T2

Probability Theory

Salvador Ruiz Correa

September 3, 2025

This lecture provides a foundational overview of key concepts in probability theory, emphasizing their relevance to statistical modeling and inference. It begins with the definition and properties of the expected value and variance, which quantify the central tendency and dispersion of random variables. The discussion then extends to joint probability distributions and joint cumulative distribution functions, capturing the behavior of multiple variables simultaneously. Concepts of conditional distribution and Bayes' theorem are introduced to formalize probabilistic reasoning under uncertainty and update beliefs based on observed evidence. The lecture concludes with an exploration of independence and conditional independence of random variables, highlighting their implications for model simplification and factorization in probabilistic frameworks. Together, these principles form the backbone of modern statistical analysis and machine learning, enabling rigorous interpretation and decision-making in complex systems.

Agenda:
1 Random vectors.
2 Joint probability distribution.
3 Marginal distribution
4 Examples.
5 Stocastic process definition.

Expected Value and Variance of a Discrete Random Variable

For a discrete random variable X, its expected value is defined as:

$$\mu_X = \mathrm{E}\left[X\right] = \sum_{\omega\in\Omega} X(\omega)p_X(\omega) = \sum_x xp(x).$$

For example, for the random variable K of Table 1, the expected value is

$$\mu_K = 2(\frac{1}{36}) + 3(\frac{1}{18}) + 4(\frac{1}{12}) + 5(\frac{1}{9}) + 6(\frac{5}{36}) + 7(\frac{1}{6}) + 8(\frac{5}{36}) + 9(\frac{1}{9}) + 10(\frac{1}{12}) + 11(\frac{1}{18}) + 12(\frac{1}{36})$$
$$= 7.$$

More generally, for a given function $g(\cdot)$ we define:

$$\mu_{g(X)} = \mathrm{E}\left[g(X)\right] = \sum_{\omega\in\Omega} g(X(\omega))p_X(\omega) = \sum_x g(x)p(x).$$

The variance of X is defined as:

$$\mathrm{var}\left[X\right] = \sum_{\omega\in\Omega}(X(\omega) - E\left[X\right])^2 p_X(\omega) = \sum_x (x-\mu_X)^2 p(x) = E\left[X^2\right] - (E\left[X\right])^2$$

For example, for the random variable K of Table 1, the variance is computed as follows

$$\begin{aligned}
\mathrm{E}[K^2] &= 2^2(\frac{1}{36}) + 3^2(\frac{1}{18}) + 4^2(\frac{1}{12}) + 5^2(\frac{1}{9}) + 6^2(\frac{5}{36}) + 7^2(\frac{1}{6}) + 8^2(\frac{5}{36}) + 9^2(\frac{1}{9}) + 10^2(\frac{1}{12}) + 11^2(\frac{1}{18}) + 12^2(\frac{1}{36}) \\
&= 54.83 \\
(E[K])^2 &= 49 \\
\mathrm{var}[K] &= 5.83.
\end{aligned}$$

Expected Value and Variance of a Continuous Random Variable

For a continuous random variable X with probability density $f_X(x)$, its expected value is defined as:

$$\mu_x = \mathrm{E}\left[X\right] = \int_{-\infty}^{\infty} \omega f_X(\omega) d\omega.$$

More generally, for a given function $g(\cdot)$ we define:

$$\mathrm{E}\left[g(X)\right] = \int_{-\infty}^{\infty} f_X(\omega) f_X(\omega) d\omega.$$

The variance of X is defined as:

$$\mathrm{var}\left[X\right] = \int_{-\infty}^{\infty} (\omega - E\left[X\right])^2 f_X(\omega) d(\omega) = E\left[X^2\right] - (E\left[X\right])^2.$$

Joint Probability Distribution

When dealing with multiple random variables, it is sometimes useful to use vector and matrix notations. Let $X_1, X_2, \ldots, X_N$ be N discrete random variables.

- The *joint probability function* of $X_1, X_2, \ldots, X_N$ is denoted as

$$p(x_1, x_2, \ldots, x_N) := p_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N).$$

- Using vector notation we can write this distribution as

$$p(\mathbf{x}) := p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$$

  where $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$ is a column (random) vector having $X_1, X_2, \ldots, X_N$ as its components. Similarly $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. Note the abuse of notation in defining $p(\mathbf{x})$ and $p(x_1, x_2, \ldots, x_N)$ above.
- Let $X_1, X_2, \ldots X_N$ be a set of discrete random variables with joint distribution $p(x_1, x_2, \ldots, x_N)$. Variable X_n can be *marginalized* from the joint distribution function as follows.

$$p(\mathbf{x}_{\neg n}) = p(x_1, x_2, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N) \quad = \sum_{x_n} p(x_1, x_2, \ldots, x_N).$$

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	θ_1
0	0	1	θ_2
0	1	0	θ_3
0	1	1	θ_4
1	0	0	θ_5
1	0	1	θ_6
1	1	0	θ_7
1	1	1	θ_8

x_1	x_2	$p(x_1, x_2) = \sum_{x_3} p(x_1, x_2, x_3)$
0	0	$\theta_1 + \theta_2$
0	1	$\theta_3 + \theta_4$
1	0	$\theta_5 + \theta_6$
1	1	$\theta_7 + \theta_8$

- Let X_1, X_2 be two discrete random variables. The *conditional distribution function* of X_1 given X_2 is defined as

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)},$$

$$p(x_1, x_2) = p(x_1 \mid x_2)p(x_2) = p(x_2 \mid x_1)p(x_1),$$

with $p(x_2) > 0$.

x_1	x_2	$p(x_1, x_2)$
0	0	$\theta_1 + \theta_2$
0	1	$\theta_3 + \theta_4$
1	0	$\theta_5 + \theta_6$
1	1	$\theta_7 + \theta_8$

x_2	$p(x_2) = \sum_{x_1} p(x_1, x_2)$
0	$\theta_1 + \theta_2 + \theta_5 + \theta_6$
1	$\theta_3 + \theta_4 + \theta_7 + \theta_8$

x_1	x_2	$p(x_1 \mid x_2)$
0	0	$(\theta_1 + \theta_2)/(\theta_1 + \theta_2 + \theta_5 + \theta_6)$
0	1	$(\theta_3 + \theta_4)/(\theta_3 + \theta_4 + \theta_7 + \theta_8)$
1	0	$(\theta_5 + \theta_6)/(\theta_1 + \theta_2 + \theta_5 + \theta_6)$
1	1	$(\theta_7 + \theta_8)/(\theta_3 + \theta_4 + \theta_7 + \theta_8)$

$\sum_{x_1} p(x_1 \mid x_2) = 1$

- Let X_1, X_2, and X_3 be three discrete random variables. Compute $p(x_2 \mid x_3)$ from the table given below. Compute $\sum_{x_2} p(x_2 \mid x_3, x_1)$.

x_2	x_1	x_3	$p(x_2 \mid x_3, x_1)$
0	0	0	θ_1
0	0	1	θ_2
0	1	0	θ_3
0	1	1	θ_4
1	0	0	$1 - \theta_1$
1	0	1	$1 - \theta_2$
1	1	0	$1 - \theta_3$
1	1	1	$1 - \theta_4$

x_2	x_3	$p(x_2 \mid x_3)$
0	0	$\theta_1 + \theta_3$
0	1	$\theta_2 + \theta_4$
1	0	$1 - (\theta_1 + \theta_3)$
1	1	$1 - (\theta_2 + \theta_4)$

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ be two sets of discrete random variables. The *conditional distribution function* of $\mathbf{X}$

given $\mathbf{Y}$ is defined as

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}$$

with $p(\mathbf{y}) > 0$.

- Let X_1, X_2 be two discrete random variables. The Bayes rule establishes that

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{p(x_2 \mid x_1)p(x_1)}{p(x_2)} = \frac{p(x_2 \mid x_1)p(x_1)}{\sum_{x_1} p(x_2 \mid x_1)p(x_1)}$$

  with $p(x_2) > 0$.

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ be two sets of discrete random variables. The *Bayes rule* establishes that

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} = \frac{p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \mathbf{x})p(\mathbf{y})}{\sum_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})}$$

  with $p(\mathbf{y}) > 0$.

- Let X_1, X_2 be two discrete random variables. These variables are independent if

$$p(x_1, x_2) = p(x_1)p(x_2),$$

  or alternatively,

$$p(x_1 \mid x_2) = p(x_1).$$

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ be two sets of discrete random variables. These sets of variables are *independent* if

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}),$$

  or alternatively,

$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x}).$$

- Let X_1, X_2, X_3 be three discrete random variables. We say that X_1 is *conditionally independent* of X_2 given X_3 (denoted by $X_1 \perp\!\!\!\perp X_2 \mid X_3$) if

$$p(x_1, x_2 \mid x_3) = p(x_1 \mid x_3)p(x_2 \mid x_3)$$

  or alternatively,

$$p(x_1 \mid x_2, x_3) = p(x_1 \mid x_3).$$

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ and $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_L\}$ be three sets of discrete random variables. We say that $\mathbf{X}$ is *conditionally independent* of $\mathbf{Y}$ given $\mathbf{Z}$ (denoted by $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$) if

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$$

  or alternatively,

$$p(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}).$$

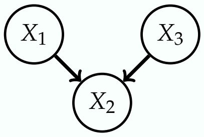
- *Bayesian networks in brief* A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks are directed acyclic graphs (DAGs) whose nodes represent variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters, or hypotheses. Each edge represents a direct conditional dependency. Any pair of nodes that are not connected (i.e. no path connects one node to the other) represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. Bayesian networks having $X_1, X_2, \ldots, X_N$ variables factorize as

$$p(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} p(x_n|\mathsf{pa}(X_n)),$$

where $\mathsf{pa}(\mathsf{x_n})$ represent the parents of variable X_a in the associated DAG.

Box 1: Bayesian Network Example

Consider the example of a Bayesian network having three random variables X_1, X_2, X_3.



$$p(x_1, x_2, x_3) = p(x_2 \mid x_1, x_3) p(x_3) p(x_1)$$

x_2	x_1	x_3	$p(x_2 \mid x_3, x_1)$
0	0	0	θ_1
0	0	1	θ_2
0	1	0	θ_3
0	1	1	θ_4
1	0	0	$1 - \theta_1$
1	0	1	$1 - \theta_2$
1	1	0	$1 - \theta_3$
1	1	1	$1 - \theta_4$

x_1	$p(x_1)$
0	θ_5
1	$1 - \theta_5$

x_3	$p(x_3)$
0	θ_6
1	$1 - \theta_6$

Definition 1: A Bayesian network structure

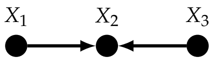
A Bayesian network structure G is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$. Let $\mathsf{pa}_G(X_i)$ denote the parents of X_i in G, and let $\mathsf{nd}(X_i)$ denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of conditional independence assumptions, called the *local independencies*:

$$\forall X_i \in \{X_1, \ldots, X_n\}, \quad X_i \perp\!\!\!\perp \mathsf{nd}(X_i) \mid \mathsf{pa}_G(X_i)$$

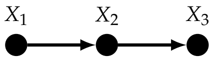
- *Bayesian Network Examples*
  In other words, the local independencies state that each node X_i is conditionally independent of its nondescendants given its parents. Solution.

  Draw the DAG associated with the following probability distributions:

1. $p(x_1, x_2, x_3) = p(x_2)p(x_3)p(x_2 \mid x_1, x_3)$. Show that $X_1 \perp\!\!\!\perp X_3$.

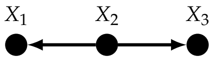


2. $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$. Show that $X_1 \perp\!\!\!\perp X_3|X_2$.



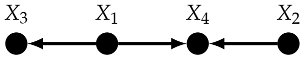
$$\begin{aligned}
p(x_1, x_3 \mid x_2) &= \frac{p(x_1, x_2, x_3)}{p(x_2)} \\
&= \frac{p(x_1)p(x_2|x_1)p(x_3|x_2)}{p(x_2)} \\
&= \frac{p(x_1, x_2)p(x_3|x_2)}{p(x_2)} \\
&= \frac{p(x_1 \mid x_2)p(x_2)p(x_3|x_2)}{p(x_2)} \\
&= p(x_1|x_2)p(x_3|x_2) \\
&\implies X_1 \perp\!\!\!\perp X_3|X_2
\end{aligned}$$

3. $p(x_1, x_2, x_3) = p(x_2)p(x_1|x_2)p(x_3|x_2)$. Show that $X_1 \perp\!\!\!\perp X_3|X_2$.

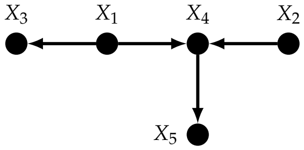


$$\begin{aligned} p(x_1, x_3 \mid x_2) &= \frac{p(x_1, x_2, x_3)}{p(x_2)} \\ &= \frac{p(x_2)p(x_1|x_2)p(x_3|x_2)}{p(x_2)} \\ &= p(x_1|x_2)p(x_3|x_2) \\ &\implies X_1 \perp\!\!\!\perp X_3|X_2 \end{aligned}$$

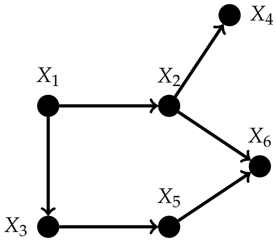
4. $p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2)$. Show that $X_3 \perp\!\!\!\perp X_4|X_1$.



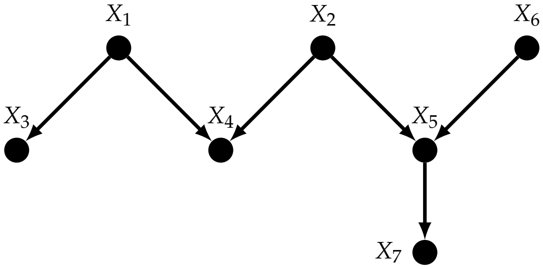
5. $p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2)p(x_5|x_4)$. Show that $X_1 \perp\!\!\!\perp X_5|X_4$.



6. $p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$. Show that $X_2 \perp\!\!\!\perp X_3|X_1$.



7. Write down the factorization of $p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ encoded in the following DAG.



Box 2: Application Example (Was it the burglar?)

Mary lives in San Francisco City. One afternoon, she is driving back home and receives a phone call from her neighbor Jane. She told her that her house alarm was set off (A). While driving, she also heard on the radio (R) that a small earthquake (E) hit the city. Small earthquakes sometimes activate the alarm, and perhaps this is the reason why the alarm was sounding.

- What is the probability that a burglar (B) broke into the house?
- What is the probability that a burglar broke into the house given that the alarm was set off?
- What is the probability that a burglar broke into the house given that the alarm was set off and a small earthquake hit the city?

To solve this problem we make several considerations.

1. Define the random variables $R(\mathsf{R}) = 1$, $R(\mathsf{R}^c) = 0$, $A(\mathsf{A}) = 1$, $A(\mathsf{A}^c) = 0$, $E(\mathsf{E}) = 1$, $E(\mathsf{E}^c) = 0$, and $B(\mathsf{B}) = 1$, $B(\mathsf{B}^c) = 0$.
2. Assume that the joint distribution of random variables R, A, E, and B factorizes as

$$p(R, A, E, B) = p(E = e)p(B = b)p(A = a \mid B = b, E = e)p(R = r \mid E = e)$$

our using our simplified notation

$$p(r, a, e, b) = p(e)p(b)p(a \mid b, e)p(r \mid e).$$

The factors of the probability distribution correspond to the following tables.

	b	$p(b)$
B^c	0	$b_1 = .9$
B	1	$1 - b_1$

	e	$p(e)$
E^c	0	$e_1 = 0.95$
E	1	e_2

	r	$p(r)$
R^c	0	$c_1 = 0.05$
R	1	c_2

		b	e	$p(b, e)$
B^c	E^c	0	0	b_1e_1
B^c	E	0	1	b_1e_2
B	E^c	1	0	b_2e_1
B.	E	1	1	b_2e_2

		r	e	$p(r \mid e)$
R^c	E^c	0	0	$f_1 = 0.99$
R^c	E	0	1	$f_2 = 0.01$
R	E^c	1	0	$f_3 = 1 - f_1$
R.	E	1	1	$f_4 = 1 - f_2$

			a	b	e	$p(a \mid b,e)$
A^c	B^c	E^c	0	0	0	q_1
A^c	B^c	E	0	0	1	q_2
A^c	B	E^c	0	1	0	q_3
A^c	B	E	0	1	1	q_4
A	B^c	E^c	1	0	0	q_5
A	B^c	E	1	0	1	q_6
A	B	E^c	1	1	0	q_7
A	B	E	1	1	1	q_8

$$q_1 + q_5 = 1,\ q_2 + q_6 = 1,$$
$$q_3 + q_5 = 1,\ q_4 + q_8 = 1,$$
$$b_1 + b_2 = 1,\ e_1 + e_2 = 1,$$
$$c_1 + c_2 = 1,$$
$$f_1 + f_3 = 1,$$
$$f_2 + f_4 = 1.$$

The table representing the joint probability function is

				r	a	b	e	$p(r,a,e,b)$
R^c	A^c	B^c	E^c	0	0	0	0	$p_1 = b_1e_1f_1p_1$
R^c	A^c	B^c	E	0	0	0	1	$p_2 = b_1e_2f_1p_2$
R^c	A^c	B	E^c	0	0	1	0	$p_3 = b_2e_1f_1p_3$
R^c	A^c	B	E	0	0	1	1	$p_4 = b_2e_2f_1p_4$
R^c	A	B^c	E^c	0	1	0	0	$p_5 = b_1e_1f_2p_5$
R^c	A	B^c	E	0	1	0	1	$p_6 = b_1e_2f_2p_6$
R^c	A	B	E^c	0	1	1	0	$p_7 = b_2e_1f_2p_7$
R^c	A	B	E	0	1	1	1	$p_8 = b_2e_2f_2p_8$
R	A^c	B^c	E^c	1	0	0	0	$p_9 = b_1e_1f_3p_1$
R	A^c	B^c	E	1	0	0	1	$p_{10} = b_1e_2f_3p_2$
R	A^c	B	E^c	1	0	1	0	$p_{11} = b_2e_1f_3p_3$
R	A^c	B	E	1	0	1	1	$p_{12} = b_2e_2f_3p_4$
R	A	B^c	E^c	1	1	0	0	$p_{13} = b_1e_1f_4p_5$
R	A	B^c	E	1	1	0	1	$p_{14} = b_1e_2f_4p_6$
R	A	B	E^c	1	1	1	0	$p_{15} = b_2e_1f_4p_7$
R	A	B	E	1	1	1	1	$p_{16} = b_2e_2f_4p_8$

$$q_1 + q_5 = 1,$$
$$q_2 + q_6 = 1,$$
$$q_3 + q_5 = 1,$$
$$q_4 + q_8 = 1,$$
$$b_1 + b_2 = 1,$$
$$e_1 + e_2 = 1,$$
$$c_1 + c_2 = 1,$$
$$f_1 + f_3 = 1,$$
$$f_2 + f_4 = 1.$$

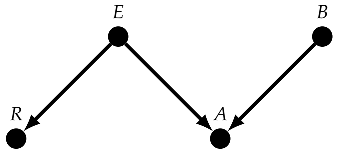


Figure 1: Directed Acyclic Graph encoding the probability distribution $p(r,a,e,b) = p(e)p(b)p(a \mid b,e)p(r \mid e)$. Dark circles represent random variables R, A, E, and B.

- The *graphical model* (a Directed Acyclic Graph or DAG) encoding the probability distribution factorization is shown in Figure 21.
- Assume that $q_1 = 0.98$, $q_2 = 0.9$, $q_3 = 0.1$, $q_4 = 0.01$.
- What is the probability that a burglar broke into the house?

$$P(B = 1) = P(\mathsf{B}) =$$
$$= p_3 + p_4 + p_7 + p_8 + p_{11} + p_{12} + p_{15} + p_{16},$$

$$P(B = 1) = \sum_a \sum_r \sum_e p(r, a, B = 1, e) = 0.1.$$

- What is the probability that a burglar broke into the house given that the alarm was set off?

$$\begin{aligned} P(B=1 \mid A=1) &= \frac{P(B=1, A=1)}{P(A=1)} \\ &= \frac{\mathsf{p}_7+\mathsf{p}_8+\mathsf{p}_{15}+\mathsf{p}_{16}}{\mathsf{p}_5+\mathsf{p}_6+\mathsf{p}_7+\mathsf{p}_8+\mathsf{p}_{13}+\mathsf{p}_{14}+\mathsf{p}_{15}+\mathsf{p}_{16}}, \end{aligned}$$

$$P(B=1 \mid A=1) = \frac{\sum_r \sum_e p(r, A=1, B=1, e)}{\sum_r \sum_b \sum_e p(r, A=1, b, e)} \qquad = 0.81.$$

- What is the probability that a burglar broke into the house given that the alarm was set off and a small earthquake hit the city?

$$\begin{aligned} P(B=1 \mid A=1, E=1) &= \frac{P(B=1, A=1, E=1)}{P(A=1, E=1)} \\ &= \frac{\mathsf{p}_8+\mathsf{p}_{16}}{\mathsf{p}_6+\mathsf{p}_8+\mathsf{p}_{14}+\mathsf{p}_{16}}, \end{aligned}$$

$$\begin{aligned} P(B=1 \mid A=1, E=1) &= \frac{\sum_r p(r, A=1, B=1, E=1)}{\sum_r \sum_b p(r, A=1, b, E=1)} \\ &= 0.51. \end{aligned}$$

Box 3: Solution Summary

- What is the probability that a burglar broke into the house?

$$P(B=1) = 0.1.$$

- What is the probability that a burglar broke into the house given that the alarm was set off?

$$P(B=1 \mid A=1) = 0.81.$$

- What is the probability that a burglar broke into the house given that the alarm was set off and a small earthquake hit the city?

$$P(B=1 \mid A=1, E=1) = 0.51.$$

- The *a priori* probability that the burglar broke into the house is 0.1. The probability that a burglar broke into the house given that the alarm was set off is much higher (0.81). However, knowing that a small earthquake hit the city *explains away* the observation that the alarm was set off, diminishing the probability that a burglar broke (0.51).

- Note that the factorization features of the probability distribution and the sum-product distributive property help us reduce the computational complexity.

$$\begin{aligned}
p(b \mid a) = \frac{p(a,b)}{p(a)} &= \frac{\sum_r \sum_e p(r,a,e,b)}{\sum_b \sum_r \sum_e p(r,a,e,b)} \\
&= \frac{\sum_r \sum_e p(e)p(b)p(a \mid e,b)p(r \mid e)}{\sum_b \sum_r \sum_e p(e)p(b)p(a \mid e,b)p(r \mid e)} \\
&= \frac{p(b) \sum_e p(e)p(a \mid e,b) \sum_r p(r \mid e)}{\sum_b p(b) \sum_e p(e)p(a \mid e,b) \sum_r p(r \mid e)} \\
&= \frac{p(b) \sum_e p(e)p(a \mid e,b)\phi_1(e)}{\sum_b p(b) \sum_e p(e)p(a \mid e,b)\phi_1(e)} \\
&= \frac{p(b)\phi_2(a,b)}{\sum_b p(b)\phi_2(a,b)} \\
&= \frac{p(b)\phi_2(a,b)}{\phi_3(a)}
\end{aligned}$$

Box 4: Application Example (Pairs of dice)

You are told that there are two pairs of coins. The first pair is fair,

$$\theta_1(\mathsf{H}) = \theta_1(\mathsf{T}) = \frac{1}{2}.$$

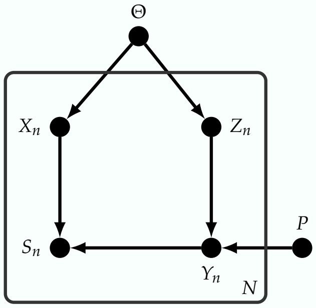
The second pair is biased as both coins have probability

$$\theta_2(\mathsf{H}) = \frac{2}{3}, \; \theta_2(\mathsf{T}) = \frac{1}{3}$$

of producing heads and tails, respectively. One of the two pairs is chosen at random with probability 0.5 and thrown 10 times. The sum of the coins is recorded for each throw.

1. If the throw results in the sequence "1010101010"? Which pair of dice was picked for the throw?
2. Repeat (1) for the sequence "2222200000".
3. Assume that the result of the second coin is flipped (F) with probability $\mathsf{p} \;=\; 0.5$ before recording the result of the throw with probability. Repeat 1 and 2 under this assumption. Consider the case for which $\mathsf{p} = 0.2$. Repeat exercises 1 and 2.

The graphical model associated with this problem is as follows.



- X_n represents the n-th result from the throw of the first coin; $X_n(\mathsf{H}) = 1$ and $X_n(\mathsf{T}) = 0$, and

$$p(x_n \mid \theta) = \theta^{x_n}(1-\theta)^{(1-x_n)}.$$

- Z_n epresents the unobserved result from the n-th throw of the second coin; $Z_n(\mathsf{F}) = 1$ and $Z_n(\mathsf{F}^c) = 0$, and

$$p(z_n \mid \theta) = \theta^{z_n}(1-\theta)^{(1-z_n)}.$$

- Y_n represents the recorded result from the n-th throw of the second coin; $Y_n(\mathsf{H}) = 1$ and $Y_n(\mathsf{T}) = 0$, and

$$p(y_n \mid z_n, p) = p^{I(y_n \neq z_n)}(1-p)^{I(y_n = z_n)}.$$

- S_n is a deterministic function of X_n and Y_n,

$$S_n = X_n + Y_n,$$

$$p(s_n \mid x_n, y_n) = \mathrm{I}(s_n == x_n + y_n),$$

and

$$I(e) = \begin{cases} 1, & \text{if } e = \mathsf{true}; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$, $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$, $\mathbf{S} = (S_1, S_2, \ldots, S_N)$, $f_n = I(y_n \neq z_n) = 1 - I(y_n = z_n)$, and $\mathbf{f} = (f_1, f_2, \ldots, f_N)$. ($f_n$ indicates whether the second was flipped or not.) The joint probability distribution for $\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s}$ given θ and p is

$$\begin{aligned}
p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s}, \theta, p) &= p(\theta)p(p)\prod_{n=1}^{N} p(x_n \mid \theta)p(z_n \mid \theta)p(y_n \mid z_n, p)p(s_n \mid x_n, y_n) \\
p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) &= \frac{p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s}, \theta, p)}{p(\theta)p(p)} \\
p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) &= \prod_{n=1}^{N} p(x_n \mid \theta)p(z_n \mid \theta)p(y_n \mid z_n, p)p(s_n \mid x_n, y_n) \\
&= \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{(1-x_n)}\theta^{z_n}(1-\theta)^{(1-z_n)}p^{I(y_n \neq z_n)}(1-p)^{I(y_n = z_n)}I(s_n = x_n + y_n). \\
&= \theta^{\sum_{n=1}^{N} x_n}(1-\theta)^{\sum_{n=1}^{N}(1-x_n)}\theta^{\sum_{n=1}^{N} z_n}(1-\theta)^{\sum_{n=1}^{N}(1-z_n)}p^{\sum_{n=1}^{N} I(y_n \neq z_n)}(1-p)^{\sum_{n=1}^{N} I(y_n = z_n)}
\end{aligned}$$

Notice that

$$\begin{aligned}
x_n &= s_n - y_n, \\
y_n &= z_n(1 - f_n) + (1 - z_n)f_n,
\end{aligned}$$

and

$$x_n = s_n - z_n + 2z_n f_n - f_n.$$

We define N_x, and N_z as follows,

$$N_x = \sum_{n=1}^{N}(s_n - z_n + 2z_n f_n - f_n) = N_s - N_z + 2N_{zf} - N_f \geq 0,$$

$$\begin{aligned}
N_s &= \sum_{n=1}^{N} s_n, \\
N_z &= \sum_{n=1}^{N} z_n, \\
N_{zf} &= \sum_{n=1}^{N} z_n f_n,
\end{aligned}$$

and

$$N_f = \sum_{n=1}^{N} f_n.$$

Therefore we can write the probability distribution function as follows,

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) &= \theta^{\sum_{n=1}^{N} x_n} (1-\theta)^{\sum_{n=1}^{N}(1-x_n)} \theta^{\sum_{n=1}^{N} z_n} (1-\theta)^{\sum_{n=1}^{N}(1-z_n)} p^{\sum_{n=1}^{N} f_n} (1-p)^{\sum_{n=1}^{N}(1-f_n)} \\ &= \theta^{N_x}(1-\theta)^{(N-N_x)} \theta^{N_z} (1-\theta)^{(N-N_z)} p^{N_f} (1-p)^{N-N_f} \\ &= \theta^{N_x+N_z}(1-\theta)^{(2N-(N_x+N_z))} p^{N_f}(1-p)^{N-N_f} \\ &= p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p) \end{aligned}$$

Observe that $N_x = N_x(\mathbf{s}, \mathbf{z}, \mathbf{f})$, $N_z = N_z(\mathbf{z})$, and therefore

$$p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) = p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p).$$

Also,

$$p(\mathbf{s} \mid \theta, p) = \sum_{\mathbf{x}} \sum_{\mathbf{z}} \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) = \sum_{\mathbf{f}} \sum_{\mathbf{z}} p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p),$$

and

$$\begin{aligned} p(\theta \mid \mathbf{s}, p) = \frac{p(\mathbf{s}, \theta, p)}{p(\mathbf{s}, p)} &= \frac{p(\mathbf{s} \mid \theta, p) p(\theta, p)}{p(\mathbf{s}, p)} \\ &= \frac{p(\mathbf{s} \mid \theta, p) p(\theta \mid p) p(p)}{p(\mathbf{s} \mid p) p(p)} \\ &= \frac{p(\mathbf{s} \mid \theta, p) p(\theta \mid p)}{p(\mathbf{s} \mid p)} \\ &= \frac{p(\mathbf{s} \mid \theta, p) p(\theta)}{p(\mathbf{s} \mid p)}. \end{aligned}$$

since $\theta \perp\!\!\!\perp p$.

Box 5: Decision Rule

$$\mathrm{I}\left(\frac{p(\theta_1 \mid, \mathbf{s}, p)}{p(\theta_2 \mid, \mathbf{s}, p)} > 1\right) \Rightarrow \text{choose } \theta_1,$$

which is equivalent to

$$\mathrm{I}\left(\frac{p(\mathbf{s} \mid \theta_1, p)p(\theta_1)}{p(\mathbf{s} \mid \theta_2, p)p(\theta_2)} > 1\right) \Rightarrow \text{choose } \theta_1,$$

since

$$p(\theta \mid, \mathbf{s}, p) \propto p(\mathbf{s} \mid \theta, p)p(\theta) = \sum_{\mathbf{f}}\sum_{\mathbf{z}} p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p)p(\theta).$$

Notice that

$$p(\mathbf{s}|p) = \sum_{\theta}\sum_{\mathbf{f}}\sum_{\mathbf{z}} p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p)p(\theta),$$

with $\theta \in \{\theta_1, \theta_2\}$.

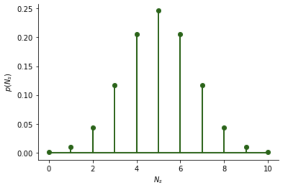


Figure 2: Probability distribution $p(N_s)$ in Box for $\theta = \frac{1}{2}$ for problem 3 in Box 9. The expected value $\mu_{N_s} = 5$ and the entropy is $E(N_s) = 2.71$.

These equations lead to the following results regarding question 3.

$$\frac{p([2,2,2,2,2,0,0,0,0,0] \mid \frac{2}{3}, \frac{1}{2})p(\frac{2}{3})}{p([2,2,2,2,2,0,0,0,0,0] \mid \frac{1}{2}, \frac{1}{2})p(\frac{1}{2})} = 1.80 \Longrightarrow \theta_2 = \frac{2}{3}.$$

$$\frac{p([1,0,1,0,1,0,1,0,1,0] \mid \frac{2}{3}, \frac{1}{2})p(\frac{2}{3})}{p([1,0,1,0,1,0,1,0,1,0] \mid \frac{1}{2}, \frac{1}{2})p(\frac{1}{2})} = 0.13 \Longrightarrow \theta_1 = \frac{1}{2}.$$

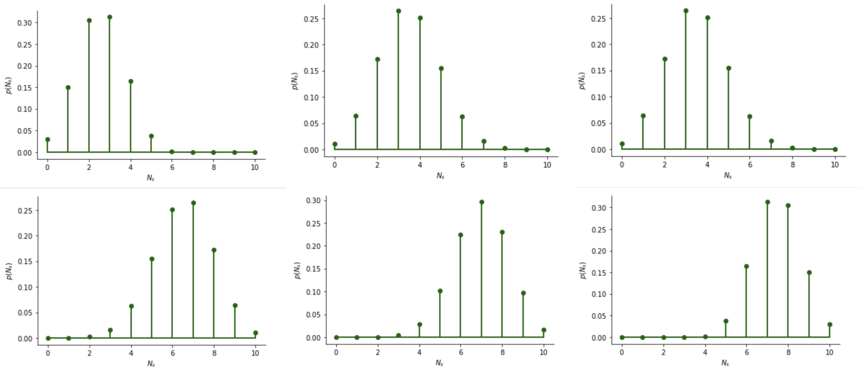


Figure 3: From left to right and from top to bottom, probability distributions $p(N_s)$ in Box for $\theta = \frac{1}{100}$, θ are $\frac{1}{9}$, $\frac{1}{5}$, $\frac{99}{100}$, $\frac{8}{9}$, $\frac{4}{5}$ for problem 3 in Box 9. The corresponding mean values, μ_{N_s}, are 2.55, 3.06, 3.5, 6.5, 6.9, respectively. and $\mu_{N_s} = 7.45$, respectively. The corrresponding entropies, $E(N_s)$, are 2.44, 2.56, 2.56, 2.44, and 2.22, respectively.

Also, notice that

$$\mathrm{E}[\mathbf{s} \mid \theta, p] = \sum_{\mathbf{s}}\sum_{\mathbf{f}}\sum_{\mathbf{z}} \mathbf{s}p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p),$$

$$\mathrm{E}[N_s \mid \theta, p] = \sum_{N_s} N_s p(N_s \mid \theta, p).$$

The *effect* of θ on N_s is

$$e(\theta_1, \theta_2 \mid p) = \mathrm{E}[N_s \mid \theta_1, p] - E[N_s \mid \theta_2, p]$$

Definition 2: Entropy

The average amount of information of a discrtete random variable X is the expectation of $\mathrm{I}(x) = -\log_2(p(x))$ with respect to the distribution $p(x)$ and is given by

$$H(X) = \mathrm{E}[\mathrm{I}(x))] = \mathrm{E}[-\log_2(p(x))] = -\sum_x p(x)\log_2(p(x)).$$

The entropy of a Bernoulli random variable $X \sim \mathsf{Bernoulli}(\mathsf{x} \mid \mathsf{p})$ is

$$H(X \mid p) = -p\log_2(p) - (1-p)\log_2(1-p).$$

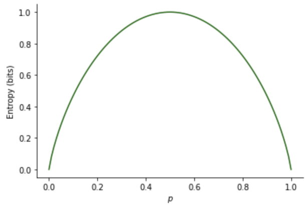


Figure 4: Entropy of the random variable $X \sim \mathsf{Bernoulli}(\mathsf{x} \mid \mathsf{p})$.

Kullback–Leibler Divergence

The Kullback–Leibler divergence (also called KL-divergence, relative entropy, and I-divergence) is a measure of how one probability distribution P_1 is different from a second, reference probability distribution P_2. For discrete probability distributions, P_1 and P_2 defined on the same sample space, the relative entropy from P_2 to P_1 is defined to be

$$D_{KL}(P_1 \| P_2) = -\sum_x p_1(x)\log_2(\frac{p_2}{p_1}).$$

In other words, it is the expectation of the logarithmic difference between the probabilities P_1 and P_2, where the expectation is taken using the probabilities P_1.

Conjugate Priors

Let X and Θ be two random variables. In Bayesian probability theory, if the posterior distribution $p(\theta|x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a *conjugate prior* for the likelihood function $p(x|\theta)$. Recall that:

$$\underbrace{p(\theta \mid x)}_{\textit{posterior}} = \frac{\overbrace{p(x \mid \theta)}^{\textit{likelihood}}\ \overbrace{p(\theta)}^{\textit{prior}}}{\underbrace{p(x)}_{\textit{evidence}}},$$

where

$$p(x) = \underbrace{\int p(x \mid \theta)p(\theta)d\theta}_{\textit{marginalization}}.$$

For example, if $p(\theta)$ has a Beta distribution, and $p(x \mid \theta)$ has a binomial distribution, then $p(\theta \mid x)$ also has a Beta distribution. More specifically,

$$\begin{aligned} &p(x \mid \theta) \sim \mathsf{Binomial}(x, \mid N, p) \\ &p(\theta \mid \alpha, \beta) \sim \mathsf{Beta}(\theta \mid \alpha, \beta) \\ &p(\theta \mid x) \sim \mathsf{Beta}(x \mid \alpha + \sum_{n=1}^{N} x_i, \beta + N - \sum_{n=1}^{N} x_i) \end{aligned}$$

Consider the probability distribution from Box 9:

$$p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p) = \theta^{N_x + N_z}(1-\theta)^{(2N-(N_x+N_z))} p^{N_f}(1-p)^{N-N_f}$$

Clearly,

$$p(\theta, p \mid \mathbf{s}, \mathbf{z}, \mathbf{f}) = \mathsf{Beta}(\theta \mid \alpha + \alpha_1, \beta + \beta_1)\mathsf{Beta}(p \mid \gamma + \gamma_1, \delta + \delta_1)$$

where $\alpha_1 = N_x + N_z$, $\beta_1 = N - (N_x + N_z)$, $\delta_1 = N_f$, $\gamma_2 = N - N_f$, $p(\theta \mid \alpha, \beta) = \mathsf{Beta}(\theta \mid \alpha, \beta)$ and $p(s \mid \gamma, \delta) = \mathsf{Beta}(s \mid \gamma, \delta)$.

Marginalizing with respect to p, $\mathbf{z}$, and $\mathbf{f}$ we obtain:

$$\begin{aligned} p(\theta \mid \mathbf{s}) &= \sum_{\mathbf{z}} \sum_{\mathbf{f}} \mathsf{Beta}(\theta \mid \alpha + \alpha_1, \beta + \beta_1) \int_p \mathsf{Beta}(p \mid \gamma + \gamma_1, \delta + \delta_1) dp, \\ &= \sum_{\mathbf{z}} \sum_{\mathbf{f}} \mathsf{Beta}(\theta \mid \alpha + \alpha_1, \beta + \beta_1). \end{aligned}$$

Expected Value and Covariance Matrix of Random Vectors

Let $X_1, X_2, \ldots, X_N$ be N discrete random variables with joint probability distribution $p(x_1, x_2, \ldots, x_N)$. Let $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$

- The expected value of $\mathbf{X}$ is given by

$$\mathrm{E}[\mathbf{X}] = \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}) = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_N]]^T.$$

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

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X - \mu_x)(Y - \mu_y)] = \mathrm{E}[XY] - \mu_x \mu_y,$$

where

$$\mathrm{E}[XY] = \sum_{\omega_x \in \Omega_x} \sum_{\omega_y \in \Omega_y} X(\omega_x) Y(\omega_y) P(X = X(\omega_x), Y = Y(\omega_y)).$$

Using the notation defined in the previous section we can write

$$\mathrm{E}[XY] = \sum_x \sum_y xy p(x, y).$$

- The covariance of vector $\mathbf{X}$ is defined as

$$\mathrm{cov}(\mathbf{X}) = \begin{bmatrix} \mathrm{cov}(X_1) & \mathrm{cov}(X_1, X_2) & \dots & \mathrm{cov}(X_1, X_N) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \dots & \mathrm{cov}(X_2, X_N) \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}(X_1, X_N) & \mathrm{cov}(X_2, X_N) & \dots & \mathrm{var}(X_N) \end{bmatrix}$$

- The covariance of the random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{cov}(\mathbf{X}, \mathbf{Y}) = \begin{bmatrix} \mathrm{cov}(X_1, Y_1) & \mathrm{cov}(X_1, Y_2) & \dots & \mathrm{cov}(X_1, Y_N) \\ \mathrm{cov}(X_2, Y_1) & \mathrm{cov}(X_2, Y_2) & \dots & \mathrm{cov}(X_2, Y_N) \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}(X_N, Y_1) & \mathrm{cov}(X_N, Y_2) & \dots & \mathrm{cov}(X_N, Y_N) \end{bmatrix}$$

The Multivariate Gaussian Distribution

The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x, the Gaussian distribution (probability density function) can be written in the form

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\},$$

where μ is the mean and σ^2 is the variance. For a N-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\|\boldsymbol{\Sigma}\|^{\frac{1}{2}}} \exp\{-(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\},$$

where $\boldsymbol{\mu}$ is a N-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $N \times N$ covariance matrix, and $\|\boldsymbol{\Sigma}\|$ denotes the determinant of Σ.

The Gaussian distribution arises in many different contexts and can be motivated from a variety of different perspectives. For example, the Gaussian distribution arises when we consider the sum of multiple random variables. The central limit theorem (due to Laplace) tells us that, subject to certain mild conditions, the sum of a set of random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases. We can illustrate this by considering N variables $N_1, \dots, X_N$ each of which has a uniform distribution over the interval $[0, 1]$, and then considering the density of the mean $(X_1 + \cdots + X_N)/N$. For large N, this density tends to be a Gaussian density (Figures 31 and 32).

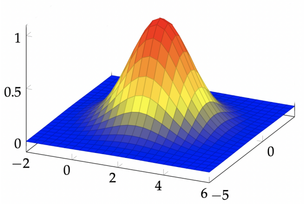


Figure 5: Surface plot of a two-dimensional Gaussian density.

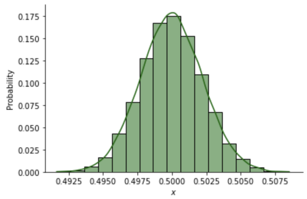


Figure 6: Histogram and KDE plots from 20,000 samples of $(X_1 + \cdots + X_N)/N$ for $N = 50,000$, with $X_n \sim$ Bernoulli$(x \mid p)$. The estimated mean and variance are 0.49, and $5.04e - 06$, respectively.

The Central Limit Theorem

In practice, the convergence to a Gaussian as N increases can be rapid. One consequence of this result is that the binomial distribution, which is the sum of N observations of the Bernoulli random variable, will tend to a Gaussian as $N \to \infty$ (Figure 31).

Theorem 1: The Central Limit Theorem (CLT)

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with expected value $\mathrm{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$Z_N = \frac{\left[\sum_{n=1}^{N} X_n\right] - \mu}{\sqrt{N}\sigma}$$

converges to the standard Gaussian (Normal) random variable as $N \to \infty$,

$$\lim_{N\to\infty} F_{Z_N}(z) = \lim_{N\to\infty} P(Z_N \leq z) = \Phi(z) \quad ,\forall z \in \mathbb{R},$$

where $\Phi(z)$ is the standard Gaussian cummulative distribution function

$$\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(\alpha \mid, 0, 1) d\alpha.$$

That is,

$$Z_\infty \sim \mathcal{N}(z \mid, 0, 1).$$

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with expected value $\mathrm{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$\bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n,$$

i.e. the average of $X_1, X_2, ..., X_N$, has a Gaussian probability density with mean μ and variance $\frac{\sigma^2}{N}$:

$$\bar{X}_N \sim \mathcal{N}(x \mid \mu, \frac{\sigma^2}{N}) \text{ as } N \to \infty.$$

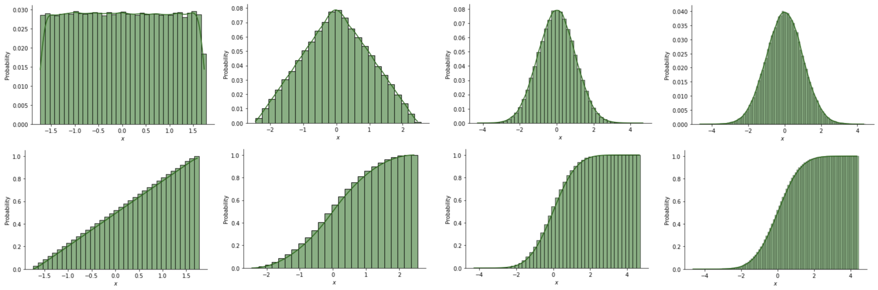


Figure 7: Top: Histogram and KDE plots of Z_N for various values of N: $1, 2, 10$, and 100 (see Definition 15). We observe that as N increases, the density tends towards a Gaussian density. Bottom: Corresponding cumulative distribution functions. Here $X_n \sim \mathrm{Uniform}(x_n \mid 0, 1)$ for wich $\mu = 0.5$ and $\sigma^2 = \frac{1}{12}$.

Clearly, $\lim_{N\to\infty} \bar{X}_N = \mu$, so the average is an *unbiased estimator* of the mean.

Chebyshev's Inequality

Let X be a random variable with a finite expected value μ and finite non-zero variance σ^2. Then for any real number $k > 0$,

$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}.$$

Only the case $k > 1$ is useful. When $k \leq 1$, right-hand side $\frac{1}{k^2} \geq 1$ and the inequality is trivial as all probabilities are ≤ 1.

Box 6: Example

Suppose that an unbiased coin is thrown 100 times. What is the bound that the number of heads will be greater than 70 or less than 30?

- Let K be the number of heads. Because K has a binomial distribution with $\mu = 0.5$:
- $\mathrm{E}[K] = N\mu = 50$.
- $\mathrm{var}[K] = N\mu(1-\mu) = 25$.
- The standard deviation is $\mathrm{std}[K] = \sqrt{\mathrm{var}[K]} = \sqrt{25} = 5$.
- The values 70 and 30 are 20 units from the average, which is 4 standard deviations (i.e., $20/5$).

$$P(\mid K - \mathrm{E}[K] \mid \geq 4\mathrm{var}[K]) \leq \frac{1}{4^2} = 0.0625.$$

- Repeat the previous exercise for a) $\mu = 0.6$ and b) $\mu = 0.4$.