HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 11 — Basic probability distributions

Jorge N. Tendeiro

Hiroshima University

Random variables (review)

A random variable is a function that allows converting any type of event into numbers.

Tossing a die:

$$X = \begin{cases} 1, \text{ if face 1 lands up} \\ 2, \text{ if face 2 lands up} \\ 3, \text{ if face 3 lands up} \\ 4, \text{ if face 4 lands up} \\ 5, \text{ if face 5 lands up} \\ 6, \text{ if face 6 lands up} \end{cases}$$

This is an example of a discrete random variable.

Height in *cm*:

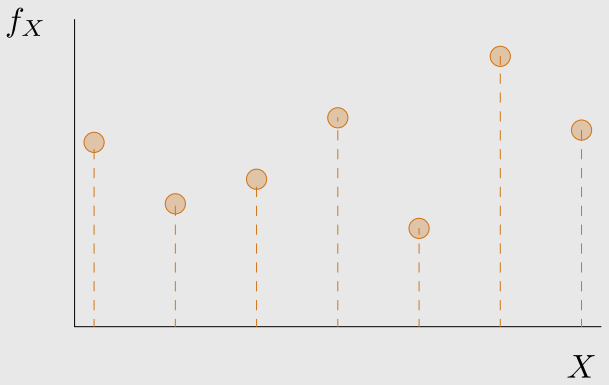
$$Y \geq 0$$

This is an example of a continuous random variable.

Probability distributions (review)

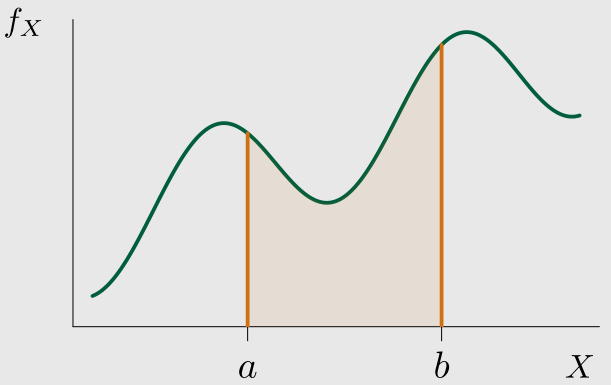
Probability distribution: Correspondence between the values of a random variable and their probabilities.

Discrete random variable
The probability distribution is known as the probability mass function.



$$P(X = x_k) = f_X(x_k)$$

Continuous random variable
The probability distribution is known as the probability density function.



$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

Probability distributions are tools to understand phenomena.
Probabilities can be computed from the probability mass (or density) function.

Today

Learn how to use common probability distributions to calculate the probability of various phenomena.
For example:

- What is the probability that 10 people resign out of 100 workers?

- What is the probability of making 5 typos within one page when writing a report?

- What is the probability of a student scoring at most 50 in an exam?

We will learn how to do computations with the following probability distributions using Excel:

- Binomial distribution
- Poisson distribution
- Normal distribution

Binomial distribution (discrete distribution)

Binomial experiment

A binomial experiment is an experiment verifying the following four conditions:

1. There is a fixed number of repeated trials, say n.

2. Each trial has only two possible outcomes (yes/no, heads/tails, live/die).
   We usually refer to one of the outcomes as a *success* (e.g., yes, heads, live) and the other as a *failure* (no, tails, die).

3. Trials are independent from one another.

4. The probability of success, say p, is constant across trials.

Binomial distribution

The binomial distribution is used to model $X =$ *number of successes* in a binomial experiment:

$$X \sim B(n, p),$$

where

- $n =$ number of trials
- $p =$ probability of success in each trial.

Binomial distribution — Examples

> *Throw a fair coin 5 times. Count the number of heads.*

Here,

- Trial: Throw a fair coin.
- Possible outcomes: Heads or tails.
- Success: Heads.
- Independence between throws is assumed.
- $n = 5, p = 0.5$.

Denoting $X =$ *number of heads* we have that

$$X \sim B(n = 5, p = 0.5).$$

Other examples:

- The number of people X clicking an advertisement out of n people, when an advertisement of a webpage is clicked with probability p for each person.
- The number of people X that cancel a contract out of n users, when each person cancels a contract with probability p.

Binomial distribution — Probability mass function

If $X \sim B(n, p)$ then the probability mass function, known as the binomial distribution, is given by

$$\begin{aligned} f_X(k) &= P(X = k) \\ &= {}_nC_k p^k (1-p)^{n-k}, \end{aligned}$$

where:

- ${}_nC_k = \frac{n!}{k!(n-k!)}$ is the so-called binomial coefficient
- $n!$ (read: 'n factorial') is equal to $n(n-1)(n-2) \times \cdots \times 3 \times 2 \times 1$
- $k = 0, 1, \ldots, n$.

Binomial distribution — Properties

$$\boxed{X \sim B(n, p)}$$

Properties of the binomial distribution:

- $\mathbb{E}[X] = np$
- $V[X] = np(1-p)$
- Special case when $n = 1$:
  In this case the binomial distribution is known as the Bernoulli distribution with success rate p:

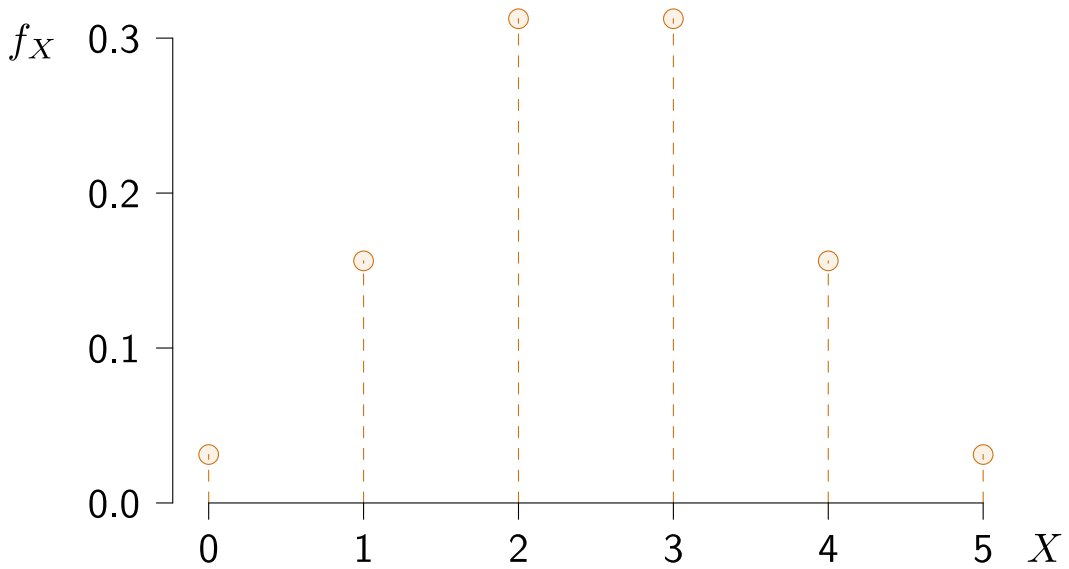
$$X \sim Be(p).$$

Binomial distribution

> *Throw a fair coin 5 times. Count the number of heads.*

In this case,

$$X \sim B(5, 0.5)$$

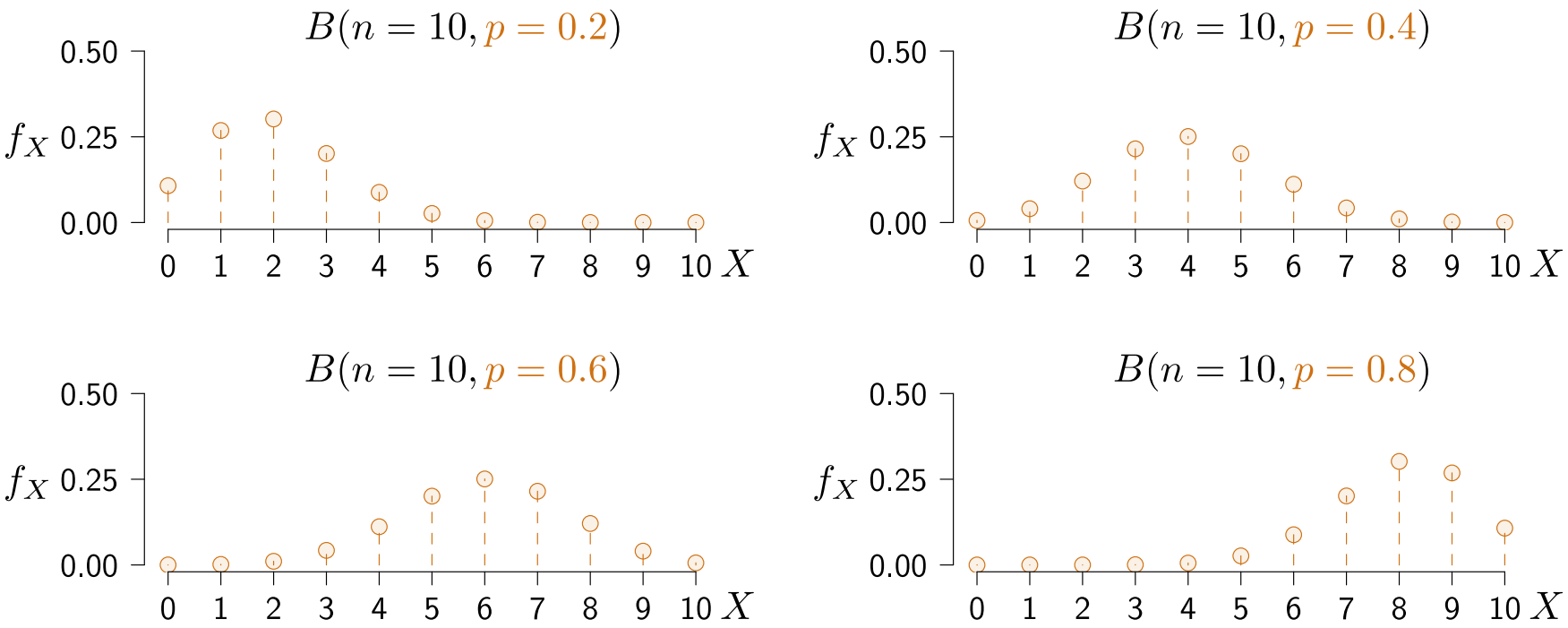


The most likely number of heads is $X = 2$ and $X = 3$.

Binomial distribution — Parameters

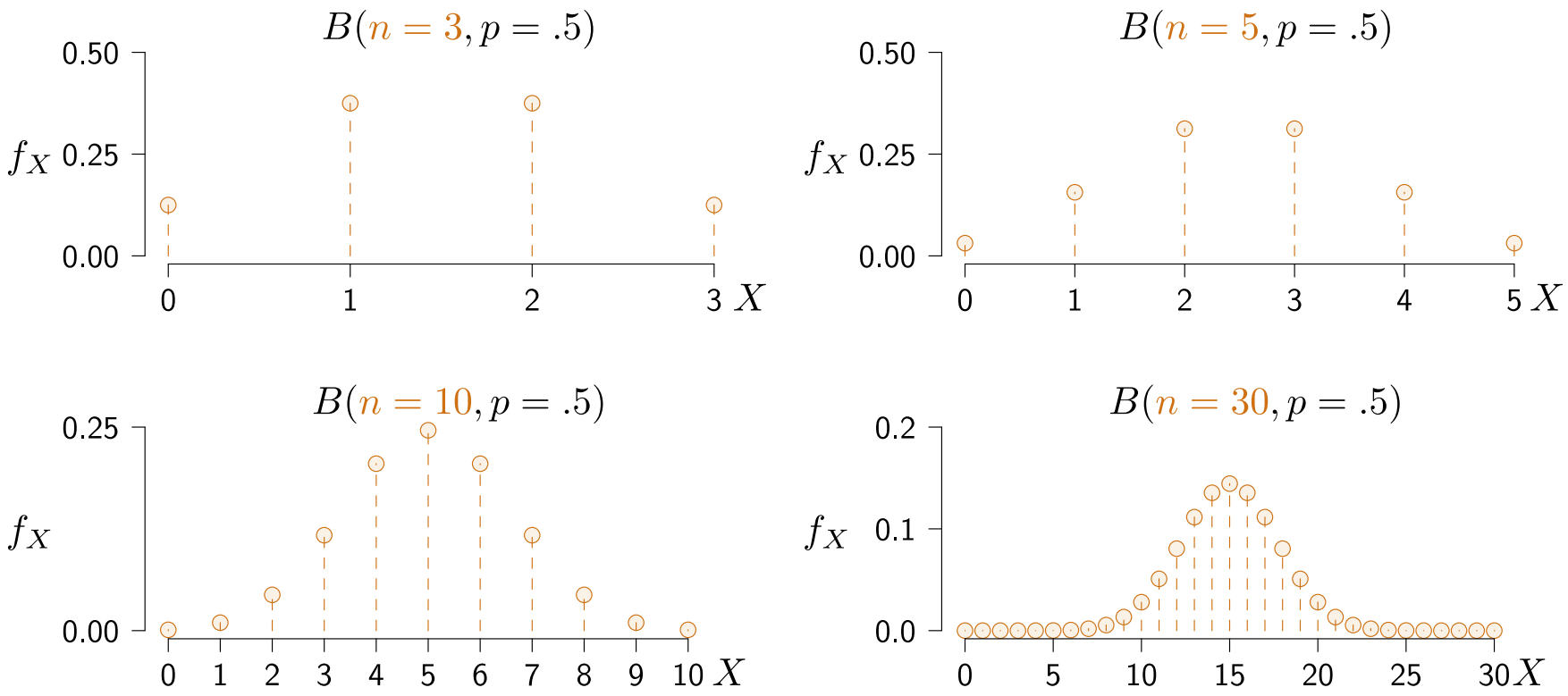
n and p are called the parameters of the binomial distribution $B(n, p)$.
Knowing the values of n and p completely determines the shape of the binomial distribution.

> *Increasing p moves the distribution to the right:*



Binomial distribution — Parameters

> *Increasing n approximates the binomial distribution to the normal distribution:*

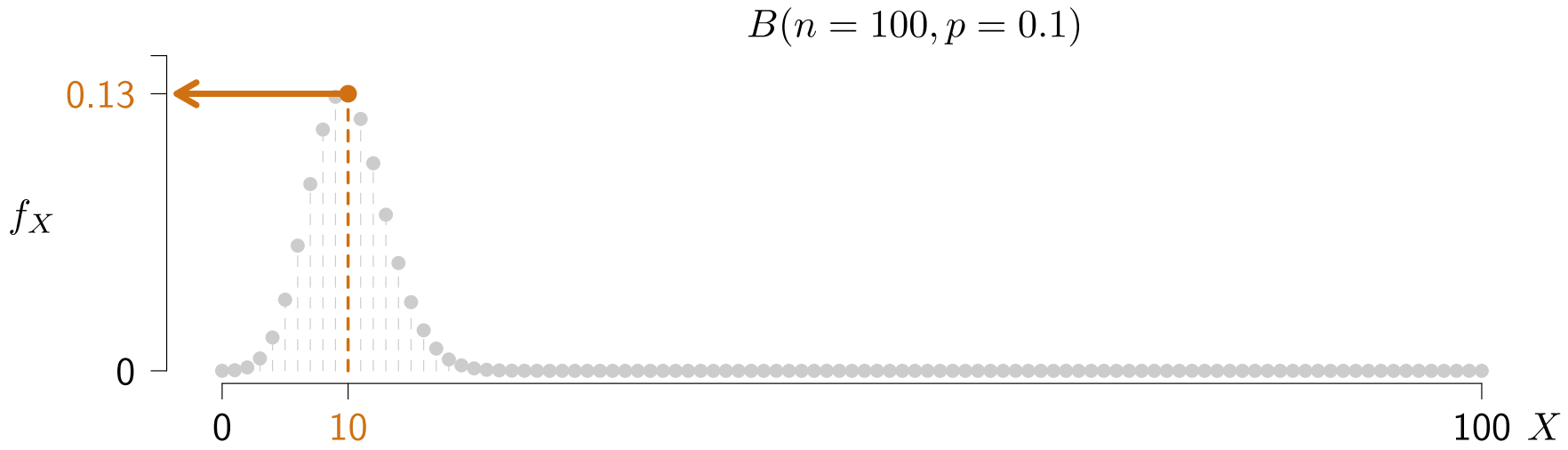


Binomial distribution — Computing probabilities

> *Out of 100 workers, each person resigns with probability 0.1.*
> *In this case, what is the probability that 10 people resign?*

Denote $X =$ *number of people resigning*.
Under the assumptions of a binomial experiment, $X \sim B(100, 0.1)$, from which we can find that.

$$P(X = 10) = f_X(10) = 0.13.$$



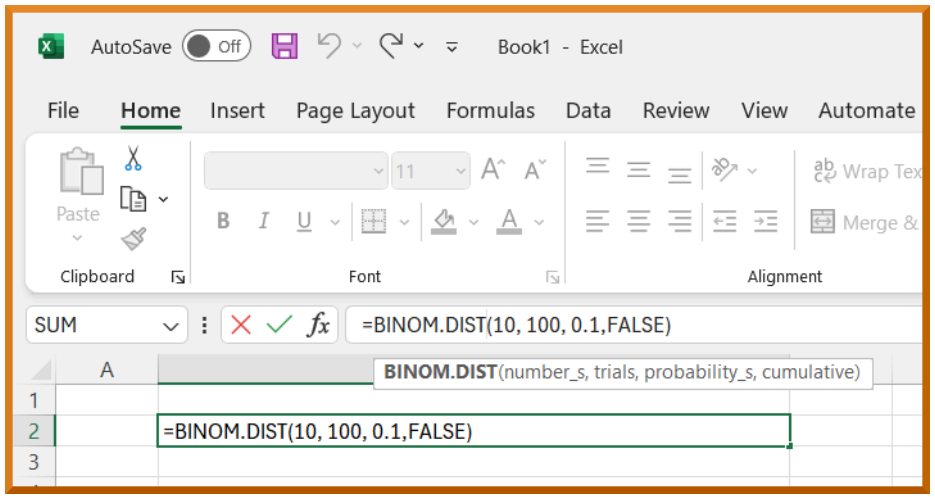
Binomial distribution — Computing probabilities

Given a binomial distribution with parameters n (the number of trials) and p (probability of success in each trial), the probability of observing x successes can be calculated in Excel by the following command:

Input

```
=BINOM.DIST(x, n, p, FALSE)
```

Note:
The command `BINOM.DIST(x, n, p, TRUE)` would compute $P(X \leq x)$ instead of $P(X = x)$.



Exercise (1)

The germination rate of a flower seed at 20°C is 80%.
When there are 500 seeds, what is the probability that 400 seeds among them sprout under 20°C?

Hint:
Assume that the number of seeds X which sprout under 20°C is modeled by the following binomial distribution:

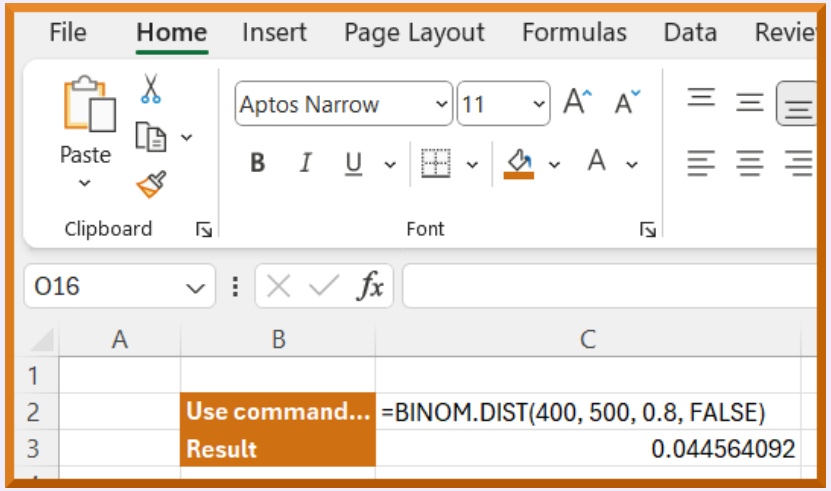
$$X \sim B(500, 0.8).$$

Exercise (1) — ANSWER

The germination rate of a flower seed at 20°C is 80%.
When there are 500 seeds, what is the probability that 400 seeds among them sprout under 20°C?

Hint:
Assume that the number of seeds X which sprout under 20°C is modeled by the following binomial distribution:

$$X \sim B(500, 0.8).$$



Poisson distribution (discrete distribution)

Poisson distribution

The Poisson distribution is used to model the number of rare events occurring in a fixed time or space interval.

Assumptions required to use the Poisson distribution:

- Events occur at a constant mean rate.
  In particular, the mean rate of occurrence is independent of previous occurrences.
- Events occur independently of each other.
  In particular, the time since the last event is not relevant.
- Two events **cannot** occur at exactly the same instant.

Poisson distribution — Examples

Below is the first example in which the Poisson distribution was used:

> *The number of soldiers X who died by being kicked by a horse within 20 years.*

- Time interval: 20 years
- Rare event: Dying by a horse kick.

Other examples:

- The number of defective products produced at a factory within one month.
- The number of fatalities from traffic accidents in one year.
- The number of typos on one page.

Poisson distribution — Probability mass function

Random variable X follows a Poisson distribution with parameter λ ($\lambda > 0$), denoted $X \sim Po(\lambda)$, when the probability mass function, known as the Poisson distribution, is given by

$$\begin{aligned} f_X(k) &= P(X = k) \\ &= \frac{\lambda^k}{k!} e^{-\lambda}, \end{aligned}$$

for $k = 0, 1, 2, \ldots$ and $e = 2.71828\ldots$ (Euler's, or Napier's, number).

Properties of the Poisson distribution:

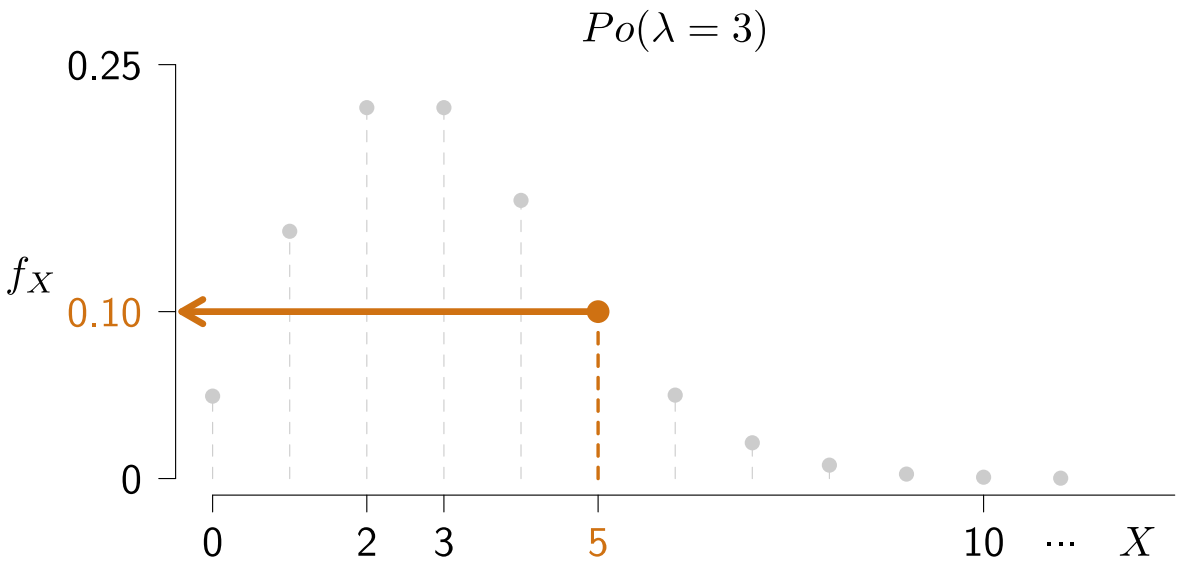
- $\mathbb{E}[X] = \lambda$
- $V[X] = \lambda$
- Interpretation of λ (based on the fact that $\mathbb{E}[X] = \lambda$):

> *Expected number of occurrences in the given fixed time or space interval.*

Poisson distribution — Computing probabilities

> *Suppose that you make 3 typos per page on average when typing. In this case, what is the probability of making 5 typos on a page?*

Let $X =$ *number of typos on one page*.
Assume that the assumptions required to use the Poisson distribution hold.
Then $X \sim Po(3)$ and the required probability is $f_X(5) = P(X = 5)$.



The most likely number of typos is $X = 2$ and $X = 3$.

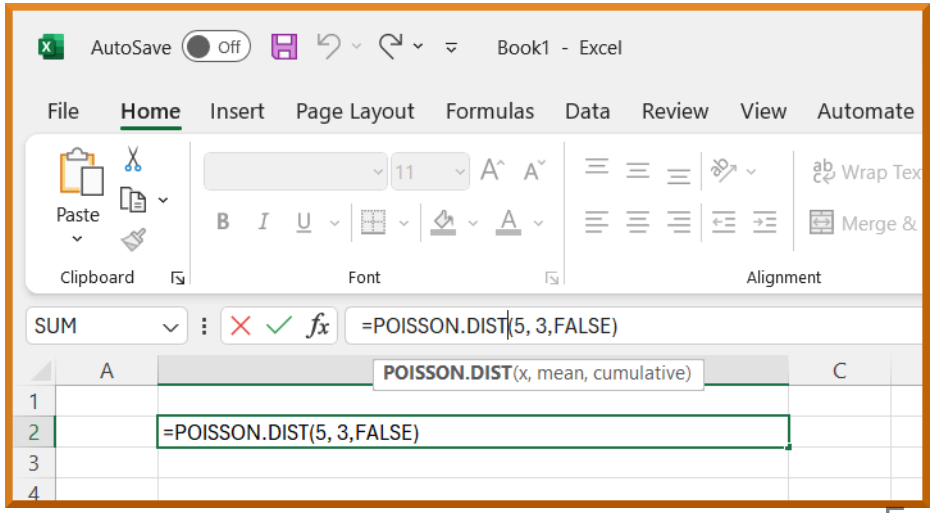
Poisson distribution — Computing probabilities

Given a Poisson distribution with parameter (=mean) λ, the probability of having x occurrences of the rare event can be calculated in Excel by the following command:

Input

```
=POISSON.DIST(x, mean, FALSE)
```

Note:
The command `POISSON.DIST(x, mean, TRUE)` would compute $P(X \leq x)$ instead of $P(X = x)$.



Exercise (2)

A factory produces an average of five defective products per day.
What is the probability that only one defective product will be produced in one day?

Hint:
Assume that the number of defective products per day X is modeled by the following Poisson distribution:

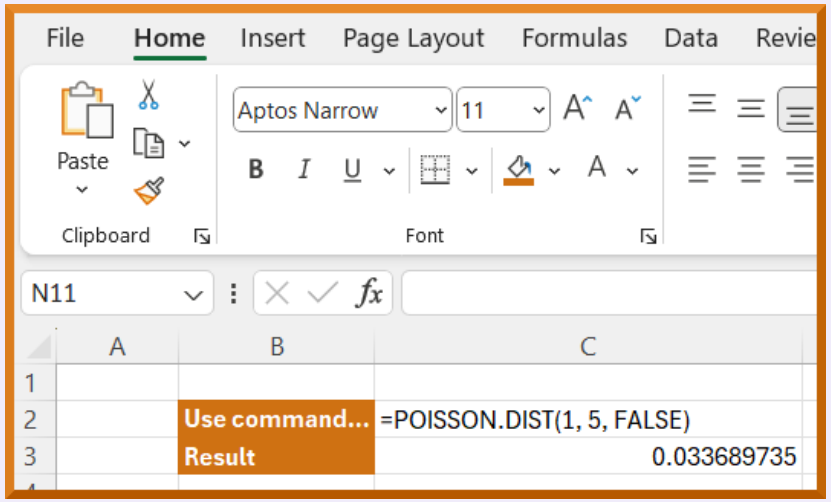
$$X \sim Po(5).$$

Exercise (2) — ANSWER

A factory produces an average of five defective products per day.
What is the probability that only one defective product will be produced in one day?

Hint:
Assume that the number of defective products per day X is modeled by the following Poisson distribution:

$$X \sim Po(5).$$



Binomial distribution converging to the Poisson distribution

Given binomial distributions $X \sim B(n, p)$ such that:

- n is increasingly large
- p is increasingly small
- np remains constant,

then the following holds:

$$f_{X;B(n,p)} \underset{n}{\longrightarrow} f_{X;Po(\lambda=np)}$$

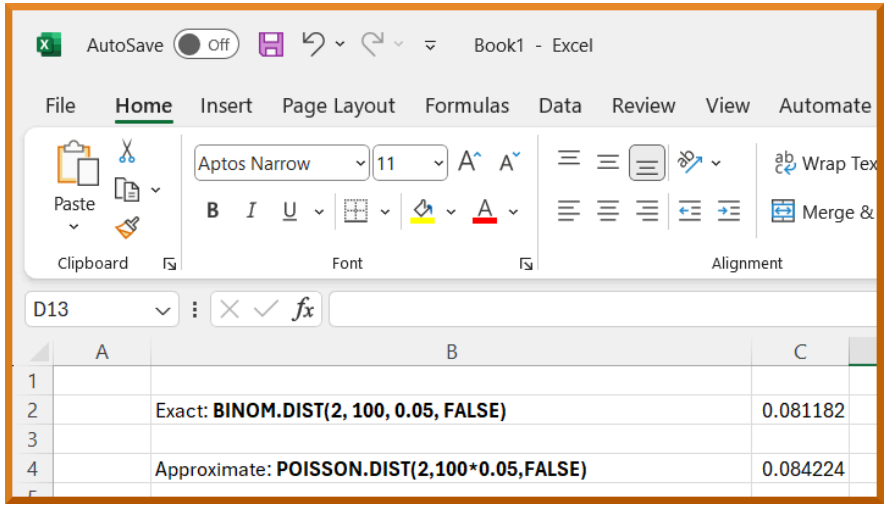
What this means is that we can use the Poisson distribution to approximate probabilities of the binomial distribution.

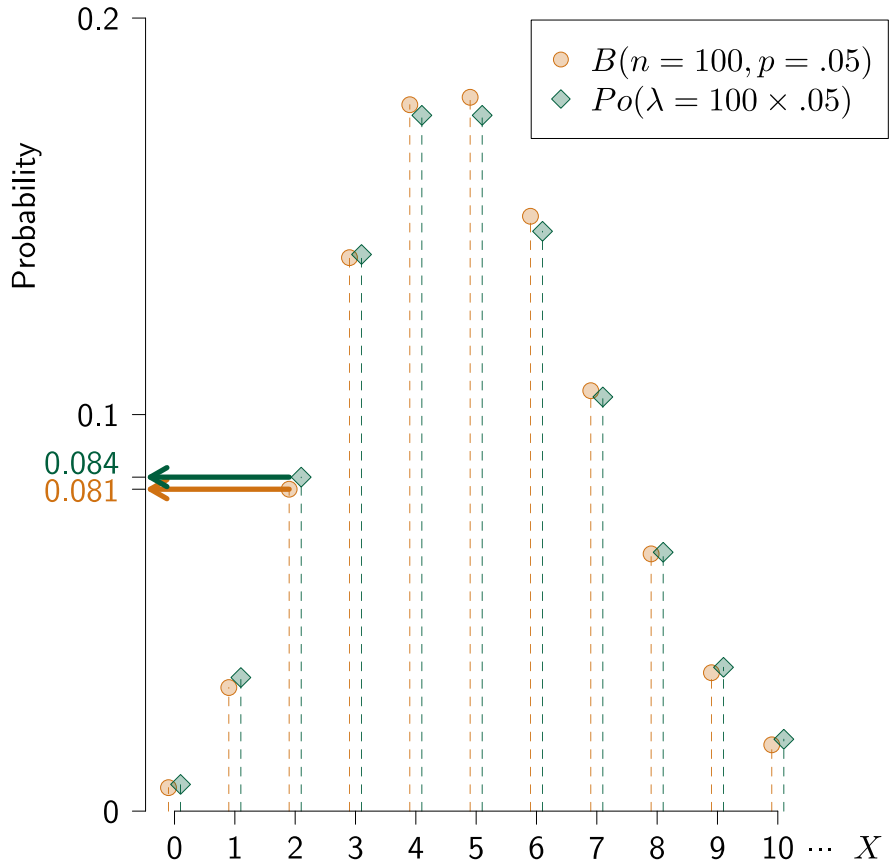
Example:

Let $X \sim B(100, 0.05)$. Compute $P(X = 2)$ in Excel in two ways:

- Exactly (via `BINOM.DIST`).
- Approximately (via `POISSON.DIST`).

Binomial distribution converging to the Poisson distribution





Normal distribution (continuous distribution)

Normal distribution

Random variable X follows a normal distribution with parameters μ (mean) and σ^2 (variance; $\sigma^2 > 0$), denoted $X \sim N(\mu, \sigma^2)$, when the probability density function, known as the normal distribution, is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

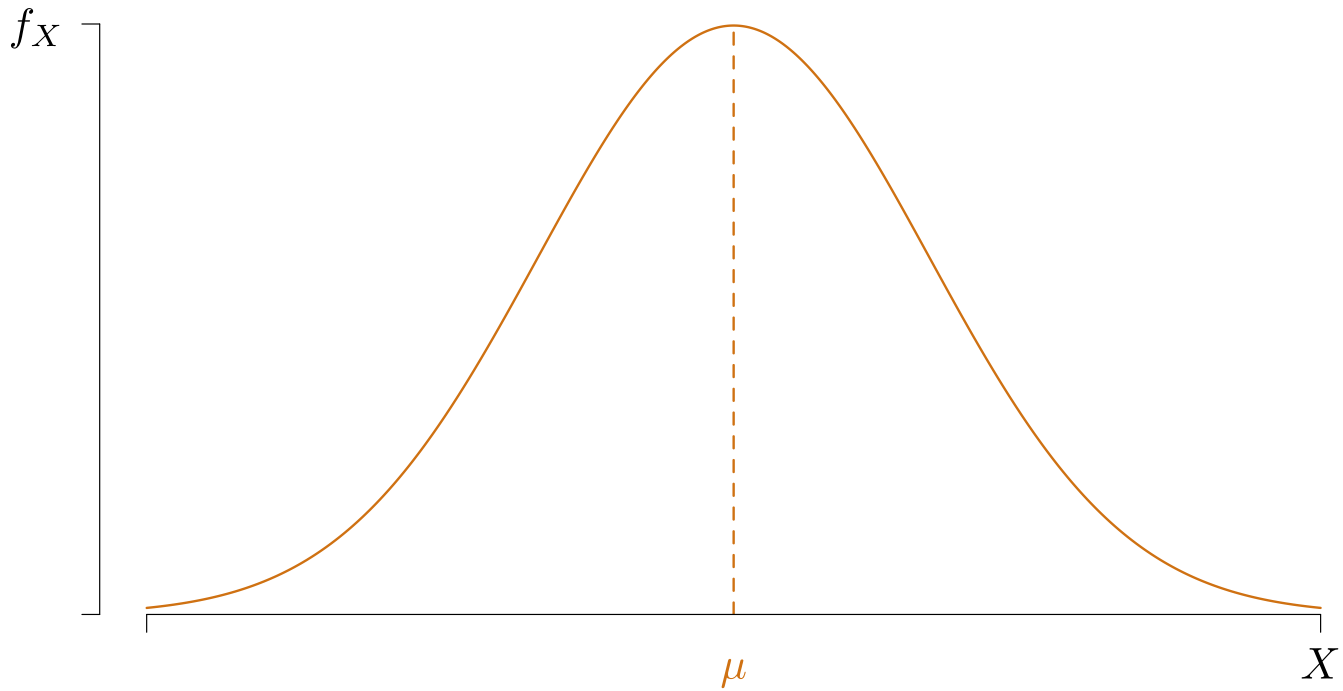
for any real number x.

Properties of the normal distribution:

- $\mathbb{E}[X] = \mu$
- $V[X] = \sigma^2$

Normal distribution

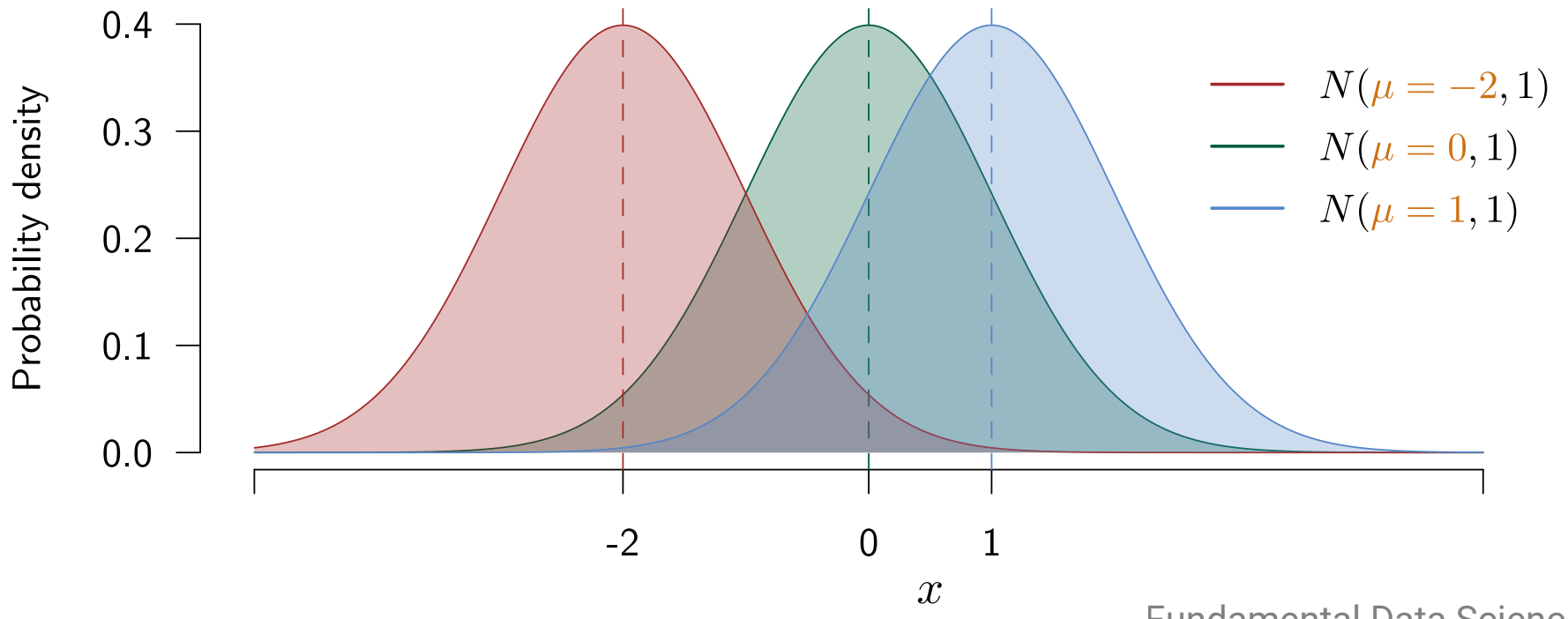
The normal probability density function is symmetric around the mean μ.
The closer to the mean, the higher the probability density funtion f_X is.
It has a well-known 'bell' shape.



Normal distribution — Mean parameter μ

μ is the center (location) of the normal distribution.

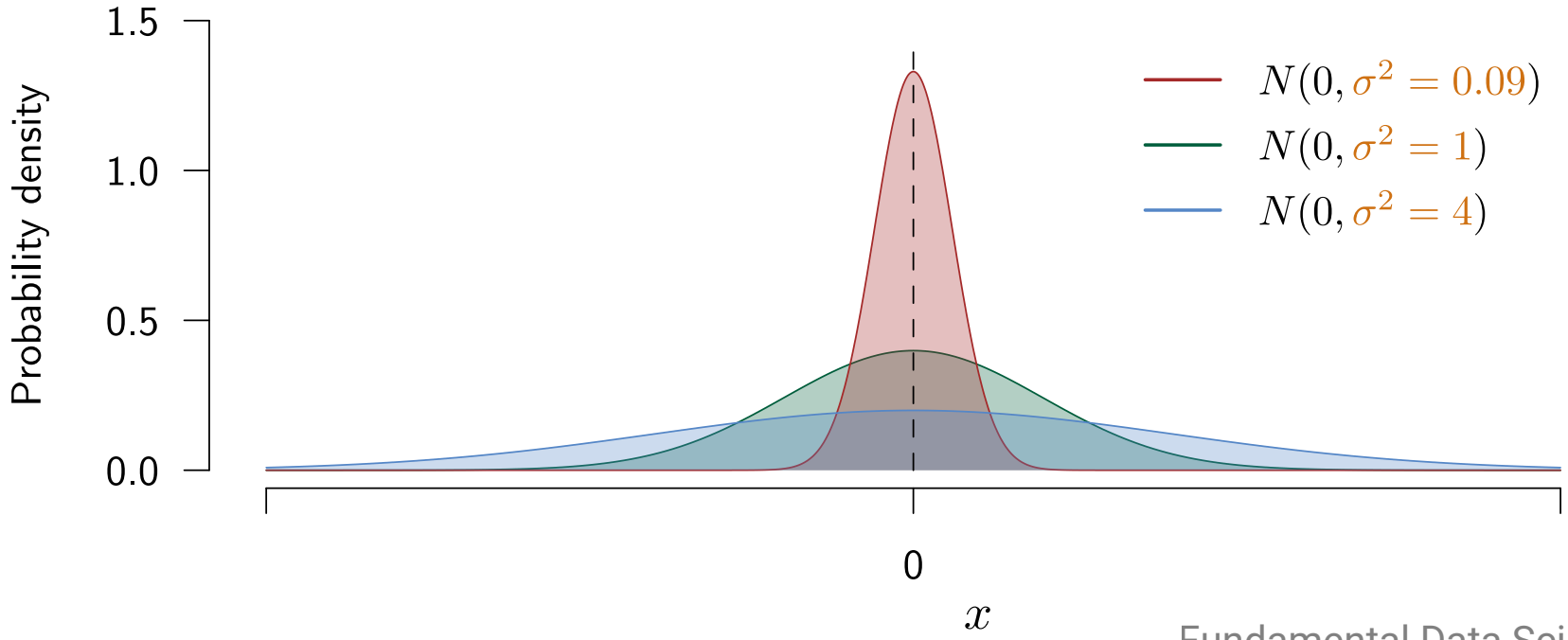
> *Changing μ alone moves the distribution to the right (when μ increases) or to the left (when μ decreases).*



Normal distribution — Variance parameter σ^2

σ^2 describes the width (spread) of the normal distribution.

> *Changing σ^2 alone spreads the distribution away from the mean (when σ^2 increases) or shrinks the distribution towards the mean (when σ^2 decreases).*



Normal distribution — Examples

Normal distributions are useful to model a process that adds together random values from a common distribution. Here's a very cool example.

Random walk

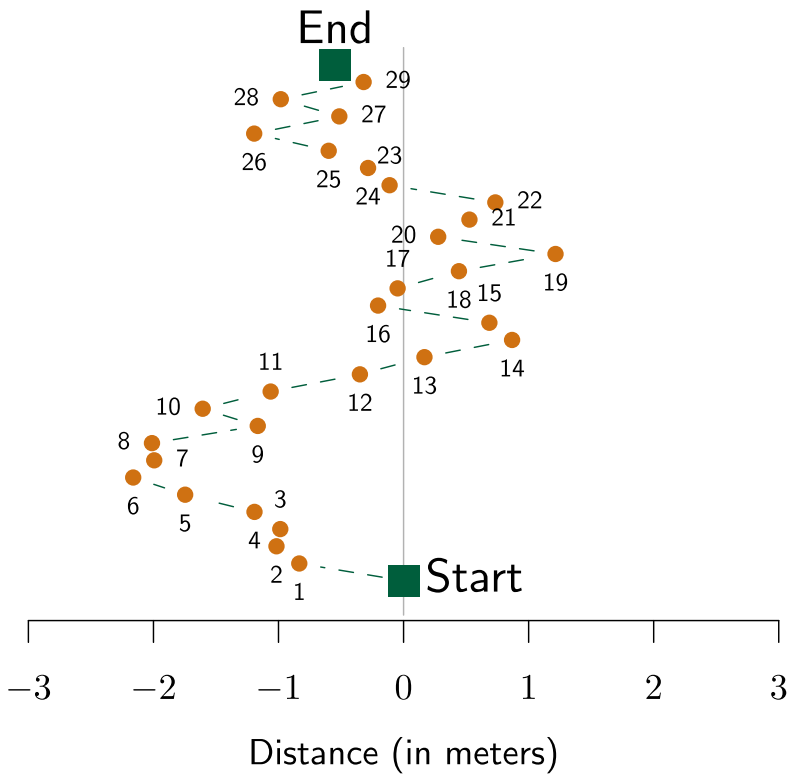
Suppose you stand on the halfway line of a football pitch.

You repeat the following procedure 30 times:

- Generate a random number between -1 and $+1$, say, d.
- You move d meters to the left in case $d < 0$, otherwise you move d meters to the right.

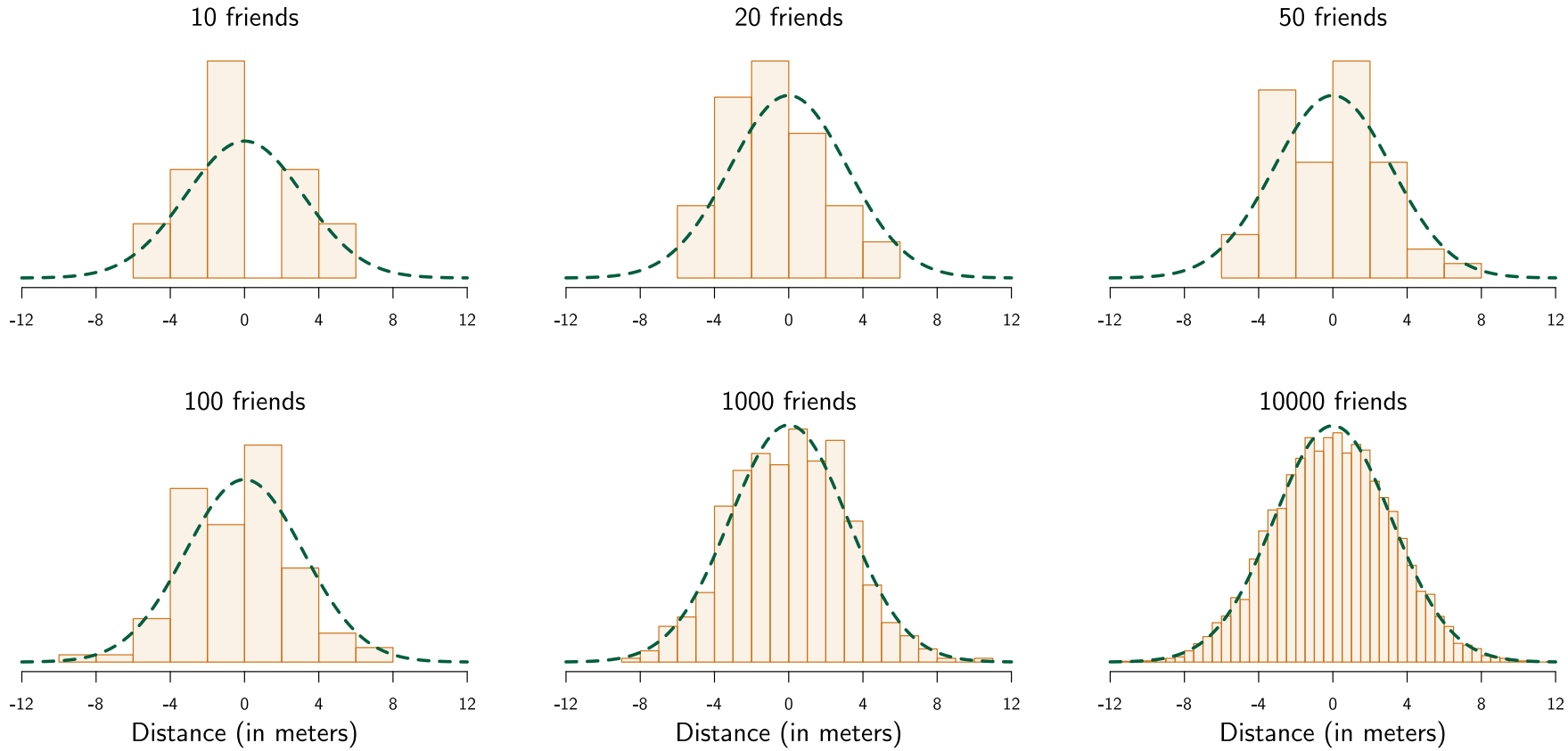
You record your final position.

Now, clearly, it is difficult to predict where you will end up...



Normal distribution — Examples

However, suppose you invite some friends to repeat the same process...



Yes, the distances are normally distributed!

Normal distribution — Examples

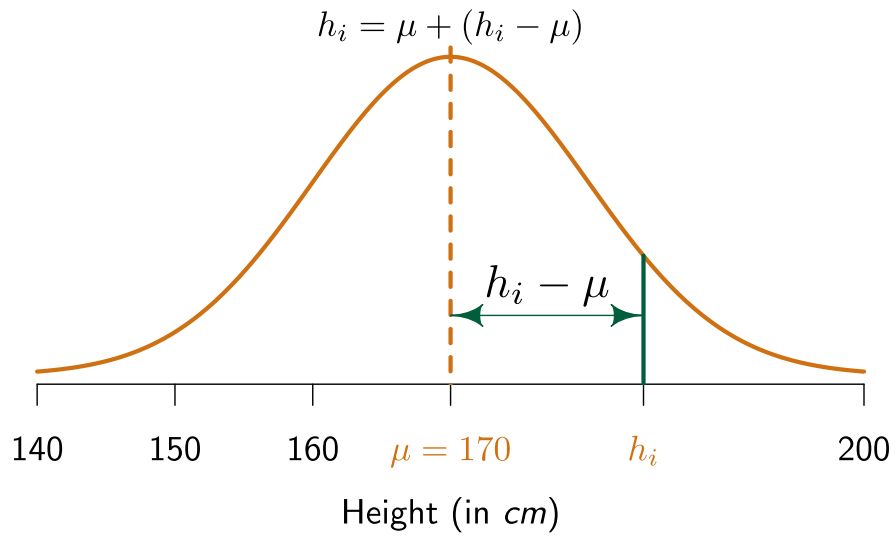
Another example:

> Height in *cm*.

The idea is that the height of a person can be regarded as being equal to the population's mean height $+$ a quantity that *looks as if it were a random value*.
In statistics we refer to this quantity as being a residual or error.

For example, assuming that $\mu = 170\ cm$ in the population, then the height of the i-th person can be decomposed as follows:

$$h_i = \underbrace{170}_{\mu} + \underbrace{h_i - 170}_{\substack{\text{error for the} \\ i\text{-th person}}}.$$



Normal distribution — Computing probabilities

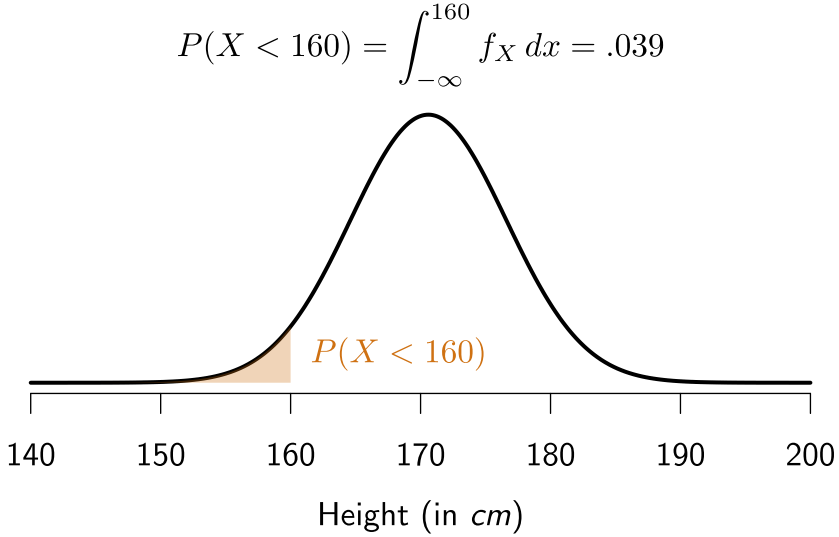
Recall that, for continuous random variables, we must compute probabilities for intervals of values:

$$P(a \leq X \leq b) = \text{area under function } f_X.$$

Also, the probability at any value is equal to 0: $P(X = a) = 0$.

Here is an example:

> *Suppose that the heights (in cm) of boys in a group follow a normal distribution with mean* 170.6 *and variance* 36.
>
> *What is the probability that a boy in this group is shorter than* $160\ cm$*?*



Normal distribution — Computing probabilities

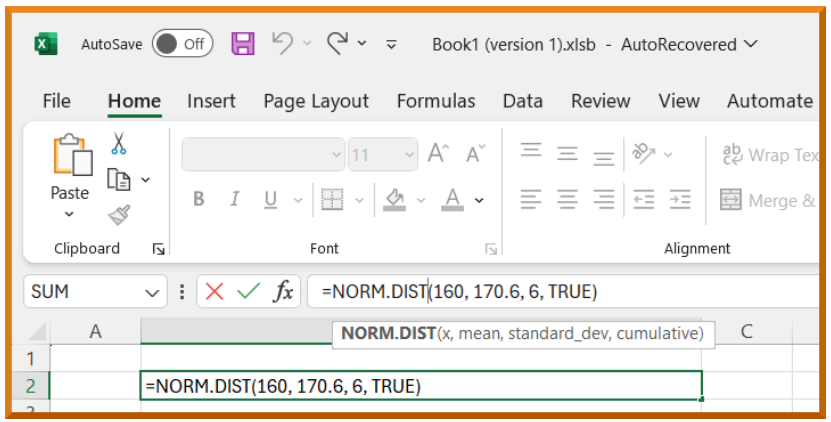
Given random variable X following a normal distribution with mean μ and variance σ^2, the probability $P(X < x)$ can be calculated in Excel by the following command:

Input

```
=NORM.DIST(x, mean, standard deviation, TRUE)
# Note: "FALSE" would give the probability *density* instead.
```

Note:
The standard deviation, σ, is equal to the square root of the variance:

$$\sigma = \sqrt{\sigma^2}.$$

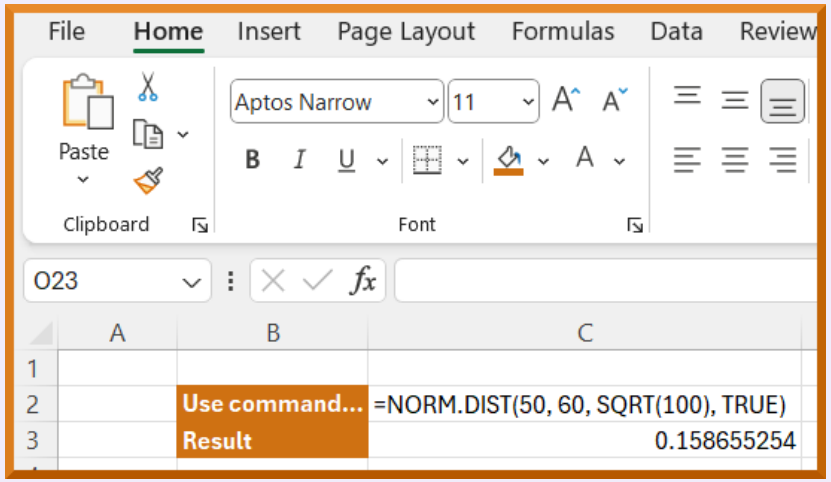


Exercise (3)

Scores of a test follow a normal distribution with mean 60 and variance 100.
Suppose Mr. A scored 50 points in the test.
How does Mr. A rank percentage-wise, counting from the bottom?

Exercise (3) — ANSWER

Scores of a test follow a normal distribution with mean 60 and variance 100.
Suppose Mr. A scored 50 points in the test.
How does Mr. A rank percentage-wise, counting from the bottom?



Relationships between distributions: Binomial, Poisson, and normal

Relationships between distributions

We already learned about the relationship between the binomial and the Poisson distributions:

$$B(n,p) \underset{n\to\infty}{\longrightarrow} Po(\underbrace{np}_{\lambda}),$$

as long as p decreases in such a way that (np) remains constant.

There are two more interesting limits:

- The binomial distribution converges to the normal distribution:

$$B(n,p) \underset{n\to\infty}{\longrightarrow} N(\underbrace{np}_{\mu}, \underbrace{np(1-p)}_{\sigma^2})$$

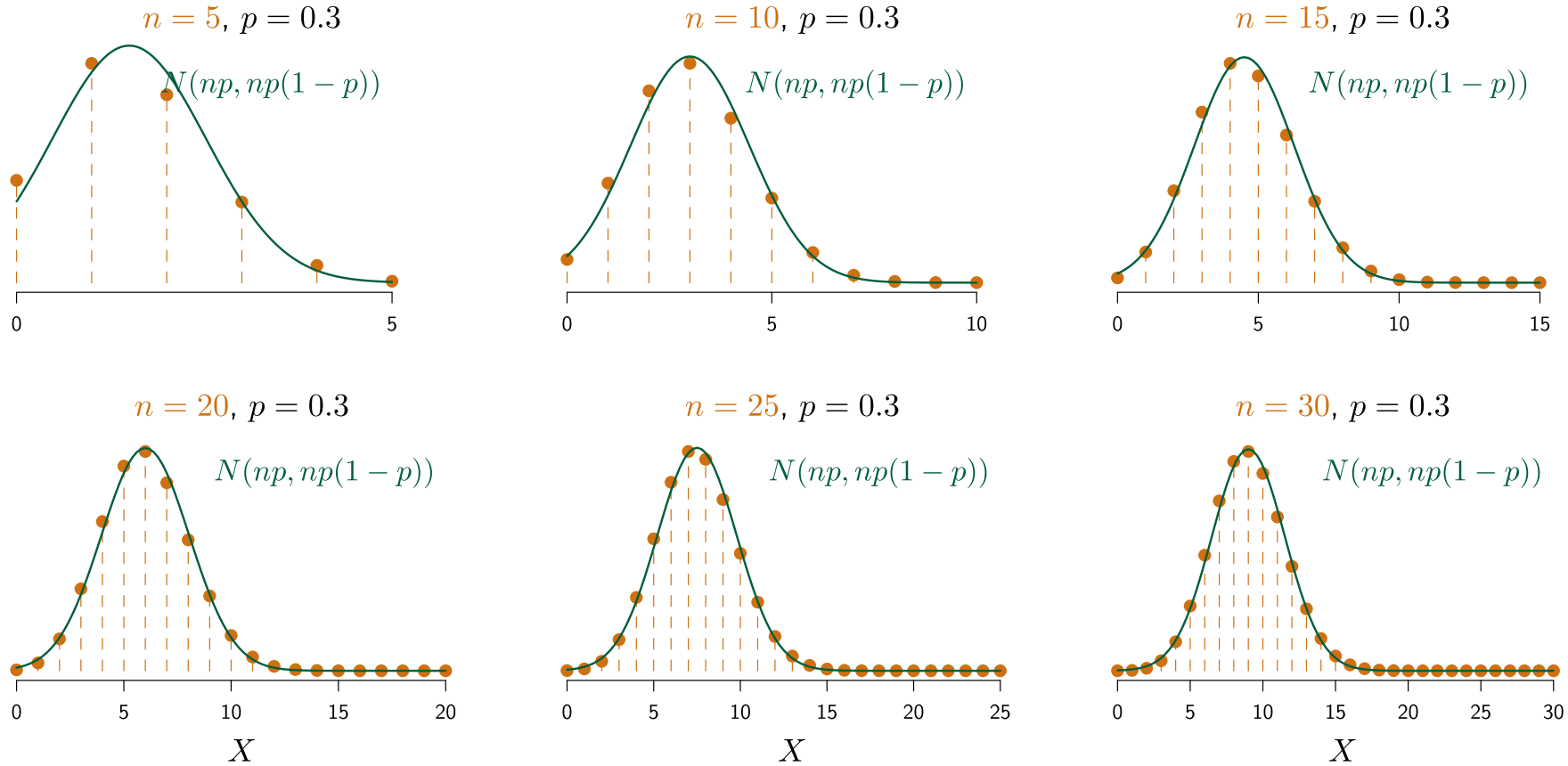
- The Poisson distribution converges to the normal distribution:

$$Po(\lambda) \underset{\lambda\to\infty}{\longrightarrow} N(\underbrace{\lambda}_{\mu}, \underbrace{\lambda}_{\sigma^2})$$

Let's visualize each limit.

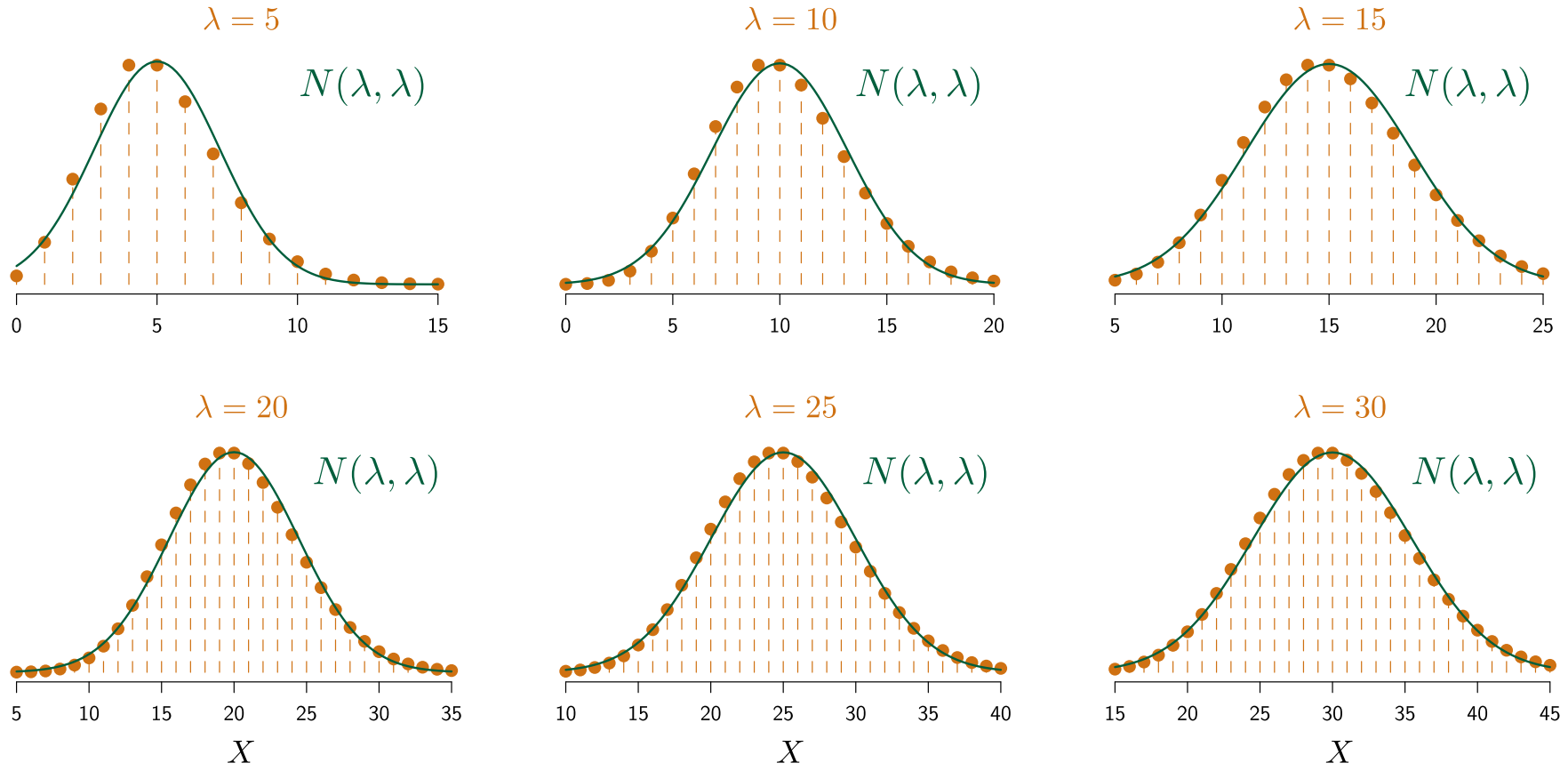
The binomial distribution converges to the normal distribution

Let's compare $B(n, p)$ with $N(np, np(1-p))$ as n increases and $p = 0.3$:

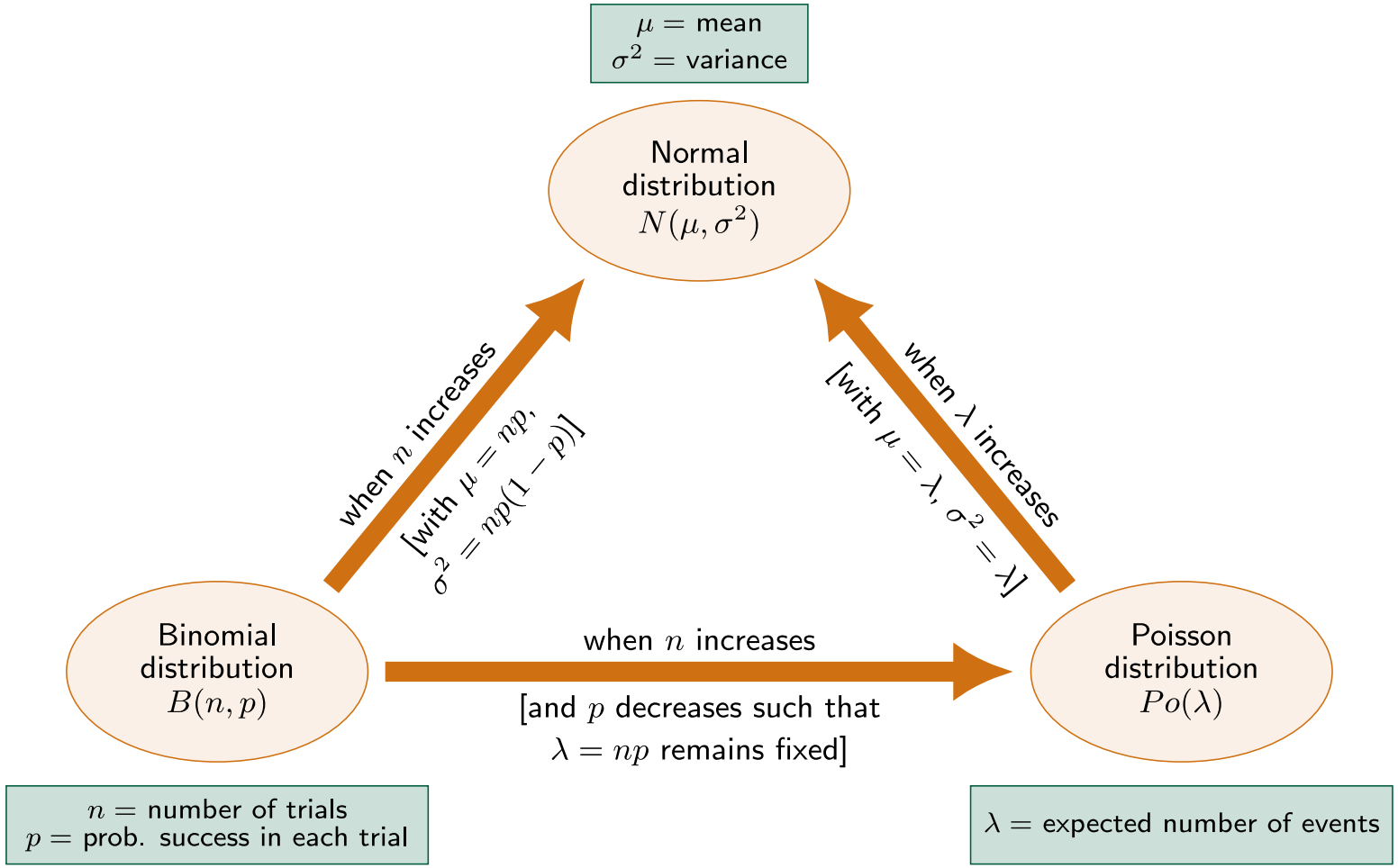


The Poisson distribution converges to the normal distribution

Let's compare $Po(\lambda)$ with $N(\lambda, \lambda)$ as λ increases:



Relationships between distributions



Summary

We learned how to calculate probabilities of various events by fitting a probability distribution to real-life phenomena.

Discrete distributions:

- Binomial distribution
- Poisson distribution

Continuous distributions:

- Normal distribution

We use probabilities as a tool to understand real-life phenomena.