# Probability - Session 3

## Some important discrete distributions — the Binomial and Poisson distributions

Elizabeth Williamson
with thanks to Jennifer Rogers

Foundations of Medical Statistics

# Session objectives

By the end of the session you should be able to:

- calculate probabilities using permutations and combinations
- state key properties of the Bernouilli, binomial and Poisson distributions.
- explain the assumptions of the binomial and Poisson distributions.
- explain the assumptions underlying the derivation of the Poisson distribution

# Outline

# Reminder: random variables

A **random variable** $X$ is a variable which takes a numerical value which depends on the outcome of the random experiment under consideration.

A discrete random variable can be characterised by its **probability distribution function.**

- $P(X = x)$ for all the $x$ values which $X$ can take.

Simple example — the **Bernouilli** distribution:

- A single binary outcome (1=success, 0=failure)
- $P(X = 1) = \pi$

Today: obtaining probability distribution functions for some important discrete distributions

# Calculating probabilities

Recall that the probability of an event can be defined as the relative frequency of that event in repeated experiments, e.g.

$$P(\text{roll a six}) = \lim_{n\to\infty} \frac{\text{number of 6s in } n \text{ rolls}}{n}$$

# Calculating probabilities

If:

- finite number of possible outcomes, and
- each outcome is equally probable.

Then

$$P(\text{event } A \text{ occurs}) = \frac{\text{number of outcomes in which event } A \text{ occurs}}{\text{total number of outcomes}}$$

# Simple example: rolling a die

To do this, we must count how many possible outcomes there are, and how many of these for which the event of interest occurs.

For example, consider rolling a die:

- We are interested in $P(\text{rolling an even number})$.
- Outcomes where we roll an even number: $\{2, 4, 6\}$
- Outcomes in total: $\{1, 2, 3, 4, 5, 6\}$

$$P(\text{rolling an even number}) = \frac{3}{6} = 0.5.$$

However... in more complex situations it is not so easy to calculate the number of outcomes.

# Outline

# Permutations and combinations

*Combinatorics* is the branch of maths concerned with counting.

A *permutation* is an arrangement **with** regard to order, whereas in a *combination* the order does not matter.

For example, the letters AB and BA are two different permutations, but the same combination of letters.

# Permutations

If we have 3 distinct objects, labelled A, B and C, how many permutations are there? There are 6 permutations:

ABC
ACB
BAC
BCA
CAB
CBA

More generally, suppose we have $n$ different objects. In how many ways can $n$ different objects be arranged? Each arrangement (i.e. respecting different orders) is a permutation.

# Permutations of $n$ objects

Consider a box with $n$ compartments that we are going to fill with the $n$ different objects.

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | . | . | . | n |

To find the number of permutations, we consider the number of possibilities of filling each box.

- There are $n$ ways of filling the first box.
- For each of these $n$ ways, there are then $n-1$ ways of filling the second.
- $\ldots$
- At the end, there is only 1 way of filling the last box.

So the total number of ways of filling the boxes is 'n-factorial':

$$n \times (n-1) \times \ldots \times 3 \times 2 \times 1 = n!$$

# Permutations of $x$ objects from $n$

Now suppose we are only selecting $x$ objects from the $n$ available.

How many permutations are there are of $x$ objects from $n$?

- There are $n$ objects to choose from to begin with.
- Then $n - 1$ to choose from.
- $\dots$
- For the $x$th object, there are $n - x + 1$ to choose from.

So there are $n \times (n-1) \dots \times (n - x + 1)$ ways in total.

We write ${}^nP_x$ for the number of permutations of $x$ objects from $n$.

# Permutations of $x$ objects from $n$

It can be useful to express ${}^nP_x$ in a more concise manner.

$$
\begin{aligned}
{}^nP_x &= n \times (n-1) \ldots \times (n-x+1) \\
&= \frac{n(n-1)(n-2)...(n-x+1)(n-x)(n-x-1)... \times 3 \times 2 \times 1}{(n-x)(n-x-1)... \times 3 \times 2 \times 1} \\
&= \frac{n!}{(n-x)!}.
\end{aligned}
$$

# Example

How many 4 digit bank pin numbers are there in which we can only use each digit at most once?

# Example

How many 4 digit bank pin numbers are there in which we can only use each digit at most once?

$$\begin{aligned} {}^{10}P_4 &= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1} \\ &= 10 \times 9 \times 8 \times 7 = 5040. \end{aligned}$$

# Combinations of $x$ objects from $n$

Suppose now we are not concerned with the order of the $x$ objects.

We write ${}^nC_x$ for the number of combinations of $x$ objects from $n$ objects.

- The number of ways of permuting $x$ objects from $n$ different objects is
$$ {}^nP_x = \frac{n!}{(n-x)!} $$
- There are $x!$ ways of permuting $x$ objects
- i.e. each combination of $x$ objects has $x!$ permutations.

$\Longrightarrow$ Therefore,
$$ {}^nC_x = \frac{{}^nP_x}{x!} = \frac{n!}{x!(n-x)!} $$

${}^nC_x$ is also sometimes written $\binom{n}{x}$, which we call "n choose x".

# Example: the UK National Lottery

- you select 6 numbers from 1 to 59.
- balls labelled 1 to 59 are jumbled up; 6 are selected at random.
- you win the jackpot if your 6 choices match the 6 drawn from the machine (irrespective of the order).

What is the probability of winning the jackpot?

- There are ${}^{59}C_6$ ways of choosing 6 balls from 59

$$ {}^{59}C_6 = 45,057,474 $$

- Each of these outcomes is equally likely.
- So the probability that your selected 6 are drawn is

$$ 1/45,057,474 = 0.00000002219 $$

# Outline

# The Bernoulli distribution

A random variable $X$ follows a Bernoulli distribution if it takes two possible values 0 and 1, and has probability distribution:

$$\begin{aligned} P(X=1) &= \pi \\ P(X=0) &= 1-\pi \end{aligned}$$

The expectation and variance are:

$$E(X)=\pi, \qquad Var(X)=\pi(1-\pi)$$

# Outline

# The binomial distribution

The binomial distribution is very important in medical statistics.

A binomial distribution counts the number of successes in a series of $n$ independent Bernoulli trials.

If $X$ follows a binomial distribution we write $X \sim bin(n, \pi)$

For example,

- The number of boys in a 4-child family, $X \sim bin(4, 0.51)$.
- If 10 patients are given an experimental drug, the number who recover follows a binomial distribution, $X \sim bin(10, \pi)$

# The binomial distribution

As an example,

- Consider families with 4 children.
- Let $X$ denote how many boys there are (of 4 children).
- $X$ can take values $x = 0, 1, 2, 3, 4$.
- What is $P(X = x)$?
  - Assume that the probability that a child is male is 0.51.
  - Assume that the gender of each child is independent of the gender of the other children.

# The binomial distribution

Let us consider some particular outcomes

- The probability of no boys:

$$P(GGGG) = 0.49 \times 0.49 \times 0.49 \times 0.49 = 0.49^4$$

- The probability that the first child is a boy, but no others:

$$P(BGGG) = 0.51 \times 0.49 \times 0.49 \times 0.49 = 0.51 \times 0.49^3$$

- The probability that the second child is a boy, but no others:

$$P(GBGG) = 0.49 \times 0.51 \times 0.49 \times 0.49 = 0.51 \times 0.49^3$$

More generally, the probability of a family having a particular order of $x$ boys and $(4 - x)$ girls is: $0.51^x \times 0.49^{(4-x)}$

# Number of boys in a family of 4 children

Note that the different possible outcomes are not equally likely.

However, the different ways of having $x$ boys in a family all have the same probability.

Thus,

$$P(X = x) = \{\text{Number of ways of having } x \text{ boys}\} \times 0.51^x \times 0.49^{(4-x)}$$

# Number of ways of having x boys in a family of 4 children

Consider the number of ways of having 2 boys in a family of 4

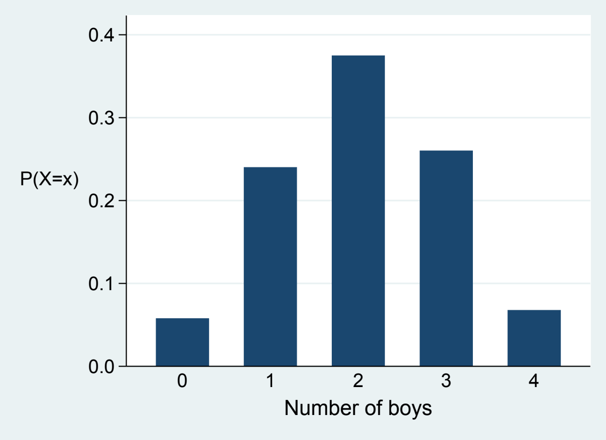
- There are 4 births
- Each way of having 2 boys in the family corresponds to a combination of two of the birth numbers.
  - E.g. the combination 1,2, corresponds to BBGG.
  - E.g. the combination 1,4, corresponds to BGGB.
- We know there are ${}^4C_2$ combinations of 2 objects from 4.

There are ${}^4C_x$ ways of having $x$ boys in a family of 4. So,

$$P(X = x) = {}^4C_x \times 0.51^x \times 0.49^{(4-x)}$$

# The binomial distribution

Graph of the probability distribution function for the number of boys in a 4-child family



# The binomial distribution

More generally, suppose we conduct $n$ independent trials, each of which results in a 'success' or 'failure', with constant probability of success $\pi$.

Then letting $X$ denote the number of successes, $X$ follows a *binomial distribution*, $X \sim \textit{binomial}(n, \pi)$, and

$$\begin{aligned} P(X = x) &= {}^{n}C_{x}\pi^{x}(1-\pi)^{n-x} \\ &= \binom{n}{x}\pi^{x}(1-\pi)^{n-x} \end{aligned}$$

for $x = 0, 1, \ldots, n$.

# Expectation and variance of the Binomial distribution

If $X \sim Bin(n, \pi)$ then $X$ represents the number of successes in a series of $n$ independent Bernoulli trials.

- Let $X_i$, $i = 1, .., n$ be independent Bernoulli trials.
- Then we can write $X = \sum_{i=1}^{n} X_i$.

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} E(X_i) \\ &= \sum_{i=1}^{n} \pi \\ &= n\pi. \end{aligned}$$

$$Var(X) =$$

# Expectation and variance of the Binomial distribution

If $X \sim Bin(n, \pi)$ then $X$ represents the number of successes in a series of $n$ independent Bernoulli trials.

- Let $X_i$, $i = 1, .., n$ be independent Bernoulli trials.
- Then we can write $X = \sum_{i=1}^{n} X_i$.

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} E(X_i) \\ &= \sum_{i=1}^{n} \pi \\ &= n\pi. \end{aligned}$$

$$\begin{aligned} Var(X) &= Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} Var(X_i) \\ &= n\pi(1-\pi). \end{aligned}$$

# Outline

# The hypergeometric distribution

Suppose we are estimating prevalence by taking a sample of size $n$ from a population of size $N$.

Assume that $M$ subjects in the population have the disease of interest.

What is the distribution of $X =$ number of sampled subjects with the disease?

# The hypergeometric distribution

There are:

- ${}^{N}C_n$ ways of choosing $n$ subjects from a population of size $N$.
- ${}^{M}C_x$ ways of choosing $x$ subjects with the disease from the $M$ with the disease in the population.
- ${}^{N-M}C_{n-x}$ ways of choosing the $n-x$ remaining non-diseased subjects.

Each selection of $n$ subjects occurs with equal probability. Thus,

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

# Example: the lottery

Returning to the lottery example, the number of matched balls follows a hypergeometric distribution.

To find $P(X = 3)$, think of the 6 numbers which are drawn from the machine, and the 53 which are not drawn.

There are ${}^{6}C_3$ ways of choosing the 3 matches that we get.

For each of these, there are ${}^{53}C_3$ ways of choosing the 3 that do not match.

So we have:

$$
\begin{aligned}
P(X = 3) &= \frac{{}^{6}C_3 \times {}^{53}C_3}{{}^{59}C_6} \\
&= 0.010
\end{aligned}
$$

# Outline

# The Poisson distribution

- The Poisson distribution is used to model the **number of events** occurring in a fixed time interval $T$ when:
  - events occur randomly in time,
  - they occur at a constant rate $\lambda$ per unit time,
  - they occur independently of each other.

- Examples where this might arise:
  - Emissions from a radioactive source,
  - The number of deaths in a large cohort of people over a year,
  - The number of car accidents occurring in a city over a year.

We are now going to derive the probability distribution function of the Poisson distribution

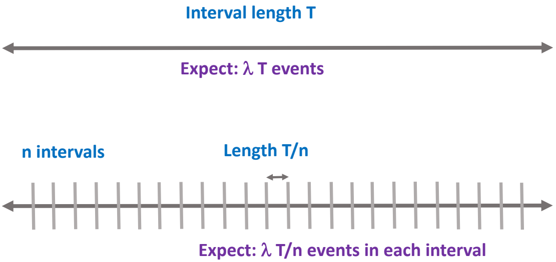
# Derivation of the Poisson distribution - I

Events occur a constant rate of $\lambda$ per unit of time.

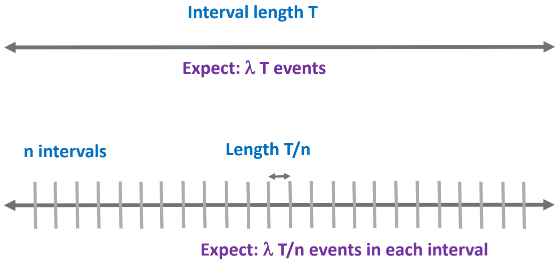
- Over time $T$ we expect $E(X) = \lambda T$ events.

Imagine dividing the interval $T$ into $n$ sub-intervals.

- Each sub-interval is length $T/n$.
- We expect $\lambda T/n$ events in each sub-interval



# Derivation of the Poisson distribution - II




- Increasing $n$ gives more, shorter, intervals
- Increase $n$ until each interval contains at most one event

Then,

- within each sub-interval we have a Bernoulli outcome
- the total number of intervals in which an event occurs is Binomial

# Derivation of the Poisson distribution - III

We let $X$ = total number of intervals in which an event occurs

- $X \sim Bin(n, \pi)$ as $n \to \infty$ where $n$ is the number of intervals

- The probability of an event in each small interval is $\pi = \lambda T/n$
- We write $\mu = \lambda T$, giving $\pi = \mu/n$

- The probability distribution function of $X$ is:

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x}$$

# Derivation of the Poisson distribution - IV

$$P(X = x) \quad = \quad \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x}$$

# Derivation of the Poisson distribution - IV

$$
\begin{aligned}
P(X = x) &= \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x}
\end{aligned}
$$

# Derivation of the Poisson distribution - IV

$$
\begin{aligned}
P(X=x) &= \binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{x!(n-x)!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{n^x(n-x)!}\frac{\mu^x}{x!}\left(1-\frac{\mu}{n}\right)^{n-x}
\end{aligned}
$$

# Derivation of the Poisson distribution - IV

$$
\begin{aligned}
P(X = x) &= \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{n^x(n-x)!} \frac{\mu^x}{x!} \left(1 - \frac{\mu}{n}\right)^{n-x}
\end{aligned}
$$

Now, as $n \to \infty$,

$$
\frac{n!}{n^x(n-x)!} = \frac{n(n-1)...(n-x+1)}{n^x} \to 1
$$

# Derivation of the Poisson distribution - IV

$$
\begin{aligned}
P(X=x) &= \binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{x!(n-x)!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{n^x(n-x)!}\frac{\mu^x}{x!}\left(1-\frac{\mu}{n}\right)^{n-x}
\end{aligned}
$$

Now, as $n \to \infty$,

$$\frac{n!}{n^x(n-x)!} = \frac{n(n-1)...(n-x+1)}{n^x} \to 1$$

and

$$\left(1-\frac{\mu}{n}\right)^{n-x} \approx \left(1-\frac{\mu}{n}\right)^n \to e^{-\mu},$$

# Derivation of the Poisson distribution - IV

$$
\begin{aligned}
P(X = x) &= \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{n!}{n^x(n-x)!} \frac{\mu^x}{x!} \left(1 - \frac{\mu}{n}\right)^{n-x}
\end{aligned}
$$

Now, as $n \to \infty$,

$$\frac{n!}{n^x(n-x)!} = \frac{n(n-1)...(n-x+1)}{n^x} \to 1$$

and

$$\left(1 - \frac{\mu}{n}\right)^{n-x} \approx \left(1 - \frac{\mu}{n}\right)^n \to e^{-\mu},$$

So we have

$$P(X = x) \to \frac{\mu^x}{x!} e^{-\mu} \text{ as } n \to \infty.$$

# Properties of the Poisson distribution

- The number of events occurring in a fixed interval $T$ at a constant rate $\lambda$ follows a Poisson distribution with parameter $\mu = \lambda T$,

$$X \sim \text{Poisson}(\mu = \lambda T)$$

if

$$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}, \text{ for } x = 0, 1, 2, ...$$

In the practical session, you will prove that:

- $E(X) = \mu$
- $Var(X) = \mu$

# Outline

# Summary

- Combinatorics
  - A *permutation* is an arrangement **with** regard to order,
  - In a *combination* the order does not matter.
- The binomial distribution:
  - used when we have a sequence of independent binary (Bernouilli) trials with constant probability of success.
  - has expectation $n\pi$ and variance $n\pi(1-\pi)$.
- The Poisson distribution:
  - used to model the number of events occurring independently at a constant rate in a fixed time interval
  - has expectation equal to its variance ($= \mu$)