independent?
Note that

$$f_1(x)f_2(y) = \left(x+\frac{1}{2}\right)\cdot\left(y+\frac{1}{2}\right) = xy + \frac{x+y}{2} + \frac{1}{4} \neq x+y = f(x,y)$$

Then X,Y are not independent. ■

Theorem 2.5.3 — Factorization Theorem of Independence.
Let X,Y are random variables with joint pdf/pmf $f(x,y)$ and joint support A with support of x,y: A_1,A_2 respectively, then

X and Y are independent $\iff \exists\, g(x) \geq 0,\ \exists\, h(y) \geq 0$ **s.t.** $f(x,y) = g(x)\cdot h(y) \quad \forall\, (x,y) \in A_1 \times A_2$

Notes for Theorem 2.5.3:
1. If $\implies$ holds, then

$$\textbf{marginal pdf/pmf of } X\textbf{:}\ f_1(x) \propto g(x) \quad \forall\, x \in A_1$$

and

$$\textbf{marginal pdf/pmf of } Y\textbf{:}\ f_2(y) \propto h(x) \quad \forall\, y \in A_2$$

2. **If A is not rectangular, then X,Y must be dependent.**

Proof: $\exists (x,y) \in \mathbb{R}^2$ s.t. $x \in A_1$ and $y \in A_2$ but $(x,y) \notin A$, so $f_1(x), f_2(y) > 0$ with $f(x,y) = 0$, this is a contradiction.

■ **Example 2.8**

$$f(x,y) = \begin{cases} \dfrac{\theta^{x+y}e^{-2\theta}}{x!y!} & x,y \in A = \mathbb{N} \cup \{0\} \\ 0 & \textbf{otherwise} \end{cases}$$

Are X,Y independent?
Note that

$$f(x,y) = \frac{\theta^{x+y}e^{-2\theta}}{x!y!} = \frac{\theta^x}{x!}e^{-\theta} \cdot \frac{\theta^y}{y!}e^{-\theta} \qquad \forall\ (x,y) \in A \times A$$

It's easy to check

$$f_1(x) = \frac{\theta^x}{x!}e^{-\theta} \qquad \textbf{and} \qquad f_2(y) = \frac{\theta^x}{x!}e^{-\theta}$$

for $x,y \in A$. Then we have

$$f(x,y) = f_1(x) \cdot f_2(y) \qquad \Longleftrightarrow \qquad \textbf{X and Y are independent}$$

■

■ **Example 2.9** Let X,Y be continuous random variables with joint pdf

$$f(x,y) = \begin{cases} \dfrac{3}{2}y(1-x^2) & -1 < x < 1, 0 < y < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Are X and Y independent? By **Theorem 2.5.3**

$$f(x,y) = \underbrace{\frac{3}{2}y}_{h(y)}\underbrace{(1-x^2)}_{g(x)} \qquad \Longrightarrow \qquad X \textbf{ and } Y \textbf{ are independent}$$

Note that $\exists k_1, k_2 \in \mathbb{R}$ s.t. $f_1(x) = k_1 g(x)$ and $f_2(y) = k_2 h(y)$. This gives us that

$$1 = k_1 \cdot \int_{-1}^{1}(1-x^2)dx = k_1 \cdot \frac{4}{3} \qquad \Longleftrightarrow \qquad k_1 = \frac{3}{4}$$

and

$$1 = k_2 \cdot \int_0^1 \frac{3}{2}ydy = k_2 \cdot \frac{3}{2} \cdot \frac{1}{2} \qquad \Longleftrightarrow \qquad k_1 = \frac{4}{3}$$

■

■ **Example 2.10** Let X,Y be continuous random variables with joint pdf

$$f(x,y) = \begin{cases} \frac{2}{\pi} & 0 < x < \sqrt{1-y^2}, -1 < y < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Are X,Y independent?
Note that the support

$$A = \left\{(x,y) \in \mathbb{R} : 0 < x < \sqrt{1-y^2}, -1 < y < 1\right\}$$

is not rectangular, then X,Y are not independent. ■

Lemma 2.5.4
Let g,h be functions, if X,Y are independent, then $g(X)$ and $h(Y)$ are independent.

Note: The converse is not true.

Definition 2.5.2 — Joint Expectation.
Let X,Y be bivariate discrete and let $h(x,y)$ be a real valued function. If

$$\sum_{(x,y)\in A} |h(x,y)| f(x,y) < \infty$$

Then

$$E[h(X,Y)] = \sum_{(x,y)\in A} h(x,y) f(x,y)$$

Otherwise, we say that $E[h(X,Y)]$ **DNE**.

Let X,Y be bivariate continuous and let $h(x,y)$ be a real valued function. If

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,y)| f(x,y) < \infty$$

Then

$$E[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f(x,y)$$

Otherwise, we say that $E[h(X,Y)]$ **DNE**.

■ **Example 2.11**

$$E[XY] = \begin{cases} \sum\limits_{(x,y)\in A} xyf(x,y) & \textbf{if } X,Y \textbf{ are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy & \textbf{if } X,Y \textbf{ are continuous} \end{cases}$$

provided that $E[XY]$ exists. ■

■ **Example 2.12** Let X, Y be continuous random variables and $E[X]$ exists, then

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dydx = \ displaystyle \int_{-\infty}^{\infty}()x\cdot[\int_{-\infty}^{\infty} f(x,y)dy]dx = \int_{-\infty}^{\infty} x\cdot f_1(x)dx$$

■

Proposition 2.5.5 — Linearity.
Let $a, b \in \mathbb{R}$ and X, Y be random variables, then

$$E[ag(X,Y)+bh(X,Y)] = aE[g(X,Y)]+bE[h(X,Y)]$$

Let $a_1, ..., a_n \in \mathbb{R}$ and $X_1, ..., X_n$ be random variables, then

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

Proposition 2.5.6
Let $X_1, ..., X_n$ be independent random variables and $g_1, ..., g_n$ be functions, then

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

Definition 2.5.3 — Covariance of X and Y.

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

where $\mu_x = E[X]$ and $\mu_y = E[Y]$. If $Cov(X,Y) = 0$, we say X **and** Y **are uncorrelated**

Proposition 2.5.7
1. $Cov(X,Y) = E[XY] - E[X]\cdot E[Y]$
2. X and Y are independent $\implies Cov(X,Y) = 0$
3. $Cov(X,X) = \text{Var}[X]$
4. $\text{Var}[aX+bY] = a^2 \cdot \text{Var}[X] + b^2 \cdot \text{Var}[Y] + 2ab \cdot Cov(X,Y)$
5.
$$\text{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \text{Var}[X_i] + \sum_{i\neq j} a_i a_j \cdot Cov(X_i, X_j)$$
6. If $X_1, ..., X_n$ are independent,
$$\text{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \text{Var}[X_i]$$
7. $Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z)$

■ **Example 2.13** Let X, Y be random variables with joint pdf

$$f(x,y) = \begin{cases} x+y & 0<x<1, 0<y<1 \\ 0 & \textbf{otherwise} \end{cases}$$

Since

$$\text{Var}[X+Y] = E\left[(X+Y-E[X+Y])^2\right]$$

Then by properties of variance we have

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2Cov(X,Y)$$

Note that

$$f_1(x) = \begin{cases} x+\frac{1}{2} & 0<x<1 \\ 0 & \textbf{otherwise} \end{cases} \quad \textbf{and} \quad f_2(y) = \begin{cases} y+\frac{1}{2} & 0<t<1 \\ 0 & \textbf{otherwise} \end{cases}$$

then

$$E[X] = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \frac{7}{12} = E[Y] \qquad \textbf{and} \qquad E[X^2] = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \frac{5}{12} = E[Y^2]$$

so we have

$$\text{Var}[X] = \text{Var}[Y] = E[X^2] - (E[X])^2 = \frac{11}{144}$$

Note that

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy = \int_0^1\int_0^1 xy(x+y)dxdy = \frac{1}{3}$$

Then we have

$$Cov(X,Y) = E[XY] - E[X]E[Y] = -\frac{1}{144}$$

Hence,

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2Cov(X,Y) = \frac{20}{144}$$

as desired. ■

Definition 2.5.4 — Correlation Coefficient.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}}$$

*** The measure of linear association between X and Y**

Theorem 2.5.8

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

Note:
1. If $\rho(X,Y) = 1$, then $Y = aX + b$ for some $a > 0$ and $b \in \mathbb{R}$
2. If $\rho(X,Y) = -1$, then $Y = aX + b$ for some $a < 0$ and $b \in \mathbb{R}$

■ **Example 2.14** Let $Z \sim N(0,1)$, $X = Z$ and $Y = Z^2$, then $\rho(X,Y) = 0$ because there is **no linear relationship** between X and Y. ■

■ **Example 2.15** Let X, Y be random variables with joint pdf

$$f(x,y) = \begin{cases} x+y & 0 < x < 1, 0 < y < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Since

$$\text{Var}[X] = \text{Var}[Y] = \frac{11}{144} \qquad \textbf{and} \qquad Cov(X,Y) = -\frac{1}{144}$$

Then we have

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}} = -\frac{1}{11}$$

as desired. ■

2.6 Conditional Distributions

Definition 2.6.1 — Discrete Conditional Distributions.
Let X, Y be bivariate discrete random variables with joint pdf $f(x,y)$, then

1. The conditional pdf of X given $Y = y$ is

$$f_1(x \mid y) = \frac{f(x,y)}{f_2(y)} = P(X = x \mid Y = y) \qquad \textbf{where } f_2(y) \textbf{ is the pmf of } y \textbf{ with } f_2(y) > 0$$

2. Conditional pdf of Y given $X = x$ is

$$f_2(y \mid x) = \frac{f(x,y)}{f_1(x)} = P(Y = y \mid X = x) \qquad \textbf{where } f_1(x) \textbf{ is the pmf of } x \textbf{ with } f_1(x) > 0$$

Proposition 2.6.1

$$f_1(x \mid y) \geq 0 \quad , \sum_x f_1(x \mid y) = 1 \qquad \textbf{and} \qquad f_2(y \mid x) \geq \quad , \sum_y f_2(y \mid x) = 1$$

Definition 2.6.2 — Continuous Conditional Distributions.
Let X, Y be bivariate continuous random variables with joint pdf $f(x,y)$, then

1. The conditional pdf of X given $Y = y$ is

$$f_1(x \mid y) = \frac{f(x,y)}{f_2(y)} \qquad \textbf{where } f_2(y) \textbf{ is the pmf of } y \textbf{ with } f_2(y) > 0$$

2. Conditional pdf of Y given $X = x$ is

$$f_2(y \mid x) = \frac{f(x,y)}{f_1(x)} \qquad \textbf{where } f_1(x) \textbf{ is the pmf of } x \textbf{ with } f_1(x) > 0$$

Note:

$$P(Y \le y \mid X = x) = \int_{-\infty}^{y} f_2(t \mid x)dt$$

$$P(X \le x \mid Y = y) = \int_{-\infty}^{y} f_1(t \mid y)dt$$

Proposition 2.6.2

$$f_1(x \mid y) \ge 0 \quad , \int_{-\infty}^{\infty} f_1(x \mid y)dx = 1 \qquad \textbf{and} \qquad f_2(y \mid x) \ge 0 \quad , \int_{-\infty}^{\infty} f_2(y \mid x)dy = 1$$

■ **Example 2.16**

$$f(x,y) = \begin{cases} 8xy & 0 < y < x < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Find $f_1(x \mid y)$ and $f_2(y \mid x)$
Note that

$$f_1(x) = \int_0^x 8xy\, dy = 4x^3 \quad \textbf{with } 0 < x < 1 \qquad \textbf{and} \qquad f_2(y) = \int_y^1 8xy\, dx = 4y - 4y^3 \quad \textbf{with } 0 < y < 1$$

Then we have

$$f_1(x \mid y) = \frac{f(x,y)}{f_2(y)} = \frac{8xy}{4y - 4y^3} = \frac{8x}{4 - y^3} \quad \textbf{with } 0 < y < x < 1$$

$$f_2(y \mid x) = \frac{f(x,y)}{f_1(x)} = \frac{8xy}{4x^3} = \frac{2y}{x^2} \quad \textbf{with } 0 < y < x < 1$$

■

■ **Example 2.17**

$$f(x,y) = \begin{cases} (1-p)^2 p^{x+y} & x, y \in A = \mathbb{N} \cup \{0\} \\ 0 & \textbf{otherwise} \end{cases}$$

Is X and Y independent?
Note that

$$f_1(x) = (1-p)p^x \quad \textbf{with } x \in A \qquad \textbf{and} \qquad f_2(y) = (1-p)p^y \quad \textbf{with } y \in A$$

and we also have

$$f_1(x \mid y) = \frac{f(x,y)}{f_2(y)} = (1-p)p^x \quad \textbf{with } x \in A \qquad \textbf{and} \qquad f_2(y \mid x) = \frac{f(x,y)}{f_1(x)} = (1-p)p^y \quad \textbf{with } y \in A$$

Then we have

$$f_1(x \mid y) \cdot f_2(y \mid x) = (1-p)^2 p^{x+y} = f(x,y) = \frac{f(x,y)}{f_1(x)} \cdot \frac{f(x,y)}{f_2(y)} \quad \Longrightarrow \quad f_1(x) \cdot f_2(y) = f(x,y) \quad \forall \ (x,y) \in A \times A$$

Hence, X and Y are independent. ■

Theorem 2.6.3
Let X and Y are random variables with marginal pdfs/pmfs $f_1(x)$, $f_2(y)$, marginal support A_1 and A_2, conditional pmfs/pdfs $f_1(x \mid y)$ and $f_2(y \mid x)$. Then

$$\textbf{X and Y are independent} \iff f_1(x \mid y) = f_1(x) \quad \textbf{and} \quad f_2(y \mid x) = f_2(y) \quad \forall\ x \in A_1\ ,\ y \in A_2$$

Proposition 2.6.4 — Product Rule.

$$f(x,y) = f_1(x \mid y) \cdot f_2(y) = f_2(y \mid x) \cdot f_1(x)$$

■ **Example 2.18** Let $Y \sim Poi(\lambda)$ and $X \mid Y = y \sim Bin(y,p)$, what's the distribution of X?

Let $A = \mathbb{N} \cup \{0\}$, note that

$$f_2(y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad \textbf{with } y \in A \qquad \textbf{and} \qquad f_1(x \mid y) = \binom{y}{x} p^x (1-p)^{y-x} \quad \textbf{with } x = 0,1,...,y$$

Then we have

$$f(x,y) = f_1(x \mid y) \cdot f_2(y) = \frac{(1-p)^{y-x}\lambda^y}{(y-x)!} \quad \textbf{with } y \in A \textbf{ and } x = 0,1.,,,,y$$

This gives us that

$$\begin{aligned} f_1(x) = \sum_y f(x,y) = \sum_{y=x}^{\infty} \frac{p^x e^{-\lambda}}{x!} \cdot \frac{(1-p)^{y-x}\lambda^y}{(y-x)!} &= \frac{p^x e^{-\lambda}}{x!} \cdot \sum_{y=x}^{\infty} \frac{(1-p)^{y-x}\lambda^y}{(y-x)!} \\ &= \frac{p^x e^{-\lambda}\lambda^x}{x!} \cdot \sum_{y=x}^{\infty} \frac{(1-p)^{y-x}\lambda^{y-x}}{(y-x)!} \\ &= \frac{p^x e^{-\lambda}\lambda^x}{x!} \cdot e^{\lambda(1-p)} \\ &= \frac{(p\lambda)^x e^{-p\lambda}}{x!} \end{aligned}$$

Therefore, we have $X \sim Poi(p\lambda)$ ■

■ **Example 2.19** Let $Y \sim Gamma(\alpha, 1)$, $f_2(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)} e^{-y}$ with $y > 0$ and $f_1(x \mid y) = ye^{-xy}$ with $x, y > 0$.

Find $f_1(x)$

Note that

$$f(x,y) = f_1(x \mid y) = f_2(y) = \frac{y^{\alpha}}{\Gamma(\alpha)} e^{-(x+1)y}$$

Since

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

then we have

$$\begin{aligned} f_1(x) = \int_0^{\infty} \frac{y^{\alpha}}{\Gamma(\alpha)} e^{-(x+1)y} dy &= \int_0^{\infty} \left(\frac{t}{x+1}\right)^{\alpha} \cdot \frac{1}{\Gamma(\alpha)} e^{-t} \cdot \frac{1}{x+1} dt \\ &= \frac{1}{(x+1)^{\alpha+1}\Gamma(\alpha)} \underbrace{\int_0^{\infty} t^{\alpha} e^{-t} dt}_{=\Gamma(\alpha+1)} \\ &= \frac{\alpha}{(x+1)^{\alpha+1}} \end{aligned}$$

for $x > 0$ as desired. ■

2.7 Appendix 2 (provided by Lucy Gao)

Relationship Between Covariance and Independence

Relationship Between Uncorrelated and Independence

We saw in Lecture 7 that if X and Y are independent, then $Cov(X, Y) = 0$, i.e. X and $Y$are uncorrelated. This is true because:

$$Cov(X, Y) = E[XY] - E[X]E[Y],$$

and when X and Y are independent, $E[XY] = E[X]E[Y]$. This may lead you to wonder about the converse – are uncorrelated random variables independent? Unfortunately, this is another *non-property* – not all uncorrelated random variables are independent, as the following example demonstrates.

A Cautionary Example – Uncorrelated May Not Mean Independent

Let X be a non-constant continuous random variable with a symmetric probability density function (pdf), i.e. $f(x) = f(-x)$ for all $x \in \mathbb{R}$. For example, you could imagine that $X \sim N(0, 1)$ or $X \sim \text{Uniform}(-1, 1)$. Let $Y = X^2$.

First, we will show that X and $Y = X^2$ are uncorrelated. We write:

$$Cov(X, Y) = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2]. \tag{1}$$

Since X has a symmetric pdf, for p odd, we have

$$\begin{aligned} E[X^p] &= \int_{-\infty}^{\infty} x^p f(x)dx \\ &= \int_{-\infty}^{0} x^p f(x)dx + \int_{0}^{\infty} x^p f(x)dx \\ &= \int_{-\infty}^{0} x^p f(-x)dx + \int_{0}^{\infty} x^p f(x)dx \\ &= -\int_{0^{\infty}} x^p f(x)dx + \int_{0}^{\infty} x^p f(x)dx \\ &= 0. \end{aligned}$$

Thus, $E[X^3] = 0$ and $E[X] = 0$ in (1), which means that $Cov(X, Y) = 0$.

However, X and Y are *not* independent – they are dependent! One way to see this is to observe that their joint support is $\{(x, x^2) : x \in A\}$, where A is the support of X, which is not a rectangle parallel to the x and y axes when $A \neq \emptyset$.

Correctly Characterizing Independence With Expectations

We have demonstrated that $Cov(X, Y) = 0$ does not imply that X and Y are independent. That is, $E[XY] = E[X]E[Y]$ does not imply that X and Y are independent. You might now ask – is there a correct way to chacterize independence between two random variables with expectations? As it turns out, there is.

Theorem. *X and Y are independent if and only if $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for all functions $g(\cdot)$ and $h(\cdot)$.*

The ($\Rightarrow$) direction follows from the fact that X and Y are independent if and only if their joint pdf factorizes into the product of their marginal pdfs. We can prove the ($\Leftarrow$) direction by contradiction. Suppose that

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] \text{ for all functions } g(\cdot) \text{ and } h(\cdot), \tag{2}$$

but X and Y are dependent. Then, there exists some set A in the support of X and some set B in the support of Y such that

$$P(X \in A, Y \in B) \neq P(X \in A)P(Y \in B). \tag{3}$$

But if we consider $g(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A \end{cases}$ and $h(y) = \begin{cases} 1, & y \in B, \\ 0, & y \notin B \end{cases}$, then we have by (2) that

$$P(X \in A, Y \in A) = P(X \in A)P(Y \in A).$$

(If this is surprising to you, please see the first recap session on Friday.) This contradicts (3), so X and Y are independent.

Conclusion

The key takeaway here is that uncorrelated random variables may not be independent. This is because whether two random variables are uncorrelated relies solely on whether $E[XY] = E[X]E[Y]$. This is not enough to ensure independence – we would need $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for all functions $g(\cdot)$ and $h(\cdot)$ in order to guarantee independence, which is a much stronger condition.

2.8 Conditional Expectation

Definition 2.8.1
For a function g, the conditional expectation of $g(Y)$ given $X = x$ is

$$E[g(Y) \mid X = x] = \begin{cases} \sum\limits_{\textbf{all y}} g(y) \cdot f_2(y \mid x) & \textbf{if } Y \textbf{ is discrete random variable} \\ \int_{-\infty}^{\infty} g(y) \cdot f_2(y \mid x) & \textbf{if } Y \textbf{ continuous random variable} \end{cases}$$

unless $\sum\limits_{\textbf{all y}} g(y) \cdot f_2(y \mid x)$ does not converge in which case $E[g(Y) \mid X = x]$ is **DNE** $\left(\int_{-\infty}^{\infty}\right)$. The $E[g(X) \mid Y = y]$ is similarly.

Note:
1. $g(y) = y$, $E[Y \mid X = x]$ is called conditional mean
2. $g(y) = (y - E[Y \mid X = x])^2$ implies

$$\text{Var}[Y \mid X = x] = E\Big[(y - E[Y \mid X = x])^2 \mid X = x\Big] = E[Y^2 \mid X = x] - [E[Y \mid X = x]]^2$$

is called conditional variance.

■ **Example 2.20**

$$f(x,y) = \begin{cases} 8xy & 0 < y < x < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

and we have

$$f_1(x \mid y) = \frac{f(x,y)}{f_2(y)} = \frac{8xy}{4y - 4y^3} = \frac{8x}{4 - y^3} \quad \textbf{with } 0 < y < x < 1$$

$$f_2(y \mid x) = \frac{f(x,y)}{f_1(x)} = \frac{8xy}{4x^3} = \frac{2y}{x^2} \quad \textbf{with } 0 < y < x < 1$$

Find $E[X \mid Y = y]$ and $\text{Var}[X \mid Y = y]$

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_1(x \mid y)dx = \int_y^1 \frac{8x^2}{4 - 4y^2}dx = \frac{2}{3} \cdot \frac{1 - y^3}{1 - y^2} \quad \textbf{for } 0 < y < 1$$

$$E[X^2 \mid Y = y] = \int_{-\infty}^{\infty} x^2 \cdot f_1(x \mid y)dx = \int_y^1 \frac{2x^2}{1 - y^2}dx = \frac{1}{2}(1 + y^2)$$

Therefore, we have

$$\text{Var}[X \mid Y = y] = E[X^2 \mid Y = y] - (E[X \mid Y = y])^2 = \frac{1}{2}(1 + y^2) - \frac{4}{9}\left(\frac{1 - y^3}{1 - y^2}\right)^2 \quad \textbf{for } 0 < y < 1$$

as desired. ■

Proposition 2.8.1
If X, Y are independent, then for all functions g, h:

$$E[g(X) \mid Y = y] = E[g(X)] \qquad \textbf{and} \qquad E[h(Y) \mid X = x] = E[h(Y)]$$

Proposition 2.8.2 — Substitution Rule.

$$E[h(X,Y) \mid X = x] = E[h(x,Y) \mid X = x]$$

For example:

$$E[X+Y \mid X = x] = E[x+Y \mid X = x] = x + E[Y \mid X = x]$$
$$E[XY \mid X = x] = E[xY \mid X = x] = xE[Y \mid X = x]$$

Theorem 2.8.3 — Double-Expectation Formula.
Let g be a function and X, Y be random variables, then

$$E[E[g(X) \mid Y]] = E[g(X)]$$

Proof (continuous case):

$$\begin{aligned}
E[E[g(X) \mid Y]] = E\left[\int_{-\infty}^{\infty} g(x) f_1(x \mid y)dx\right] &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x) f_1(x \mid y)dx\right] f_2(y)dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f_1(x \mid y) f_2(y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)dydx \\
&= \int_{-\infty}^{\infty} g(x)\left[\int_{-\infty}^{\infty} f(x,y)dy\right]dx \\
&= \int_{-\infty}^{\infty} g(x) f_1(x)dx \\
&= E[g(X)]
\end{aligned}$$

■ **Example 2.21** Let $Y \sim Poi(\lambda)$ and $X \mid Y = y \sim Bin(y, p)$. Then

$$E[X \mid Y = y] = yp \quad \Longrightarrow \quad E[X \mid Y] = pY$$

By **Double-Expectation Formula**

$$E[E[X \mid Y]] = E[pY] = pE[Y] = p\lambda = E[X]$$

Then we have $X \sim Poi(p\lambda)$ ■

Proposition 2.8.4

$$\text{Var}[Y] = E[\text{Var}[Y \mid X]] + \text{Var}[E[Y \mid X]]$$

■ **Example 2.22** Let $Y \sim X = x \sim Bin(x, p)$ and $X \sim Poi(\lambda)$. Then

$$\text{Var}[Y \mid X = x] = xp(1-p) \implies \text{Var}[Y \mid X] = Xp(1-p) \implies E[\text{Var}[Y \mid X]] = E[Xp(1-p)] = p(1-p)\lambda$$

and

$$E[Y \mid X = x] = xp \implies E[Y \mid X] = Xp \implies \text{Var}[E[Y \mid X]] = \text{Var}[Xp] = p^2\text{Var}[X] = p^2\lambda$$

Therefore, we have

$$\text{Var}[Y] = E[\text{Var}[Y \mid X]] + \text{Var}[E[Y \mid X]] = p(1-p)\lambda + p^2\lambda = \lambda p$$

■

2.9 Joint MGFs

Definition 2.9.1
Let $X_1 \ldots . X_n$ be random variables, if $E\left[e^{\Sigma_{i=1}^n t_i X_i}\right]$ exists $\forall t_i \in (-h_i, h_i)$ with some $h_i > 0$. Then

$$M(t_1, \ldots., t_n) = E\left[e^{\Sigma_{i=1}^n t_i X_i}\right] \quad \forall\, t_1, \ldots. \, t_n \ \textbf{s.t.}\ E\left[e^{\Sigma_{i=1}^n t_i X_i}\right] \ \textbf{exists}$$

is called the joint MGF

Proposition 2.9.1
Given $M(t_1, t_2)$, then

$$M_X(t_1) = M(t_1, 0) = E[e^{t_1 X + 0 \cdot Y}] = E[e^{t_1 X}]$$

and

$$M_Y(t_2) = M(0, t_2) = E[e^{0 \cdot X + t_2 Y}] = E[e^{t_2 Y}]$$

Proposition 2.9.2
Let $X_1, \ldots., X_n$ be random variables with joint MGF $M(t_1, t_2)$, then

$$X_1, \ldots., X_n \textbf{ are independent} \iff M(t_1, \ldots., t_n) = \prod_{i=1}^n M_{X_i}(t_i)$$

■ **Example 2.23** Let X,Y be continuous random variables with joint pdf:

$$f(x,y) = \begin{cases} e^{-x-y} & x>0, y>0 \\ 0 & \textbf{otherwise} \end{cases}$$

Then

$$E[e^{t_1X+t_2Y}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x+t_2y} f(x,y)dxdy = \underbrace{\int_{-\infty}^{\infty} e^{y(t_2-1)}dy}_{<\infty \text{ when } t_2<1} \underbrace{\int_{-\infty}^{\infty} e^{x(t_1-1)}dx}_{<\infty \text{ when } t_1<1}$$

Then, $E[e^{t_1X+t_2Y}]$ exists, so

$$M(t_1,t_2) = E[e^{t_1X+t_2Y}] = \left[\int_{-\infty}^{\infty} e^{y(t_2-1)}dy\right]\left[\int_{-\infty}^{\infty} e^{x(t_1-1)}dx\right] = \frac{1}{(1-t_1)(1-t_2)}$$

for all $t_1,t_2<1$. Then we have

$$M_X(t_1) = M(t_1,0) = \frac{1}{1-t_1} \quad \forall\, t_1<1$$

and

$$M_Y(t_2) = M(0,t_2) = \frac{1}{1-t_2} \quad \forall\, t_2<1$$

This gives us that

$$M(t_1,t_2) = M_X(t_1)M_Y(t_2) \quad \Longrightarrow \quad \textbf{X and Y are independent}$$

■

■ **Example 2.24** Let $X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$ and X,Y are independent. Show what $X+Y \sim Poi(\lambda_1+\lambda_2)$

Note that

$$M(t_1,t_2) = M_X(t_1)M_Y(t_2) = e^{\lambda_1(e^{t_1}-1)}e^{\lambda_2(e^{t_2}-1)} = e^{\lambda_1(e^{t_1}-1)+\lambda_2(e^{t_2}-1)}$$

Also we have

$$M(t,t) = E[e^{tX+tY}] = E[e^{t(X+Y)}] = e^{\lambda_1(e^t-1)+\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

which is the MGF of $Poi(\lambda_1+\lambda_2)$. Since MGF are uniquem then $X+Y \sim Poi(\lambda_1+\lambda_2)$ ■

3. Multivariate Random Variables - II

3.1 Multinomial Distribution

Definition 3.1.1 — Multinomial Distribution.
Let $(X_1, ..., X_k)$ be discrete random variables with joint pmf

$$f(x_1, ..., x_k) = \begin{cases} \dfrac{n!}{x_1!x_2!...x_l!} p_1^{x_1} ....p_k^{x_k} & \textbf{where } \sum_{i=1}^{k} x_i = n \\ 0 & \textbf{otherwise} \end{cases}$$

for $0 < p_i < 1$ with $\sum_{i=1}^{k} p_i = 1$. We say $(X_1, ..., X_k) \sim Mult(n, p_1, ..., p_k)$

Proposition 3.1.1
Let $(X_1, ..., X_k) \sim Mult(n, p_1, ..., p_k)$
1. Joint MGF:

$$M(t_1, ..., t_k) = E\left[e^{t_1X_1 + ... + t_kX_k}\right] = \left(p_1e^{t_1} + ... + p_ke^{t_k}\right)^n \quad \forall\, (t_1, ..., t_k) \in \mathbb{R}^k$$

$$\textbf{To prove:} \quad (x_1 + ... + x_m)^n = \sum_{k_1 + ... + k_m = n} \frac{n!}{k_1!...k_m!} x_1^{k_1} ....x_m^{k_m}$$

2. $X_i \sim Bin(n, p_i)$ for $i = 1, ..., k$
Proof:

$$M_{X_i}(t_i) = M(0, 0, ..., t_i, .., 0, 0) = \left(p_ie^{t_i} + \sum_{j \neq i} p_j\right)^n = (p_ie^{t_i} + 1 - p_i)^n$$

which is the MGF of $Bin(n, p_i)$

3. $T = X_i + X_j$ with $i \neq j$, then $T \sim Bin(n, p_i + p_j)$.
Reason:

$$M_T(t) = E\left[e^{t(X_i+X_j)}\right] = M(0, ..., t_i, ..., t_j, ..., 0)$$

is the MGF of $X_i + X_j$.

4. $E[X_i] = np_i$ and $\text{Var}[X_i] np_i(1-p_i)$, $Cov(X_i, X_j) = -np_ip_j$.
Proof for $Cov(X_i, X_j)$**:** Note that $X_i + X_j \sim Bin(n, p_i + p_j)$ with $i \neq j$ and

$$\text{Var}[X_i + X_j] = n(p_ip_j)(1 - p_j - p_i)$$

. Then

$$\begin{aligned} n(p_ip_j)(1-p_j-p_i) = \text{Var}[X_i+X_j] &= Cov(X_i+X_j, X_i+X_j) \\ &= Cov(X_i, X_j) + Cov(X_j, X_i) + Cov(X_i, X_i) + Cov(X_j, X_j) \\ &= 2Cov(X_j + X_i) + \text{Var}[X_i] + \text{Var}[X_j] \\ &= 2Cov(X_j + X_i) + np_i(1-p_i) + np_j(1-p_j) \end{aligned}$$

so we rearrange it to get $Cov(X_i, X_j) = -np_ip_j$

5. $X_i \mid X_j = x_j \sim Bin(n - x_j, \frac{p_i}{1-p_j})$

6. $X_i \mid X_j + X_i = t \sim Bin(t, \frac{p_i}{p_i+p_j})$

3.2 Bivariate Normal Distribution

Definition 3.2.1 — Bivariate Normal Distribution.
The **Bivariate Normal Distribution** is defined as $\vec{x} \sim BVN(\vec{\mu}, \sum)$ where

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where $\mu_i \in \mathbb{R}$, $\sigma_i > 0$ for $i = 1, 2$ and $-1 < \rho < 1$

$$\begin{aligned} f(\vec{x}) = f(x_1, x_2) &= \frac{1}{2\pi(|\sum|)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \sum^{-1} (\vec{x}-\vec{\mu})^T} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]} \qquad \forall\, \vec{x} \in \mathbb{R}^2 \end{aligned}$$

Proposition 3.2.1

1. X_1 and X_2 have joint MGF:

$$M(t_1,t_2) = E[e^{t_1x_1+t_2x_2}] = e^{\vec{t}^T\vec{\mu}+\frac{1}{2}\vec{t}^T\sum\vec{t}} \qquad \forall\, t=[t_1,t_2]^T \in \mathbb{R}^2$$

2. $X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$

3.

$$X_2 \mid X_1 = x_1 \sim N(\mu_2+\rho\sigma_2\sigma_1^{-1}(x_1-\mu_1),\sigma_2^2(1-\rho^2))$$
$$X_1 \mid X_2 = x_2 \sim N(\mu_1+\rho\sigma_2\sigma_2^{-1}(x_2-\mu_2),\sigma_1^2(1-\rho^2))$$

4.

$$\forall\, 0\neq\vec{c}\in\mathbb{R}^2 \quad \vec{c}^T\vec{X} = c_1X_1+c_2X_2 \sim N\left(\vec{c}^T\vec{\mu},\vec{c}^T\sum\vec{c}\right)$$

$$\forall\, A\in\mathbb{R}^{2\times 2} \ \textbf{with}\ |A|\neq 0,\ \forall\,\vec{b}\in\mathbb{R}^2 \qquad A\vec{X}+\vec{b} \sim BVN\left(A\vec{\mu}+\vec{b},A\sum A^T\right)$$

5. $E[X_i]=\mu_i$, $\mathrm{Var}[X_i]=\sigma_i^2$ for $i=1,2$. $Cov(X_1,X_2)=\rho\sigma_1\sigma_2$, $Corr(X_1,X_2)=\rho$.
Proof for $Cov(X_1,X_2)$: Note that

$$E[X_1X_2 \mid X_1=x_1] = E[x_1X_2 \mid X_1=x_1] = x_1E[X_2 \mid X_1=x_1] = x_1(\mu_2+\rho\sigma_2\sigma_1^{-1}(x_1-\mu_1))$$

Then we have $E[X_1X_2 \mid X_1] = X_1(\mu_2+\rho\sigma_2\sigma_1^{-1}(X_1-\mu_1))$, so that

$$\begin{aligned} E[X_1X_2] = E[E[X_1X_2 \mid X_1]] &= E[X_1(\mu_2+\rho\sigma_2\sigma_1^{-1}(X_1-\mu_1))] \\ &= \mu_2E[X_1]+\rho\sigma_2\sigma_1^{-1}E[X_1^2]-\rho\sigma_2\sigma_1^{-1}E[X_1]\mu_1 \\ &= \mu_1\mu_2+\rho\sigma_1\sigma_2 \end{aligned}$$

Therefore, we have

$$Cov(X_1,X_2) = E[X_1X_2]-E[X_1]E[X_2] = \mu_1\mu_2+\rho\sigma_1\sigma_2-\mu_1\mu_2 = \rho\sigma_1\sigma_2$$

as desired.

6. $Cov(X_1,X_2)=0 \iff \rho=0 \iff X_1$ and X_2 are independent. (in BVN only)

Proposition 3.2.2
Let $X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$, if X_1 and X_2 are independent, then

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim BVN\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right)$$

Proof:

$$f(x_1,x_2) = f_1(x_1)f_2(x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}}e^{-\frac{1}{2\sigma_1}(x_1-\mu_1)^2-\frac{1}{2\sigma_2}(x_2-\mu_2)^2} = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}}e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

3.3 Appendix 3 (provided by Lucy Gao)

Is Marginally Normally Distributed Enough to Conclude That Uncorrelated = Independent?

Recall from the week 4 reading that in general, if all you know is that two random variables are uncorrelated, you cannot safely conclude that they are independent. However, we saw in Lecture 11 that the bivariate normal distribution has a very special property: if the joint distribution of X_1 and X_2 is bivariate normal, and $Cov(X_1, X_2) = 0$, then X_1 and X_2 are independent.

If all you remember is that "in the special case of normality, uncorrelated and independent are the same", then it is unfortunately easy to fall into the trap of a non-property:

Non-Property. *Suppose that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Then, $Cov(X_1, X_2) = 0$ implies that X_1 and X_2 are independent.*

We will proceed to debunk this non-property through a cautionary example.

A Cautionary Example

Suppose that $X_1 \sim N(0, 1)$. Define $P(W = 1) = 1/2$ and $P(W = -1 = 1/2)$. (This is called the Rademacher distribution.) Suppose that W and X_1 are independent, and let $X_2 = X_1 W$.

We will first show that $X_2 \sim N(0, 1)$. Observe that for any $t \in \mathbb{R}$,

$$\begin{aligned} P(X_2 \leq t) &= P(X_1 W \leq t) \\ &= P(X_1 \leq t, W = 1) + P(X_1 \geq -t, W = -1) \\ &= P(X_1 \leq t)P(W = 1) + P(X_1 \geq -t)P(W = -1) \\ &= (P(X_1 \leq t) + P(X_1 \geq -t))/2 \\ &= 2P(X_1 \leq t)/2 = P(X_1 \leq t), \end{aligned}$$

where the third equality follows from independence of W and X_1, and the last equality follows from the symmetry of the normal pdf. Thus, the CDF of X_2 is the CDF of a $N(0, 1)$ distribution, and so $X_2 \sim N(0, 1)$.

We will now show that $Cov(X_1, X_2) = 0$. Observe that

$$Cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2] = E[X_1^2 W] - 0 = E[X_1^2]E[W] = E[X_1^2]0 = 0,$$

where the third equality follows from independence of W and X_1.

Finally, we will show that X_1 and X_2 are **dependent**. Recalling the definition of independence

as $P(X_1 \in A, X_2 \in B) = P(X_1 \in A)P(X_2 \in B)$ for any set $A, B \subseteq \mathbb{R}$, it suffices to show that

$$P(X_1 \leq t, X_2 \leq t) \neq P(X_1 \leq t)P(X_2 \leq t)$$

for any $t > 0$. We now do so. Let $t > 0$. Then,

$$\begin{aligned}
&P(X_1 \leq t, X_2 \leq t) \\
&= P(X_1 \leq t, X_1 W \leq t) \\
&= P(-t \leq X_1 \leq t, W = -1) + P(X_1 \leq t, W = 1) \\
&= P(W = -1)P(-t \leq X_1 \leq t) + P(W = 1)P(X_1 \leq t) \\
&= \frac{1}{2}\left(P(-t \leq X_1 \leq t) + P(X_1 \leq t)\right) \\
&= \frac{1}{2}\left(2P(X_1 \leq t) - P(X_1 \leq -t)\right) \\
&= P(X_1 \leq t) - \frac{1}{2}P(X_1 \leq -t) \\
&= P(X_1 \leq t) - \frac{1}{2}(1 - P(X_1 \leq t)) \quad \text{(by symmetry of normal pdf)} \\
&= \frac{3}{2}P(X_1 \leq t) - \frac{1}{2}.
\end{aligned}$$

However,

$$P(X_1 \leq t)P(X_2 \leq t) = \frac{3}{2}P(X_1 \leq t) - \frac{1}{2}.$$

Since we showed that $X_1, X_2 \sim N(0, 1)$, $P(X_1 \leq t) = P(X_2 \leq t)$ and so we have

$$P(X_1 \leq t)^2 - \frac{3}{2}P(X_1 \leq t) + \frac{1}{2} = 0.$$

This is a quadratic equation in $P(X_1 \leq t)$ which has two roots: one at $P(X_1 \leq t) = 1/2$, and one at $P(X_1 \leq t) = 1$. The latter cannot be satisfied for any $t < \infty$, so this is only satisfied for $P(X_1 \leq t) = 1/2$, which cannot be the case for our choice of t, since we assumed that $t > 0$. Thus, we have shown that X_1 and X_2 are dependent!

Conclusion

The key takeaway here is that you really need **bivariate** normality in order to conclude that uncorrelated implies independent – marginal normality is not enough.

3.4 Finding Distribution of Multivariate Random Variables

Three Methods to Find the Distribution of $Y = h(X_1, ..., X_n)$ where $X_1, ..., X_n$ are Random Variables:

1. CDF technique
2. One to one Transformation Theorem (continuous only)
3. MGF technique

Method 1 - CDF Techniques Goal: Find CDF and/or pdf/pmf of $Y = h(X_1, ..., X_n)$

Discrete Case:
For all $y \in \mathbb{R}$,

$$\begin{aligned} f_Y(y) = P(Y = y) &= P(h(X_1, ..., X_n) = y) \\ &= P((x_1, ..., x_n) \in \{(x_1, ..., x_n) : h(x_1, ..., x_n) = y\}) \\ &= \sum_{(x_1, ..., x_n): h(x_1, ..., x_n) = y} P(X_1 = x_1, ..., X_n = x_n) \end{aligned}$$

so for all $y \in \mathbb{R}$,

$$F_Y(y) = P(Y \leq y) = \sum_{t: t \leq y} f_Y(y)$$

■ **Example 3.1** Let $Y = X^2$

$$f_X(x) = \begin{cases} \frac{1}{4} & \textbf{if } |x| = 1 \\ \frac{1}{2} & \textbf{if } x = 0 \\ 0 & \textbf{otherwise} \end{cases}$$

then we have

$$P(Y = y) = P(X^2 = y) = \begin{cases} P(X = \sqrt{y}) + P(X = -\sqrt{y}) & \textbf{if } y > 0 \\ P(X = 0) & \textbf{if } y = 0 \\ 0 & \textbf{if } y < 0 \end{cases} = \begin{cases} \frac{1}{2} & \textbf{if } y = 1 \\ \frac{1}{2} & \textbf{if } y = 0 \\ 0 & \textbf{otherwise} \end{cases}$$

Therefore, $Y \sim Bernouli(\frac{1}{2})$ ■

Continuous Case:
1. $\forall\, y \in \mathbb{R}$, find $R_y = \{(x_1,...,x_n) : h(x_1,...,x_n) \leq y\}$
2. Find CDF pf Y: $\forall\, y \in \mathbb{R}$,

$$F_Y(y) = P(Y \leq y) = P(h(X_1,...,X_n) \leq y) = P((X_1,...,X_n) \in R_y) = \int_{R_y} f(x_1,....,x_n)$$

3. Find the pdf of Y: $f_Y(y) = F_Y'(y)$

■ **Example 3.2** Let $X \sim Exp(\theta)$ and

$$Y = F(X) = \begin{cases} 0 & \textbf{if } x \leq 0 \\ 1 - e^{-\frac{x}{\theta}} & \textbf{if } x > 0 \end{cases}$$

1. $\forall\, y \in \mathbb{R}$, find $R_y = \{x : F(x) \leq y\}$:

$$R_y = \begin{cases} \emptyset & \textbf{if } y \leq 0 \\ \{x : x \leq 0\} & \textbf{if } y = 0 \\ \mathbb{R} & \textbf{if } y \geq 1 \\ \{x : x \leq 0\} \cup \{x > 0 : x \leq -\theta \log(1-y)\} & \textbf{if } 0 < y < 1 \end{cases}$$

Then we have

$$F_y(y) = \begin{cases} 0 & \textbf{if } y \leq 0 \\ 0 & \textbf{if } y = 0 \\ 1 & \textbf{if } y \geq 1 \\ P(X \leq -\theta \log(1-y)) & \textbf{if } 0 < y < 1 \end{cases} = \begin{cases} 0 & \textbf{if } y \leq 0 \\ y & \textbf{if } 0 < y < 1 \\ 1 & \textbf{if } y \geq 1 \end{cases}$$

so we see that $Y \sim Unif(0,1)$, the pdf is trivial. ■

■ **Example 3.3** Let $X \sim N(0,1)$ and $Y = X^2$

1. $\forall\, y \in \mathbb{R}$

$$R_y = \left\{x : x^2 \leq y\right\} = \begin{cases} \emptyset & \textbf{if } y < 0 \\ \left[-\sqrt{y}, \sqrt{y}\right] & \textbf{if } y \geq 0 \end{cases}$$

2.

$$F_Y(y) = P(X \in R_y) = \begin{cases} 0 & \textbf{if } y < 0 \\ P\left(X \in \left[-\sqrt{y}, \sqrt{y}\right]\right) & \textbf{if } y \geq 0 \end{cases}$$

3.

$$\begin{aligned} f_Y(y) = F_Y'(y) &= F_X'(\sqrt{y})\frac{1}{2\sqrt{y}} - F_X'(-\sqrt{y})\left(-\frac{1}{2\sqrt{y}}\right) \\ &= \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) \\ &= \frac{1}{2\sqrt{y}}\frac{2}{\sqrt{2\pi}}e^{-\frac{y}{2}} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-\frac{y}{2}}}_{\textbf{pdf of } \chi_1^2} \quad \textbf{for } y > 0 \end{aligned}$$

so we have $y \sim \chi_1^2$ ■

■ **Example 3.4** Let $X_1, X_2 \sim Unif(0,1)$ and i.i.d, so

$$f(x_1,x_2) = \begin{cases} 1 & \textbf{if } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \textbf{otherwise} \end{cases}$$

Now we let $Y = X_1 + X_2$ and $A = (0,1) \times (0,1)$

Then, we can immediately get that

$$F_Y(y) = \begin{cases} 0 & \textbf{if } y \le 0 \\ \frac{y^2}{2} & \textbf{if } 0 < y \le 1 \\ 1 - \frac{(2-y)^2}{2} & \textbf{if } 1 \le y < 2 \\ 1 & \textbf{if } y \ge 2 \end{cases} \qquad \textbf{and} \qquad f_Y(y) = \begin{cases} y & \textbf{if } 0 < y \le 1 \\ 2-y & \textbf{if } 1 \le y < 2 \\ 0 & \textbf{otherwise} \end{cases}$$

■

■ **Example 3.5** Let $X_i \sim Unif(0,\theta)$ be i.i.d for $1 \le i \le n$. Find the distribution of $X_{(n)} = \max_{1 \le i \le n} X_i$ and $X_{(1)} = \min_{1 \le i \le n} X_i$

$$F_{X_{(n)}} = P(X_{(n)} \le y) = P(X_i \le y \ , \ \forall \ 1 \le i \ne n) = \prod_{i=1}^{n} P(X_i \le y) = [F(y)]^n = \begin{cases} 0 & \textbf{if } y \le 0 \\ \left(\frac{y}{\theta}\right)^n & \textbf{if } 0 < y \le \theta \\ 1 & \textbf{if } y > \theta \end{cases}$$

and

$$f_{X_{(n)}}(y) = \begin{cases} \frac{n}{\theta^n}y^{n-1} & \textbf{if } 0 < y < \theta \\ 0 & \textbf{otherwise} \end{cases}$$

Note that

$$P(X_{(1)} \le y) = 1 - P(X_{(1)} > y) = 1 - P(X_i > y \ \ \forall \ 1 \le i \le n) = 1 - \prod_{i=1}^{n} P(X_i > y) = 1 - \prod_{i=1}^{n}(1 - P(X_i \le y))$$

so similarly, we have

$$F_{X_{(1)}} = \begin{cases} 0 & \textbf{if } y \le 0 \\ 1 - \left(1 - \frac{y}{\theta}\right)^n & \textbf{if } 0 < y \le \theta \\ 1 & \textbf{if } y > \theta \end{cases} \qquad \textbf{and} \qquad f_{X_{(1)}}(y) = \begin{cases} \frac{n}{\theta}\left(1 - \frac{y}{\theta}\right)^{n-1} & \textbf{if } 0 < y < \theta \\ 0 & \textbf{otherwise} \end{cases}$$

■

3.5 One to One Transformation

Special Case:
1. $n = 1$, X is continuous with support A
2. h is $1-1$ on A, that is for all $x_1, x_2 \in A, h(x_1) = h(x_2) \implies x_1 = x_2$

Theorem 3.5.1 — (Univariate) One to One Transformation Theorem.
If X is continuous with support A and $h(x)$ is a one-to-one function on A, then the pdf of $Y = h(X)$ is

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \cdot \left| \frac{d}{dy} h^{-1}(y) \right| & \textbf{if } y \in \{h(x) : x \in A\} \\ 0 & \textbf{otherwise} \end{cases}$$

where h^{-1} satisfies $h^{-1}(h(x)) = x$ for all $x \in A$

Proof: Apply CDF technique

■ **Example 3.6** Let X be a random variable with

$$f_X(x) = \begin{cases} \frac{\theta}{x^{\theta+1}} & \textbf{if } x > 1 \\ 0 & \textbf{otherwise} \end{cases}$$

and $A = (1, \infty)$, $h(x) = \log(x)$, $Y = \log(X)$
It's obviously that h is one-to-one. By the **Theorem** above, we have

$$f_Y(y) = \begin{cases} \theta e^{-\theta y} & \textbf{if } y > 0 \\ 0 & \textbf{otherwise} \end{cases}$$

■

3.6 Appendix 4 (provided by Lucy Gao)

Convolution of Probability Distributions

You may have come across *convolutions* of two functions in electrical engineering, physics, or other fields. A *convolution* is a mathematical operation on two functions f and g defined as follows:

$$f * g(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau.$$

This operation appears in applications across many fields, including signal/image processing, physics, and statistics. In this reading, we will show that the probability density function (pdf) of the sum of two independent continuous random variables X and Y is given by the convolution of the pdf of X and the pdf of Y.

Derivation of the Convolution Formula

Suppose that X and Y are two independent continuous random variables. Let $f_X(t)$ denote the pdf of X and $f_Y(t)$ denote the pdf of Y. We will show that the pdf of $X+Y$ is given by:

$$f_{X+Y}(t) = f_X * f_Y(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx.$$

We will use the CDF technique. Observe that

$$\begin{aligned} F_{X+Y}(t) = P(X+Y \leq t) &= \int_{-\infty}^{\infty}\int_{-\infty}^{t-x} f_X(x)f_Y(y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{t} f_X(x)f_Y(u-x)du\ dx \\ &= \int_{-\infty}^{t}\left[\int_{-\infty}^{\infty} f_X(x)f_Y(u-x)dx\right]\ du, \end{aligned}$$

where in the first line, we have applied the independence of X and Y to factorize their joint pdf, in the second line, we have applied the change of variables $u = x + y$ to the inner integral, and in the last line, we have used the fact that the integral bounds do not depend on x or u to interchange the order of integration.

We can now differentiate the CDF to get the pdf of $X+Y$. Applying the fundamental theorem of calculus yields

$$f_{X+Y}(t) = \frac{d}{dt}F_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx.$$

Note that applying the convolution formula with the roles of X and Y swapped says that the pdf of $Y+X = X+Y$ is given by:

$$f_{Y+X}(t) = f_Y * f_X(t) - \int_{-\infty}^{\infty} f_Y(y)f_X(t-y)dy.$$

Thus, to find the pdf of $X+Y$, you can either do the convolution $f_X * f_Y$ or the convolution $f_Y * f_X$. Either works.

Application of the Convolution Formula

In Lecture 14, we show that if X and Y are independent Exponential(1) random variables, then $X + Y \sim \text{Gamma}(2, 1)$ via the bivariate one-to-one transformation theorem. This fact can also be shown using the convolution theorem. We have:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x) dx, \tag{1}$$

where

$$f_Y(u) = f_X(u) = \begin{cases} e^{-u}, & u > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Observe that for any $t > 0$, we have $f_X(x) f_Y(t-x) > 0$ if and only if $x > 0$ and $t - x > 0$, which in turn is the case if and only if $0 < x < t$. So when $t > 0$, we can simplify (1) as:

$$f_{X+Y}(t) = \int_0^t e^{-x} e^{x-t} dx = \int_0^t e^{-t} dx = te^{-t}.$$

For any $t \leq 0$, we have $f_X(x) f_Y(t-x) = 0$ when $x \leq 0$ (because $f_X(x) = 0$ for $x \leq 0$) and $f_X(x) f_Y(t-x) = 0$ for $x > 0$ (because then $t - x < 0$ and so $f_Y(t-x) = 0$). Thus, when $t \leq 0$, we can simplify (1) as:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} 0 dx = 0.$$

Putting the pieces together, we get

$$f_{X+Y}(t) = \begin{cases} te^{-t}, & t > 0, \\ 0, & \text{otherwise,} \end{cases}$$

which is indeed the pdf of the Gamma$(2, 1)$ distribution.

Conclusion

The convolution formula is a useful tool for calculating the pdf of sums of independent and continuous random variables, although you do have to be careful when working with piecewise defined pdfs. The ideas behind the convolution formula can also be extended to derive similar formulas for the difference, product, and quotient of X and Y for independent and continuous X and Y.

Theorem 3.6.1 — Bivariate One to One Transformation Theorem.
If $U = h_1(X,Y)$ and $V = h_2(X,Y)$ defined a one-to-one transform on the joint support.
Let $A = \{(x,y) \in \mathbb{R}^2 : f(x,y) > 0\}$, then the joint pdf of U and V is

$$g(u,v) = \begin{cases} f(w_1(u,v), w_2(u,v)) \cdot \left| \frac{\partial(w_1,w_2)}{\partial(u,v)} \right| & \textbf{if } (u,v) \in \{(h_1(x,y), h_2(x,y)) : (x,y) \in A\} \\ 0 & \textbf{otherwise} \end{cases}$$

where

$$\frac{\partial(w_1,w_2)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial w_1}{\partial u} & \frac{\partial w_1}{\partial v} \\ \frac{\partial w_2}{\partial u} & \frac{\partial w_2}{\partial v} \end{vmatrix} = \frac{\partial w_1}{\partial u} \cdot \frac{\partial w_2}{\partial v} - \frac{\partial w_1}{\partial v} \cdot \frac{\partial w_2}{\partial u}$$

■ **Example 3.7** Let $U = X+Y$, $V = X-Y$ and

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \qquad f(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \qquad A \in \mathbb{R}^2$$

Then we have $X = \frac{1}{2}(U+V) = w_1(U,V)$ and $Y = \frac{1}{2}(U-V) = w_2(U,V)$, so by the **Theorem** above we have

$$g(u,v) = \frac{1}{4\pi} e^{-\frac{1}{4}(u^2+v^2)} \quad \forall\ (u,v) \in \mathbb{R}^2$$

■

■ **Example 3.8** Let X,Y be independent and $X,Y \sim Exp(1)$ with

$$f(x,y) = \begin{cases} e^{-x}e^{-y} & \textbf{if } 0 < x < \infty, 0 < y < \infty \\ 0 & \textbf{otherwise} \end{cases}$$

We will show $X+Y \sim Gamma(2,1)$. Let $U = X+Y$ and $V = X$, so $X = V = w_1(U,V)$ and $Y = w_2(U,V) = U-V$, then by the **Theorem** above we have

$$g(u,v) = \begin{cases} e^{-u} & \textbf{if } 0 < v < u < \infty \\ 0 & \textbf{otherwise} \end{cases}$$

so that

$$g_1(u) = \begin{cases} \int_0^u e^{-u} dv & \textbf{if } 0 < u < \infty \\ 0 & \textbf{otherwise} \end{cases} = \begin{cases} ue^{-u} & \textbf{if } 0 < u < \infty \\ 0 & \textbf{otherwise} \end{cases}$$

which is a pdf of $Gamma(2,1)$, so $X+Y \sim Gamma(2,1)$ ■

■ **Example 3.9** Let the support of X,Y to be $\{(x,y) \in \mathbb{R}^2 : x > 0,\ y > 0\}$, then the support of

$$U = X+Y \qquad \textbf{and} \qquad V = \frac{X}{X+Y}$$

will be

$$\{(u,v) \in \mathbb{R}^2 : 0 < v < 1,\ u > 0\}$$

■

3.7 MGF Technique and Distributions defined by Transformations

Proposition 3.7.1
Let $X_1, ..., X_n$ be i.i.d and each $X_i \sim N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \frac{X_i}{n} \sim N\left(\sum_{i=1}^{n} \frac{\mu}{n}, \sum_{i=1}^{n} \frac{\sigma^2}{n^2}\right) = N\left(\mu, \frac{\sigma^2}{n}\right)$$

Definition 3.7.1 — Distributions defined by Transformations.

1. If $\exists Z_1, ..., Z_K \sim N(0,1)$ are i.i.d such that $X = \sum_{i=1}^{n} Z_i^2$, then

$$X \sim \chi^2_{(k)} \qquad \textbf{Chi-Squared Distribution}$$

2. If $\exists Z \sim N(0,1)$, $Y \sim \chi^2_{(n)}$ and they are independent such that $X = \frac{Z}{\sqrt{\frac{Y}{n}}}$, then

$$X \sim t_{(n)} \qquad \textbf{t Distribution}$$

3. If $\exists Y_1 \sim \chi^2_{(n_1)}$ and $Y_2 \sim \chi^2_{(n_2)}$ and they are independent such that $X = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}}$, then

$$X \sim F(n_1, n_2)$$

Proposition 3.7.2
1. $X \sim \chi^2_{(k)}$ has MGF:

$$M(t) = \frac{1}{(1-2t)^{\frac{k}{2}}} \qquad \forall\, t < \frac{1}{2}$$

2. Let $X_1, ..., X_n$ be independent and each $X_i \sim \chi^2_{(k_i)}$, then

$$\sum_{i=1}^{n} X_i \sim \chi^2_{(\sum_{i=1}^{n} k_i)}$$

Proof for 2: Apply **Proposition 2.9.2**

Theorem 3.7.3
If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are i.i.d, let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)} \qquad \textbf{where} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Proof: See **Cochran's theorem** and **Lemma 3.7.4**

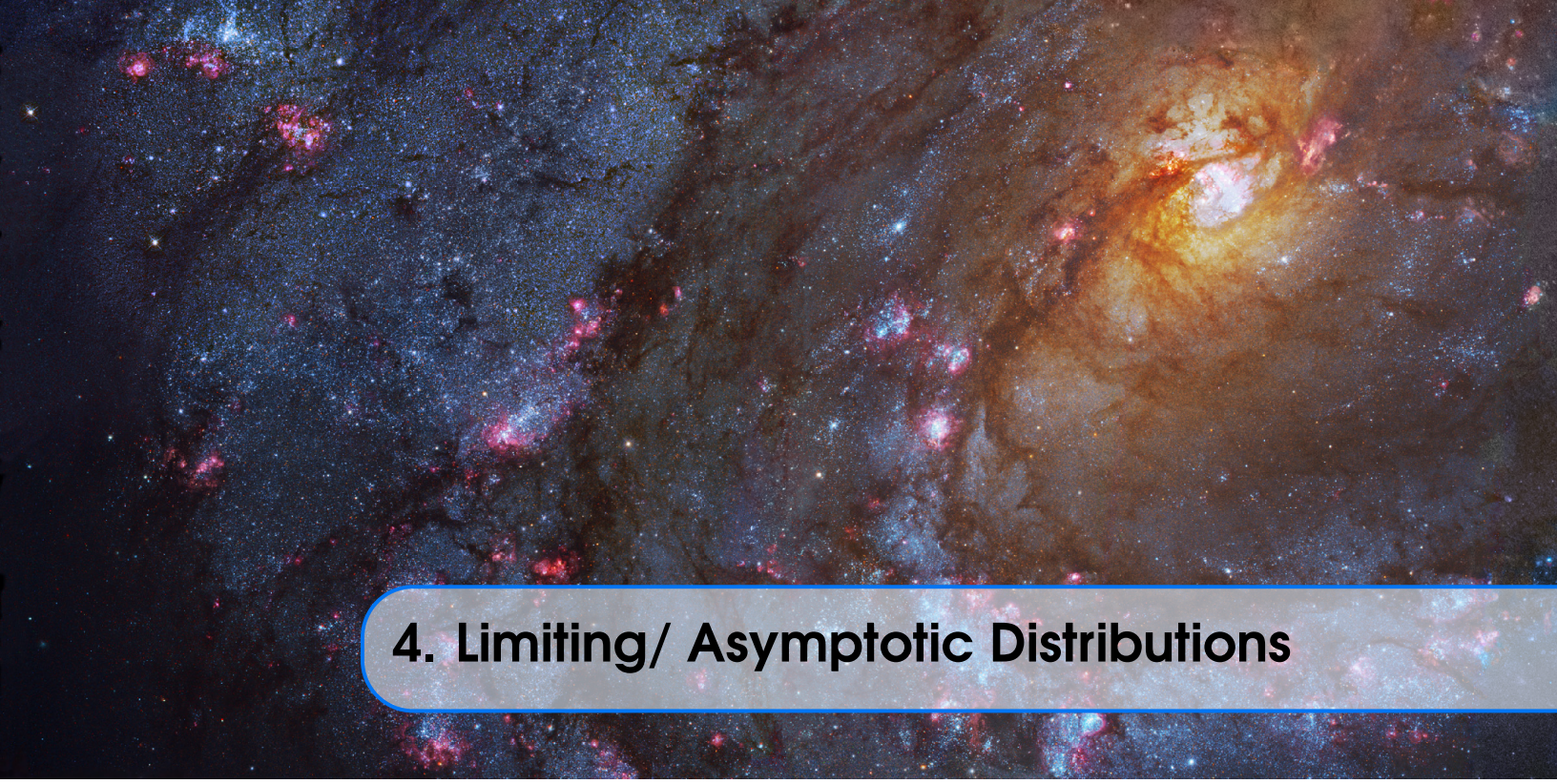
Lemma 3.7.4
If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are i.i.d, then $\overline{X}$ and s^2 are independent.

Theorem 3.7.5
If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are i.i.d, then

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

Theorem 3.7.6
If $X_1, ..., X_n \sim N(\mu_1, \sigma_1^2)$ are i.i.d and $Y_1, ..., Y_n \sim N(\mu_2, \sigma_2^2)$ are also i.i.d where each X_i,Y_j are independent for all i, j, then

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$



4. Limiting/ Asymptotic Distributions

4.1 Convergence in Distributions

Definition 4.1.1 — Convergence in Distributions - CDF Converges Pointwise.
Let $X_1, X_2, \ldots$ be a sequence of random variables where X_n have CDF $F_n(x)$ for all $n \in \mathbb{N}$ and X has a CDF $F(x)$, we say

$$X_n \xrightarrow{d} X \qquad (X_n \textbf{ converges in distributions to } X)$$

If

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \forall\, x \quad \textbf{where } F \textbf{ is continuous at } x$$

We call F is the limiting/asymptotic distribution of X_n

■ **Example 4.1** Let $W \sim Bernoulli(\frac{1}{2})$ where $X_n = W$ for all $n \in \mathbb{N}$ and $X = 1 - W$. Hence we have

$$X_n \xrightarrow{d} X$$

but is $\lim_{n\to\infty} X_n = X$? Nope, because $|X_n - X| = 1$ for all $n \in \mathbb{N}$ ■

■ **Example 4.2** Let $X_n \sim Unif(0, \frac{1}{n})$ and $X = 0$, show $X_n \xrightarrow{d} X$
Note that

$$F_n(x) = \begin{cases} 0 & \textbf{if } x \le 0 \\ nx & \textbf{if } 0 < x < \frac{1}{n} \\ 1 & \textbf{if } x \ge \frac{1}{n} \end{cases} \qquad \Longrightarrow \qquad \lim_{n\to\infty} F_n(x) = 0 \quad \forall\, x \le 0 \quad \textbf{and} \quad \lim_{n\to\infty} F_n(x) = 1 \quad \forall\, x > 0$$

Then let F be the cdf of X, we have

$$\lim_{n\to\infty} F_n(x) = F(x) = \begin{cases} 0 & \textbf{if } x < 0 \\ 1 & \textbf{if } x \ge 0 \end{cases} \qquad \Longrightarrow \qquad X_n \xrightarrow{d} X$$

■

Proposition 4.1.1
Let $b, c \in \mathbb{R}$ and $\varphi(n)$ s.t. $\lim_{n\to\infty} \varphi(n) = 0$, then

$$\lim_{n\to\infty} \left[1 + \frac{b}{n} + \frac{\varphi(n)}{n}\right]^{cn} = e^{bc}$$

idea: consider the case $\varphi(n) = 0$ for all $n \in \mathbb{N}$, then applying the limit:

$$\lim_{n\to\infty} \left[1 + \frac{b}{n}\right]^{cn} = e^{bc}$$

■ **Example 4.3** Let $X_1, ..., X_n \sim Unif(0,1)$ are i.i.d, $X_{(1)} = \min\{X_1, ..., X_n\}$ and $X_{(n)} = \max\{X_1, ..., X_n\}$

1. Find the asymptotic distribution of $nX_{(1)}$

$$\begin{aligned} F_n(x) = P(nX_{(1)} \le x) = P\left(X_{(1)} \le \frac{x}{n}\right) &= \begin{cases} 0 & \textbf{if } \frac{x}{n} \le 0 \\ 1 - \left(1 - \frac{x}{n}\right)^n & \textbf{if } 0 < \frac{x}{n} \le 1 \\ 1 & \textbf{if } \frac{x}{n} > 1 \end{cases} \\ &= \begin{cases} 0 & \textbf{if } x \le 0 \\ 1 - \left(1 - \frac{x}{n}\right)^n & \textbf{if } 0 < x \le n \\ 1 & \textbf{if } x > n \end{cases} \\ &\xrightarrow{n\to\infty} \begin{cases} 0 & \textbf{if } x \le 0 \\ 1 - e^{-x} & \textbf{if } x > 0 \end{cases} \end{aligned}$$

That is $nX_{(1)} \xrightarrow{d} Exp(1)$

2. Find asymptotic of $n(1 - X_{(n)})$

$$\begin{aligned} F_n(x) = P(n(1 - X_{(n)}) \le x) = P\left(X_{(n)} \ge 1 - \frac{x}{n}\right) &= 1 - P\left(X_{(n)} \le 1 - \frac{x}{n}\right) \\ &= 1 - F_{X_{(n)}}(x)\left(1 - \frac{x}{n}\right) \\ &= \begin{cases} 1 & \textbf{if } 1 - \frac{x}{n} \le 0 \\ 1 - (1 - \frac{x}{n})^n & \textbf{if } 0 < 1 - \frac{x}{n} \le 1 \\ 0 & \textbf{if } 1 - \frac{x}{n} > 1 \end{cases} \\ &= \begin{cases} 0 & \textbf{if } x \le 0 \\ 1 - (1 - \frac{x}{n})^n & \textbf{if } 0 < x \le n \\ 1 & \textbf{if } x > n \end{cases} \\ &\xrightarrow{n\to\infty} \begin{cases} 0 & \textbf{if } x \le 0 \\ 1 - e^{-x} & \textbf{if } x > 0 \end{cases} \end{aligned}$$

That is $n(1 - X_{(n)}) \xrightarrow{d} Exp(1)$ ■

4.2 Convergence in Probability

Definition 4.2.1 — Convergence in Probability - Converge in Probability Measure.
Let $X_1, X_2, \ldots$ be a sequence of random variables and X be a random variable. We say that $X_n \xrightarrow{p} X$ (X_n converges to X in prob) if

$$\forall\ \varepsilon > 0 \quad \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

or equivalently

$$\forall\ \varepsilon > 0 \quad \lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$$

■ **Example 4.4** Let $W \sim Unif(0,1)$, $X = 0$ and

$$X_n = \begin{cases} 1 & \textbf{if } 0 < w < \frac{1}{n} \\ 0 & \textbf{otherwise} \end{cases}$$

Then for all $\varepsilon > 0$:

$$P(|X_n - X| \geq \varepsilon) = P(X_n \geq \varepsilon) = P(X_n = 1) = P\left(0 < W < \frac{1}{n}\right) = \lim_{n\to\infty} \int_0^{\frac{1}{n}} 1dx = \lim_{n\to\infty} \frac{1}{n} = 0$$

That is $X_n \xrightarrow{p} X$ as desired. ■

■ **Example 4.5** Let $X_n \sim Bernouli(1 - \frac{1}{n})$ and $X = 1$, then for $\varepsilon > 0$:

$$P(|X_n - X| \geq \varepsilon) = \begin{cases} \frac{1}{n} & \textbf{if } 0 < \varepsilon \leq 1 \\ 0 & \textbf{if } \varepsilon > 1 \end{cases} \qquad \Longrightarrow \qquad \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

or

$$P(|X_n - X| \in \{0,1\}) = P(|X_n - X| = 1) + P(|X_n - X| = 0) = P(X_n = 1) + P(X_n = 0) = 1 - \frac{1}{n} + \frac{1}{n} = 1$$

Therefore, we have $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$, that is $X_n \xrightarrow{p} X = 1$ ■

4.3 Probability Limits Theorems

Theorem 4.3.1
If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$. (Converse is not always true, but if X is constant random variable, this will be true)
Proof: Proofs of Convergence of Random Variables

■ **Example 4.6** Let $W \sim Bernoulli(\frac{1}{2})$ and $X_n = W$ for all $n \in \mathbb{N}$. Let $X = 1 - W$, so $X \sim Bernoulli(\frac{1}{2})$ and $X_n \xrightarrow{d} X$ but $P(|X_n - X| \geq 1) = P(|X_n - X| = 1) = 1$, so $\lim_{n\to\infty} P(|X_n - X| \geq 1) = 1 \neq 0$, so $X_n \not\xrightarrow{p} X$ ■

Theorem 4.3.2
Let $c \in \mathbb{R}$, if $X_n \xrightarrow{d} c$, then $X_n \xrightarrow{p} c$
Proof: Since $X_n \xrightarrow{d} c$, that is for any $\varepsilon > 0$, we have $P(|X_n - X| \geq \varepsilon) \geq 0$ and

$$\begin{aligned} P(|X_n - c| \geq \varepsilon) &= P(X_n \geq c + \varepsilon) + P(X_n \leq c - \varepsilon) \\ &= 1 - P(X_n < c + \varepsilon) + F_n(c - \varepsilon) \\ &\leq 1 - P\left(X_n \leq c + \frac{\varepsilon}{2}\right) + F_n(c - \varepsilon) \\ &= 1 - F_n\left(c + \frac{\varepsilon}{2}\right) + F_n(c - \varepsilon) \end{aligned}$$

Note that the CDF for c:

$$F(x) = \begin{cases} 0 & \textbf{if } x < c \\ 1 & x \geq c \end{cases}$$

and $X_n \xrightarrow{d} c$ implies $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \neq c$, then

$$0 \leq \lim_{n\to\infty} P(|X_n - c| \geq \varepsilon) \leq \lim_{n\to\infty} 1 - F_n\left(c + \frac{\varepsilon}{2}\right) + F_n(c - \varepsilon) = 1 - 1 + 0 = 0$$

By Squeeze theorem, we have $\lim_{n\to\infty} P(|X_n - c| \geq \varepsilon) = 0$, that is $X_n \xrightarrow{p} c$ as desired.

Proposition 4.3.3 — Markov Inequality.
Let X be a random variable, then for all $k > 0$, $c > 0$:

$$P(|X| \geq c) \leq \frac{E[|X|^k]}{c^k}$$

(we usually take $k = 2$)

Proposition 4.3.4 — Chebyshev's Inequality.
Let X be a random variable, the for all $k > 0$:

$$P(|X - E[X]| > k\sqrt{\text{Var}[X]}) \leq \frac{1}{k^2}$$

■ **Example 4.7** Let $Y \sim Unif(0,1)$ and $X_n = Y^n$, show $X_n \xrightarrow{p} X$ for $X = 0$.
Note that

$$E[X_n] = E[Y^n] = \int_0^1 y^n dy = \frac{1}{n+1}$$

Let $\varepsilon > 0$, then

$$P(|X_n - X| \geq \varepsilon) \leq \frac{E[|X_n - X|]}{\varepsilon} = \frac{E[X_n]}{\varepsilon} = \frac{1}{(n+1)\varepsilon} \to 0$$

By Squeeze Theorem, we have $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$, that is $X_n \xrightarrow{p} X$ as desired. ■

Proposition 4.3.5
Let $X_1, X_2, ..$ be a sequence of random variables with $E[X_n] = \mu$, $\text{Var}[X_n] = \sigma_n^2 > 0$ and $\lim_{n\to\infty} \sigma_n^2 = 0$, then $X_n \xrightarrow{p} \mu$
Proof: Let $\varepsilon > 0$, then

$$0 \leq P(|X_n - \mu| \geq \varepsilon) \leq \frac{E[(X_n - \mu)^2]}{\varepsilon^2} = \frac{\text{Var}[X_n]}{\varepsilon^2} = \frac{\sigma_n^2}{\varepsilon^2} \to 0$$

Then by Squeeze theorem, we have $\lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) = 0$, that is $X_n \xrightarrow{p} \mu$ as desired.

Proposition 4.3.6 — Weak Law of Large Numbers (WLLN).
Let X_n be a sequence of i.i.d random variables with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$, then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

Proof: Note that

$$E[\overline{X}_n] = \frac{1}{n} \cdot \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot \sum_{i=1}^{n} \mu = \mu$$

and

$$\text{Var}[\overline{X}_n] = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \text{Var}[X_i] = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n} \to 0$$

Then we have

$$0 \leq P(|X_n - E[\overline{X}_n]| \geq \varepsilon) \leq \frac{E[(X_n - E[\overline{X}_n])^2]}{\varepsilon^2} = \frac{\text{Var}[X_n]}{\varepsilon^2} \to 0$$

That is $\overline{X}_n \xrightarrow{p} \mu$, which completes the proof.

■ **Example 4.8** Let X_n be a sequence of i.i.d random variables such that $X_i \sim Bernouli(p)$, so we have $E[X_i] = p$ and $\text{Var}[X_i] = p(1-p) < \infty$, then by **WLLN**: $\overline{X}_n \xrightarrow{p} p$ ■

■ **Example 4.9** Let X_n be a sequence of i.i.d random variables such that $X_i \sim Bernouli(p)$. We define

$$W_n = \frac{1}{n}\sum_{i=1}^{n} 2^{X_i}$$

Find $W_n \xrightarrow{p} ?$:
First we define $Y_i = 2^{X_i}$, so Y_i are also i.i.d. Then we have

$$E[Y_i] = 2^0(1-p) + 2^1 p = 1 + p \qquad \textbf{and} \qquad E[Y_i^2] = 2^{2\cdot 0}(1-p) + 2^{2\cdot 1} p = 1 + 3p$$

This gives us that $\text{Var}[Y_i^2] = E[Y_i^2] - (E[Y_i])^2 < \infty$, then by **WLLN** $W_n \xrightarrow{p} 1 + p$. ■

Lemma 4.3.7
Let X_n be a sequence of random variables with MGFs $M_n(t)$ and X is a random variable with MGF $M(t)$. If $\exists h > 0$ s.t. $\lim_{n\to\infty} M_b(t) = M(t)$ for all $|t| < h$, then $X_n \xrightarrow{d} X$
Proof: omitted, too much math for this class.

Theorem 4.3.8 — Central Limit Theorem (CLT).
Let $X_1, ...X_n$ be i.i.d random variables with $E[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2 < \infty$, then

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

where

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Proof:
Let $M_n(t)$ be the MGF of $\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}$, so we have

$$\begin{aligned} M_n(t) = E\left[\exp\left\{t\cdot\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\right\}\right] &= E\left[\exp\left\{\frac{t}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right\}\right] \qquad \textbf{Let } Y_i = \frac{X_i-\mu}{\sigma} \\ &= \prod_{i=1}^{n} E\left[\exp\left\{\frac{tY_i}{\sqrt{n}}\right\}\right] \qquad \textbf{since } Y_i \textbf{ are i.i.d} \\ &= \left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \end{aligned}$$

and MGF of $N(0,1)$ is $M(t) = \exp\left\{\frac{t^2}{2}\right\}$ for all $t \in \mathbb{R}$
Now we will show that $\lim_{n\to\infty} M_n(\frac{t}{\sqrt{n}}) = M(t)$. By Taylor series:

$$\begin{aligned} M_Y\left(\frac{t}{\sqrt{n}}\right) &= M_Y(0) + M_Y'(0)\cdot\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2}M_Y(0)''\cdot\left(\frac{t}{\sqrt{n}}\right)^2 + O\left(\left(\frac{t}{\sqrt{n}}\right)^2\right) \\ &= 1 + \frac{t^2}{2n} + O\left(\left(\frac{t}{\sqrt{n}}\right)^2\right) \end{aligned}$$

then we have

$$\lim_{n\to\infty}\left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right] = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + O\left(\left(\frac{t}{\sqrt{n}}\right)^2\right)\right]^n = \exp\left\{\frac{t^2}{2}\right\}$$

Then by **Lemma 4.3.7** we have $X_n \xrightarrow{d} X$ as desired.

■ **Example 4.10** Let $X_1, ..., X_n \sim Poi(\lambda)$ be i.i.d , then

$$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}} \xrightarrow{d} N(0,1) \qquad \textbf{By CLT}$$

■

■ **Example 4.11** Let $X_n \sim \chi^2_{(n)}$, then $E[\chi^2_{(n)}] = 1$ and $\text{Var}\left[\chi^2_{(1)}\right] = 2$, then by **CLT**

$$\frac{\sqrt{n}\left(\frac{\sum_{i=1}^n \chi^2_{(1)}}{n} - 1\right)}{\sqrt{2}} \xrightarrow{d} N(0,1)$$

Then we have

$$\frac{X_n - n}{\sqrt{2n}} \xrightarrow{d} N(0,1)$$

■

■ **Example 4.12** Let $X_1, X_2, ... \sim Unif(0,1)$ are i.i.d, then

$$\frac{\sqrt{n}\left(\sum_{i=1}^n \frac{X_i}{n} - \frac{1}{2}\right)}{\sqrt{\frac{1}{12}}} \xrightarrow{d} N(0,1)$$

Let $Y_i = X_i^3$ so Y_i are i.i.d so that

$$E[Y_i] = \int_0^1 x^3 dx = \frac{1}{4} \quad \textbf{and} \quad E[Y_i^2] = \int_0^1 x^6 dx = \frac{1}{7} \quad \Longrightarrow \quad \text{Var}[Y_i] = \frac{1}{7} - \left(\frac{1}{4}\right)^2 < \infty$$

Then by **CLT**:

$$\frac{\sqrt{n}\left(\sum_{i=1}^n \frac{X_i^3}{n} - \frac{1}{4}\right)}{\sqrt{\text{Var}[Y_1]}} = \frac{\sqrt{n}\left(\overline{Y}_n - E[Y_1]\right)}{\sqrt{\text{Var}[Y_1]}} \xrightarrow{d} N(0,1)$$

■

Theorem 4.3.9 — Continuous Mapping Theorem.
Let g be a continuous function, X_n be a sequence of random variables and X be a random variable

1. If $X_n \xrightarrow{p} c$, then $g(X_n) \xrightarrow{p} c$

2. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$

Theorem 4.3.10 — Slutsky's Theorem.
If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

1. $X_n + Y_n \xrightarrow{d} X + c$ and $X_n + Y_n \xrightarrow{p} X + c$

2. $X_n Y_n \xrightarrow{d} cX$ and $X_n Y_n \xrightarrow{p} cX$

3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ and $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{c}$ (when $c \neq 0$)

Note: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, does not always implies that $X_n + Y_n \xrightarrow{d} X + Y$

■ **Example 4.13** If $X_n \geq 0$ and $c \geq 0$, then by **Continuous Mapping Theorem**:

$$X_n \xrightarrow{p} c \quad \Longrightarrow \quad \sqrt{X_n} \xrightarrow{p} \sqrt{c} \qquad \textbf{and} \qquad X_n^2 \xrightarrow{p} c^2$$

If $X_n \xrightarrow{d} X \sim N(0,1)$, then by **Continuous Mapping Theorem**:

$$2X_n + 1 \xrightarrow{d} 2X + 1 \sim N(1,4) \qquad \textbf{and} \qquad X_n^2 \xrightarrow{d} X^2 \sim \chi^2_{(1)}$$

If $X_n \xrightarrow{d} X \sim N(0,1)$ and $Y_n \xrightarrow{p} c$ with $c \neq 0$, then

$$X_n + Y_n \xrightarrow{d} X + c \sim N(c,1) \qquad X_n Y_n \xrightarrow{d} CX \sim N(0,c^2) \qquad \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} \sim N(0, \frac{1}{c^2})$$

■

■ **Example 4.14** Let $X_n \sim Poi(\lambda)$ be a sequence of i.i.d random variables, define $U_n = \sqrt{n}(\overline{X}_n - \lambda)$ and $Z_n = \frac{U_n}{\sqrt{\overline{X}_n}}$. By **CLT**:

$$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}} \xrightarrow{d} N(0,1)$$

Let $g(t) = \sqrt{\lambda} t$ and we see that

$$U_n = \sqrt{n}(\overline{X}_n - \lambda) \cdot \frac{\sqrt{\lambda}}{\sqrt{\lambda}} = g\left(\sqrt{n}(\overline{X}_n - \lambda) \cdot \frac{1}{\sqrt{\lambda}}\right)$$

By **Comtinuous Mapping Theorem** we have

$$g\left(\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}}\right) \xrightarrow{d} g(N(0,1)) = \lambda \cdot N(0,1) = N(0,\lambda)$$

By **WLLN** $\overline{X}_n \xrightarrow{p} \lambda$ and define $h(x) = \sqrt{x}$, then $g(\overline{X}_n) \xrightarrow{p} g(\lambda) = \sqrt{\lambda}$
Since $\sqrt{\overline{X}_n} \xrightarrow{p} \sqrt{\lambda}$ and $U_n \xrightarrow{d} N(0,\lambda)$, then we have

$$Z_n = \frac{U_n}{\sqrt{\overline{X}_n}} \xrightarrow{d} N(0,1)$$

■

■ **Example 4.15** Let $X_n \sim Unif(0,1)$ be i.i.d and $U_n = \max_{1 \leq i \leq n} X_i$ and $V_n = e^{-n(1-U_n)}$. Then it's easy to see that $n(1-U_n) \xrightarrow{d} Exp(1)$. Now we let $T = e^{-y}$ so and $Y \sim Exp(1)$

$$F_T(t) = P(e^{-y} \leq t) = \begin{cases} 0 & \textbf{if } t \leq 0 \\ P(Y \geq -\log(t)) & \textbf{if } t > 0 \end{cases} = \begin{cases} 0 & \textbf{if } t \leq 0 \\ t & \textbf{if } 0 < t < 1 \\ 1 & \textbf{if } t \geq 1 \end{cases}$$

Then we have

$$V_n = e^{-n(1-U_n)} \xrightarrow{d} Unif(0,1)$$

Now define $W_n = \frac{n(1-U_n)}{\overline{X}_n^2}$, we see that $\overline{X}_n \xrightarrow{p} E[X_i] = \frac{1}{2}$ by **WLLN** and $\text{Var}[X_i] = \frac{1}{12} < \infty$, so we have continuous $g(t) = t^2$ s.t.

$$\overline{X}_n^2 = g(\overline{X}_n) \xrightarrow{p} \left(\frac{1}{2}\right)^2 = \frac{1}{4} \neq 0$$

Then by **Slutsky's Theorem** we have $W_n = \frac{n(1-U_n)}{\overline{X}_n^2} \xrightarrow{d} \frac{Y}{\frac{1}{4}} = 4 \cdot Y = Exp(4)$ where $Y \sim Exp(1)$, ■

Theorem 4.3.11 — Delta Method.
Let X_n be a sequence of random variables such that $a_n(X_n-\theta)\xrightarrow{d}N(0,\sigma^2)$ with $\lim_{n\to\infty}a_n=\infty$ and $g(x)$ is differentiable at $x=\theta$ and $g'(\theta)\neq 0$, then

$$a_n(g(x_n)-g(\theta))\xrightarrow{d}N\Big(0,\big[g'(\theta)\big]^2\sigma^2\Big)$$

Intuition:

$$g(X_n)\approx g(\theta)+g'(\theta)(X_n-\theta)\qquad\Longrightarrow\qquad a_n(g(X_n)-g(\theta))\approx a_ng'(\theta)(X_n-\theta)$$

Since for large n we have $a_n(X_n-\theta)\approx N(0,\sigma^2)$, which implies

$$a_n(g(x_n)-g(\theta))\approx N(0,\sigma^2)\cdot g'(\theta)=N(0,\sigma^2\big[g'(\theta)\big]^2)$$

■ **Example 4.16** Let $X_n\sim Poi(\lambda)$ be a sequence of random variables, find the limiting distribution of $Z_n=\sqrt{n}(\sqrt{\overline{X}_n}-\sqrt{\lambda})$

By previous result we have $\sqrt{n}(\overline{X}_n-\lambda)\xrightarrow{d}N(\lambda)$, Let $Z_n=\sqrt{n}(g(\overline{X}_n)-g(\lambda))$ for $g(t)=\sqrt{t}$. Since $\lambda>0$, so $g'(\lambda)$ exists and $g'(\lambda)=\frac{1}{2\sqrt{\lambda}}\neq 0$. Now by **Delta Method**:

$$Z_n\xrightarrow{d}N(0,\lambda\big[g'(\lambda)\big]^2)=N\left(0,\frac{1}{4}\right)$$

■

■ **Example 4.17** Let $X_n\sim Exp(\theta)$ be a sequence of random variables and $Z_n=\sqrt{n}(\log\big(\overline{X}_n\big)-\log(\theta))$

By CLT:

$$\frac{\sqrt{n}(\overline{X}_n-\theta)}{\theta}\xrightarrow{d}N(0,1)\qquad\Longrightarrow\qquad\sqrt{n}(\overline{X}_n-\theta)\xrightarrow{d}N(0,\theta^2)$$

Let $g(t)=\log(t)$, so by **Delta method**:

$$Z_n\xrightarrow{d}N\Big(0,\theta^2\big[g'(\theta)\big]^2\Big)=N\left(0,\frac{1}{\theta^2}\cdot\theta^2\right)=N(0,1)$$

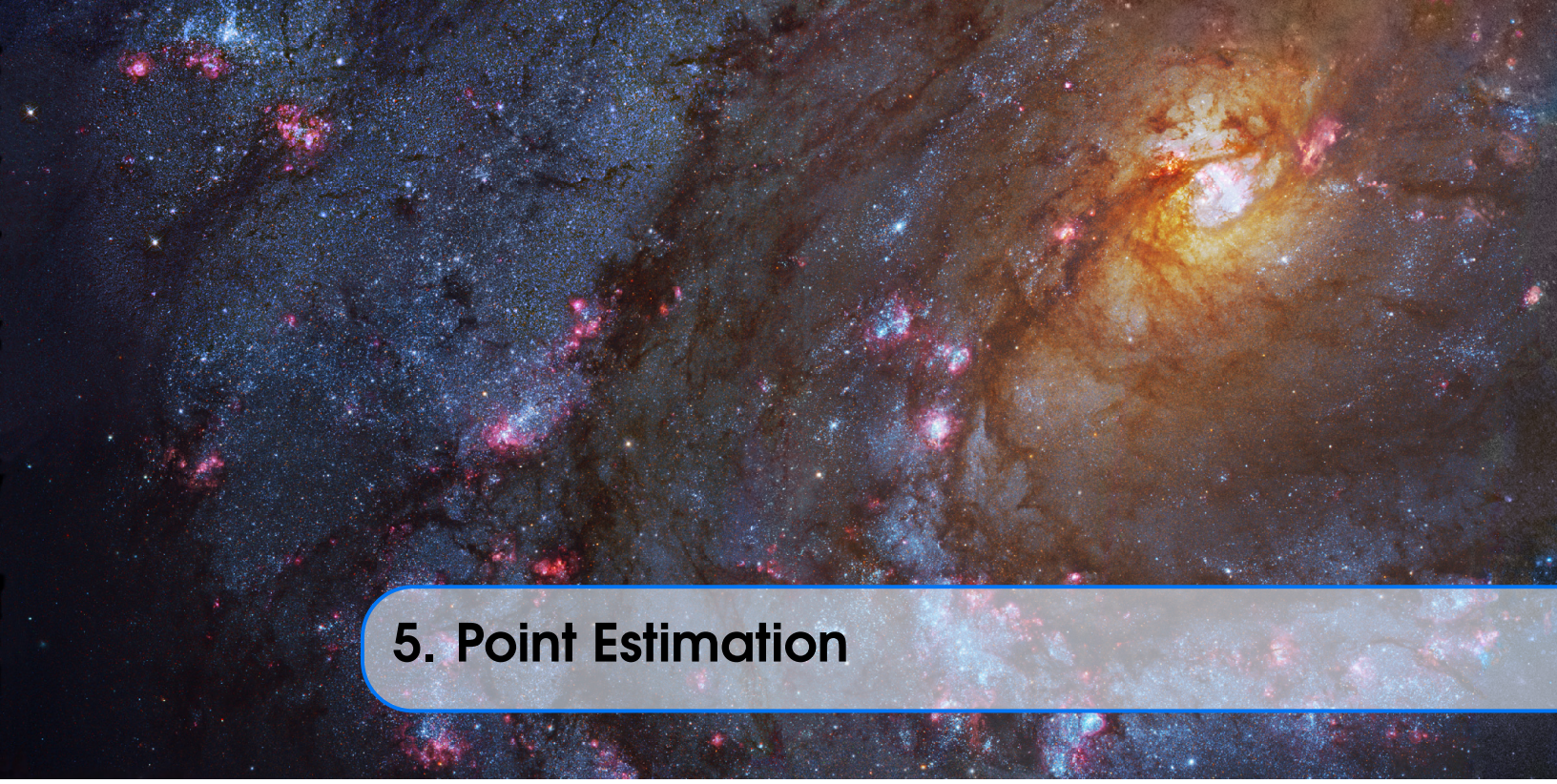
■

Let X_n be a sequence of random variables with $E[X_i]=0$ and $\text{Var}[X_i]=\sigma^2<\infty$ for all i. Find the approximate distribution of $\overline{X}_n^2$
By CLT:

$$\frac{\sqrt{n}(\overline{X}_n-0)}{\sigma}\xrightarrow{d}N(0,1)\qquad\Longrightarrow\qquad\sqrt{nX}\xrightarrow{d}N(0,\sigma^2)$$

Let $g(t)=t^2$ and $Z=N(0,1)$, by continuous mapping:

$$\frac{n\overline{X}_n^2}{\sigma^2}\xrightarrow{d}Z^2=\chi^2_{(1)}\qquad\Longrightarrow\qquad\overline{X}_n^2\xrightarrow{d}\frac{\sigma^2}{n}\chi^2_{(1)}$$



5. Point Estimation

5.1 Introduction

We define the **statistics** $T(\vec{X})$ of $\vec{X}$ only, does not contain $\vec{\theta}$.

■ **Example 5.1**

$$T(\vec{X}) = \frac{X_1 + ... + X_n}{n} = \overline{X}_n \quad \textbf{is a statistic} \qquad \overline{X}_n - \mu \quad \textbf{is not}$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \textbf{is a statistic} \qquad \frac{\overline{X}_n - \mu}{\sqrt{S_n^2}} \quad \textbf{is not}$$

■

Want to estimate $\vec{\theta}$ or $g(\vec{\theta})$ for some function g, we call a statistic $T(\vec{X})$ an **estimator** of $\vec{\theta}$ if we use it to estimate $\vec{\theta}$. We call $T(\vec{x})$ an estimate of $\vec{\theta}$.

■ **Example 5.2** $\overline{X}_n$ is an estimator for μ and $\overline{x}_n$ is an estimate ■

We will often use $\hat{\theta}(x_1, x_n)$ to indicate estimator of θ, but we often omit $(...)$. That is we use $\hat{\theta}$ to denote estimator of θ.

5.2 Method of Moments

Definition 5.2.1
Let $X_1, ..., X_n$ be i.i.d with pdf/pmf $f(x;\vec{\theta})$ where $\vec{\theta}$ is a p-dimension parameter. such that

$$\mu_1 = E[X_i] = g_1(\vec{\theta}) \qquad \mu_2 = E[X_i^2] = g_2(\vec{\theta}) \qquad ..... \qquad \mu_p = E[X_i^p] = g_p(\vec{\theta})$$

Idea: Substitute $\mu_1, \mu_2, ...$ with $\hat{\mu}_1, \hat{\mu}_2, ...$ where

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{X_i}{n} \qquad \hat{\mu}_2 = \sum_{i=1}^{n} \frac{X_i^2}{n} \qquad ..... \qquad \hat{\mu}_p = \sum_{i=1}^{n} \frac{X_i^p}{n}$$

We define the **Method of Moments (MME)** of θ to be the solution to

$$\hat{\mu}_1 = g_1(\vec{\theta}) \qquad \hat{\mu}_2 = g_2(\vec{\theta}) \qquad ..... \qquad \hat{\mu}_p = g_p(\vec{\theta}) \qquad \textbf{p unknowns and p equations}$$

■ **Example 5.3** Let $X_1, ..., X_n \sim Poi(\lambda)$ be i.i.d, we see that $\mu_1 = E[X_i] = \lambda$ so that

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \lambda$$

so $\hat{\lambda} = \hat{\mu}_1$ is the MME of λ.

Is $E[\hat{\lambda}] = \lambda$ ($\hat{\lambda}$ unbiased??)

$$E[\hat{\lambda}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n\lambda = \lambda \qquad \Longrightarrow \qquad \hat{\lambda} \quad \textbf{is unbiased}$$

Is $\hat{\lambda} \xrightarrow{p} \lambda$? ($\hat{\lambda}$ consistent ??)
By **WLLN**: $\hat{\lambda} = \overline{X}_n \xrightarrow{p} \lambda$, so $\hat{\lambda}$ is consistent ■

■ **Example 5.4** Let $X_1, ..., X_n \sim Unif(0, \theta)$ be i.i.d, so $\mu_1 = E[X_i] = \frac{\theta}{2}$, so **MME** $\hat{\theta}$ solves $\hat{\mu}_1 = \frac{\theta}{2}$, so $\hat{\theta} = 2\hat{\mu}_1 = 2\overline{X}_n$
Then

$$E[\hat{\theta}] = E[2\overline{X}_n] = 2E[\overline{X}_n] = 2\cdot\frac{\theta}{2} = \theta$$

and $\hat{\theta} = 2\overline{X}_n \xrightarrow{p} \theta$ because by **WLLN** $\overline{X}_n \xrightarrow{p} \frac{\theta}{2}$, then $\hat{\theta}$ is unbiased and consistent. ■

■ **Example 5.5** Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ be i.i.d, so $\mu_1 = E[X_i] = \mu$, $\mu_2 = E[X_i^2] = \text{Var}[X_i] + E[X_i]^2 = \sigma^2 + \mu^2$, then MMEs $\hat{\mu}, \hat{\sigma}^2$ solve:

$$\begin{cases} \hat{\mu}_1 = \mu \\ \hat{\mu}_2 = \sigma^2 + \mu^2 \end{cases} \qquad \Longleftrightarrow \qquad \begin{cases} \mu = \hat{\mu}_1 \\ \sigma^2 = \hat{\mu}_2 - \hat{\mu}_1^2 \end{cases}$$

Then $\hat{\mu} = \overline{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$, so $\hat{\mu}$ is unbiased and consistent. Also we see that

$$E[\hat{\sigma}^2] = E[X_i^2] - (\text{Var}[\overline{X}_n] + E[\overline{X}_n]^2) = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2$$

$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \sigma^2+\mu^2$ **and** $\overline{X}_n \xrightarrow{p} \mu \implies \hat{\sigma}^2 \xrightarrow{p} \sigma^2$ **by slutsky's theotem and continuous mapping**

so $\hat{\sigma}^2$ is biased but consistent ■

■ **Example 5.6** Let $X_1,...,X_n \sim Unif(-\theta,\theta)$ be i.i.d with $\theta > 0$. Then we have $\mu_1 = E[X_i] = \frac{-\theta+\theta}{2} = 0$, so we should use **higher order moments**. We see that

$$\mu_2 = E[X_i^2] = \text{Var}[X_i] + E[X_i]^2 = \frac{(\theta-(-\theta))^2}{12} + 0^2 = \frac{\theta^2}{3}$$

so **MME** solve

$$\frac{\hat{\theta}^2}{3} = \hat{\mu}_2 \qquad \implies \qquad \hat{\theta} = \sqrt{\frac{3\sum_{i=1}^{n} X_i^2}{n}}$$

■

5.3 Maximum Likelihood

Let $X_1,...,X_n$ be i.i.d with pdf/pmf $f(x;\vec{\theta})$ and $\vec{\theta} \in \Omega$ (parameter space), we observe $(x_1,....,x_n)$ from random variables $(X_1,....,X_n)$

Definition 5.3.1 — Likelihood Function.

$$L(\vec{\theta};\vec{x}) = \prod_{i=1}^{n} f(x_i\,;\vec{\theta})$$

where $f(x_i\,;\vec{\theta})$ is joint pdf/pmf of X_i's and $L:\Omega \to [0,\infty)$. The joint pdf is function of data $\vec{x}$ by parameter $\vec{\theta}$, likelihood is function of parameter $\vec{\theta}$ indexed by data $\vec{x}$.

Note: It's not true that $\int L(\theta\,;\vec{x})d\theta = 1$ or $\sum_{\theta} L(\theta;\,\vec{x}) = 1$ n general

Likelihood idea: pick $\vec{\theta}$ such that it maximizes $L(\vec{\theta})$, we call $\hat{\theta} = \arg\max_{\theta\in\Omega} L(\theta\,;\vec{x})$ the maximum likelihood estimate (MLE) of θ. Max likelihood estimator replaces $\vec{x}$ with $\vec{X}$.

Review: How to maximize function $L(\theta)$ over Ω?

Maximizer solve: $\frac{d}{d\theta}L(\theta)$ or look at the boundary of Ω.

Note:

$$\arg\max_{\theta\in\Omega} L(\theta\,;\vec{x}) = \arg\max_{\theta\in\Omega} \log(L(\theta\,;\vec{x}))$$

The log likelihood function:

$$\ell(\theta;\vec{x}) = \log(L(\theta;\vec{x})) = \log\left(\prod_{i=1}^{n} f(x_i;\vec{\theta})\right) = \sum_{i=1}^{n} \log\Big(f(x_i;\vec{\theta})\Big)$$

this is other easier to work with because ℓ is easier to diiferentiable.

■ **Example 5.7** Let $X_1,...,X_n \sim Exp(\theta)$ be i.i.d, find the **MLE** of θ. Since $f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}$, then

$$L(\theta;\vec{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \theta^{-n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}} \qquad \Longrightarrow \qquad \ell(\theta;\vec{x}) = -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

Then we see that

$$\frac{d\ell}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i$$

MLE $\hat{\theta}$ solve $\frac{d\ell}{d\theta} = 0$, we get

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}_n$$

■

■ **Example 5.8** Let $X_1,...,X_n \sim Poi(\lambda)$ be i.i.d, find the MLE of λ. Since $f(x;\theta) = \frac{\lambda^x e^{-\lambda}}{x!}$, then

$$L(\lambda;\vec{x}) = \prod_{i=1}^{n} f(x_i;\lambda) = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i!} =\propto \lambda^{\sum x_i} e^{-n\lambda}$$

so that

$$\ell(\lambda;\vec{x}) = \left(\sum_{i=1}^{n} x_i\right)\log\lambda - n\lambda - \sum_{i=1}^{n}\log(x_i!) \propto \left(\sum_{i=1}^{n} x_i\right)\log\lambda - n\lambda - c \qquad \textbf{for some } c \in \mathbb{R}$$

Then we have

$$\frac{d\ell}{d\theta} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$$

so **MLE** $\hat{\lambda}$ solves $\frac{d\ell}{d\theta} = 0$ we get

$$\frac{1}{\hat{\lambda}}\sum_{i=1}^{n} x_i - n = 0 \qquad \Longrightarrow \qquad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} = \overline{x}_n$$

which is the same as **MME** ■

■ **Example 5.9** Let $X_1,...,X_n \sim Unif(0,\theta)$ be i.i.d and it has the pdf

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \textbf{if } 0 < x \le \theta \\ 0 & \textbf{otherwise} \end{cases}$$

Then we have

$$L(\theta;\vec{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \begin{cases} \theta^{-n} & \textbf{if } \theta \ge \max\limits_{1\le i\le n} x_i \\ 0 & \textbf{otherwise} \end{cases}$$

Then we have $\hat{\theta} = \max\limits_{1\le i\le n} X_i$ is the **MLE**, which is different from the **MME** $\hat{\theta} = 2\overline{X}_n$ ■

■ **Example 5.10** Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be i.i.d, then

$$\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\log\left(2\pi\sigma^2\right)$$

This gives us that

$$\frac{d\ell}{d\mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) \qquad \textbf{and} \qquad \frac{d\ell}{\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2\sigma^2}$$

solve it to get the **MLEs**

$$\begin{cases} \frac{d\ell}{d\mu} = 0 \\ \frac{d\ell}{\sigma^2} = 0 \end{cases} \quad \Longrightarrow \quad \begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \end{cases}$$

■

5.4 Properties of MLEs

Theorem 5.4.1 — Invariances of MLEs.
If $\hat{\theta}$ is the MLE of θ, then for any function g, $g(\hat{\theta})$ is the MLE of $g(\theta)$

■ **Example 5.11** Let $X_1, \ldots, X_n \sim Poi(\lambda)$, by previous we have $\hat{\lambda}_{MLE} = \overline{X}_n$ and MLE of $E[X_i^2] = \lambda(\lambda + 1)$
What's the **MLE** of $P(X_i = 0) = e^{-\lambda}$? It's $e^{-\overline{X}_n}$ by invariance.
What's the **MLE** of

$$\mathbb{I}_{\lambda \leq 10} = \begin{cases} 1 & \lambda \leq 10 \\ 0 & \lambda > 10 \end{cases}$$

By **Invariances**, the **MLE** is

$$\mathbb{I}_{\overline{X}_n \leq 10} = \begin{cases} 1 & \overline{X}_n \leq 10 \\ 0 & \overline{X}_n > 10 \end{cases}$$

■

From now on: θ is a scalar

Definition 5.4.1
Score Function:

$$S(\theta; \vec{x}) = \frac{d}{d\theta}\ell(\theta; \vec{x})$$

Information Function:

$$I(\theta; \vec{x}) = -\frac{d}{d\theta^2}\ell(\theta; \vec{x})$$

Expected Information Function:

$$J(\theta) = E[I(\theta;\vec{x})] = E\left[-\frac{d}{d\theta^2}\ell(\theta;\vec{x})\right] = E\left[-\frac{d}{d\theta^2}\sum_{i=1}^{n}\log f(x_i;\theta)\right]$$

Theorem 5.4.2 — Asymptotic Normality and Consistency of MLE.
Under some regularity conditions (e.g support not depending on θ),

$$(\hat{\theta}-\theta)\cdot[J(\theta)]^{\frac{1}{2}} \xrightarrow{d} N(0,1) \qquad \textbf{and} \qquad \hat{\theta} \xrightarrow{p} \theta$$

Proof: Taylor's theorem to score function

■ **Example 5.12** Let $X_1,....,X_n \sim Poi(\lambda)$ be i.i.d, then we have

$$\ell(\lambda;\vec{x}) = \left(\sum_{i=1}^{n} x_i\right)\log(\lambda) - n\lambda 0\sum_{i=1}^{n}\log(x_i!)$$

and

$$S(\lambda;\vec{x}) = \frac{d}{d\lambda}\ell(\lambda;\vec{x}) = \left(\frac{1}{\lambda}\sum_{i=1}^{n} x_i\right) - n \qquad \textbf{and} \qquad I(\lambda;\vec{x}) = -\frac{d}{d\theta^2}\ell(\lambda;\vec{x}) = \frac{1}{\lambda^2}\sum_{i=1}^{n} x_i$$

Then we have

$$J(\lambda) = E[I(\lambda;\vec{x})] = E\left[\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i\right] = \frac{n}{\lambda}$$

this gives us that

$$(\hat{\lambda}-\lambda)[J(\lambda)]^{\frac{1}{2}} = (\overline{X}_n-\lambda)\cdot\sqrt{\frac{n}{\lambda}} \qquad \textbf{and} \qquad (\hat{\lambda}-\lambda)[J(\lambda)]^{\frac{1}{2}} \xrightarrow{d} N(0,1) \qquad \textbf{and} \qquad \hat{\lambda} = \overline{X}_n \xrightarrow{p} \lambda$$

$P(X_1=0) = e^{-\lambda}$, by invariance it has MLE: $e^{-\overline{X}_n}$, by continuous mapping $\sqrt{n}(\overline{X}_n-\lambda) \xrightarrow{d} N(0,\lambda)$.
Use delta method we het

$$\sqrt{n}(e^{-\overline{X}_n}-e^{-\lambda}) \xrightarrow{d} N(0,\lambda e^{-2\lambda})$$

so the large n estimate of $P(X_1=0)$:

$$e^{-\overline{X}_n} \xrightarrow{d} N\left(e^{-\lambda}, \frac{\lambda e^{-2\lambda}}{n}\right)$$

■