Recall

Let $X, \widetilde{X}, Y$ and Z be random variables.

Conditional expectation

Given $y \in \mathbb{R}$, $E[X|Y = y]$ is the expectation of X with respect to the conditional probability $P(X \in \cdot|Y = y)$. As y varies, we obtain a function $f\colon y \mapsto E[X|Y = y]$. Then the *conditional expectation* $E[X|Y]$ is the random variable $f(Y)$.

Some properties of $E[\,\cdot\,|Y]$ which maps a random variable to another **random variable**:

(1) ($\mathbb{R}$-linear) $\forall\, \alpha, \beta \in \mathbb{R},\ E[\alpha X + \widetilde{X} + \beta|Y] = \alpha E[X|Y] + E[\widetilde{X}|Y] + \beta$.

(2) (monotone) If $X \le Z$, then $E[X|Y] \le E[Z|Y]$.

(3) ($g(Y)$-scaling) In most cases, for a function $g\colon \mathbb{R} \to \mathbb{R}$ we have $E[g(Y)X|Y] = g(Y)E[X|Y]$ since $E[g(y)X|Y = y] = g(y)E[X|Y = y]$, $\forall\, y \in \mathbb{R}$.

(4) In particular, $E[E[X|Y]|Y] = E[X|Y]$ by (3).

(5) (towering property) $E[X] = E[E[X|Y]]$, i.e., compute expectations by conditioning.

(6) We take X in (5) to be the indicator variable χ_E for an event E. Note that $E[\chi_E] = P(E)$ and $E[\chi_E|Y = y] = P(E|Y = y)$. Then we can compute probabilities by conditioning,

$$P(E) = \begin{cases} \sum_y P(E|Y = y)P(Y = y) & \text{if } Y \text{ discrete} \\ \int_{-\infty}^{\infty} P(E|Y = y)f_Y(y)dy & \text{if } Y \text{ continuous.} \end{cases}$$

In particular, if $Y = \sum_{i=1}^{n} i\chi_{F_i}$ for some partition $F_1, \ldots, F_n$ of the sample space, then the law of total probability is recovered.

Moment generating functions

For a random variable X, the *moment generating function* (MGF) is $M_X(t) := E[e^{tX}]$ for $t \in \mathbb{R}$ whenever $E[e^{tX}]$ exists. Note $M_X(t) > 0$. The following facts make MGF useful:

- (generate moments) $E[X^n] = M_X^{(n)}(0)$ for $n \in \mathbb{N}$ (if $E[X^n] < \infty$).
- (determine distributions) If there exists $t_0 > 0$ such that $M_X(t) = M_Y(t)$ for $t \in (-t_0, t_0)$, then $F_X = F_Y$.
- (multiplicative under independent sums) If X, Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

A table about MGFs of common distributions can be found in the textbook [Ross, Ch. 7-Sec. 7].

Examples

Example 1. Let X, Y be random variables and $g\colon \mathbb{R} \to \mathbb{R}$ be a function. Show that

(i) $\operatorname{Cov}(X, E[Y|X]) = \operatorname{Cov}(X, Y)$.

(ii) $E[(X - E[X|Y])^2] = E[X^2] - E[E[X|Y]^2]$.

(iii) $E[(X - g(Y))^2] \geq E[(X - E[X|Y])^2]$.

Proof. (i) It follows from (3) that $XE[Y|X] = E[XY|X]$. Then by (5),

$$\begin{aligned}\operatorname{Cov}(X, E[Y|X]) &= E[XE[Y|X]] - E[X]E[E[Y|X]] \\ &= E[E[XY|X]] - E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \\ &= \operatorname{Cov}(X, Y).\end{aligned}$$

(ii) By (5) and (3), we have

$$E[XE[X|Y]] = E[E[XE[X|Y]|Y]] = E[E[X|Y]E[X|Y]] = E[E[X|Y]^2].$$

Hence

$$\begin{aligned}E[(X - E[X|Y])^2] &= E[X^2] - 2E[XE[X|Y]] + E[E[X|Y]^2] \\ &= E[X^2] - 2E[E[X|Y]^2] + E[E[X|Y]^2] \\ &= E[X^2] - E[E[X|Y]^2].\end{aligned}$$

(iii) By (5), it suffices to prove $E[(X - g(Y))^2|Y] \geq E[(X - E[X|Y])^2|Y]$.

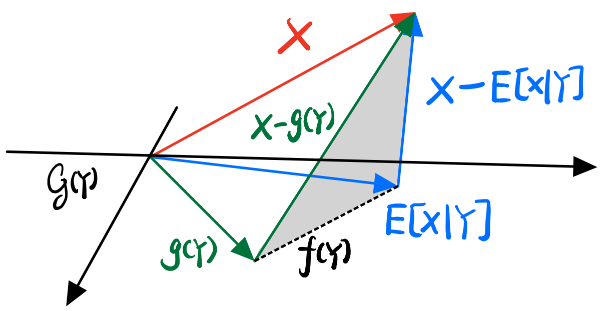


Figure 1: A possible intuition about (iii)

Based on the above intuition, we first establish that $X - E[X|Y]$ is 'orthogonal' to the plane. For any function $f\colon \mathbb{R} \to \mathbb{R}$, by (3) and (4) we have

$$\begin{aligned}E[(X - E[X|Y])f(Y)|Y] &= f(Y)E[X - E[X|Y]|Y] \\ &= f(Y)(E[X|Y] - E[E[X|Y]|Y]) \\ &= f(Y)(E[X|Y] - E[X|Y]) \\ &= 0.\end{aligned}$$

Next we focus on the shaded 'right triangle'. By viewing $E[X|Y] - g(Y)$ as $f(Y)$,

$$\begin{aligned}
&E[(X - g(Y))^2|Y] \\
&= E[(X - E[X|Y] + E[X|Y] - g(Y))^2|Y] \\
&= E[(X - E[X|Y])^2|Y] + 2E[(X - E[X|Y])(E[X|Y] - g(Y))|Y] + E[(E[X|Y] - g(Y))^2|Y] \\
&= E[(X - E[X|Y])^2|Y] + 0 + E[(E[X|Y] - g(Y))^2|Y] \\
&\geq E[(X - E[X|Y])^2|Y],
\end{aligned}$$

where the last inequality follows from (2).

□

Remark. Similar to (iii) in Example 1, it is also intuitive to use the *correlation* to indicate the linear relationship between X and Y. If $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$, we normalize X, Y (as we often do in Central Limit Theorem) to

$$\widetilde{X} := \frac{X - E[X]}{\sqrt{\mathrm{Var}(X)}} \quad \text{and} \quad \widetilde{Y} := \frac{Y - E[Y]}{\sqrt{\mathrm{Var}(Y)}}.$$

Then the *correlation coefficient* is defined as

$$\rho(X, Y) := \mathrm{Cov}(\widetilde{X}, \widetilde{Y}) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

where in the last equality recall that $\mathrm{Cov}(\alpha, Z) = 0$ for all $\alpha \in \mathbb{R}$ and r.v. Z.

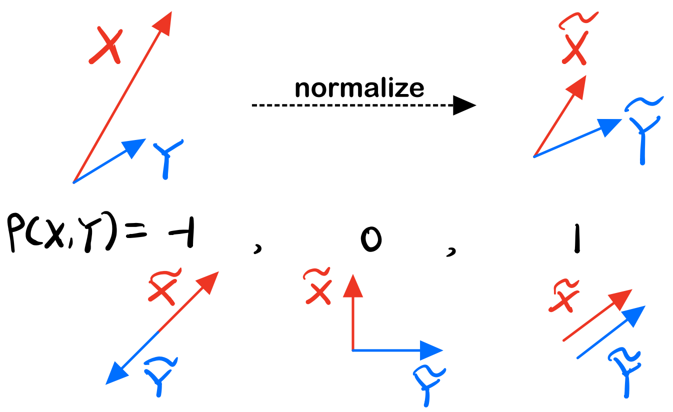


Figure 2: Special cases for $\rho(X, Y)$

If interested, we may refer to the arguments preceding [Ross, Ch. 7-Example 4d] for a basic analysis of $\rho(X, Y)$.

Example 2. Let $X \sim N(0, 1)$ and $I \sim \mathit{Bern}(1/2)$. Suppose that X, Y are independent. Define

$$Y = \begin{cases} X & \text{if } I = 0 \\ -X & \text{if } I = 1. \end{cases}$$

Find $\mathrm{Cov}(X, Y)$.

Solution. Since X, I are independent, we have X^2, I are independent. Thus $E[X^2|I = i] = E[X^2]$ for $i = 0, 1$. Note $E[X] = 0$. Then

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\
&= E[E[XY|I]] - 0 && \text{by (5)} \\
&= P(I = 0)E[XY|I = 0] + P(I = 1)E[XY|I = 1] \\
&= \frac{1}{2}E[X^2|I = 0] + \frac{1}{2}(-E[X^2|I = 1]) && \text{by def. of } Y \\
&= \frac{1}{2}(E[X^2] - E[X^2]) && \text{by independence} \\
&= 0.
\end{aligned}$$

□

Remark. By checking X, Y are NOT independent, we find another example showing $\mathrm{Cov}(X, Y) = 0 \not\Longrightarrow$ independence.

Example 3. Let $(U_i)_{i=1}^{\infty}$ be an i.i.d. sequence of random variables with common distribution $U(0, 1)$. For $x \in [0, 1]$, define $N(x) := \min\{n : \sum_{i=1}^{n} U_i > x\}$. Show that $E[N(x)] = e^x$.

Proof. Notice that $N(x)$ is a non-negative random variable. We will compute $E[N(x)]$ by layer-cake (see e.g., [Tutorial 4, Example 2]). First we prove by induction that for $n \in \mathbb{Z}_{\geq 0}$,

$$P(N(x) \geq n + 1) = \frac{x^n}{n!}. \tag{$*$}$$

When $n = 0$, we have $P(N(x) \geq 1) = P(U_1 \leq x) = x$. When $n \geq 1$, suppose ($*$) holds for $n - 1$, i.e., $P(N(x) \geq n) = \frac{x^{n-1}}{(n-1)!}$. We check ($*$) for n by conditioning on U_1,

$$\begin{aligned}
P(N(x) \geq n + 1) &= \int_{-\infty}^{\infty} P(N(x) \geq n + 1|U_1 = y)f_{U_1}(y)dy && \text{by (6)} \\
&= \int_0^1 P(y + \sum_{i=2}^{n} U_i \leq x|U_1 = y)dy && \text{by def. of } N(x),\ f_{U_1} \\
&= \int_0^1 P(\sum_{i=2}^{n} U_i \leq x - y|U_1 = y)dy \\
&= \int_0^1 P(\sum_{i=2}^{n} U_i \leq x - y)dy && \text{by } \sum_{i=2}^{n} U_i,\ U_1 \text{ independent} \\
&= \int_0^1 P(\sum_{i=1}^{n-1} U_i \leq x - y)dy && \text{by i.i.d.} \\
&= \int_0^x P(\sum_{i=1}^{n-1} U_i \leq x - y)dy && \text{by vanished integrand when } y > x \\
&= \int_0^x P(N(x - y) \geq n)dy && \text{by def. of } N(x) \\
&= \int_0^x \frac{(x - y)^{n-1}}{(n - 1)!}dy && \text{by induction hypothesis}
\end{aligned}$$

$$= \int_0^x \frac{t^{n-1}}{(n-1)!} dt \qquad \text{by change of variable } t = x - y$$
$$= \frac{x^n}{n!}.$$

Hence

$$E[N(x)] = \sum_{n=1}^{\infty} P(N(x) \geq n) = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!} = e^x.$$

□

It's good to stop here.

Limit theorems

We include this section only for a relatively complete content but without examples.

Inequalities

Proposition 4 (Markov inequality). *Let X be a non-negative random variable. Then for $\varepsilon > 0$,*

$$P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}.$$

Proposition 5 (Chebyshev inequality). *Let X be a random variable with finite mean μ and variance σ^2. Then for $\varepsilon > 0$,*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Limit theorems

Theorem 6 (Weak law of large numbers WLLN). *Let $(X_i)_{i=1}^{\infty}$ be an i.i.d. sequence of random variables with finite mean μ. Then for $\varepsilon > 0$,*

$$P\left(|\frac{X_1 + \cdots + X_n}{n} - \mu| \geq \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Theorem 7 (Strong law of large numbers SLLN). *Let $(X_i)_{i=1}^{\infty}$ be an i.i.d. sequence of random variables with finite mean μ. Then*

$$P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Theorem 8 (Central limit theorem CLT). *Let $(X_i)_{i=1}^{\infty}$ be an i.i.d. sequence of random variables with finite mean μ and variance σ^2. Then for $t \in \mathbb{R}$,*

$$P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leq t\right) \to \Phi(t) \quad \text{as } n \to \infty.$$

Some simulation experiments about the limit theorems can be played interactively by clicking here (in major browsers). It might take a 7 – 20 mins to initialize. The static PDF version is attached below for convenience.

MATH3280 Tutorial 13

Table of Contents

MATH3280 Tutorial 13

Strong Law of Large number

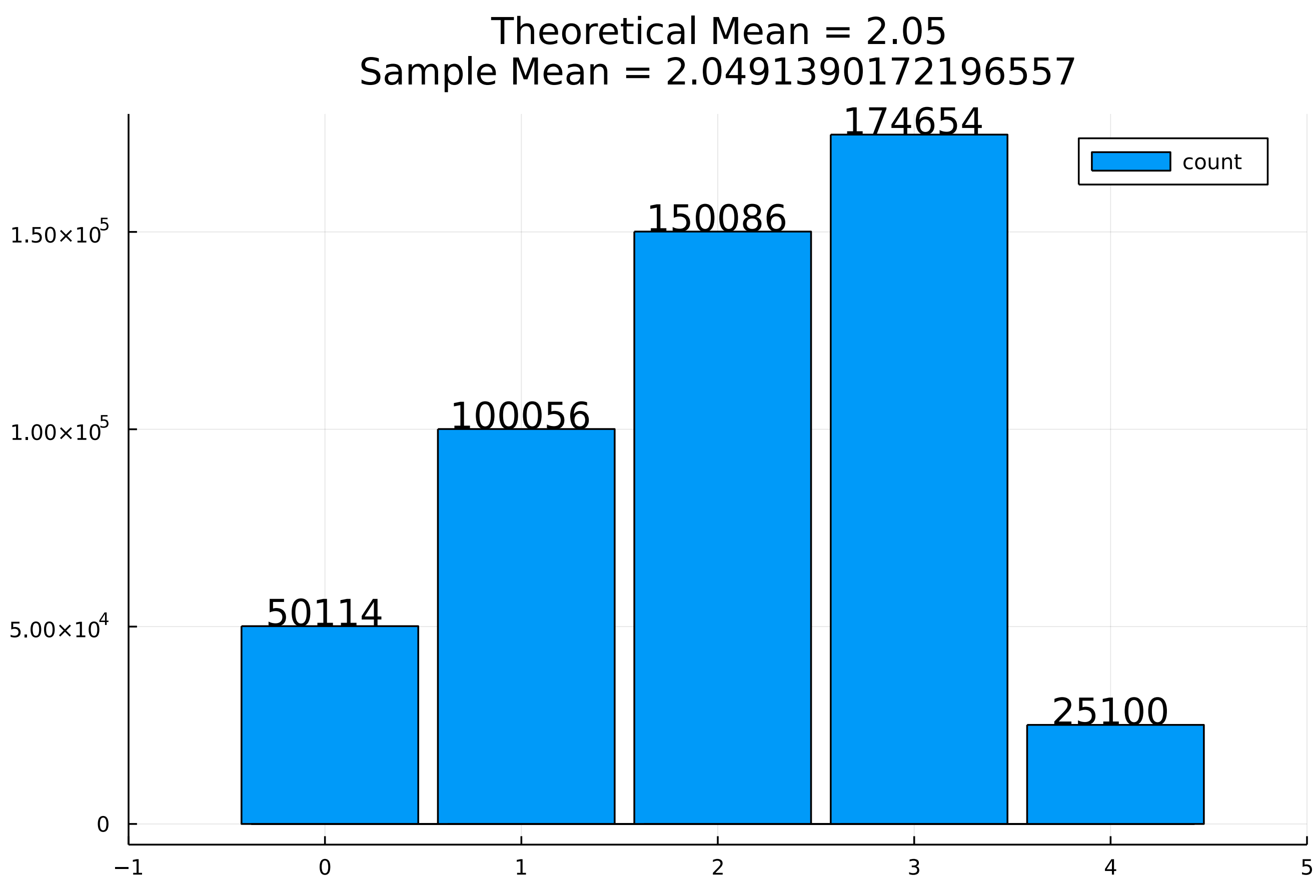
Central Limit Theorem

Normal Distributions

Strong Law of Large number

```
p = [0.1, 0.2, 0.3, 0.35, 0.05]; # make sure p is a probability vector
```

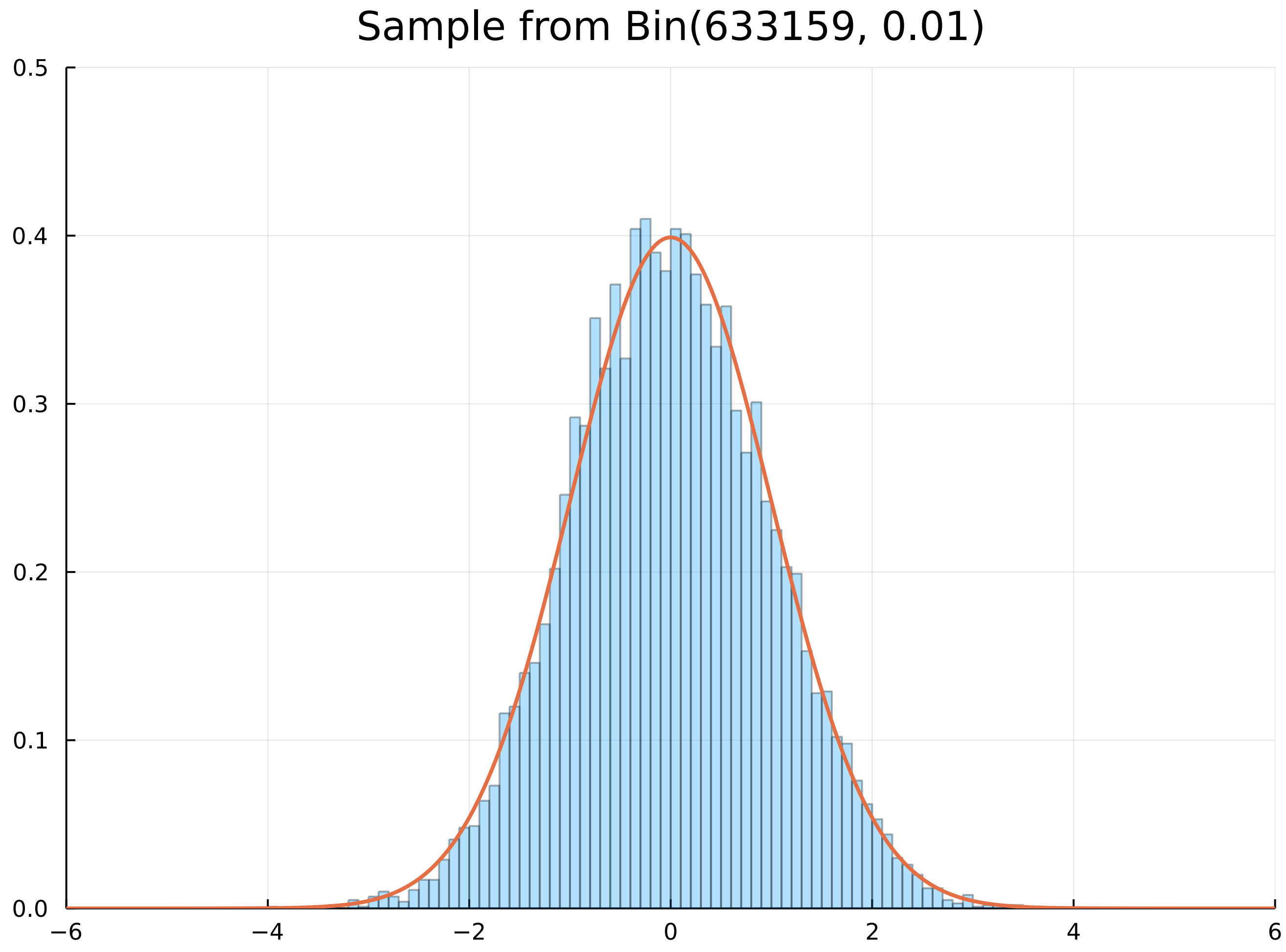
$n =$ 1000 Run! 🚀



Central Limit Theorem

$Bin(n, p)$ with $n =$ 633159 $p =$ 0.01

Sample size = 10000 Run! 🚀



Normal Distributions

$\mu =$ 0.14 $\sigma =$ 0.71