Lecture 2: Review of Probability

Zheng Tian

Contents

1 Random Variables and Probability Distributions 1

2 Expectation, Variance, and Other Moments 6

3 Two Random Variables 8

4 Four Specific Distributions 12

5 Random Sampling and the Distribution of the Sample Average 15

6 Large Sample Approximations to Sampling Distributions 16

This lecture will review the basics in probability theory. The review is by no means comprehensive. We will just refresh our mind with the concepts that will be used the lectures that follow.

1 Random Variables and Probability Distributions

1.1 Defining probabilities and random variables

Experiments, outcomes, sample space, and events

Probabilities are defined with respect to things whose occurrence are random. We use the idea of an **experiment** to symbolize the processes that generate random results. The **outcomes** of an experiment are its mutually exclusive potential results. For example, a simple experiment might be tossing a coin, the outcome of which is either getting a head(H) or a tail(T) but not both.

We can denote all the outcomes from an experiment with a set S, that is called the **sample space**. In the tossing-coin experiment, the sample space is $\{H, T\}$. Or if we toss a dice, the sample space is $\{1, 2, 3, 4, 5, 6\}$. An **event** is a subset of the sample space. Getting a head is an event, which is $\{H\} \subset \{H, T\}$.

Probability

The **probability** of an event is the proportion of the time that the event will occur in the long run. For example, we toss a coin for n times and get m heads. When n is very large, we can say that the probability of getting a head in a toss is m/n. Obviously, we cannot always repeat an experiment with infinite times. So we need a general (axiomatic) definition of probability as follows.

Definition of probability A probability of an event A in the sample space S, denoted as $\Pr(A)$, is a function that assign A a real number in $[0, 1]$, satisfying the following three conditions:

1. $0 \leq \Pr(A) \leq 1$.
2. $\Pr(S) = 1$.
3. For any disjoint sets, A and B, that is A and B have no element in common, $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

Here we use the concept of disjoint (or mutually exclusive) sets. A and B are disjoint if there is no common element between these two sets, that is, $A \cap B$ is an empty set.

Random variables

Instead of using words or a set symbol to represent an event or an outcome, we can use numeric value to do so. A **random variable** is thus a numerical summary associated with the outcomes of an experiment. You can also think of a random variable as a function mapping from an event ω in the sample space Ω to the real line, as illustrated in Figure 1.

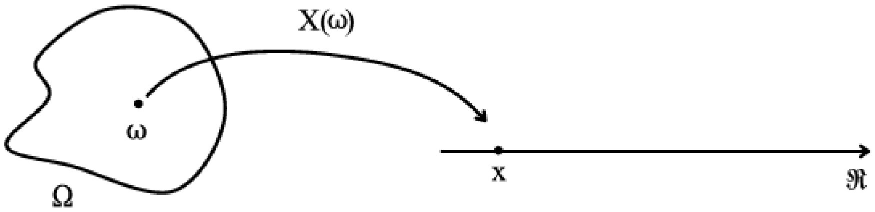


Figure 1: An illustration of random variable

Random variables can take different types of values. A **discrete** random variables takes on a discrete set of values, like $0, 1, 2, \ldots, n$ whereas a **continuous** random variable takes on a continuum of possble values, like any value in the interval (a, b).

1.2 Probability distributions

The probability distribution for a discrete random variable

The probability distribution of a discrete random variable is the list of all possible values of the variable and the probability that each value will occur. These probabilities sum to 1.

- The probability mass function

  Let X be a discrete random variable. The probability distribution of X (or the probability mass function), $p(x)$, is

  $$p(x) = \Pr(X = x)$$

  where we use X to denote the random variable and x to denote a specific value that X can take. We denote the set of all possible value of X as S.

  The axioms of probability require that (1) $0 \leq p(x) \leq 1$ and (2) $\sum_{\{\text{all } x \text{ in } S\}} p(x) = 1$.

Table 1: An illustration of the probability distribution of a discrete random variable

X	1	2	3	Sum
P(x)	0.25	0.75	0.25	1

- The cumulative probability distribution

  The **cumulative probability distribution** (or the cumulative distribution function, c.d.f.) is the probability that the random variable is less than or equal to a particular value. Let $F(x)$ be the c.d.f of X. Then $F(x) = p(X \leq x)$.

  Table 2 and Figure 2 show that the c.d.f. of a discrete random variable is a step function of x.

Table 2: An illustration of the c.d.f. of a discrete random variable

X	1	2	3	Sum
P(x)	0.25	0.50	0.25	1
C.d.f.	0.25	0.75	1	–

- Bernouli distribution

  Many experiments like tossing a coin generate two outcomes: 1 with the probability of p and 0 with the probability of $1 - p$. The random variable generated from such an experiment follows the Bernoulli distribution.

  The Bernoulli distribution

  $$G = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

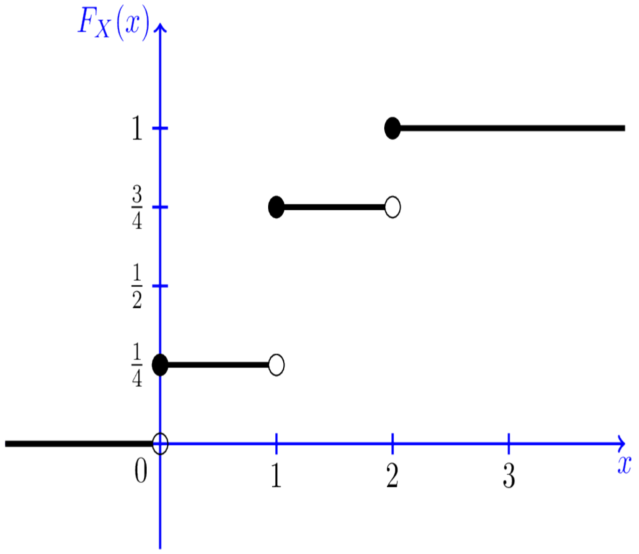


Figure 2: The c.d.f. of a discrete random variable

The probability distribution of a continuous random variable

Unlike a discrete random variable that we can enumerate its values for each corresponding event, a specific value of a continuous random variable is just a point in the real line, the probability of which is zero. Instead, we use the concept of the **probability density function (p.d.f)** as the counterpart of the probability mass function. And the definition of the p.d.f. of a continuous random variable depends on the definition of its. c.d.f.

The cumulative distribution function of a continous random variable is defined as it is for a discrete random variable. That is, for a continous random variable, X, the c.d.f. is $F(x) = \Pr(X \leq x)$. And the **p.d.f.** of X is the function that satisfies

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \text{ for all } x$$

For both discrete and continuous random variable, $F(X)$ must satisfy the following properties:

1. $F(+\infty) = 1$ and $F(-\infty) = 0$ ($F(x)$ is bounded between 0 and 1)
2. $x > y \Rightarrow F(x) \geq F(y)$ ($F(x)$ is nondecreasing)

By the definition of the c.d.f., we can conveniently calculate probabilities, such as,

- $\mathrm{P}(x > a) = 1 - \mathrm{P}(x \leq a) = 1 - F(a)$
- $\mathrm{P}(a < x \leq b) = F(b) - F(a)$.

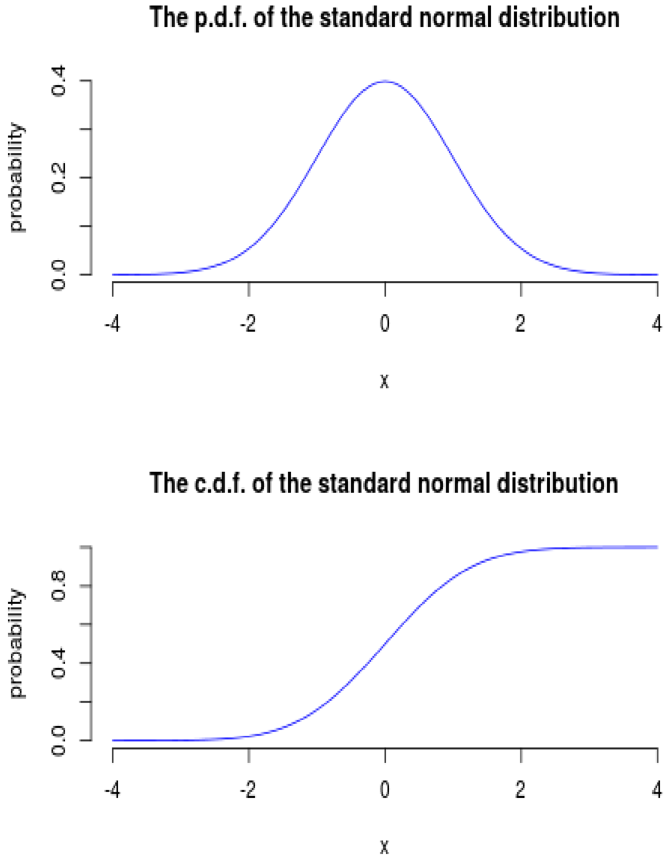


Figure 3: The p.d.f. and c.d.f. of a continuous random variable (the normal distribution)

2 Expectation, Variance, and Other Moments

2.1 The expected value of a random variable

Definition

The **expected value** of a random variable, X, denoted as $E(X)$, is the long-run average of the random variable over many repeated trials or occurrences, which is also called the **expectation** or the **mean**. The expected value measures the centrality of a random variable.

- For a discrete random variable

$$E(X) = \sum_{i=1}^{n} x_i \text{Pr}(X = x_i)$$

  e.g. The expectation of a Bernoulli random variable, G

$$E(G) = 1 \cdot p + 0 \cdot (1 - p) = p$$

- For a continuous random variable

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

2.2 The variance and standard deviation

The **variance** of a random variable X measures its average deviation from its own expected value. Let $E(X) = \mu_X$ and denote the variance of X, denoted as $\text{Var}(X)$ or σ_X^2, is then

$$\text{Var}(X) = E(X - \mu_X)^2 = \begin{cases} \sum_{i=1}^{n} (x - \mu_X)^2 \text{Pr}(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

The **standard deviation** of X is the square root of $\text{Var}(X)$ and is denoted as σ_X. That is, $\sigma_X = \sqrt{\text{Var}(X)}$

A convenient formula for calculating the variance is

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

The variance of a Bernoulli random variable, G

$$\text{Var}(G) = (1 - p)^2 p + (0 - p)^2 (1 - p) = p(1 - p)$$

and $\sigma_G = \sqrt{p(1 - p)}$.

From the definition of the expectation and variance, we can compute the expectation and variance of a linear function of X. Let $Y = a + bX$, then

- $E(Y) = a + E(X)$
- $\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$.

2.3 Moments of a random variable, skewness and kurtosis

The expectation and variance are two special cases of the **moments** of a distribution.

Definition of the moments of a distribution

k$^{\mathbf{th}}$ moment The k$^{\text{th}}$ **moment** of the distribution of X is $E(X^k)$. So, the expectation is the "first" moment of X.

k$^{\mathbf{th}}$ central moment The k$^{\text{th}}$ central moment of the distribution of X with its mean μ_X is $E(X - \mu_X)^k$. So, the variance is the second central moment of X.

It is important to remember that not all the moments of a distribution exist. This is especially true for continuous random variables, for which the integral to compute the moments may not converge.

Skewness and kurtosis

We also use the third and fourth central moments to measure how a distribution looks like asymmetric and how thick are its tails.

- Skewness

  The skewness of a distribution provides a mathematical way to describe how much a distribution deviates from symmetry. It is defined as

  $$\text{Skewness} = E(X - \mu_X)^3/\sigma_X^3$$

  - A symmetric distribution has a skewness of zero.
  - The skewness can be either positive or negative.
  - That $E(X - \mu_X)^3$ is divided by σ_X^3 is to make the skewness measure unit free. That is, changing the units of Y does not change its skewness.

- Kurtosis

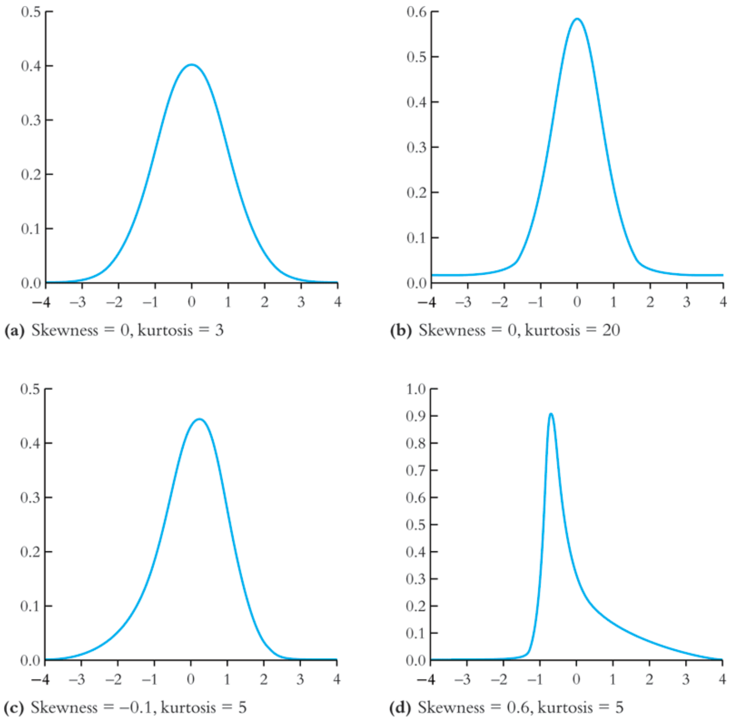
  The kurtosis of the distribution of a random variable X measures how much of the variance of X arises from extreme values, which makes the distribution have "heavy" tails.

The kurtosis of the distribution of X is

$$\text{Kurtosis} = E(X - \mu_X)^4 / \sigma_X^4$$

- The kurtosis must be positive.
- The kurtosis of the normal distribution is 3. So a distribution that has its kurtosis exceeding 3 is called heavy-tailed, or **leptokurtic**.
- The kurtosis is also unit free.

Figure 4 displays four distributions with different skewness and kurtosis. All four distributions have a mean of zero and a variance of one, while (a) and (b) are symmetric and (b)-(d) are heavy-tailed.



(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = −0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5

Figure 4: Four distributions with different skewness and kurtosis

3 Two Random Variables

3.1 The joint and marginal distributions

The joint probability distributions

- The joint probability (density) functions
  - For two discrete random variables X and Y, the joint probability distribution of X and Y is the probability that X and Y simultaneously take on certain values, x and

y, that is

$$p(x, y) = \Pr(X = x, Y = y)$$

which must satisfy the following

1. $p(x, y) \geq 0$
2. $\sum_x \sum_y p(x, y) = 1$

- For two continuous random variables, the counterpart of $p(x, y)$ is the joint probability density function, $f(x, y)$, such that

  1. $f(x, y) \geq 0$
  2. $\int_x \int_y f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$

The marginal probability distribution

- The marginal probability (density) distribution

  The marginal probability (density) distribution function of X is computed from the joint probability (density) distribution function, $f(x, y)$ as

$$f_X(x) = \begin{cases} \sum_y p(x, y) & \text{in the discrete case} \\ \int_y f(x, y)\,\mathrm{d}y & \text{in the continuous case} \end{cases}$$

3.2 Conditional distributions

The conditional probability

For any two events A and B, the conditional probability of A given B is defined as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

The conditional probability (density) distribution

- For the discrete case, the conditional probability function is

$$p(x|y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$$

- the continuous case, the conditional density function is

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

The conditional expectation

- Definition

$$E(X|Y=y) = \begin{cases} \sum_x xp(x|y) & \text{in the discrete case} \\ \int_x xf(x|y)\,\mathrm{d}x & \text{in the continuous case} \end{cases}$$

- The law of iterated expectation: $E(X) = E\left[E(X|Y)\right]$

  Proof. we prove the law of iterated expectation for the continuous case. The proof for the discrete case is similar.

$$\begin{aligned} E(X) &= \int_x xf_X(x)\,\mathrm{d}x \\ &= \int_x \int_y xf(x,y)\,\mathrm{d}y\,\mathrm{d}x && \text{by the definition of } f_X(x)) \\ &= \int_x \int_y xf(x|y)f_Y(y)\,\mathrm{d}y\,\mathrm{dx} && \text{by the definition of } f(x|y) \\ &= \int_y \left[\int_x xf(x|y)\,\mathrm{d}x\right] f_Y(y)\,\mathrm{d}y && \text{by the property of integral} \\ &= \int_y E(X|Y=y)f_Y(y)\,\mathrm{d}y \\ &= E\left[E(X|Y)\right] \end{aligned}$$

□

  - If $E(X|Y) = 0$, then $E(X) = E\left[E(X|Y)\right] = 0$.

Independence

Two random variables X and Y are **independent** if

$$f(x|y) = f_X(x) \text{ or } f(y|x) = f_Y(y)$$

It follows that X and Y are independent if

$$f(x,y) = f(x|y)f_Y(y) = f_X(x)f_Y(y)$$

3.3 Covariance and Correlation

Covariance

The covariance of two random variables X and Y is

$$\mathrm{Cov}(X,Y) = E(X-\mu_X)(Y-\mu_Y) \equiv \sigma_{XY}$$

The covariance can also be computed as

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

If X and Y are independent, then

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \int_x \int_y (x-\mu_x)(y-\mu_Y) f(x,y)\,\mathrm{d}(x)\,\mathrm{d}(y) \\
&= \int_x (x-\mu_X) f_X(x)\,\mathrm{d}(x) \int_y (y-\mu_Y) f_Y(y)\,\mathrm{d}(y) \\
&= [E(X)-\mu_X]\,[E(Y)-\mu_Y] \\
&= 0
\end{aligned}$$

It means that if X and Y are independent, then they are uncorrelated as well. But the opposite direction does not hold.

Correlation

The correlation coefficient between X and Y is given by

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{[\mathrm{Var}(X)\mathrm{Var}(Y)]^{1/2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \equiv \rho_{XY}$$

Some useful operations

Let X, Y be random variables, with the means μ_X and μ_Y, the variance σ_X^2 and σ_Y^2, and the covariance σ_{XY}, respectively. Then,

$$\begin{aligned}
E(a+bX+cY) &= a + b\mu_X + c\mu_Y \\
\mathrm{Var}(aX+bY) &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \\
\mathrm{Cov}(a+bX+cV,Y) &= b\sigma_{XY} + c\sigma_{VY} \\
\mathrm{Cov}(X,Y) &\leq \sigma_X\sigma_Y \text{ and } \mathrm{Corr}(X,Y) \leq 1
\end{aligned}$$

4 Four Specific Distributions

4.1 The normal distribution

Definition

The p.d.f. of a normally distributed random variable x is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

for which $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, and x is denoted as $x \sim N(\mu, \sigma^2)$

The standard normal distribution is a special case of the normal distribution, for which $\mu = 0$ and $\sigma = 0$. The p.d.f of the standard normal distribution is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

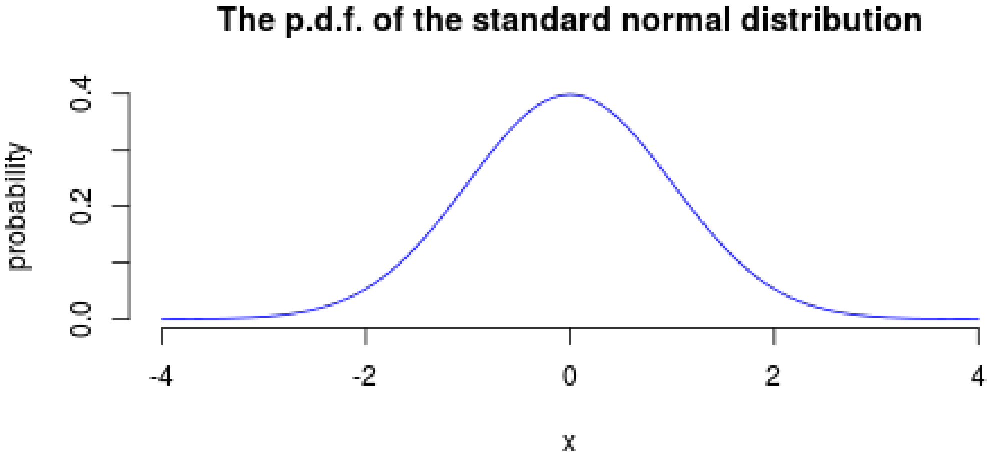
The c.d.f of the standard normal distribution is often denoted as $\varPhi(x)$.

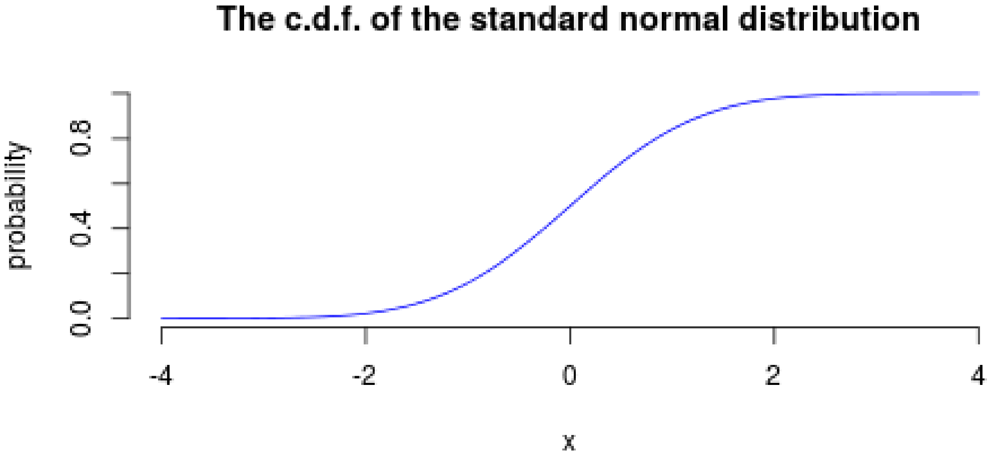
Transforming a normally distributed random variable to the standard normal distribution

Let X be a normally distributed random variable with the mean μ and the standard deviation σ, i.e., $X \sim N(\mu, \sigma^2)$. Then $Z = \frac{X-\mu}{\sigma}$ follows the standard normal distribution, $N(0,1)$.

It follows that for any two number $c_1 < c_2$ and let $d_1 = (c_1 - \mu)/\sigma$ and $d_2 = (c_2 - \mu/\sigma)$, then

$$\begin{aligned}
\Pr(X \le c_2) &= \Pr(Z \le d_2) = \varPhi(d_2) \\
\Pr(X \ge c_1) &= \Pr(Z \ge d_1) = 1 - \varPhi(d_1) \\
\Pr(c_1 \le X \le c_2) &= \Pr(d_1 \le Z \le d_2) = \varPhi(d_2) - \varPhi(d_1)
\end{aligned}$$





The multivariate normal distribution

The normal distribution can be generalized to describe the joint distribution of a set of random variables, which have the multivariate normal distribution. (See Appendix 17.1 for the p.d.f of this distribution and the special case of the bivariate normal distribution.)

- Important properties of the multivariate normal distribution

  1. If X and Y have a bivariate normal distribution with covariance σ_{XY} and a and b are two constants, then

  $$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X + b^2\sigma_Y + 2ab\sigma_{XY})$$

  More generally, if n random variables, $x_1, \ldots, x_n$, have a multivariate normal distribution, then any linear combination of these variables is normally distributed, for example, $\sum_{i=1}^{n} x_i$. For any real numbers, $\alpha_1, \ldots, \alpha_n$, a linear combination of x_i is $\sum_i \alpha_i x_i$.

2. If a set of random variables has a multivariate normal distribution, then the marginal distribution of each of the variables is normal.

3. If random variables with a multivariate normal distribution have covariances that equal zero, then these random variables are independent.

   Let X and Y be two random variables with a bivariate normal distribution. The joint p.d.f of X and Y is $f(x, y)$, with the marginal p.d.f. being $f_X(x)$ and $f_Y(y)$, respectively. Then we have

   $$\text{Cov}(X, Y) = 0 \Leftrightarrow f(x, y) = f_X(x)f_Y(y)$$

   Note: this property only holds for random variables with a multivariate normal distribution. Generally, uncorrelation does not imply independence.

4. If X and Y have a bivariate normal distribution, then

   $$E(Y|X = x) = a + bx$$

   where a and b are constants.

4.2 The chi-squared distribution

Let $Z_1, \ldots, Z_n$ be n indepenent standard normal distribution, i.e. $Z_i \sim N(0, 1)$ for all $i = 1, \ldots, n$. The random variable

$$W = \sum_{i=1}^{n} Z_i^2$$

has a chi-squared distribution with n degrees of freedom, denoted as $W \sim \chi^2(n)$, with $E(W) = n$ and $\text{Var}(W) = 2n$

e.g. If $Z \sim N(0, 1)$, then $W = Z^2 \sim \chi^2(1)$ with $E(W) = 1$ and $\text{Var}(W) = 2$.

4.3 The student t distribution

Let $Z \sim N(0, 1)$, $W \sim \chi^2(m)$, and let Z and W be independently distributed. Then the random variable

$$t = \frac{Z}{\sqrt{W/m}}$$

has a student t distribution with m degrees of freedom, denoted as $t \sim t(m)$.

As $n \to \infty$, t has a standard normal distribution.

4.4 The F distribution

Let $W_1 \sim \chi^2(n_1)$, $W_2 \sim \chi^2(n_2)$, and W_1 and W_2 are independent. Then the random variable

$$F = \frac{W_1/n_1}{W_2/n_2}$$

has an F distribution with (n_1, n_2) degrees of freedom, denoted as $F \sim F(n_1, n_2)$

- If $t \sim t(n)$, then $t^2 \sim F(1, n)$
- As $n_2 \to \infty$, the $F(n_1, \infty)$ distribution is the same as the $\chi^2(n_1)$ distribution, divided by n_1.

5 Random Sampling and the Distribution of the Sample Average

5.1 Random sampling

Simple random sampling n objects are selected at random from a **population**, and each member of the population is equally likely to be included in the sample

i.i.d. draws when $Y_1, Y_2, \ldots, Y_n$ are drawn from the same distribution and are independently distributed, they are said to be **independently and identically distributed** or **i.i.d**. This fact can be denoted as $Y_i \sim IID(\mu_Y, \sigma_Y^2)$ for $i = 1, 2, \ldots, n$.

5.2 The sampling distribution of the sample average

The sample average

The **sample average** or **sample mean**, $\overline{Y}$, of the n observation $Y_1, Y_2, \ldots, Y_n$ is

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Note that since Y_i is random, so is $\overline{Y}$.

The mean and variance of $\overline{Y}$

Suppose that $Y_i \sim IID(\mu_Y, \sigma_Y^2)$ for all $i = 1, \ldots, n$. Then, by the definition of $\overline{Y}$ and the fact that Y_i and Y_j are independent for any $i \neq j$, implying $\text{Cov}(Y_i, Y_j) = 0$, we have

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \cdot n\mu_Y = \mu_Y$$

and

$$\operatorname{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(Y_i, Y_j) = \frac{\sigma_Y^2}{n}$$

e.g. If $Y_1, \ldots, Y_n$ are i.i.d. draws from $N(\mu_Y, \sigma_Y^2)$, then $\overline{Y} \sim N(\mu_Y, \sigma_Y^2/n)$.

6 Large Sample Approximations to Sampling Distributions

6.1 The law of large numbers

Convergence in probability

Let $S_1, \ldots, S_n, \ldots$ be a sequence of random variables, denoted as $\{S_n\}$. $\{S_n\}$ is said to converge in probability to a limit μ that is, $S_n \xrightarrow{p} \mu$, if and only if

$$\Pr\left(|S_n - \mu| \geq \delta\right) \to 0$$

as $n \to \infty$ for every $\delta > 0$.

- e.g. $S_n = \overline{Y}$. That is, $S_1 = Y_1$, $S_2 = 1/2(Y_1 + Y_2)$, $S_n = 1/n\sum_i Y_i$, and so forth.

The law of large numbers

If $Y_1, \ldots, Y_n$ are i.i.d., $E(Y_i) = \mu_Y$ and $\operatorname{Var}(Y_i) < \infty$, then $\overline{Y} \xrightarrow{p} \mu_Y$

6.2 The central limit theorem

Convergence in distribution

Let $F_1, F_2, \ldots, F_n, \ldots$ be a sequence of cumulative distribution functions corresponding to a sequence of random variables, $S_1, S_2, \ldots, S_n, \ldots$. Then the sequence of random variables S_n is said to **converge in distribution** to S, denoted as $S_n \xrightarrow{d} S$, if the distribution functions $\{F_n\}$ converge to F, the distribution function of S. That is,

$$S_n \xrightarrow{d} S \text{ if and only if } \lim_{n\to\infty} F_n(t) = F(t)$$

where the limit holds at all points t at which the limiting distribution F is continuous. The distribution F is called the **asymptotic distribution** of S_n.

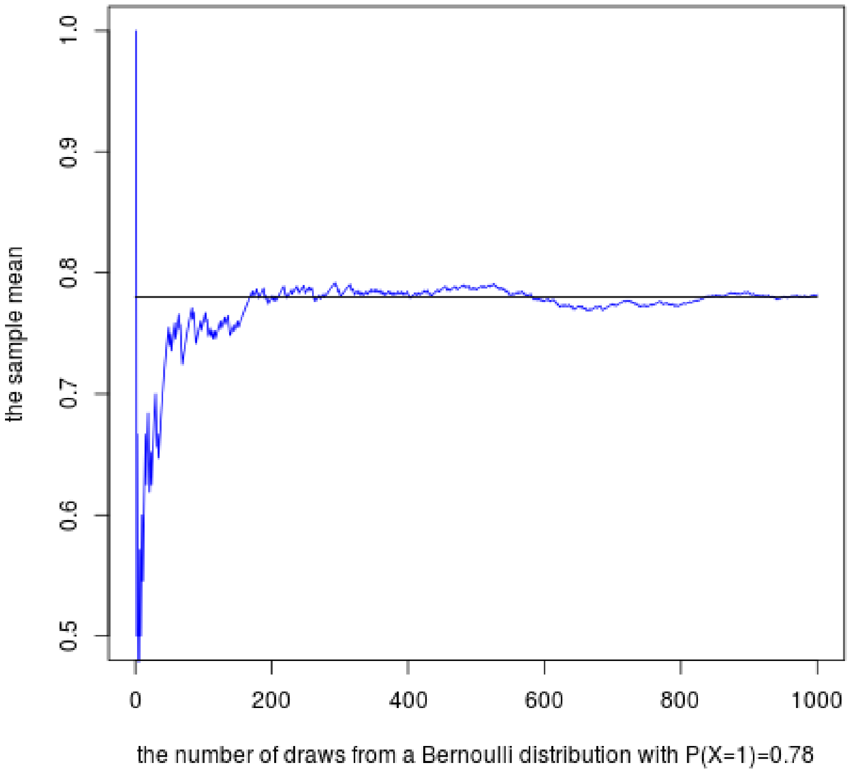


Figure 5: An illustration of the law of large numbers

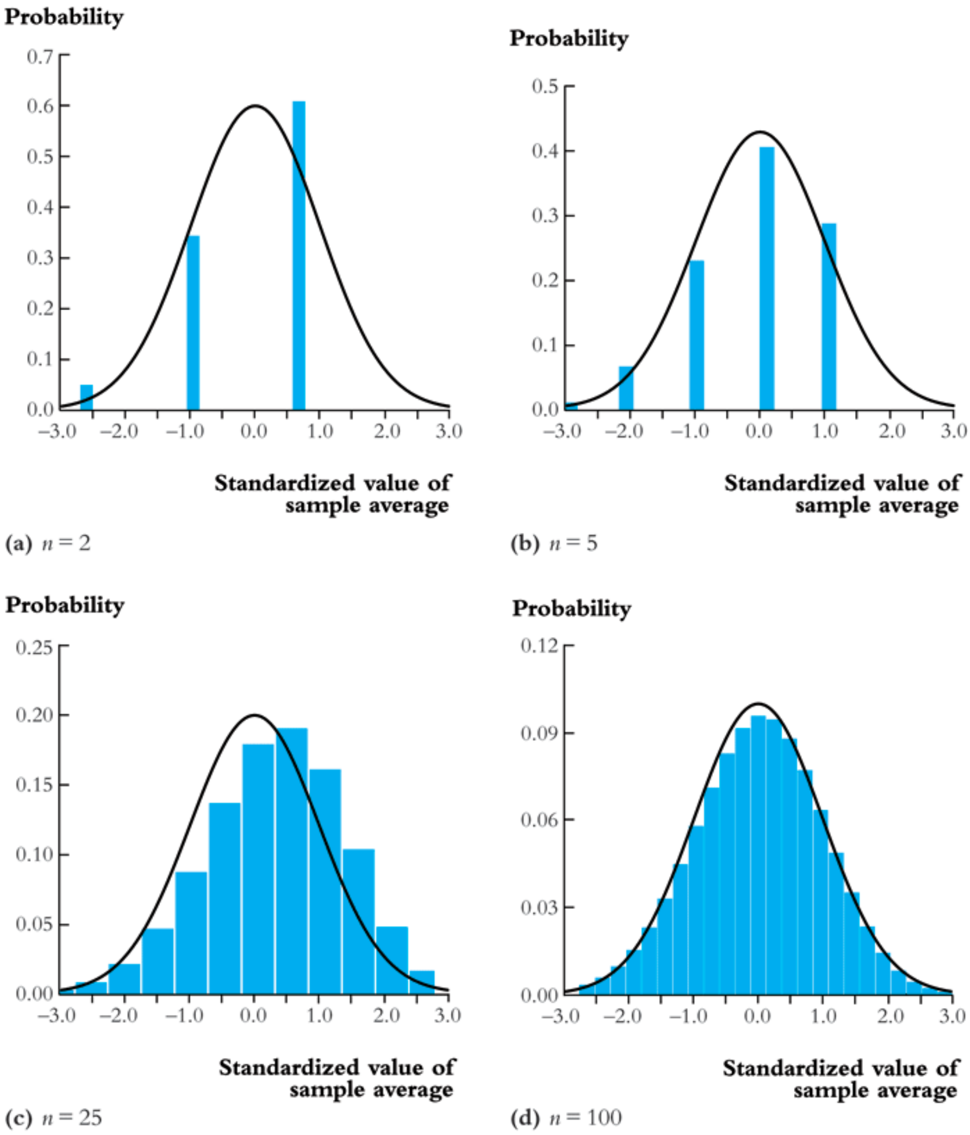
The central limit theorem (Lindeberg-Levy CLT)

If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random samples from a probability distribution with finite mean μ_Y and finite variance σ_Y^2, i.e., $0 < \sigma_Y^2 < \infty$ and $\overline{Y} = (1/n)\sum_i^n Y_i$. Then

$$\sqrt{n}(\overline{Y} - \mu_Y) \xrightarrow{d} N(0, \sigma_Y^2)$$

It follows that since $\sigma_{\overline{Y}} = \sqrt{\text{Var}(\overline{Y})} = \sigma_Y/\sqrt{n}$,

$$\frac{\overline{Y} - \mu_Y}{\sigma_{\overline{Y}}} \xrightarrow{d} N(0, 1)$$



(a) $n = 2$

(b) $n = 5$

(c) $n = 25$

(d) $n = 100$

Figure 6: Distribution of the standardized sample average of n Bernoulli random variable with p = 0.78