Statistical Computing
Week 9: 29. November 2019
Central Limit Theorem

Contents

- Random variables, random sample and independence
- The normal distribution
- Central Limit Theorem
- Workshop: Simulation and Central Limit Theorem

Random variables

- The lecturer's height is 1.92m. This is not a random value, it is fixed.
- The height of a person who will be chosen in the future is unknown and random.
- The height of a general arbitrary person is random.

In the second and third cases above we consider height as a **random variable** X.

The lecturer rolled a die. The result was a 4. This result is not random.
If we roll a die now it will show a random number. The number shown on a future die roll can be considered as a random variable D.

D is a discrete random variable.
X is a continuous random variable.

A random variable must be numeric.

Random Sample and Independence

An important statistical property of two or more random variables is dependence/independence.

Two **discrete** random Variables X_1 and X_2 are independent iff (if and only if)

$$P((X_1 = k_1) \cap (X_2 = k_2)) = P(X_1 = k_1) \cdot P(X_2 = k_2)$$

$P(X = k)$ considered as a function of k forms the **probability mass function** (pmf)

$$f_X(k) = P(X = k).$$

The independence property expressed in terms of the pmfs is

$$f_{X_1, X_2}(k_1, k_2) = f_{X_1}(k_1) f_{X_2}(k_2).$$

Suppose that tomorrow, we will sample a population. We will obtain a random sample of n observations $X_1, X_2, \ldots, X_n$.

For a random sample, the independence property is

$$f_{X_1, X_2, \ldots, X_n}(k_1, k_2, \ldots k_n) = f_{X_1}(k_1) f_{X_2}(k_2) \cdots f_{X_n}(k_n).$$

An equivalent definition of independence is using conditional probability.

Two **discrete** random Variables X_1 and X_2 are independent iff (if and only if)

$$P(X_2 = k_2 | X_1 = k_1) = P(X_2 = k_2)$$

This means that information about X_1 does not change the probabilities of X_2.

A **continuous random variable** X is characterised by its **probability density function** (pdf) $f_X(x)$.

By definition the density is the derivative of the **cumulative distribution function** (cdf) $F_X(x)$:

$$F_X(x) = P(X \leqslant x) \qquad\qquad f_X(x) = F_X'(x)$$

Two continuous random Variables X_1 and X_2 are independent iff

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2).$$

For a random sample of n discrete random variables $X_1, X_2, \ldots, X_n$, the independence property is

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

Effectively the same formula as in the discrete case.

Independent and identically distributed (iid)

Suppose that tomorrow, we will sample a population.
We will obtain a random sample of observations $X_1, X_2, \ldots, X_n$.

A practical assumption is that each random variable X_i has the same distribution:

$$f_{X_i}(x) = f_X(x)$$

Another practical assumption is that all the random variables are independent.

Putting the two together we have that $X_1, X_2, \ldots, X_n$ are **independent and identically distributed** (iid) with:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_X(x_1) f_X(x_2) \cdots f_X(x_n) = \prod_{i=1}^{n} f_X(x_i)$$

In statistics it is very common to assume that our sampled data are iid.
This is a very strong assumption but it is usually acceptable.

Expectation of a random variable

The mean value that a random variable X takes is called the **expected value** or expectation.

The definition of the expected value for X is

Discrete: $$\mathrm{E}(X) = \sum_k k f(k)$$

Continuous: $$\mathrm{E}(X) = \int_{\mathbb{R}} x f(x) dx$$

Expectation and mean: linear transformation

If X is a random variable with a_0 and a_1 constant then for $Y = a_0 + a_1 X$

$$E(Y) = a_0 + a_1 E(X)$$

This corresponds to the mean of a previously obtained sample $\overline{x}$. If $y_i = a_0 + a_1 x_i$, then

$$\overline{y} = a_0 + a_1 \overline{x}.$$

For n random variables $X_1, X_2, \ldots, X_n$ and constants $a_0, a_1 \ldots a_n$, then

$$E(a_0 + a_1 X_1 + \ldots a_n X_n) = a_0 + a_1 E(X_1) + \ldots + a_n E(X_n)\,.$$

When $X_1, X_2, \ldots, X_n$ are **iid** random variables:
Each random variable has the same Expectation (identically distributed).

$$E(X_i) = E(X) = \mu$$

The mean of a sample to be collected in the future is given by $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

Because $\overline{X}$ is a function of random variables it is itself a random variable with an expectation

$$\begin{aligned}
\mathrm{E}(\overline{X}) &= \mathrm{E}\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right) \\
&= \frac{1}{n}\,\mathrm{E}(X_1) + \frac{1}{n}\,\mathrm{E}(X_2) + \cdots + \frac{1}{n}\,\mathrm{E}(X_n) \\
&= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu \\
&= \mu
\end{aligned}$$

$$\boxed{\mathrm{E}(\overline{X}) = \mu}$$

Variance of a random variable

A measure of spread for a random variable X is its **variance**.

The definition of the variance of X is

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

Discrete:
$$\mathrm{Var}(X) = \sum_k k^2 f(k) - \left(\sum_k k f(k)\right)^2$$

Continuous:
$$\mathrm{Var}(X) = \int_{\mathbb{R}} x^2 f(x)dx - \left(\int_{\mathbb{R}} x f(x)dx\right)^2$$

Variance of a random sample

When $X_1, X_2, \ldots, X_n$ are **independent** random variables and $a_0, a_1 \ldots a_n$ are constants, then

$$\mathrm{Var}(a_0 + a_1 X_1 + \ldots a_n X_n) = a_1^2 \,\mathrm{Var}(X_1) + \ldots + a_n^2 \,\mathrm{Var}(X_n)\,.$$

When $X_1, X_2, \ldots, X_n$ are **iid** random variables:
Each random variable has the same variance (identically distributed).

$$\mathrm{Var}(X_i) = \mathrm{Var}(X) = \sigma^2$$

Because $\overline{X}$ is a function of random variables it is itself a random variable with variance.

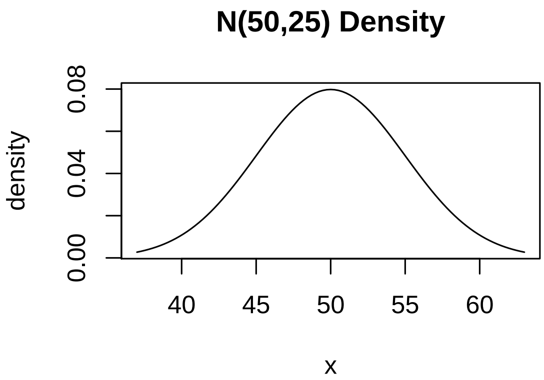
$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right) \\
&= \frac{1}{n^2}\mathrm{Var}(X_1) + \frac{1}{n^2}\mathrm{Var}(X_2) + \cdots + \frac{1}{n^2}\mathrm{Var}(X_n) \\
&= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

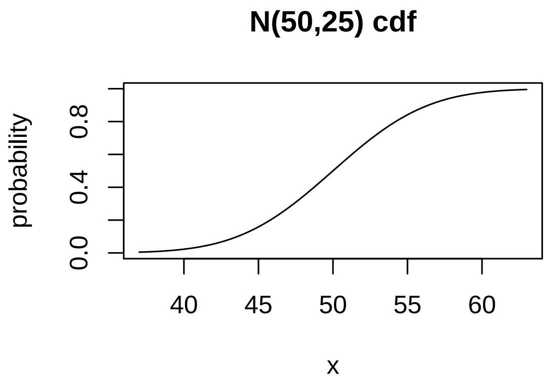
$$\boxed{\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}}$$

The normal distribution

The normal distribution is a continuous distribution with a bell shaped density function. The distribution has two parameters, μ and σ^2.

Let $X \sim N(\mu, \sigma^2)$, then $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.



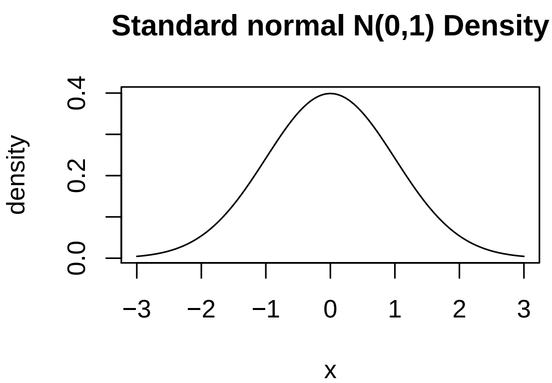


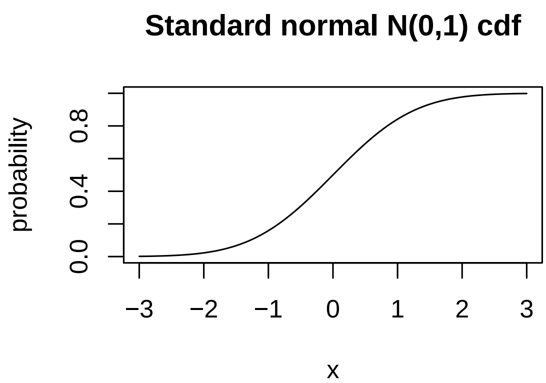
The density function of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}$$

The standard normal distribution

Z has a **standard normal distribution**, when $Z \sim N(0,1)$
i.e. has expectation 0 and variance 1





The density function of Z is

$$f_Z(x) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad x \in \mathbb{R}$$

Let $X \sim N(\mu, \sigma^2)$. For constants a_0 and a_1

$$Y = a_0 + a_1 X \qquad \Leftrightarrow \qquad Y \sim N(\mu + a_0, a_1^2 \sigma^2)$$

Choosing $a_0 = \dfrac{-\mu}{\sigma}$ and $a_1 = \dfrac{1}{\sigma}$, gives

$$Z = a_0 + a_1 X = \frac{X - \mu}{\sigma} \qquad \Leftrightarrow \qquad Z \sim N(0, 1)$$

Theorem: We can transform any normal random variable into a standard normal random variable by subtracting the expectation and dividing by the mean.

When $\phi(z)$ is the density function and $\Phi(z)$ the cdf of of Z then

$$f_X(x) = \phi\left(\frac{x - \mu}{\sigma}\right) \qquad\qquad P(X \leqslant x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Sample mean of a normal distribution

Theorem
Let $X_1, X_2, \ldots, X_n$ be $N(\mu, \sigma^2)$ **iid** random variables, then $\overline{X}$ is also normally distributed with mean μ and variance σ^2.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The implication of this is that, if our sample *originates from a normal distribution*, then the sample mean is normally distributed around the same expectation. The larger the sample size the smaller the variance.

The Central Limit Theorem goes even further, giving a similar result regardless of the distribution.

Central Limit Theorem

Theorem

Let $X_1, X_2, \ldots, X_n$ be **iid** random variables with an arbitrary distribution, then $\overline{X}$ tends to a normal distribution with mean μ and variance σ^2 as $n \to \infty$.

Formally

$$\lim_{n\to\infty} P\left(\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \leqslant x\right) = \Phi(x).$$

For large n, the distribution of the random variable $Z = \dfrac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$ is well approximated by the standard normal $N(0,1)$ distribution.

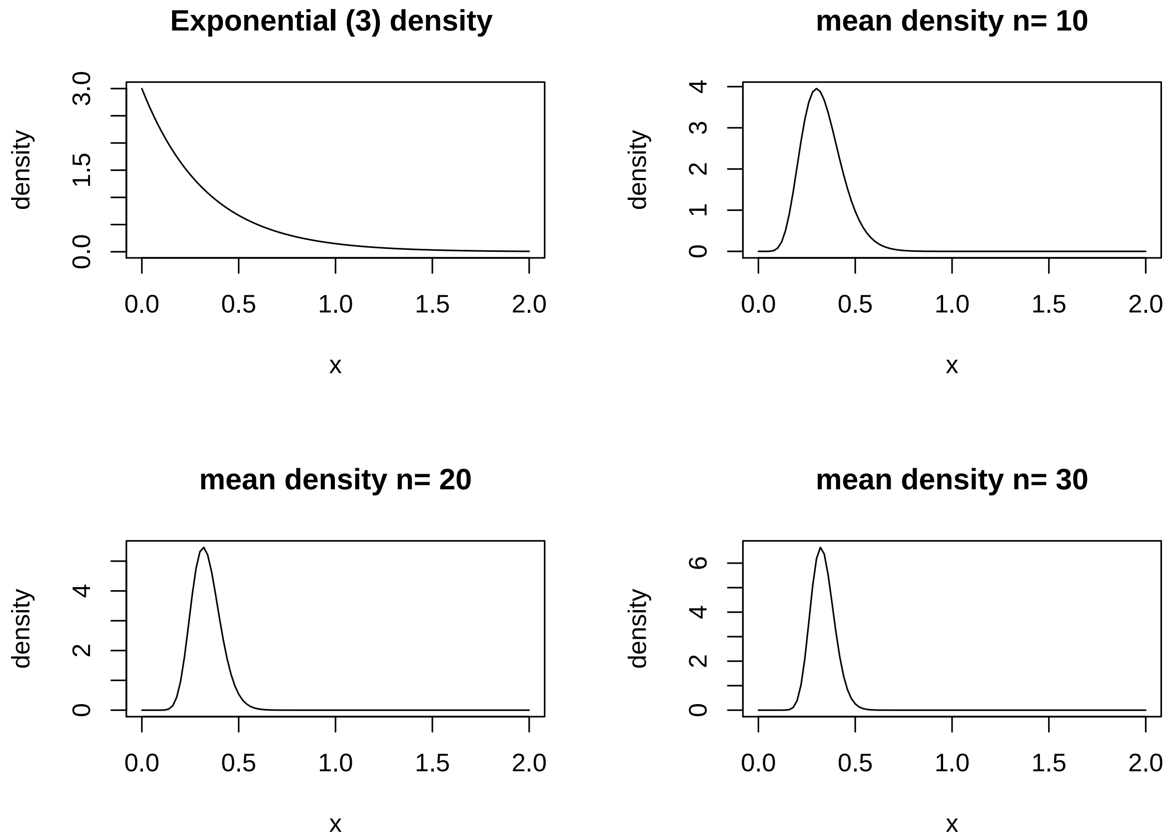
$$\Rightarrow \qquad \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \overset{a}{\sim} N(0,1) \qquad \text{or} \qquad \overline{X} \overset{a}{\sim} N(\mu, \sigma^2/n)$$

In practice $n > 30$ ist often considered “large enough” to use the central limit theorem.

The implication of the central limit theorem is very powerful. It means that the mean of a iid random sample is approximately normally distributed, for any population distribution, **including when the distribution is unknown**.

We can use the central limit theorem to get good approximations to probabilities relating to $\overline{X}$.

Example: $X \sim \text{Exp}(3)$, $E(X) = \frac{1}{3}$



Workshop: Simulation and the central limit theorem

We can **simulate** a random sample using a random number generator.

In the workshop you will learn some general principles of simulation, and how this is implemented in R. In addition you will learn about 3 other R functions used to obtain numerical values from a given probability distributions.

In the second section you will use simulation to demonstrate the central limit theorem.

Sections 3 and 4 consists of further reading and theoretical exercises (homework).