STAT1011
Introduction to Statistics

Tutorial 5

Chan Chun Ho, Benjamin*†

October 31, 2017

Abstract

It aims to introduce normal distribution which is one of the continuous probability distributions. Although the concepts of continuous probability distributions are beyond the scope of this course, Section 1 is optional yet useful to understand properties of normal distribution. Some exercises are provided for students to practice. Some materials are extracted from STAT2001 lecture note and do credit to Dr Ho Kwok Wah.

If students are interested in formal treatment of continuous probability distributions, they should consider to take STAT2001.

Notations and Definitions

- Set of real numbers: $\mathbb{R} = (-\infty, \infty)$
- Set membership: $x \in A$ means "x is an element of the set A".
- Interval: $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$
- Absolute value: $|x| = x$ if $x \geq 0$, $|x| = -x$ if $x < 0$
- Differentiation: $\frac{d}{dx} f(x)$; Integration: $\int f(x)dx$
- Expectation: $E(\cdot)$; Variance: $Var(\cdot)$; Standard deviation: $SD(\cdot)$
- Probability distribution: A tilde ($\sim$) means "has the probability distribution of".
- Parameter(s): Population characteristic(s) that can be set to different values to produce different probability distributions

*For enquiry, please email to 1155049861@link.cuhk.edu.hk.
†Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho

1 (Optional) Introduction

1.1 Continuous Random Variables

- A **random variable** is a function that maps each element in a sample space into a real number $x \in \mathbb{R}$.
- A random variable X is said to be **continuous** if the **cumulative distribution function (cdf)** $F(x) = P(X \leq x)$ is a continuous function for all $x \in \mathbb{R}$.

1.2 Continuous Probability Distributions

- For a continuous random variable X, the **probability density function (pdf)** $f(x)$ has to satisfy the following conditions:
  1. $f(x) \geq 0$
  2. $P(X = x) = 0$ for any x
  3. $P(a \leq X \leq b) = \int_a^b f(x)dx$
  4. $\int_{-\infty}^{\infty} f(x)dx = 1$

  The pdf is a non-negative function, which has no direct probability interpretation, that is $f(x) \neq P(X = x)$. From (1) and (2), $f(x)$ can be greater than 1 and the probability of X at any point x is 0. From (3), the probability of X falling in $[a, b]$ is given by integrating $f(x)$ over $[a, b]$. Intuitively, this probability is given by the area bounded by the function, the x-axis and the vertical lines $x = a$ and $x = b$. In fact, $F(X) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$. By Fundamental Theorem of Calculus, $f(x) = \frac{d}{dx}F(x)$. Informally, "density" means the rate of change in cumulative probability.

1.3 Mean and Variance of Continuous Random Variables

- The **mean**, or **expectation**, or **expected value** of X and $g(X)$ are

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

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

- The **variance** of X is

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = E(X^2) - \mu^2$$

- The **standard deviation** of X is simply $\sigma = SD(X) = \sqrt{Var(X)}$.

2 Normal Distribution

A **normal random variable** X with parameters (μ, σ^2), denoted by $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma > 0$, has a pdf as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \quad -\infty < x < \infty.$$

- $E(X) = \mu$
- $Var(X) = \sigma^2$

The normal distribution (sometimes called the Gaussian distribution) is generally considered to be the most important continuous distribution in Statistics.

Since $f(x)$ is (1) maximized at $x = \mu$, (2) symmetric about $x = \mu$ and (3) bell-shaped, the mean, mode and median are equal to μ. Intuitively, μ and σ^2 control the location and scale of $f(x)$. If μ is greater, the center of $f(x)$ is located more to the right. If σ^2 is greater, the spread is wider while the peak is lower since the total area under $f(x)$ is always 1.

2.1 Standard Normal Distribution

A **standard normal random variable** Z, denoted by $N(0, 1)$ has a pdf as

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad , \quad -\infty < z < \infty.$$

- $E(Z) = 0$
- $Var(Z) = 1$

Standard normal distribution is a special case of normal distribution with $\mu = 0$ and $\sigma^2 = 1$. The numerical values, $P(Z \leq z)$, can be obtained from many computer packages or from table.

Theorem 1. If $X \sim N(\mu, \sigma^2)$, define $Z = \frac{X-\mu}{\sigma}$. Since X is a random variable, Z is also a random variable. Then, $Z \sim N(0, 1)$.

Theorem 2. If $Z \sim N(0, 1)$, define $X = \mu + \sigma Z$. Since Z is a random variable, X is also a random variable. Then, $X \sim N(\mu, \sigma^2)$.

Theorem 3. $P(|X - \mu| \leq \sigma) = P(|Z| \leq 1) = 0.68$, $P(|X - \mu| \leq 2\sigma) = P(|Z| \leq 2) = 0.95$ and $P(|X - \mu| \leq 3\sigma) = P(|Z| \leq 3) = 0.997$.

2.2 Useful Formulas

Assume that $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$. Here $f(x)$, $f(z)$ and $F(x)$ are defined as before.

(a) $P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2)$

(b) $P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$

(c) $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$

(d) $P(Z \leq -z) = P(Z \geq z)$

For formula (a), since X is a continuous random variable, from Section 1, it is known that $P(X = x) = 0$.

For formula (b), it is by definition of $F(x) = P(X \leq x)$.

For formula (c), note that $P(-\infty < X < \infty) = 1$, hence $P(X \leq x) + P(X > x) = 1$.

For formula (d), note that $f(z)$ is symmetric about $z = 0$.

3 Exercises

Exercise 5.1. (13-14 1st Term Final, Q4, 15%)
In classifying students, three categories are used based on the scores of a test; individuals whose scores are less than 140 (Group 1), those with scores between 140 and 160 (Group 2), and those with scores over 160 (Group 3). For all students who took the test, the test scores are normally distributed with a mean equal to 124 and a standard deviation equal to 13.7. Suppose a random sample of 10 students who took the test is selected, what is the probability that six of them are in Group 1, three of them are in Group 2, and one of them is in Group 3?

Solution. Let X be a student score, so $X \sim N(124, 13.7^2)$. Let p_1, p_2 and p_3 be probabilities of a score falling in Group 1, 2 and 3 respectively.

$$\begin{aligned} p_1 &= P(X < 140) \\ &= P(Z < \frac{140 - 124}{13.7}) \\ &= P(Z < 1.17) \\ &= 0.879 \end{aligned}$$

$$\begin{aligned} p_2 &= P(140 \leq X \leq 160) \\ &= P(\frac{140 - 124}{13.7} \leq Z \leq \frac{160 - 124}{13.7}) \\ &= P(1.17 \leq Z \leq 2.63) \\ &= 0.9957 - 0.879 \\ &= 0.1167 \end{aligned}$$

$$\begin{aligned} p_3 &= P(X > 160) \\ &= 1 - P(X \leq 160) \\ &= 1 - P(Z \leq \frac{160 - 124}{13.7}) \\ &= 1 - P(Z \leq 2.63) \\ &= 1 - 0.9957 \\ &= 0.0043 \end{aligned}$$

[In general, if $X \sim N(\mu, \sigma^2)$, then $P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$, where $Z \sim N(0, 1)$.]

Note that $p_1 + p_2 + p_3 = 0.879 + 0.1167 + 0.0043 = 1$

Let Y_1, Y_2, Y_3 be number of individuals falling in Groups 1, 2 and 3. By multinomial distribution with parameters $(10, 0.879, 0.1167, 0.0043)$, (please refer to tutorial 4)

$$P(Y_1 = 6, Y_2 = 3, Y_3 = 1) = \frac{10!}{6!\ 3!\ 1!}\ (0.879)^6\ (0.1167)^3\ (0.0043)^1 = 0.002648$$

Exercise 5.2. (13-14 2nd Term Final, Q2, 10%)
Suppose that you bet \$5 each of a sequence of 50 independent fair games. (When you win, you get \$10, including the \$5 you bet. When you lose, you lose \$5). Find the probability that you will lose more than \$75.

Solution. Let X be the number of lost games. Then, $X \sim$ binomial(50,0.5).
Since $np = (50)(0.5) = 25 > 5$ and $n(1-p) = 25 > 5$, we can use normal approximation to binomial distribution. (Please refer to lecture 5 and tutorial 6.)

Remark: If you lose 32 games, you lose 5(32)-5(18)=\$70. If you lose 33 games, you lose 5(33)-5(17)=\$80.

$$\begin{aligned}
P(\text{lose more than \$75}) &= P(X \geq 33) \\
&\approx P(Z > \frac{33 - 0.5 - 50 \cdot 0.5}{\sqrt{50 \cdot 0.5 \cdot 0.5}}) \quad (-0.5 : \text{continuity correction}) \\
&= P(Z > 2.12) \\
&= 1 - P(Z \leq 2.12) \\
&= 1 - 0.9830 \\
&= 0.017
\end{aligned}$$

Exercise 5.3. A candy maker produces mints that have a label weight of 20.4 grams. Assume that the distribution of the weights of these mints is $N(21.37, 0.16)$.

(a) Let X denote the weight of a single mint selected at random from the production line. Find $P(X > 22.07)$.

(b) Suppose that 15 mints are selected at random and weighted. Let Y equal the number of these mints that weigh less than 20.857 grams. Find $P(Y \leq 2)$.

Solution.
(a)

$$P(X > 22.07) = P(\frac{X - 21.37}{\sqrt{0.16}} > \frac{22.07 - 21.37}{\sqrt{0.16}}) = P(Z > 1.75) = 1 - 0.9599 = 0.0401.$$

(b)

$$P(X < 20.857 = P(\frac{X - 21.37}{\sqrt{0.16}} > \frac{20.857 - 21.37}{\sqrt{0.16}}) \approx P(Z < -1.28) \approx 0.1.$$

So, $Y \sim$ binomial$(15, 0.1)$. (Please refer to tutorial 4.)

$$\begin{aligned} P(Y \leq 2) &= P(Y=0) + P(Y=1) + P(Y=2) \\ &= \binom{15}{0}(0.1)^0(0.9)^{15} + \binom{15}{1}(0.1)^1(0.9)^{14} + \binom{15}{2}(0.1)^2(0.9)^{13} \\ &= 0.20589 + 0.34315 + 0.26690 \\ &= 0.81594 \end{aligned}$$

Exercise 5.4. (Quality Control)
A shoe factory owns a machine that cuts pieces from slabs of compressed rubber to be used as soles on a certain brand of men's shoe. The thickness measurements of these soles are normally distributed with the standard deviation $\sigma = 0.2$ millimeters. Occasionally, for some unforeseeable reason, the mean changes from its target setting of $\mu = 25$ millimeters. To be able to take timely corrective measures, such as readjusting the machine's setting, it is important to monitor product quality by measuring the thickness of a random sample of soles taken periodically from the machine's output. Suppose that the following plan is used to monitor the product quality. The thickness measurements for a random sample of 5 soles are observed, and the sample mean $\bar{X}$ is recorded. If $\bar{X} < 24.8$ or $\bar{X} > 25.2$, the machine is considered to be out of control. Production is then halted and the machine is readjusted.

(a) When the true mean is $\mu = 25$ millimeters, what is the probability that a sample will indicate 'out of control'?

(b) When the true mean is $\mu = 25$ millimeters, what is the probability that three consecutive samples will not indicate 'out of control'?

(c) Suppose that the true mean has changed to $\mu = 25.3$ millimeters. What is the probability that a sample will indicate 'out of control'?

Solution.
(a) If $\mu = 25$, $\sigma = 0.2$,

$$\begin{aligned} P(\text{a sample will indicate 'out of control'}) &= 1 - P(24.8 \leq \bar{X} \leq 25.2) \\ &= 1 - P(\frac{24.8-25}{0.2/\sqrt{5}} \leq Z \leq \frac{25.2-25}{0.2/\sqrt{5}}) \\ &\approx 1 - P(-2.24 \leq Z \leq 2.24) \\ &= P(Z < -2.24) + P(Z > 2.24) \\ &= 2 \cdot P(Z < -2.24) \\ &= 0.025 \end{aligned}$$

[Optional: Given that the production mean is equal to the target, there is still 0.025 probability that a sample will indicate 'out of control'. In fact, it is a type I error and the probability is called α. In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis.]

(b) P(3 samples will not indicate 'out of control') $= (1 - 0.025)^3 = 0.9269$.

(c) If $\mu = 25.3$,

$$\begin{aligned} P(\text{a sample will indicate 'out of control'}) &= 1 - P(24.8 \le \bar{X} \le 25.2) \\ &= 1 - P(\frac{24.8 - 25.3}{0.2/\sqrt{5}} \le Z \le \frac{25.2 - 25.3}{0.2/\sqrt{5}}) \\ &\approx 1 - P(-5.59 \le Z \le -1.12) \\ &= P(Z < -5.59) + P(Z > -1.12) \\ &= P(Z < -5.59) + [1 - P(Z \le -1.12)] \\ &= 0.0001 + 1 - 0.1314 \\ &= 0.8687 \end{aligned}$$

[Optional: Given that the production mean is more than 1σ above the target, there is still $1 - 0.8687 = 0.1313$ probability that a sample will not indicate 'out of control'. In fact, it is a type II error and the probability is called β. In statistical hypothesis testing, a type II error is the incorrect acceptance of a false null hypothesis.]

In fact, the true mean μ is unknown and only the sample means $\bar{X}$s are observed. Due to the sampling distribution of sample means, the 'out of control' signal is probabilistic.

Interested students could take STAT4007 Statistical Quality Control in the future.

Exercise 5.5. The light bulbs used in the U.S. space shuttle have an average life expectancy of $\mu = 100$ hours. The distribution of life expectancies is normal with $\sigma = 10$ hours. For a trip that is expected to last 380 hours, the crew takes along four new bulbs (one in place and three spares). If each bulb is replaced immediately when it burns out, what is the probability that the four bulbs will be sufficient for the entire 380-hour trip?

Solution. $X_i \sim N(100, 10^2),\ \mu = 100,\ \sigma = 10.$
By the Central Limit Theorem, approximately (please refer to lecture 5 and tutorial 6)

$$\begin{aligned} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} &\sim N(0,1) \\ \Rightarrow \frac{\sum_{i=1}^4 X_i/4 - \mu}{\sigma/\sqrt{4}} &\sim N(0,1) \\ \Rightarrow \frac{\sum_{i=1}^4 X_i - 4\mu}{2\sigma} &\sim N(0,1) \end{aligned}$$

Thus,

$$P(\sum_{i=1}^4 X_i > 380) = P(\frac{\sum_{i=1}^4 X_i - 4 \cdot 100}{2 \cdot 10} > \frac{380 - 4 \cdot 100}{2 \cdot 10}) = P(Z > -1) = 1 - 0.1587 = 0.8413.$$