Quantitative Methods

Human Sciences, 2020–21

Elias Nosrati

Lecture 4: 5 November 2020

Today

Today

- Recap on random variables and probability distributions

Today

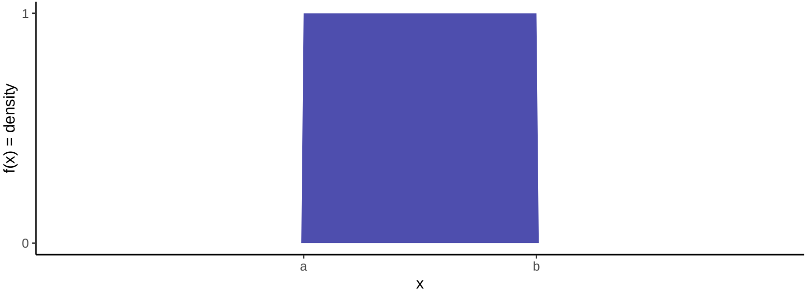
- Recap on random variables and probability distributions
- Key features of probability distributions

Today

- Recap on random variables and probability distributions
- Key features of probability distributions
- Problem sheet 3 (tutorial)

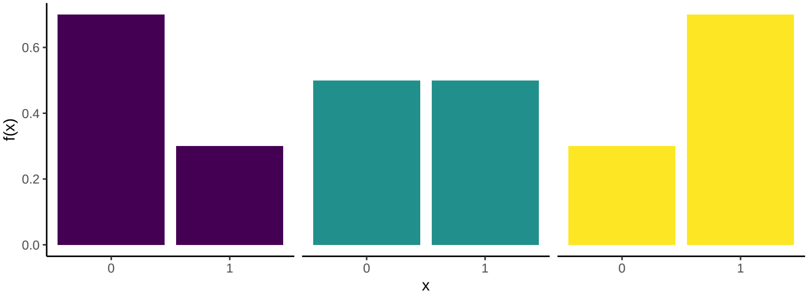
Uniform distribution

Uniform distribution



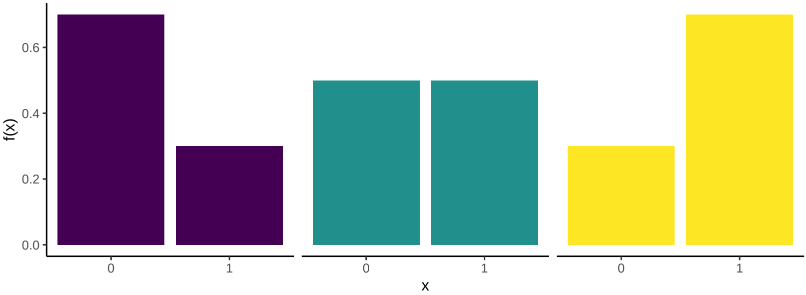
A random variable X has a *Uniform distribution* on the interval (a, b) if it takes every value within the interval (a, b) with equal likelihood and any value outside this interval with zero likelihood.

Bernoulli distribution



A random variable X has a *Bernoulli distribution* if it takes only two distinct and mutually exclusive values.

Bernoulli distribution



A random variable X has a *Bernoulli distribution* if it takes only two distinct and mutually exclusive values.

Example

X is the number of Tails obtained when flipping a coin.

Binomial distribution

A random variable X has a *Binomial distribution* if it is the sum of n independent and identically distributed Bernoulli random variables.

Binomial distribution

A random variable X has a *Binomial distribution* if it is the sum of n independent and identically distributed Bernoulli random variables.

Example

X is the number of Tails obtained when flipping a coin n times.

Poisson distribution

A random variable X follows a *Poisson distribution* if it is a count of the number of events occurring in a fixed interval of time or space and if these events occur with a known constant mean rate and independently of the time since the last event.

Poisson distribution

A random variable X follows a *Poisson distribution* if it is a count of the number of events occurring in a fixed interval of time or space and if these events occur with a known constant mean rate and independently of the time since the last event.

Example

X is the number of global pandemics occurring every decade, or the number of patients arriving in an emergency room between 10 and 11pm, or the number of meteorites striking Earth every 100 years.

Poisson distribution

A random variable X follows a *Poisson distribution* if it is a count of the number of events occurring in a fixed interval of time or space and if these events occur with a known constant mean rate and independently of the time since the last event.

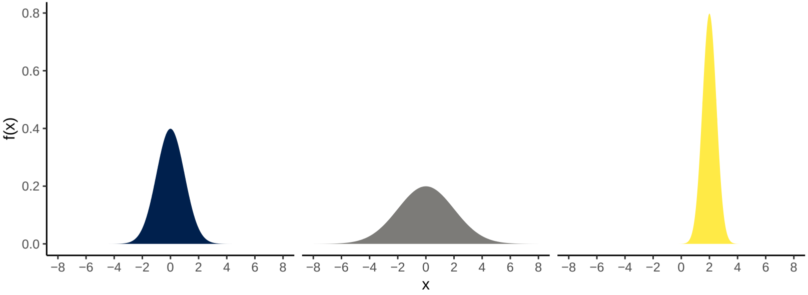
Example

X is the number of global pandemics occurring every decade, or the number of patients arriving in an emergency room between 10 and 11pm, or the number of meteorites striking Earth every 100 years.

Remark

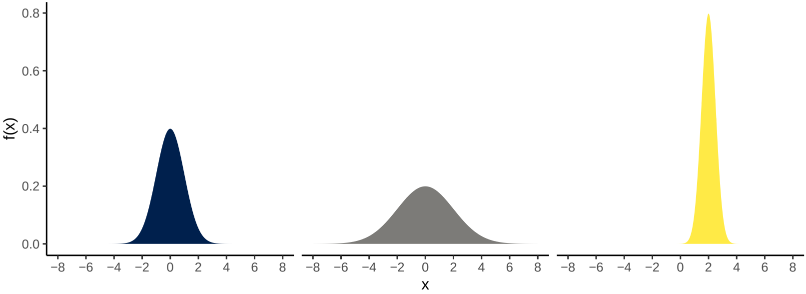
Independence assumption: the number of global pandemics per decade may not follow a Poisson distribution if the first wave of the pandemic increases the probability of follow-up waves of similar magnitude.

Normal distribution



A continuous random variable X has a *Normal distribution* if its probability density function is shaped like a bell curve.

Normal distribution



A continuous random variable X has a *Normal distribution* if its probability density function is shaped like a bell curve.

What are the differences between these three Normal distributions?

Expected value

Expected value

Definition

The *expected value* (or *population mean*) of a random variable X is a weighted average of the possible values that X can take, weighted by their probability densities:

$$\mathbb{E}(X) = \begin{cases} \sum x\, f(x) & \text{if } X \text{ is discrete,} \\ \int x\, f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Expected value

Definition

The *expected value* (or *population mean*) of a random variable X is a weighted average of the possible values that X can take, weighted by their probability densities:

$$\mathbb{E}(X) = \begin{cases} \sum x\, f(x) & \text{if } X \text{ is discrete,} \\ \int x\, f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Example

Let X be the result of rolling a fair die.

Expected value

Definition

The *expected value* (or *population mean*) of a random variable X is a weighted average of the possible values that X can take, weighted by their probability densities:

$$\mathbb{E}(X) = \begin{cases} \sum x\, f(x) & \text{if } X \text{ is discrete,} \\ \int x\, f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Example

Let X be the result of rolling a fair die. Then

$$\mathbb{E}(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5.$$

Expected value

Definition

The *expected value* (or *population mean*) of a random variable X is a weighted average of the possible values that X can take, weighted by their probability densities:

$$\mathbb{E}(X) = \begin{cases} \sum x\, f(x) & \text{if } X \text{ is discrete,} \\ \int x\, f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

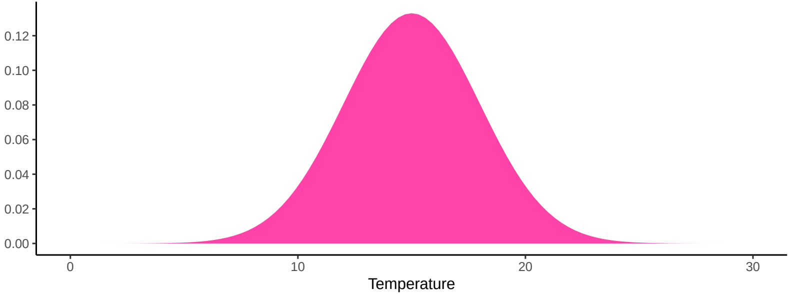
Example

Let X be the result of rolling a fair die. Then

$$\mathbb{E}(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5.$$

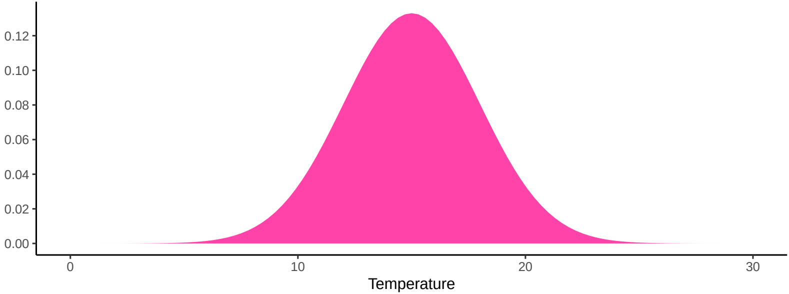
Note that, in this example, X never equals its expected value (you can never roll 3.5!).

Expected value: continuous case



What to do when X takes on an infinite number of possible values?

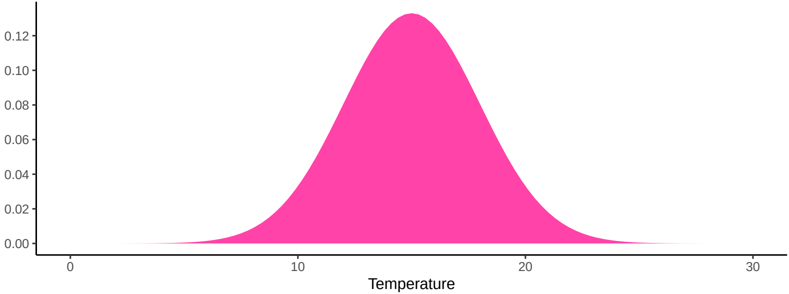
Expected value: continuous case



What to do when X takes on an infinite number of possible values?

- Solve analytically.

Expected value: continuous case



What to do when X takes on an infinite number of possible values?

- Solve analytically.
- Use the Law of Large Numbers: as the sample size increases, the sample mean converges to the population mean or expected value.

Law of Large Numbers: motivating example

Law of Large Numbers: motivating example

- Suppose we sample n days and measure the temperature.

Law of Large Numbers: motivating example

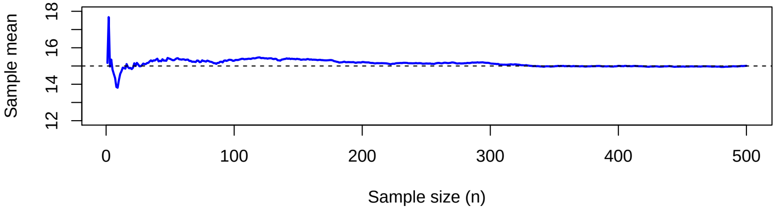
- Suppose we sample n days and measure the temperature.
- Each observation is drawn from a Normal distribution whose true but unknown mean is 15.

Law of Large Numbers: motivating example

- Suppose we sample n days and measure the temperature.
- Each observation is drawn from a Normal distribution whose true but unknown mean is 15.
- Can we use the average temperature in our sample to approximate the population mean?

Law of Large Numbers: motivating example

- Suppose we sample n days and measure the temperature.
- Each observation is drawn from a Normal distribution whose true but unknown mean is 15.
- Can we use the average temperature in our sample to approximate the population mean?



Law of Large Numbers

Law of Large Numbers

- Suppose we obtain a random sample of n independently and identically distributed observations $X_1, X_2, \ldots, X_n$ from a probability distribution with expectation $\mathbb{E}(X)$.

Law of Large Numbers

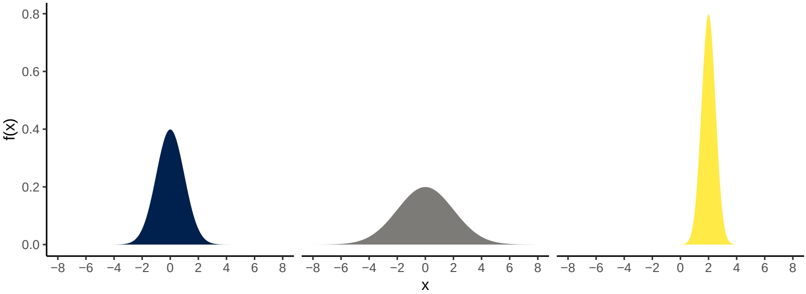
- Suppose we obtain a random sample of n independently and identically distributed observations $X_1, X_2, \ldots, X_n$ from a probability distribution with expectation $\mathbb{E}(X)$.
- *The Law of Large Numbers* states that as n becomes large, the sample average of these n random variables will approach $\mathbb{E}(X)$:

Law of Large Numbers

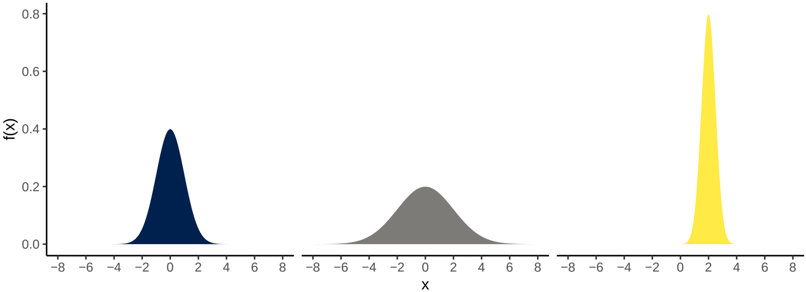
- Suppose we obtain a random sample of n independently and identically distributed observations $X_1, X_2, \ldots, X_n$ from a probability distribution with expectation $\mathbb{E}(X)$.
- *The Law of Large Numbers* states that as n becomes large, the sample average of these n random variables will approach $\mathbb{E}(X)$:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \rightsquigarrow \mathbb{E}(X).$$

Normal distribution

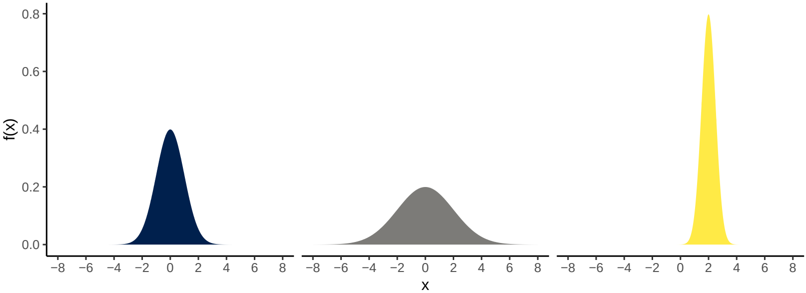


Normal distribution



- Expectation as a measure of central tendency.

Normal distribution



- Expectation as a measure of central tendency.
- Variance as a measure of variability or spread.

Variance and standard deviation

Definition

The *variance* of a random variable X is defined as

$$\mathbb{V}(X) = \mathbb{E}[X - \mathbb{E}(X)]^2$$

Variance and standard deviation

Definition

The *variance* of a random variable X is defined as

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}[X - \mathbb{E}(X)]^2 \\ &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.\end{aligned}$$

Variance and standard deviation

Definition

The *variance* of a random variable X is defined as

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}[X - \mathbb{E}(X)]^2 \\ &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2. \end{aligned}$$

The square root of $\mathbb{V}(X)$ is known as the *standard deviation* of X.

Expectation and variance in **R**

```
X <- rnorm(1000, 15, 3)
```

Expectation and variance in R

```
X <- rnorm(1000, 15, 3)
```

```
head(X)

## [1] 15.17 20.19  9.55 16.49 12.61 13.37
```

Expectation and variance in **R**

```r
X <- rnorm(1000, 15, 3)
```

```r
head(X)

## [1] 15.17 20.19  9.55 16.49 12.61 13.37
```

```r
mean(X)

## [1] 14.94125
```

Expectation and variance in **R**

```r
X <- rnorm(1000, 15, 3)
```

```r
head(X)

## [1] 15.17 20.19  9.55 16.49 12.61 13.37
```

```r
mean(X)

## [1] 14.94125
```

```r
var(X)

## [1] 9.095238
```

Expectation and variance in **R**

```r
X <- rnorm(1000, 15, 3)
```

```r
head(X)
```

```
## [1] 15.17 20.19  9.55 16.49 12.61 13.37
```

```r
mean(X)
```

```
## [1] 14.94125
```

```r
var(X)
```

```
## [1] 9.095238
```

```r
sd(X)
```

```
## [1] 3.015831
```

The Central Limit Theorem

The Central Limit Theorem

- Suppose we obtain a random sample of n independently and identically distributed random variables $X_1, X_2, \ldots, X_n$ from any probability distribution.

The Central Limit Theorem

- Suppose we obtain a random sample of n independently and identically distributed random variables $X_1, X_2, \ldots, X_n$ from any probability distribution.
- *The Central Limit Theorem* states that, as n becomes large, the sum of these random variables will follow a Normal distribution.

The Central Limit Theorem

- Suppose we obtain a random sample of n independently and identically distributed random variables $X_1, X_2, \ldots, X_n$ from any probability distribution.
- *The Central Limit Theorem* states that, as n becomes large, the sum of these random variables will follow a Normal distribution.
- Equivalently, suppose we gather a random sample of observations $X_1, \ldots, X_n$ and calculate the sample mean $\overline{X} = \frac{1}{n} \sum_i X_i$.

The Central Limit Theorem

- Suppose we obtain a random sample of n independently and identically distributed random variables $X_1, X_2, \ldots, X_n$ from any probability distribution.
- *The Central Limit Theorem* states that, as n becomes large, the sum of these random variables will follow a Normal distribution.
- Equivalently, suppose we gather a random sample of observations $X_1, \ldots, X_n$ and calculate the sample mean $\overline{X} = \frac{1}{n} \sum_i X_i$.
- If this procedure is performed many times, the Central Limit Theorem states that the probability distribution of $\overline{X}$ will be Normally distributed.

The Central Limit Theorem in **R**

In a random sample of 600 people, what is the probability that over half of them intend to vote given that the population mean is 0.5?

The Central Limit Theorem in **R**

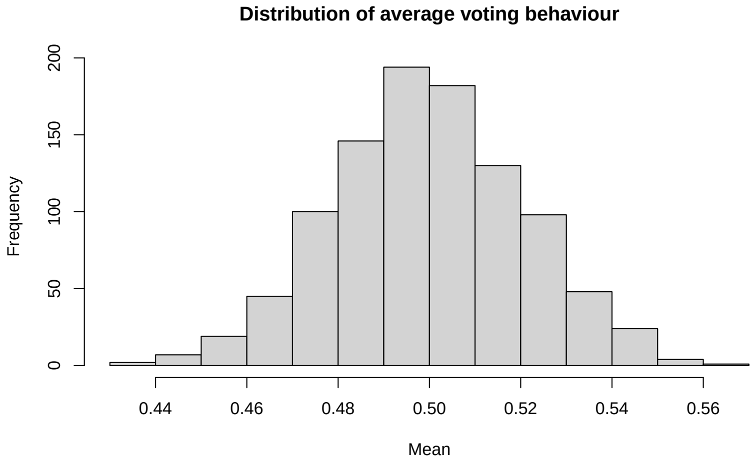
In a random sample of 600 people, what is the probability that over half of them intend to vote given that the population mean is 0.5?

```r
# To be filled
prop_vote <- rep(NA, 1000)

# Repeat experiment 1000 times
for (i in 1:1000) {
  # Sample 600 people and ask about voting
  # Yes or no: Bernoulli random variable
  sample <- rbinom(600, 1, 0.5)
  # Save proportion of voters
  prop_vote[i] <- mean(sample)
}
```

The Central Limit Theorem in **R** (cont.)

```
hist(prop_vote,
     main = "Distribution of average voting behaviour",
     xlab = "Mean")
```



Measures of association

Measures of association

- Just as the mean and the variance provide information about single distributions, the covariance between two random variables provides information about joint distributions.

Measures of association

- Just as the mean and the variance provide information about single distributions, the covariance between two random variables provides information about joint distributions.
- Roughly speaking, covariance measures a tendency of two random variables to go up or down together.

Measures of association

- Just as the mean and the variance provide information about single distributions, the covariance between two random variables provides information about joint distributions.
- Roughly speaking, covariance measures a tendency of two random variables to go up or down together.

Definition

The *covariance* between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Measures of association

- Just as the mean and the variance provide information about single distributions, the covariance between two random variables provides information about joint distributions.
- Roughly speaking, covariance measures a tendency of two random variables to go up or down together.

Definition

The *covariance* between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Note that $\mathrm{Cov}(X, X) = \mathbb{V}(X)$ and that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

Measures of association (cont.)

Two random variables that have zero covariance are *uncorrelated*:

Measures of association (cont.)

Two random variables that have zero covariance are *uncorrelated*:

Definition

The *correlation* between two random variables X and Y is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

Correlation and dependence