TECHNISCHE
UNIVERSITÄT
WIEN

Introduction to Statistics
Conditional Distributions and Independence
Inequalities
Convergence Concepts
CLT and LLN

LV Nr. 105.692
Summer Semester 2021

Multiple random variables

We defined a (univariate real) random variable X as a function from a sample space Ω to $\mathbb{R}$.
An n-dimensional random vector $(X_1, X_2, \dots, X_n)$ consists of n random variables

1. For $n = 2$, we call (X, Y) a two-dimensional (bivariate) random vector
2. Discrete random vector... has a countable number of possible values
   1. The function $p(x, y) = P(X = x, Y = y)$ is called joint probability mass function (joint pmf) of (X, Y).
      $*$ It completely defines the probability distribution of (X, Y)
      $*$ For any event $A \subset \mathbb{R}^2$ define $P((X, Y) \in A) = \sum_{(x,y) \in A} p(x, y)$
   - The marginal pmfs of X and Y are given by

$$p_X(x) = \sum_{y \in \mathbb{R}} p(x, y) \qquad \text{and} \qquad p_Y(y) = \sum_{x \in \mathbb{R}} p(x, y)$$

Multiple random variables

Continuous random vector... has uncountable possible values

1. $f(x,y)$ is a joint probability density function (joint pdf) od a continuous bivariate vector (X,Y) if for every $A \subset \mathbb{R}^2$,

$$P((X,Y) \in A) = \int\int_A f(x,y)\,dx\,dy$$

2. The marginal pdfs of X and Y are given by

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy \qquad \text{and} \qquad f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx$$

Joint pmfs

- Let (X, Y) be a two-dimensional discrete random variable.
  1. Let X take values $\{x_1, \dots, x_n\}$ and Y take values $\{y_1, \dots, y_m\}$.
     The pair (X, Y) takes values $\{(x_1, y_1), (x_1, y_2), \dots (x_n, y_m)\}$.
     1. The joint pmf is $p(x_i, y_j) = P(X = x_i, Y = y_j)$

X \ Y	y_1	$\dots$	y_j	$\dots$	y_m	
x_1	$p(x_1, y_1)$	$\dots$	$p(x_1, y_j)$	$\dots$	$p(x_1, y_m)$	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
x_i	$p(x_i, y_1)$	$\dots$	$p(x_i, y_j)$	$\dots$	$p(x_i, y_m)$	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
x_n	$p(x_n, y_1)$	$\dots$	$p(x_n, y_j)$	$\dots$	$p(x_n, y_m)$	

with $0 \leqslant p(x_i, y_i) \leqslant 1$ and $\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$

  2. The cumulative distribution function:

$$F(x, y) = P(X \leqslant x, Y \leqslant y) = \sum_{x_i \leqslant x} \sum_{y_j \leqslant y} p(x_i, y_j).$$

Joint pdfs

Let (X, Y) be a continuous two-dimensional random variable.

1. Let X take values in $[a, b]$ and Y take values in $[c, d]$

   The pair (X, Y) takes values in $[a, b] \times [c, d]$.

   1. $f(x, y)$ is the joint pdf $f(x, y)$ with

      $f(x, y) \geqslant 0$ for $(x, y) \in [a, b] \times [c, d]$ and $\int_a^b \int_c^d f(x, y)\, dxdy = 1$.

   2. The cumulative distribution function:

      $$F(x, y) = P(X \leqslant x, Y \leqslant y) = \int_a^x \int_c^y f(x, y)\, dx\, dy.$$

Expected value

- If it exists, the expected value $\mathbb{E}g(X,Y)$ is given by
  - Discrete case

$$\mathbb{E}\,g(X,Y) = \sum_{(x,y)\in\mathbb{R}} g(x,y)\cdot p(x,y).$$

  - Continuous case

$$\mathbb{E}\,g(X,Y) = \int\int_{\mathbb{R}^2} g(x,y)\cdot f(x,y)\,dxdy$$

Example: Rolling two dice

(1) Roll two fair dice.

$X =$ number on the first die and $Y =$ number on the second die

- X takes values $\{1, 2, 3, 4, 5, 6\}$ and Y takes values $\{1, 2, 3, 4, 5, 6\}$.

(a) Joint probability table:

X \ Y	1	2	3	4	5	6	
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
							1

$p(i, j) = \frac{1}{36}$ for all i, j

(1) Roll two fair dice.
$X =$ number on first die and $Y =$ number on second die

- X takes values $\{1, 2, 3, 4, 5, 6\}$ and Y takes values $\{1, 2, 3, 4, 5, 6\}$.

Joint probability table:

X \ Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
							1

(b) Marginal distribution for X: found by summing the rows

X	1	2	3	4	5	6
$p(x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(1) Roll two fair dice. Let X be the number on first die and Y the sum of numbers on both dice.

- X takes values $\{1, 2, 3, 4, 5, 6\}$ and Y takes values $\{1, 2, 3, 4, 5, 6\}$.

  Joint probability table:

X \ Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(y_j)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

(b) Marginal distribution for Y: found by summing the columns

Y	1	2	3	4	5	6
$p(y_j)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(1) Roll two fair dice.
$X =$ number on first die and $Y =$ number on second die

(c) From the joint probability table compute $F(3.5, 4)$.

Joint probability table:

X \ Y	1	2	3	4	5	6	
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	

- We compute

$$F(3.5, 4) = \mathbb{P}(X \leqslant 3.5, Y \leqslant 4) = \sum_{x_i \leqslant 3.5} \sum_{y_j \leqslant 4} p(x_i, y_j) = \frac{12}{36} = \frac{1}{3}.$$

(2) Roll two fair dice. Let X be the number on the first die and Y the sum of numbers on both dice.

(a) From joint probability table, find marginal distributions for X and for Y

(b) Calculate $\mathbb{E}(X)$, $\mathbb{E}(Y)$ and $\mathbb{E}(XY)$.

(2) Roll two fair dice. Let X be the number on first die and Y the sum of numbers on both dice.

Answer:

(a) X takes values $\{1, 2, 3, 4, 5, 6\}$ and Y takes values $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

Joint probability table:

X \ Y	2	3	4	5	6	7	8	9	10	11	12	
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
												1

Joint probability table:

X \ Y	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(y_j)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Marginal distributions:

X	1	2	3	4	5	6	
$p(x_j)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Y	2	3	4	5	6	7	8	9	10	11	12	
$p(y_j)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

(b) Joint probability table:

X \ Y	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(y_j)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

- $\mathbb{E}(XY) = \frac{1}{36}(27 + 33 \cdot 2 + 39 \cdot 3 + 45 \cdot 4 + 51 \cdot 5 + 57 \cdot 6) = \frac{987}{36} \approx 27.42$

X	1	2	3	4	5	6	
$p(x_j)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Y	2	3	4	5	6	7	8	9	10	11	12	
$p(y_j)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

- $E(X)$= $\frac{21}{6} = 3.5$, $E(Y)$= $\frac{21}{6} \approx 7.028$

(3) Suppose (X, Y) takes values in $[0,1] \times [0,1]$, with uniform density $f(x, y) = 1$. Visualize the event $'X > Y'$ and find its probability.

(4) Let X and Y be random variables and let (X, Y) takes values in $[0,1] \times [0,1]$ with $f(x, y) = \frac{3}{2}(x^2 + y^2)$.

   (a) Show that $f(x, y)$ is a valid probability density function.
   (b) Visualize the event $A =' X > 0.3 \text{ and } Y > 0.5'$. Find its probability.
   (c) Find the marginal density $f_X(x)$. Use this to find $\mathbb{P}(X < 0.5)$.
   (d) Find the cumulative distribution function $F(x, y)$.
   (e) Use the cumulative distribution function $F(x, y)$ to find the marginal cumulative distribution function $F_Y(y)$ and $\mathbb{P}(Y > 0.3)$.
   (f) Calculate $\mathbb{E}(XY)$.

Answer:

(a) First, note $f(x,y) = \frac{3}{2}(x^2 + y^2) \geqslant 0$, for all $(x,y) \in [0,1] \times [0,1]$. Then,

$$\int_0^1 \int_0^1 f(x,y) dx\, dy = \int_0^1 \int_0^1 \frac{3}{2}(x^2 + y^2)\, dx\, dy = \int_0^1 (\frac{x^3}{2} + \frac{3xy^2}{2})\Big|_0^1 dy$$
$$= \int_0^1 (\frac{1}{2} + \frac{3y^2}{2})\, dy = 1.$$

(b) The probability of event A is

$$\begin{aligned}\mathbb{P}(A) &= \int_{0.3}^1 \int_{0.5}^1 f(x,y)\, dy\, dx \\ &= \int_{0.3}^1 \int_{0.5}^1 \frac{3}{2}(x^2 + y^2)\, dy\, dx = \int_{0.3}^1 (\frac{3x^2}{4} + \frac{7}{6}) \\ &= 0.5495\end{aligned}$$

(c) The marginal density $f_X(x)$ is found by integrating out the y:

$$f_X(x) = \int_0^1 f(x,y)\,dy = \left(\frac{3x^2y}{2} + \frac{y^3}{2}\right)\Big|_0^1 = \frac{3x^2+1}{2}, \text{ for } x \in [0,1].$$

- We calculate

$$\mathbb{P}(X < 0.5) = \int_0^{0.5} f_X(x)\,dx = \int_0^{0.5} \frac{3x^2+1}{2}\,dx = \left(\frac{x^3}{2} + \frac{x}{2}\right)\Big|_0^{0.5} = \frac{5}{16}.$$

(d) The cumulative distribution function for $(x,y) \in [0,1] \times [0,1]$ equals:

$$F(x,y) = \int_0^y \int_0^x f(u,v)\,du\,dv = \int_0^y \int_0^x \frac{3}{2}(u^2+v^2)\,du\,dv = \frac{x^3y + xy^3}{2}.$$

(e) The marginal cumulative distribution function $F_Y(y)$ can be found from

$$F_Y(y) = F(1, y) = \frac{1}{2}(y^3 + y), \quad \text{for } y \in [0, 1].$$

- Then we calculate

$$\mathbb{P}(Y > 0.3) = 1 - \mathbb{P}(Y \leqslant 0.3) = 1 - F_Y(0.3) = 1 - 0.1635 = 0.8365.$$

- Another way to calculate $F_Y(y)$ is by using the marginal density $f_Y(x)$, i.e. for $y \in [0, 1]$

$$f_Y(y) = \int_0^1 f(x, y)\, dx$$

and the definition

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \int_0^y f_y(v)\, dv.$$

(f) $\mathbb{E}(XY) = \int_0^1 \int_0^1 xyf(x, y)\, dxdy = \int_0^1 \int_0^1 \frac{3}{2}(x^3y + xy^3)dxdy = \frac{1}{2}.$

Independence

- Two random variables X and Y are independent if

$$F(x,y) = F_X(x) \cdot F_Y(y).$$

1. Discrete random variables X and Y are independent if

$$p(x,y) = p_X(x) \cdot p_Y(y), \qquad \text{for all } x, y$$

2. Continuous random variables X and Y are independent if

$$f(x,y) = f_X(x) \cdot f_Y(y), \qquad \text{for all } x, y$$

- If two random variables X and Y are independent, then

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

Examples

(1) Roll two fair dice.
X = number on first die and Y = number on second die.

Are X and Y independent?

Answer: From the probability table

X \ Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(y_j)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

we conclude that X and Y are independent, since

$$p(x,y) = p(x) \cdot p(y), \quad \text{for all} \quad x,\, y$$

(2) Roll two fair dice. Let X be the number on first die and Y the sum of numbers on both dice.

Are X and Y independent?

Answer: From the probability table

X \ Y	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(y_j)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

we see, for example

$$p(x_2, y_9) = 0 \neq \frac{1}{6} \cdot \frac{4}{36} = p(x_2) \cdot p(y_9).$$

Thus, X and Y are not independent.

(4) Let X and Y be random variables and let (X, Y) takes values in $[0,1] \times [0,1]$ with $f(x,y) = \frac{3}{2}(x^2 + y^2)$.

Are X and Y independent?

Answer:

- X and Y are not independent because

$$f(x,y) = \frac{3}{2}(x^2 + y^2)$$
$$\neq \frac{3x^2+1}{2} \cdot \frac{3y^2+1}{2} = f_X(x) \cdot f_Y(y)$$

  - Another way:

$$\mathbb{E}(XY) = \frac{1}{2} \neq \frac{5}{8} \cdot \frac{5}{8} = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Conditional distributions

- The conditional probability distribution of one random variable Y given another X is the probability model for when you have seen X and know its value but have not yet seen Y and don't know its value.
- X is no longer random: Once you know its value x, it's a constant not a random variable.
- We write the density of this probability model as $f(y \mid x)$. We write expectations with respect to this model as $\mathbb{E}(Y \mid x)$, and we write probabilities as

$$\mathbb{P}(Y \in A \mid x) = \mathbb{E}(I_A(Y) \mid x) = \int_A f(y \mid x) dy$$

probability is a special case of expectation!

- the integrals are replaced by sums in the discrete case

Conditional distributions

- Let (X, Y) be a bivariate discrete random variable with joint pmf $p(x, y)$ and marginal pmfs $p_X(x)$ and $p_Y(y)$. For any x such that $\mathbb{P}(X = x) = p_X(x) > 0$ the conditional pmf of Y given that $X = x$ is the function of y defined by

$$p(y|x) = \mathbb{P}(Y = y|X = x) = \frac{p(x,y)}{p_X(x)}.$$

1. $\mathbb{E}(g(Y)|x) = \sum_y g(y)\, p(y|x)$

   conditional expected value of $g(Y)$ given $X = x$

- Let (X, Y) be a bivariate continuous random variable with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional pdf of Y given that $X = x$ is the function of y defined by

$$f(y|x) = \frac{f(x,y)}{f_X(x)}.$$

1. $\mathbb{E}(g(Y)|x) = \int_{\mathbb{R}} g(y)\, f(y|x)\, dy$

   conditional expected value of $g(Y)$ given $X = x$

Conditional distributions and independence

- If X and Y are independent, the conditional pdf of Y given $X = x$ is

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x) \cdot f_Y(y)}{f_X(x)} = f_Y(y),$$

regardless the value of x.

Lemma (Verification of independence)

Let (X, Y) be a bivariate random vector with joint pdf $f(x,y)$, Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$

$$f(x,y) = g(x) \cdot h(y).$$

Independence

- Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ be function only of y. Then

$$\mathbb{E}(g(X)\,h(Y)) = \mathbb{E}(g(X)) \cdot \mathbb{E}(h(Y)).$$

- Let X and Y be independent random variables with moment generating functions $M_X(x)$ and $M_Y(y)$. Then the moment generating function of their sum $Z = X + Y$ is given by

$$M_Z(t) = M_X(t) \cdot M_Y(t).$$

In general: Let $X_1, \ldots X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $Z = X_1 + \cdots + X_n$. Then the mgf of Z is

$$M_Z(t) = M_{X_1}(t) \cdot \ldots \cdot M_{X_n}(t)$$

In particular, if $X_1, \ldots, X_n$ are additionally identically distributed with mgf $M_X(t)$, then

$$M_Z(t) = (M_X(t))^n.$$

HW (1) Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ with fixed real constants a and b.

(a) Show that the mgf of X is of the form

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

(b) Show that $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

(2) Let $X_1, \ldots X_n$ be independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants.
Using properties of mgfs show that

$$Z = \sum_{i=1}^{n} (a_i X_i + b_i) \sim \mathcal{N}\Big(\sum_{i=1}^{n} (a_i \mu_i + b_i), \sum_{i=1}^{n} a_i^2 \, \sigma_i^2 \Big)$$

(*) Especially, if $X_1, \ldots X_n$ are independent and equally distributed with $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ then the sum $S = X_1 + \cdots + X_n$ and the mean $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ are also normally distributed with $S \sim \mathcal{N}(n\mu, n\sigma^2)$ and $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

... very important!

Properties of Conditional Expectation

1. $\mathbb{E}(Y \mid X)$ is a function of X, hence a random variable, say $g(X)$.
2. (Iterated Expectations) If Y is integrable, then

$$\mathbb{E}\left(\mathbb{E}(Y \mid X)\right) = \mathbb{E}(Y)$$

- **Random sum of random variables:** Suppose $X_1, X_2, \ldots$ is an infinite sequence of identically distributed random variables, with mean $\mathbb{E}(X_i) = \mu_X$, and suppose N is a nonnegative integer-valued random variable independent of the X_i with $\mathbb{E}(N) = \mu_N$. This implies that $\mathbb{E}(X_i \mid N) = \mathbb{E}(X_i) = \mu_X$.
- What is $\mathbb{E}(S_N) = \mathbb{E}(\sum_{i=1}^{N} X_i)$?
- Answer:

$$\begin{aligned}\mathbb{E}(S_N \mid N) &= \mathbb{E}(\sum_{i=1}^{N} X_i \mid N) \\ &= \mathbb{E}(X_1 \mid N) + \ldots + \mathbb{E}(X_N \mid N) \\ &= \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_N) \\ &= N\mu_X\end{aligned}$$

so that

$$\mathbb{E}(S_N) = \mathbb{E}\left(\mathbb{E}(S_N \mid N)\right) = \mathbb{E}(N\mu_X) = \mathbb{E}(N)\mu_X = \mu_N\mu_X$$

Properties of Conditional Expectation

1. If Y is integrable and a is any function, then

   $$\mathbb{E}\left(a(X)Y \mid X\right) = a(X)\mathbb{E}(Y \mid X)$$

   and

   $$\mathbb{E}\left(a(X)g(Y) \mid X\right) = a(X)\mathbb{E}(g(Y) \mid X)$$

   for any function g.
2. If X and Y are random variables and g and h are functions such that $g(X)$ and $h(Y)$ are integrable, then

   $$\mathbb{E}\left(g(X)h(Y)\right) = \mathbb{E}\left(g(X)\mathbb{E}(h(Y) \mid X)\right)$$

3. If Y is integrable, then

   $$\mathbb{E}\left(\mathbb{E}(Z \mid X, Y) \mid X\right) = \mathbb{E}(Z \mid X)$$

Conditional Expectation and Prediction

1. The parallel axis theorem

$$\mathbb{E}\left((X-a)^2\right) = \mathbb{V}\text{ar}(X) + (a - \mathbb{E}(X))^2$$

   has an analog for conditional expectation.
2. Just replace expectations by conditional expectations (and variances by conditional variances) to obtain the **Conditional Parallel Axis Theorem**

$$\mathbb{E}\left((Y-a(X))^2 \mid X\right) = \mathbb{V}\text{ar}(Y \mid X) + (a(X) - \mathbb{E}(Y \mid X))^2$$

3. Then,

$$\begin{aligned}\mathbb{E}\left((Y-a(X))^2\right) &= \mathbb{E}\left(\mathbb{E}\left((Y-a(X))^2 \mid X\right)\right)\\ &= \mathbb{E}(\mathbb{V}\text{ar}(Y \mid X)) + \mathbb{E}\left((a(X) - \mathbb{E}(Y \mid X))^2\right)\end{aligned}$$

4. Therefore, *for predicting a random variable Y given the value of another random variable X, the predictor function $a(X)$ that minimizes the expected squared prediction error*

$$\mathbb{E}\left((Y-a(X))^2\right)$$

   is the conditional expectation

$$a(X) = \mathbb{E}(Y \mid X)$$

Conditional Variance and Iterated Variance Formula

Conditional variance is just like variance, just replace ordinary expectation with conditional expectation.

$$\begin{aligned}\mathbb{V}\mathrm{ar}(Y \mid X) &= \mathbb{E}\left((Y - \mathbb{E}(Y \mid X))^2 \mid X\right) \\ &= \mathbb{E}(Y^2 \mid X) + \mathbb{E}\left(Y \mid X\right)\right)^2\end{aligned}$$

If Y is integrable, by taking $a(X) = \mathbb{E}(Y)$ (constant), gives

$$\mathbb{E}\left((Y - \mathbb{E}(Y))^2\right) = \mathbb{E}(\mathbb{V}\mathrm{ar}(Y \mid X)) + \mathbb{E}\left((\mathbb{E}(Y) - \mathbb{E}(Y \mid X))^2\right)$$

which obtains the iterated variance formula:

$$\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}\left(\mathbb{V}\mathrm{ar}(Y \mid X)\right) + \mathbb{V}\mathrm{ar}\left(\mathbb{E}(Y \mid X)\right)$$

has an analog for conditional expectation.

1. As an application, we can compute the variance of the random sum of random variables $S_N = \sum_{i=1}^{N} X_i$, where N is random with $\mathbb{E}(N) = \mu_N$ and $\mathbb{V}\mathrm{ar}(N) = \sigma_N^2$, and X_i are iid.
    1. We saw that $\mathbb{E}(S_N \mid N) = N\mu_X$.
    2. Similarly, $\mathbb{V}\mathrm{ar}(S_N) = N\sigma_X^2$
    3. Hence,

$$\begin{aligned} \mathbb{V}\mathrm{ar}(S_N) &= \mathbb{E}\left(\mathbb{V}\mathrm{ar}(S_N \mid N)\right) + \mathbb{V}\mathrm{ar}\left(\mathbb{E}(S_N \mid N)\right) \\ &= \mathbb{E}(N\sigma_X^2) + \mathbb{V}\mathrm{ar}(N\mu_X) \\ &= \mu_N\sigma_X^2 + \mu_X^2\sigma_N^2 \end{aligned}$$

Covariance

- The covariance of two random variables X and Y is given by

$$\mathbb{C}\text{ov}(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))\right).$$

1. Covariance measures the linear relationship between X and Y. Variance is a special case of covariance. For $X = Y$ we obtain
$$\mathbb{C}\text{ov}(X,X) = \mathbb{V}\text{ar}(X).$$

1. Properties
    1. $\mathbb{C}\text{ov}(X,Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$
    2. $\mathbb{C}\text{ov}(aX + b, cY + d) = ac\,\mathbb{C}\text{ov}(X,Y)$ for the constants a, b, c and d
    3. $\mathbb{C}\text{ov}(X_1 + X_2, Y) = \mathbb{C}\text{ov}(X_1, Y) + \mathbb{C}\text{ov}(X_2, Y)$
    4. $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$ for any X and Y
        - $*$ Especially when X and Y are independent then $\mathbb{V}\text{ar}(X+Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)$
    5. If X and Y are independent, then $Cov(X,Y) = 0$.
        - $*$ The converse is not true! If covariance is 0 the random variables X and Y might not be independent.

Correlation

The correlation between X and Y is defined by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}.$$

1. Properties

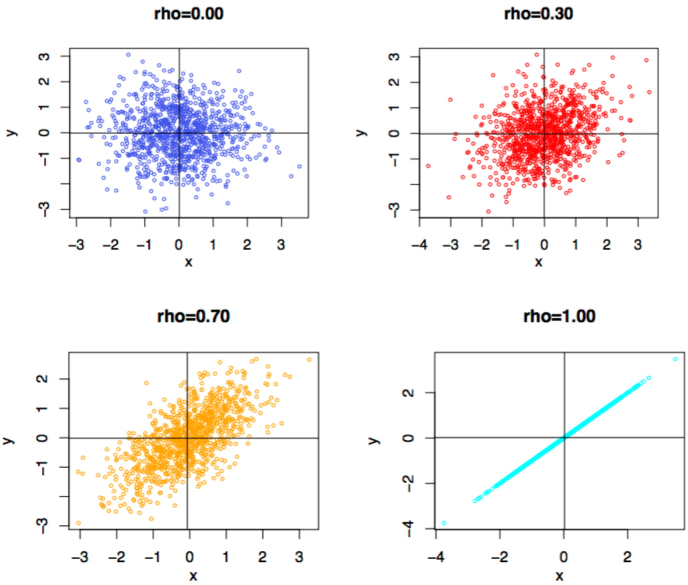
   1. $\rho(X,Y)$ is the covariance of the standardized versions of X and Y.
   2. $-1 \leqslant \rho(X,Y) \leqslant 1$

      $\rho(X,Y) = 1$ if and only if $Y = aX + b$ with $a > 0$
      $\rho(X,Y) = -1$ if and only if $Y = aX + b$ with $a < 0$

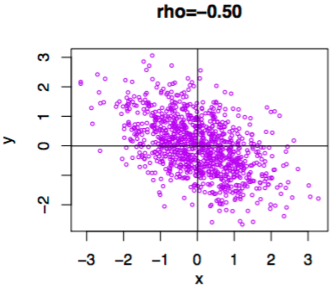
   HW: prove these properties

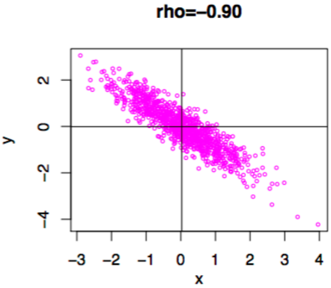
   - Compute $0 \leqslant \mathbb{V}\text{ar}\left(\frac{X}{sd(X)} \pm \frac{Y}{sd(Y)}\right)$
   - Compute $0 \leqslant \mathbb{V}\text{ar}\left(Y - \text{corr}(X,Y)\frac{sd(Y)}{sd(X)}X\right)$

Correlation



Correlation


rho=−0.50


rho=−0.90

Example

- We flip a fair coin three times. Let X be the number of heads in the first two flips and let Y be the number of heads on the last two flips. Compute $Cov(X,Y)$ and $\rho(X,Y)$.

  Answer:

  1. $\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}, \quad |\Omega| = 8.$
  2. Joint probability table:

X \ Y	0	1	2	$\mathbb{P}(X=i) = p(x_i)$
0	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{4}$
1	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$\mathbb{P}(y=j) = p(y_j)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

  3. From marginal distributions we compute $\mathbb{E}(X) = 1$ and $\mathbb{E}(Y) = 1$.
  4. $\mathbb{E}(XY) = 1 \cdot \frac{2}{8} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 4 \cdot \frac{1}{8} = \frac{5}{4}$
  5. $\mathbb{C}\text{ov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{1}{4}$, $\mathbb{V}\text{ar}(X) = \frac{1}{2}$, $\mathbb{V}\text{ar}(Y) = \frac{1}{2}$
  6. $\rho(X,Y) = \frac{\mathbb{C}\text{ov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{1}{2}$

- Second way: using the properties of covariance

  1. Let X_i be the result of ith flip, $i = 1, 2, 3$
  2. $X_i \sim ber(\frac{1}{2})$
     1. $\mathbb{E}(X_i) = \frac{1}{2}$ and $\mathbb{V}\text{ar}(X_i) = \frac{1}{4}$
     2. Also, the different tosses are independent. Thus,

$$\mathbb{C}\text{ov}(X_1, X_2) = \mathbb{C}\text{ov}(X_1, X_3) = \mathbb{C}\text{ov}(X_2, X_3) = 0$$

  3. Then, $X = X_1 + X_2$ und $Y = X_2 + X_3$,

$$\begin{aligned} \mathbb{C}\text{ov}(X, Y) &= \mathbb{C}\text{ov}(X_1 + X_2, X_2 + X_3) \\ &= \mathbb{C}\text{ov}(X_1, X_2) + \mathbb{C}\text{ov}(X_1, X_3) + \mathbb{C}\text{ov}(X_2, X_2) + Cov(X_2, X_3) \\ &= \mathbb{C}\text{ov}(X_2, X_2) = \mathbb{V}\text{ar}(X_2) = \frac{1}{4}. \end{aligned}$$

HW

HW Let X be a random variable that takes values $-2, -1, 0, 1, 2$, each with probability $\frac{1}{5}$. Let $Y = X^2$. Show that $\mathrm{Cov}(X, Y) = 0$, but X and Y are not independent.

HW There are two hospitals in a city. 55 babies are born every day in the larger hospital, while 18 babies are born in the smaller hospital every day. Over a year, wach hospital recorded the days when more than 65% of babies born were boys.

1. Let G_i be the Bernoulli random variable that takes the value 1 if more than 65% of babies born on the ith day in the larger hospital were boys. Let K_i be the Bernoulli random variable that takes the value 1 if more than 65% of babies born on the ith day in the smaller hospital were boys. Find the distributions of G_i and K_i.
2. Let G be the number of days on which more than 65% of babies born in the larger hospital were boys. What is the distribution of G? Compute the expected value and the variance of G.
3. Let K be the number of days that more than 65% of the babies born in the smaller hospital were boys. Find the distribution of K.
4. Use the CLT to approximate the 0.74-quantile of G.
5. Find the correlation between G and K.

Covariance Matrices

The covariance matrix of an m-dimensional random vector $\mathbf{X}$ and an n-dimensional random vector Y is the nonrandom matrix $\mathbf{C} = \mathbb{C}\mathrm{ov}(\mathbf{X}, \mathbf{Y}) : m \times n$ with elements

$$c_{ij} = \mathbb{C}\mathrm{ov}(X_i, Y_j)$$

1. If $\mathbf{X}$ and $\mathbf{Y}$ are random vectors with means $\boldsymbol{\mu}_\mathbf{X}$, $\boldsymbol{\mu}_\mathbf{Y}$, $\mathbf{a}$ and $\mathbf{c}$ constant vectors and $\mathbf{B}$ a constant matrix, then

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\mathbf{X}, \mathbf{Y}) &= \mathbb{E}\left((\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})'\right) \\
\mathbb{V}\mathrm{ar}(\mathbf{X}) &= \mathbb{C}\mathrm{ov}(\mathbf{X}, \mathbf{X}) = \mathbb{E}\left((\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})'\right) \\
\mathbb{V}\mathrm{ar}(\mathbf{a} + \mathbf{B}\mathbf{X}) &= \mathbf{B}\mathbb{V}\mathrm{ar}(\mathbf{X})\mathbf{B}' \\
\mathbb{V}\mathrm{ar}(a + \mathbf{c}'\mathbf{X}) &= \mathbf{c}'\mathbb{V}\mathrm{ar}(\mathbf{X})\mathbf{c}
\end{aligned}$$

2. The variance matrix of any random vector is symmetric and positive semi-definite.
   - A matrix $\mathbf{A}$ is said to be positive semi-definite if

$$\mathbf{c}'\mathbf{A}\mathbf{c} \geqslant 0, \quad \text{for every nonzero vector } \mathbf{c}$$

Correlation Matrices

The correlation matrix of an n-dimensional random vector $\mathbf{X}$ with no constant components, that is $\mathbb{V}\mathrm{ar}(X_i) > 0$ for all i, is the nonrandom matrix $\mathbf{R} = \mathrm{corr}(\mathbf{X}, \mathbf{Y}) : n \times n$ with elements

$$r_{ij} = \frac{\mathbb{C}\mathrm{ov}(X_i, Y_j)}{\sqrt{\mathbb{V}\mathrm{ar}(X_i)\mathbb{V}\mathrm{ar}(X_j)}} = \mathrm{corr}(X_i, X_j)$$

1. Every correlation matrix is positive semi-definite.
2. The correlation matrix of a random vector $\mathbf{X}$ is positive definite if and only the variance matrix of $\mathbf{X}$ is positive definite.
3. If $\mathbf{C}$ is the variance matrix and $\mathbf{D}$ is a diagonal matrix having the same diagonal elements as $\mathbf{C}$, then the correlation matrix is

   $$\mathbf{D}^{-1/2}\mathbf{C}\mathbf{D}^{-1/2}$$

   from which we see that a correlation matrix, like a variance matrix, is positive semidefinite.

Hölder’s and Cauchy-Schwarz Inequalities

Theorem

(Hölder’s Inequality) Let X and Y be any two random variables, and let p and q satisfy

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$|\mathbb{E}XY| \leqslant \mathbb{E}|XY| \leqslant \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} \left(\mathbb{E}|Y|^q\right)^{\frac{1}{q}}. \tag{1}$$

A special case of Hölder's Inequality, when $p = q = 2$, is the Cauchy-Schwarz Inequality.

Theorem

(Cauchy-Schwarz Inequality) For any two random variables X and Y

$$|\mathbb{E}XY| \leqslant \mathbb{E}|XY| \leqslant \left(\mathbb{E}|X|^2\right)^{\frac{1}{2}} \left(\mathbb{E}|Y|^2\right)^{\frac{1}{2}}. \tag{2}$$

(Covariance Inequality 1) Let X and Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively. By applying the Cauchy-Schwarz Inequality we obtain

$$|\mathbb{E}(X-\mu_X)(Y-\mu_Y)| \leqslant \left(\mathbb{E}(X-\mu_X)^2\right)^{\frac{1}{2}} \left(\mathbb{E}(X-\mu_X)^2\right)^{\frac{1}{2}},$$

which after squaring both sides leads to

$$\left(\mathbb{C}\mathrm{ov}(X,Y)\right)^2 \leqslant \sigma_X^2\, \sigma_Y^2.$$

Liapunov’s and Minkowski’s Inequalities

Liapunov’s Inequality is a special case of Hölder’s Inequality for $Y = 1$. Then

$$\mathbb{E}|X| \leqslant \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}}, \quad 1 < p < \infty \tag{3}$$

and

$$\left(\mathbb{E}|X|^r\right)^{\frac{1}{r}} \leqslant \left(\mathbb{E}|X|^s\right)^{\frac{1}{s}}, \quad 1 < r < s < \infty, \tag{4}$$

Theorem

(Minkowski’s Inequality) Let X and Y be any two random variables. Then for $1 \leqslant p < \infty$

$$\left(\mathbb{E}|X+Y|^p\right)^{\frac{1}{p}} \leqslant \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} \left(\mathbb{E}|Y|^p\right)^{\frac{1}{p}}. \tag{5}$$

Jensen's Inequality

Definition

A function $g(x)$ is *convex* if $g(\lambda x + (1-\lambda)y) \leqslant \lambda g(x) + (1-\lambda)g(y)$ for all x and y and $0 < \lambda < 1$. The function $g(x)$ is *concave* if $-g(x)$ is convex.

Theorem

(Jensen's Inequality) For any random variable X, if $g(x)$ is a convex function, then

$$\mathbb{E}(g(X)) \geqslant g\left(\mathbb{E}(X)\right). \tag{6}$$

Equality holds if and only if, for every line $ax + b$ that is tangent to $g(x)$ at $x = \mathbb{E}(X)$, $\mathbb{P}(g(X) = aX + b) = 1$.

Application

1. Jensen's Inequality shows that

$$\mathbb{E}(X^2) \geqslant (\mathbb{E}(X))^2,$$

since $g(x) = x^2$ is convex.

2. Also, if x is positive than $g(x) = \frac{1}{x}$ is convex and

$$\mathbb{E}\left(\frac{1}{X}\right) \geqslant \frac{1}{\mathbb{E}(X)}.$$

3. Jensen's Inequality can be used to prove an inequality between three different kind of means. For positive numbers $a_1, \dots, a_n$ we define

(i)	arithmetic mean	$a_A = \frac{1}{n}(a_1 + a_2 + \cdots + a_n)$
(ii)	geometric mean	$a_G = \left(a_1 a_2 \cdot \cdots \cdot a_n\right)^{\frac{1}{n}}$
(iii)	harmonic mean	$a_H = \dfrac{1}{\frac{1}{n}\left(\frac{1}{a_1} + \frac{1}{a_2} \cdots + \frac{1}{a_n}\right)}$

Then,

$$a_H \leqslant a_G \leqslant a_A.$$

Covariance Inequalities

Covariance inequalities that allow us to bound an expectation without using higher-order moments.

Theorem

(Covariance Inequality 2) Let X be any random variable and $g(x)$ and $h(x)$ any functions such that $Eg(X)$, $Eh(X)$ and $E\left(g(X)h(X)\right)$ exist.

1. If $g(x)$ is a nondecreasing function and $h(x)$ is a nonincreasing function, then
$$E\left(g(X)h(X)\right) \leqslant \left(Eg(X)\right)\left(Eh(X)\right). \tag{7}$$
2. If $g(x)$ and $h(x)$ are either both nondecreasing or both nonincreasing, then
$$E\left(g(X)h(X)\right) \geqslant \left(Eg(X)\right)\left(Eh(X)\right). \tag{8}$$

Large sample theory: Convergence Concepts

Convergence concepts widely used in statistics are convergence in distribution and its special case convergence in probability to a constant.

1. **Convergence in Distribution:** A sequence of random variables $X_1, \ldots, X_n$, with X_n having distribution function F_n, converges in distribution to a random variable X with distribution function F if

   $$F_n(x) \to F(x), \quad \text{as } n \to \infty$$

   for every real number x that is a continuity point of F. We indicate this by writing

   $$X_n \xrightarrow{d} X, \quad \text{as } n \to \infty$$

2. "Continuity point" means a point x such that F is continuous at x (a point where F does not jump). If the limiting random variable X is continuous, then every point is a continuity point. If X is discrete or of mixed type, then $F_n(x) \to F(x)$ must hold at points x where F does not jump but it does not have to hold at the jumps.

Convergence in Distribution

Theorem (Helley-Bray)

A sequence of random variables $X_1, \dots, X_n$ converges in distribution to a random variable X if and only if

$$\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$$

for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$.

- $\mathbb{E}(I_A(X_n)) = \mathbb{P}(X_n \in A)$ may fail to converge to $\mathbb{E}(I_A(X)) = \mathbb{P}(X \in A)$ because indicator functions, though bounded, are not continuous.
- How does one establish that a sequence of random variables converges in distribution? By writing down the distribution functions and showing that they converge? No:
  - In common applications in statistics, convergence in distribution is a consequence of the central limit theorem or the law of large numbers.
  - **[HW]** Let $X_1, X_2, \dots$ be i.i.d uniform $(0, 1)$ and $X_{(n)} = \max_{1 \leqslant i \leqslant n} X_i$. Show that the sequence $Y_n = n(1 - X_{(n)})$ converges in distribution to an exponential $\exp(1)$ random variable.

Convergence in Probability

Definition

A sequence of random variables $X_1, X_2, \ldots$ *converges in probability* to a random variable X if for every $\epsilon > 0$ it holds

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1.$$

- We write $X_n \xrightarrow{\text{p}} X$.
- Note that the definition says nothing about the convergence of the random variables X_n to the random variable X in the sense in which it is understood in real analysis.
- The definition deals only with the convergence of probabilities $\mathbb{P}(|X_n - X| \geqslant \epsilon)$ to 0 (or equivalently the convergence of probabilities $\mathbb{P}(|X_n - X| < \epsilon)$ to 1).

Convergence in Probability to a Constant

Definition

A sequence of random variables $X_1, X_2, \ldots$ *converges in probability* to a constant c if for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - c| \geqslant \epsilon) \to 0 \qquad \text{as. } n \to \infty$$

or, equivalently,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - c| < \epsilon) = 1.$$

We write $X_n \xrightarrow{\text{p}} c$, or

$$X_n - c = o_p(1)$$

Application: Weak Law of Large Numbers (WLLN)

(**Weak Law of Large Numbers (WLLN)**) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(X_1) = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$. Then, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ.
Proof: For every $\epsilon > 0$

$$\begin{aligned}\mathbb{P}(|\bar{X}_n - \mu| \geqslant \epsilon) = \mathbb{P}(\left(\bar{X}_n - \mu\right)^2 \geqslant \epsilon^2) &\leqslant \frac{\mathbb{E}\left(\bar{X}_n - \mu\right)^2}{\epsilon^2} \\ &= \frac{\mathbb{V}\mathrm{ar}(\bar{X}_n)}{\epsilon^2} = \frac{\mathbb{V}\mathrm{ar}(X_1)}{n\varepsilon^2} \\ &= \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

We applied Chebyshev's inequality,

$$\mathbb{P}\left(|X - \mu| \geqslant k\sigma\right) \leqslant \frac{1}{k\sigma^2},$$

and used $\mathbb{V}\mathrm{ar}(\bar{X}_n) = \frac{\mathbb{V}\mathrm{ar}(X_1)}{n}$.

- The LLN also holds if second moments do not exist but it is much harder to prove.

Convergence in Probability and in Distribution

We now prove that convergence in probability to a constant and convergence in distribution to a constant are the same concept:

$$X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c$$

Proof: Let F_n denote the cdf of X_n and suppose $X_n \xrightarrow{d} c$. Then,

$$\mathbb{P}\left(|X_n - c| > \epsilon\right) \leqslant F_n(c-\epsilon) + 1 - F_n(c+\epsilon) \to 0$$

so $X_n \xrightarrow{p} c$.
Conversely, suppose $X_n \xrightarrow{p} c$. Then, for $x < c$,

$$F_n(x) \leqslant \mathbb{P}\left(|X_n - c| > \frac{c-x}{2}\right) \to 0,$$

and for $x > c$,

$$F_n(x) \geqslant 1 - \mathbb{P}\left(|X_n - c| > \frac{x-c}{2}\right) \to 1,$$

so $X_n \xrightarrow{d} c$ □.

It is not true that convergence in distribution to a random variable is the same as convergence in probability to a random variable!

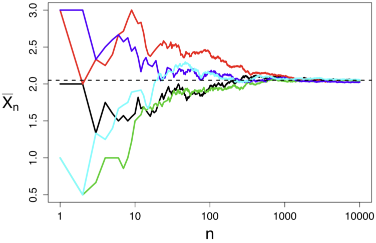
WLLN: Example

Consider a discrete random variable with the pmf

x	0	1	2	3	4
$p(x)$	0.1	0.2	0.3	0.35	0.05

Then, $\mathbb{E}(X) = \sum_{x=0}^{4} x\, p(x) = 2.05$ and $\mathbb{V}\mathrm{ar}(X) = \sum_{x=0}^{4} (x - 2.05)^2\, p(x) = 1.1475$.
Since the variance of X exists, the WLLN says

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{p}} E(X) = 2.05 \quad \text{when } n \to \infty$$



Simulation Example

The following code in R illustrates the LLN. Run the code and comment on the commands provided.

```
n=50
r=30
alpha=5
beta=1

plot(1,type="n",bty="n",xlim=c(0,n),ylim=c(0,3*alpha*beta),
xlab="n",yaxt="n", ylab="", main="Law␣of␣Large␣Numbers")

axis(2,at=c(0,alpha*beta,10*alpha*beta),
labels=c("0",paste("EW=",alpha*beta),""),las=1)

co <- rainbow(n=r)
for(i in 1:r){
m <- cumsum(rgamma(n,shape=alpha, scale=beta))/(1:n)
lines(1:n,m,col=co[i])

abline(h=(alpha*beta),lwd=4)
```

The Continuous Mapping Theorem

Theorem (**Continuous Mapping**)

If g is a function continuous at all points of a set A, if $X_n \xrightarrow{d} X$, and if $\mathbb{P}(X \in A) = 1$, then

$$g(X_n) \xrightarrow{d} g(X)$$

The main point of the theorem is the following two corollaries.

Corollary

If g is everywhere continuous and $X_n \xrightarrow{d} X$, then

$$g(X_n) \xrightarrow{d} g(X)$$

Here the set A in the theorem is the whole real line hence $\mathbb{P}(X \in A) = 1$ holds trivially.

Corollary

If the function g is continuous at the point c and $X_n \xrightarrow{p} c$, then

$$g(X_n) \xrightarrow{p} g(c)$$

Consistency

Definition

An estimator T_n of a parameter θ is called *consistent* if it converges in probability to the true value of the parameter, i.e. if $T_n \xrightarrow{p} \theta$ as $n \to \infty$.

[HW] (**Consistency**) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(X_1) = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$.

1. Show that sample mean is a *consistent* estimator of μ.
2. What is a sufficient condition for the sample variance S^2 to converge in probability to σ^2?

- If we know that $X_n \xrightarrow{p} a$ then by the previous corollary we also have

$$X_n^2 \xrightarrow{p} a^2, \qquad e^{X_n} \xrightarrow{p} e^a \qquad \text{and} \qquad \sqrt{X_n} \xrightarrow{p} \sqrt{a}, \text{ for } a \geqslant 0.$$

- Note that, $X_n \xrightarrow{p} X$ in general does not imply that also moments converge $\mathbb{E}(X_n^k) \xrightarrow{p} \mathbb{E}(X^k)$, for $k > 0$.
- (**Consistency of sample standard deviation**) If sample variance is a consistent estimator of σ^2, then by the continuous mapping theorem, the sample standard deviation is a consistent estimator of σ.
- (**Moments**) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(|X_1|^k) < \infty$ for some positive integer k. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{p} \mathbb{E}(X_1^k) \quad \text{as } n \to \infty.$$

Thus, if $\mathbb{E}(X_1^2) < \infty$ then $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}(X_1^2)$. By Tthe continuous mapping theorem, $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{p} \left(\mathbb{E}(X_1)\right)^2$ and it follows

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\Big) \xrightarrow{p} \mathbb{E}(X_1^2) - \big(\mathbb{E}(X_1)\big)^2 = \mathbb{V}\mathrm{ar}(X_1).$$

Central Limit Theorem

When second moments exist, we actually have something much stronger than the WLLN:

$$\bar{X}_n - \mu = O_p(n^{-1/2})$$

If $X_1, X_2, \ldots$ are iid random variables having mean μ and variance σ^2, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$$

This fact is called the central limit theorem (CLT). The CLT is hard to prove in full generality. We'll show a simpler version, next.

- $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$ means asymptotic normality
- $\bar{X}_n - \mu = O_p(n^{-1/2})$ means root n rate

1. If the random variables $X_1, X_2, \dots$ are i.i.d., then the CLT is a *stronger* result than the WLLN in that the former provides an estimate of the probability $P(|\bar{X}_n - \mu| > \epsilon)$.
2. Namely,

$$P(|\bar{X}_n - \mu| > \epsilon) = P(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{\epsilon\ \sqrt{n}}{\sigma}) \approx 1 - P(|Z| \leqslant \frac{\epsilon\ \sqrt{n}}{\sigma}) = 1 - \Phi(\frac{\epsilon\ \sqrt{n}}{\sigma}),$$

where Z is $\mathcal{N}(0,1)$, and WLLN follows.

3. Also, WLLN does not require the existence of the second moment.

Central Limit Theorem

Theorem

(**Central Limit Theorem (CLT)**) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables whose moment generating functions (mgfs) $M_{X_i}(t)$ exist for $|t| < h$ for some positive h. Let $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2 > 0$. Let $\bar{X}_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ and $G_n(x)$ be the cdf of $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n} \cdot \frac{(\bar{X}_n - \mu)}{\sigma}$. Then, for any x

$$\lim_{n \to \infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy,$$

i.e. $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ has a limiting standard normal distribution.

Proof

We are going to show that the mgf of $\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}$ converges to the mgf of a standard normal random variable, for t in a neighnothood of zero, by using a Taylor series around zero. First we rewrite $\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}$ as follows

$$\begin{aligned}\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} &= \frac{\frac{1}{n}\sum_{i=1}^{n} X_i-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{1}{n}(\sum_{i=1}^{n} X_i-n\mu)}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i,\end{aligned}$$

where $Y_i = \frac{X_i-\mu}{\sigma}$ are i.i.d. random variables with $\mathbb{E}(Y_i)=0$ and $\mathbb{V}\text{ar}(Y_i)=1$. The mgf of the Y_is,

$$M_{Y_i}(t) = e^{-\frac{\mu}{\sigma t}} \cdot M_{X_i}\left(\frac{t}{\sigma}\right)$$

exists for $|t| < h\sigma$.

Proof

From properties of mgf we obtain

$$M_{\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}}(t) = M_{\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} Y_i}(t) = M_{\sum_{i=1}^{n} Y_i}\left(\frac{t}{\sqrt{n}}\right)$$
$$\overset{\text{indep.}}{=} \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Next, we expand $M_{Y_i}(\frac{t}{\sqrt{n}})$ in a Taylor series around zero

$$M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} \frac{M_{Y_i}^{(k)}(0)}{k!} \cdot \left(\frac{t}{\sqrt{n}}\right)^k$$

for $|t| < h\sigma\sqrt{n}$.
The kth moment of Y_i is equal to the kth derivative of the mgf M_{Y_i} evaluated at $t = 0$, i.e. $M_{Y_i}^{(k)}(0) = \frac{d^k}{dt^k} M_{Y_i}(t)\Big|_{t=0} = E(Y_i^k)$.

Proof

Thus, from $M_{Y_i}^{(0)}(0) = 1$, $M_{Y_i}^{(1)}(0) = \mathbb{E}(Y_i) = 0$, $M_{Y_i}^{(2)}(0) = E(Y_i^2) = \mathbb{V}\mathrm{ar}(Y_i) + (\mathbb{E}(Y_i))^2 = 1$ we obtain

$$\begin{aligned} M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) &= M_{Y_i}^{(0)}(0) + \frac{M_{Y_i}^{(1)}(0)}{1!} \cdot \frac{t}{\sqrt{n}} + \frac{M_{Y_i}^{(2)}}{2!} \cdot (\frac{t}{\sqrt{n}})^2 + R_{Y_i}(\frac{t}{\sqrt{n}}) \\ &= 1 + \frac{t^2}{2n} + R_{Y_i}(\frac{t}{\sqrt{n}}) \end{aligned}$$

and for fixed $t \neq 0$, the remainder satisfies $\lim_{n\to\infty} \frac{R_{Y_i}(\frac{t}{\sqrt{n}})}{\left(\frac{t}{\sqrt{n}}\right)^2} = 0$.

Proof

Since t is fixed, $\lim\limits_{n\to\infty} \frac{R_{Y_i}(\frac{t}{\sqrt{n}})}{\left(\frac{1}{\sqrt{n}}\right)^2} = \lim_{n\to\infty} n\, R_{Y_i}(\frac{t}{\sqrt{n}}) = 0$. Thus, for fixed t we can write

$$\lim_{n\to\infty}\left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n = \lim_{n\to\infty}\left(1+\frac{t^2}{2n}+R_{Y_i}(\frac{t}{\sqrt{n}})\right)^n$$

$$= \lim_{n\to\infty}\left(1+\frac{1}{n}\cdot\left(\frac{t^2}{2}+n\,R_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)\right)^n = \lim_{n\to\infty}\left(1+\frac{a_n}{n}\right)^n,$$

where $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty}\left(\frac{t^2}{2}+n\,R_{Y_i}(\frac{t}{\sqrt{n}})\right) = \frac{t^2}{2}$. In the last step we use that for a sequence $a_1, a_2, \ldots$ of real numbers converging to a, $\lim\limits_{n\to\infty}(1+\frac{a_n}{n}) = e^a$.
Thus,

$$\lim_{n\to\infty} M_{\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}}(t) = \lim_{n\to\infty}\left(M_{Y_i}(\frac{t}{\sqrt{n}})\right)^n = e^{\frac{t^2}{2}},$$

the mgf of a standard normal distribution.

The CLT and Addition Rules

Any distribution that has second moments and appears as the distribution of the sum of iid random variables (an “addition rule”) is approximately normal when the number of terms in the sum is large.

- $\text{Bin}(n, p)$ is approximately normal when n is large and neither np or $n(1-p)$ is near zero.
- $\text{NegBin}(r, p)$ is approximately normal when r is large and neither rp or $r(1-p)$ is near zero.
- $\text{Poi}(\lambda)$ is approximately normal when λ is large.
- $\text{Gammma}(\alpha, \lambda)$ is approximately normal when α is large.

The CLT and Addition Rules

Suppose $X_1, X_2, \dots$ are iid Ber(p) and $Y = X_1 + \dots + X_n$, so that $Y \sim Bin(n, p)$. Applying the CLT obtains

$$\sqrt{n}(\bar{X}_n - p) \overset{\text{d}}{\to} \mathcal{N}(0, p(1-p))$$

so that

$$Y = n\bar{X}_n \approx \mathcal{N}(np, np(1-p))$$

by the continuous mapping theorem, which also allows us to deduce from

$$\sqrt{n}(\bar{X}_n - \mu) \overset{\text{d}}{\to} Y$$

where $Y \sim \mathcal{N}(0, \sigma^2)$ to get

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \overset{\text{d}}{\to} \frac{Y}{\sigma}$$

from the continuity of $z \to z/\sigma$, and since $Y/\sigma \sim \mathcal{N}(0, 1)$,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \overset{\text{d}}{\to} \mathcal{N}(0, 1)$$

Correction for Continuity

If X is an integer-valued random variable whose distribution is approximately that of Y, a normal random variable with the same mean and variance as X, and F is the cdf of X and G is the cdf of Y, then the correction for continuity says for integer x

$$\mathbb{P}(X \leqslant x) = F(x) \approx G(x + 1/2)$$

and

$$\mathbb{P}(X \geqslant x) = 1 - F(x - 1) \approx 1 - G(x - 1/2)$$

so for integer a and b

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) &\approx \mathbb{P}(a - 1/2 < Y < b + 1/2) \\ &= G(b + 1/2) - G(a - 1/2) \end{aligned}$$

We always use correction for continuity when the random variable being approximated is integer-valued.

Normal approximation of binomial distribution

Let $X_1, \dots, X_n$ be i.i.d. $bin(1, p)$ random variables (Bernulli trials). Then the sum $S_n = X_1 + \cdots + X_n \sim bin(n, p)$ and $\mathbb{E}(S_n) = np$ and $\mathbb{V}\mathrm{ar}(S_n) = np(1-p)$. Thus,

$$\frac{S_n - np}{\sqrt{np(1-p)}} \approx \mathcal{N}(0, 1).$$

Since binomial random variables are discrete, the normal approximation can be improved by applying the *continuity correction*. Then,

$$\begin{aligned} P(a \leqslant S_n \leqslant b) &= P(a - 0.5 \leqslant S_n < b + 0.5) \\ &= P\Big(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leqslant \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\Big) \\ &\approx \Phi\Big(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\Big), \end{aligned}$$

where $a < b$ are integers.
Assume for example $n = 100$, $p = 0.1$. Then
$P(S_n = 7) = P(6.5 \leqslant S_n < 7.5) = 0.0889$. Using the normal approximation with continuity correction we have

$$P(S_n = 7) = P(6.5 \leqslant S < 7.5) \approx \Phi(\frac{7.5 - 10}{3}) - \Phi(\frac{6.5 - 10}{3}) = 0.0807.$$

The rule of thumb is to use continuity correction, and use normal approximation whenever $\min\{np, np(1-p)\} \geqslant 10$.

Slutsky's Theorem

Theorem

(**Slutsky's Theorem**) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$, a is a constant, then

(a) $Y_n X_n \xrightarrow{d} aX$

(b) $X_n + Y_n \xrightarrow{d} X + a$

(c) $X_n - Y_n \xrightarrow{d} X - a$

(d) $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{a}$ if $a \neq 0$.

Example: As an example of Slutsky's theorem, we show that convergence in distribution does not imply convergence of moments: Let X have the standard normal distribution and Y have the standard Cauchy distribution, and define

$$Z_n = X + \frac{Y}{n}$$

By Slutsky's theorem,

$$Z_n \xrightarrow{d} X$$

but Z_n does not have first moments and X has moments of all orders.

Example

Let $X_1, \ldots, X_n$ be i.i.d. random variables with common $\mathcal{N}(0,1)$ distribution. We are going to determine the limiting distributions of

$$T_n = \sqrt{n} \cdot \frac{X_1 + \cdots + X_n}{X_1^2 + \cdots + X_n^2} \qquad \text{and} \qquad U_n = \frac{X_1 + \cdots + X_n}{\sqrt{X_1^2 + \cdots + X_n^2}}.$$

First we express T_n in the form

$$T_n = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i}{\frac{1}{n} \sum_{i=1}^n X_i^2} = \frac{A_n}{B_n}, \quad \text{where } A_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \text{ and } B_n = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

Here $A_n \sim \mathcal{N}(0,1)$ and $X_1^2, \ldots, X_n^2$ are i.i.d. $\chi^2(1)$ random variables. Also $nB_n = X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$ and $\mathbb{V}\mathrm{ar}(nB_n) = 2n$ and thus $\mathbb{V}\mathrm{ar}(B_n) = \frac{2}{n}$.

Comparison of the LLN and the CLT

1. The CLT implies the LLN: By Slutsky's theorem, if $X_1, X_2, \ldots$ is an iid sequence of random variables with finite variance, the CLT implies

$$\bar{X}_n - \mu = \frac{1}{\sqrt{n}} \cdot \sqrt{n}(\bar{X}_n - \mu) \overset{d}{\to} 0 \cdot Y = 0$$

   where $Y \sim \mathcal{N}(0, \sigma^2)$.

2. Because convergence in distribution to a constant and convergence in probability to a constant are the same thing, this implies the LLN.
3. But the CLT gives much more information than the LLN: It says that the size of the estimation error $\bar{X}_n - \mu$ is about $\sigma/\sqrt{n}$ and also gives us the shape of the error distribution (i. e., normal).

Example

By Chebyshev's inequality we have for any $\epsilon > 0$

$$P(|B_n - 1| > \epsilon) \leqslant \frac{\mathbb{V}\mathrm{ar}(B_n)}{\epsilon^2} = \frac{2}{n\,\epsilon^2} \to 0 \quad \text{as } n \to \infty.$$

We obtained $A_n \overset{\mathrm{d}}{\to} Z$ and $B_n \overset{\mathrm{p}}{\to} 1$, where Z is a standard normal random variable. By Slutsky's theorem it follows that

$$T_n = \frac{A_n}{B_n} \overset{\mathrm{d}}{\to} Z.$$

Similarly we can show that

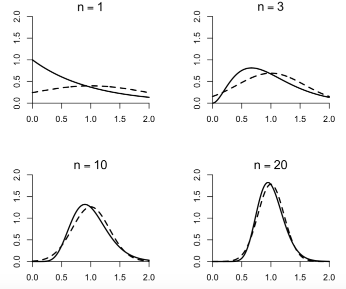
$$U_n = \frac{A_n}{\sqrt{B_n}} \overset{\mathrm{d}}{\to} Z.$$

Normal approximation-CLT for exponential distribution

Let $X_1, X_2, \dots$ be i.i.d. $\exp(1)$ random variables. Then, $\mathbb{E}(X_1) = \mu = 1$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 = 1 > 0$. By the CLT,

$$\bar{X}_n \;\approx\; \mathcal{N}(1, \frac{1}{n}).$$

Since the exponential distribution is very skewed, large samples are needed for the normal approximation to hold. In the following figure, the pdf of $\bar{X}_n$ for $n = 1, 3, 10, 20$ is presented. The dotted lines correspond to the normal approximation according to the CLT.



Example:simulation

[HW] (**Simulation**) As an illustation of the CLT in case of i.i.d. exp(1) random variables, consider the following R code. Comment on commands used.

```
for (n in c(1,2,3,5,10,50,100,500,1000,2000)){
result <- c()
mu <- sigma <- 1
for (i in 1:5000) {
X <-  rexp(n, 1/mu)
result[i] <- (mean(X)-mu)/(sigma/sqrt(n))
}
his <- hist(result, breaks=seq(min(result)-1, max(result)+1,
by=0.25), plot=FALSE)

ylim <- range(his$density, dnorm(0))
hist(result, breaks=seq(min(result)-1, max(result)+1, by=0.25),
prob=TRUE, ylim=ylim, main=paste("n␣=",n), col="yellow")
x <- seq(-4, 4, by=0.1)
lines(x, dnorm(x), lty=1, lwd=2, col="red")
Sys.sleep(1)
```

Limit distribution of T statistics

Assume $X_1, X_2, \ldots$ are i.i.d. random variables such that $\mathbb{E}(X_1) = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\tilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Consider

$$\tilde{T}_n = \frac{\bar{X}_n - \mu}{\frac{\tilde{S}_n}{\sqrt{n}}}.$$

By letting

$$A_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{and} \qquad B_n = \frac{\sigma}{\tilde{S}_n}$$

obtains

$$\tilde{T}_n = A_n B_n.$$

Limit distribution of T statistics

Since

$$A_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

by the CLT and

$$B_n \xrightarrow{\mathrm{p}} 1$$

by the WLLN, we can apply Slutsky's theorem, which then implies that

$$\tilde{T}_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

In other words, T statistics are asymptotically normal (for increasing degrees of freedom the pdf of the Student distribution converges to the pdf of a standard normal distribution).
Note that $\tilde{S}_n^2$ has denominator n. The same limiting distribution is obtained when considering the sample variance S^2 with $n-1$ as the denominator.

Limit distribution of T statistics: Variation

Let $X_1, X_2, \ldots$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Consider the random variable

$$T_n = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}},$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. We know that

$$\frac{(n-1)\,S^2}{n} \xrightarrow{\text{p}} \sigma^2,$$

which implies $S \xrightarrow{\text{p}} \sigma$, and thus $\frac{\sigma}{S} \xrightarrow{\text{p}} 1$. From the CLT we know that $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$. Hence, using Slutsky's theorem

$$T_n = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{\sigma}{S} \xrightarrow{\text{d}} \mathcal{N}(0, 1). \tag{9}$$

- Thus for sufficiently large n ($n \geqslant 30$) we can approximate, for example, $P(T_n \leqslant t)$ by $\Phi(t)$.
- Actually, for this result to hold the Xs need not to be normally distributed. It suffices that they are i.i.d. random variables with common mean μ and variance σ^2. Note that, in general, the rule of thumb that n should be at least 30 for the normal approximation to apply, may not be useful.
- The statistic T_n is a well-known random variable in statistics and we will return to it later in the course. The result (9) will be used to construct large sample confidence interval estimator for μ of the form

$$\left(\bar{X}_n - z\frac{S}{\sqrt{n}}, \bar{X}_n + z\frac{S}{\sqrt{n}}\right)$$

  where z is quantile of the standard normal determined by the desired confidence level, and also in hypotheses testing.

[HW] (Distribution of (modified) sample variance) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $EX_1 = \mu$, $\mathit{Var}\, X_1 = \sigma^2$ and $\mathit{Var}\,(X_1 - \mu)^2 = \tau^2 < \infty$. Consider modified sample variance $\tilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Show that

$$\frac{\tilde{S}_n^2 - \sigma^2}{\frac{1}{\sqrt{n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \tau^2).$$

Note that the above could be formulated in terms of sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Namely, under the same assumptions, one can prove

$$\frac{S^2 - \sigma^2}{\frac{1}{\sqrt{n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \tau^2).$$

The assumptions made in the CLT are crucial. For example, if $X_1, X_2, \ldots$ are i.i.d. standard Cauchy random variables, then the requirement that $\mathbb{V}\mathrm{ar}(X_1) < \infty$ is violated (mgf does not exist), and in fact $\bar{X}_n$ is itself standard Cauchy.

almost sure convergence

Definition

A sequence of random variables $X_1, X_2, \dots$ converges *almost surely* to a random variable X if for every $\epsilon > 0$

$$P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1.$$

Convergence almost surely is stronger than convergence in probability. Moreover,

$$X_n \overset{\text{a.s.}}{\to} X \quad \text{then also } X_n \overset{\text{p}}{\to} X.$$

Theorem

(**Strong Law of Large Numbers (SLLN)**) Let $X_1, X_2, \dots$ be i.i.d. random variables with $EX_1 = \mu$ and $VarX_1 = \sigma^2 < \infty$. Then, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to μ.