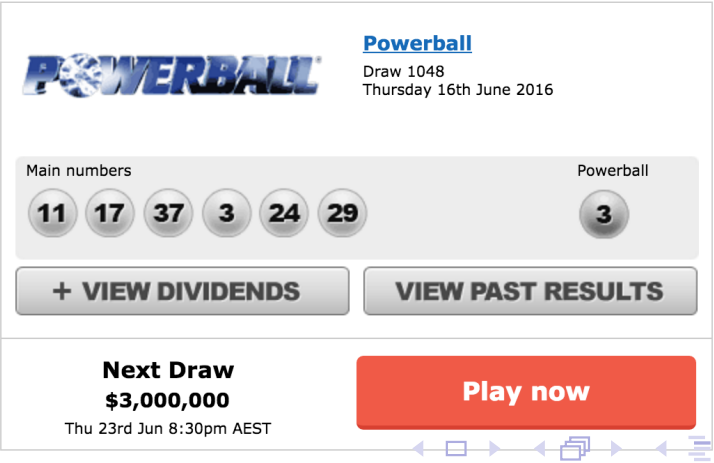
Topic5: Discrete Random Variables

Outline

Example1: Powerball

Powerball is a lottery in Australia with prize money of up to $80 million dollars (2009). Most jackpot wins are not shared by multiple tickets. The Powerball consists of drawing 6 numbers from a machine containing balls numbered 1 to 40, and then drawing 1 number (the Powerball) from a separate machine containing balls numbered 1 to 20.

A 'Division 1' win means picking all 6 main winning numbers and the Powerball. **What is the probability of winning Powerball (Division 1)?**



Example2: Tasmanian Fruit flies

Tasmania is currently free of fruit fly which adds several million dollars to the annual export income earned by the horticultural industries. However the 14 species of fruit fly on the Australian mainland are a constant economic threat. South Australia remains the only Australian mainland state that is fruit fly free, with prevention, detection and eradication measures costing about $5 million annually.

TasmaniaGovt Biosecurity

Suppose there are 100 fruit flies buzzing around a lime tree. The flies have a 20% chance on landing on the tree and act independently.

What is the probability that exactly 20 flies land on the lime tree?

Note: If there were only 2 flies, we could use simple probability.

- The set of all possible values: $x = 0, 1, 2$.
- The likelihood of each value (discrete):
  $P(X = 0) = P(\text{no flies land}) = 0.8 \times 0.8 = 0.64$
  $P(X = 1) = P(\text{1 fly lands}) = 0.2 \times 0.8 + 0.8 \times 0.2 = 0.32$
  $P(X = 2) = P(\text{both flies land}) = 0.2^2 = 0.04$

But we can't use this approach for a realistic number of flies, like $n = 100$. So we need to develop a general formula for $P(X = x)$.

Example3: Calcium deficiency in orchards

While magnesium deficiency occurs in most districts in New South Wales, calcium deficiency is rarely seen in citrus orchards. A deficient range is below 1.6 percent of dry leaf matter and satisfactory range is 3-5.5.

A small size orchard is considering ordering an expensive fertilizer. **What is the chance that more than 1 tree will be calcium deficient?**

NSW agriculture

Discrete Distributions

Definition (Discrete Distribution)

For any discrete distribution X, we have a sample space Ω with values $x = \{x_1, x_2, \ldots\}$ and associated probabilities $\{p_1, p_2 \ldots\}$, where $\{p_i = P(X = x_i)\}$.

Properties:

- there is a countable number of possible values;
- $\sum_i p_i = 1$

Definition (Probability Distribution Function)

The probability distribution function (or probability distribution) of X is the set of $\{x, P(X = x)\}$.

Definition (Cumulative Distribution Function (CDF))

The cumulative distribution function (CDF) of X is

$$F(x) = P(X \leq x)$$

This is a step function. Statistical tables are often presented in terms of the CDF.

The Mean of a Discrete Distribution

Definition (Mean or Expectation)

The mean of X is

$$\mu = E(X) = \sum_x xP(X = x)$$

Definition (Expectation of a Function)

The expectation of $g(X)$ is

$$E(g(X)) = \sum_x g(x)P(X = x)$$

For example: $E(X^2) = \sum_x x^2 P(X = x)$.

The Variance of a Discrete Distribution

Definition (Variance)

The variance of X is

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

What does the 'Mean' mean?

There are 4 main statistical usages of the word 'mean':

- The mean of a sample: $\bar{x}$
- The mean of a population: μ
- The mean of a random variable X (describing a population): $E(X)$
- The distribution of the sampling mean: $\bar{X}$

Similarly, there are 4 main usages of the word 'variance'.

Example

Mean and variance of 5 tosses of a coin

Let X = the number of heads in 5 tosses of a coin, $x = 0, 1, \ldots, 5$, with probability distribution function:

x	0	1	2	3	4	5
$P(X = x)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

Find the mean and variance of X.

$$\mu = E(X) = \sum_x xP(X = x) = 0 \times \frac{1}{32} + 1 \times \frac{5}{32} + \ldots 5 \times \frac{1}{32} = 2.5$$

$$E(X^2) = \sum_i x^2 P(X = x) = 0^2 \times \frac{1}{32} + 1^2 \times \frac{5}{32} + \ldots 5^2 \times \frac{1}{32} = 7.5$$

Hence

$$Var(X) = E(X^2) - E(X)^2 = 7.5 - (2.5)^2 = 1.25$$

In R

```
x=c(0,1,2,3,4,5)
p=c(1,5,10,10,5,1)/32
sum(x*p)

## [1] 2.5

sum(x^2*p)

## [1] 7.5

sum(x^2*p)-(sum(x*p))^2

## [1] 1.25

sum((x-sum(x*p))^2*p)

## [1] 1.25
```

Have a try

In a certain game, 5 coins are tossed, where X denotes the number of heads. It costs \$8.00 to play, and the player receives $\$2^X$ as prize money. Show that the expected loss for 1 game is $\$0.41$.

```
#Check your answer
x=c(0,1,2,3,4,5)
p=c(1,5,10,10,5,1)/32
sum((2^x)*p)-8.00

## [1] -0.40625
```

Your Turn

Suppose you toss a fair coin 1000 times: when it lands heads you receive $1, and when it lands tails you pay me $1. What is your expected profit or loss?

```
x=c(1,-1)
p=c(0.5,0.5)
sum(x*p)

## [1] 0
```

Common Types of Discrete Distributions

There are an infinite number of discrete distributions.

We will concentrate on 2 special examples:

- The Hypergeometric Distribution (Urn model);
- The Binomial Distribution;

For your reference, the Poisson Distribution is also considered as an Appendix.

Example1: Hypergeometric Distribution

Definition (General Hypergeometric Model)

The Hypergeometric distribution models a context described by an urn model.

Suppose an urn contains N balls, with N_1 of type 1, N_2 of type 2, $\ldots$ N_k of type k, where $\sum_{i=1}^{k} N_i = N$.

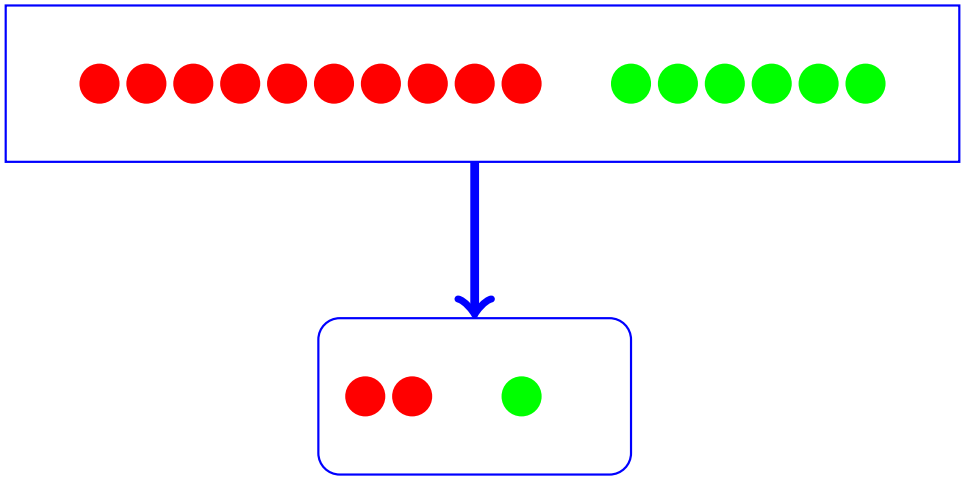
We select a random sample (without replacement) of n balls (where $n \leq N$.)

The probability that we select exactly n_i of type i is

$$P(\text{Select } n_i \text{ balls of each type } i) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_k}{n_k}}{\binom{N}{n}}$$

Example (Packet of M & Ms)

Suppose a small Christmas packet of M & Ms contains 16 chocolates, of which 10 are red and 6 are green. We select a random sample of 3 chocolates. What is the probability of selecting exactly 2 red ones?



The probability that we select exactly $n = 2$ red balls is

$$\frac{\binom{10}{2}\binom{6}{1}}{\binom{16}{3}} \approx 0.48$$

```r
choose(10,2)*choose(6,1)/choose(16,3)

## [1] 0.4821429
```

```r
dhyper(2,10,6,3)    # dhyper(x,N1,N2,n)

## [1] 0.4821429
```

Example (Powerball)

What is the probability of winning Powerball (Division 1)?

```
(choose(6,6)*choose(34,0)/choose(40,6))*
  (choose(1,1)*choose(19,0)/choose(20,1))

## [1] 1.302633e-08
```

```
dhyper(6,6,34,6)*dhyper(1,1,19,1)

## [1] 1.302633e-08
```

Does this match the claim on the powerball website? 'The chance of winning a Division One prize in Powerball is 1 in 76,767,600'.

Powerball Rules

Example2: Binomial Distribution

Definition (Binomial Distribution)

The Binomial distribution models a context in which we have:

- a fixed number n of independent Binary trials;
- a fixed likelihood of a success at each trial $p = P(\text{success})$.

If $X =$ the number of successes in n trials, then $X \sim Bin(n, p)$ with

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, 2, \ldots, n.$$

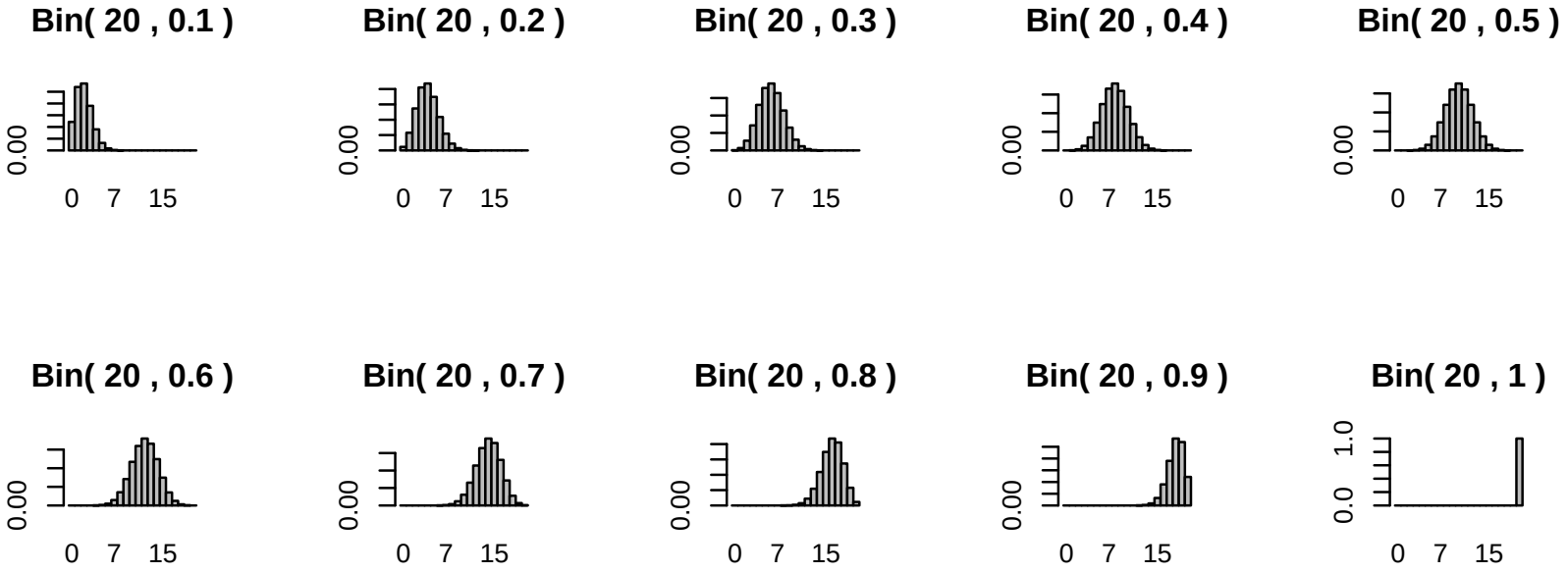
Note: $\sim$ reads 'is distributed as'. Factorials BinomialCoefficients

Notes:

- A *Binary* (or Bernoulli) trial is an event where there can only be 2 options: success or failure. For example, 1 fruit fly is buzzing around a fruit tree: will it land or not?
- *Success* designates the event we are interested in counting, which may not be good. For example, $p = P(\text{fruit fly lands on fruit tree})$.
- The Binomial distribution has 2 parameters: n and p. Parameters represent the numerical inputs needed for the model.
- It can be shown (by algebra) that the Binomial distribution has mean $E(X) = np$ and variance $Var(X) = np(1-p)$.

Example of Binomial Distribution with changing p

$X \sim Bin(n = 20, p)$, for different $p = 0.1, 0.2, \ldots, 1$.



Example

Fruit Flies

There are 10 fruit flies buzzing around a lime tree. The flies have a 20% chance on landing on the tree and act independently. If X represents the number of flies that land on the tree, what is the distribution of X? What is the chance that no flies land on the tree? What is the chance that less than 2 flies land on the tree?

1. Identify the model:
   $X =$ the number of flies that land $\sim Bin(n, p)$, where
   $n =$ number of flies $= 10$; $p = P(\text{fruit fly lands}) = 0.2$.
2. Calculate probability:
   $P(\text{no flies land}) = P(X = 0) = \binom{10}{0}(0.2)^0(0.8)^{10} = \frac{10!}{0!(10-0)!}(0.8)^{10} = (0.8)^{10} \approx 0.11$.

Calculate probability:
$P(\text{less than 2 flies land}) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.1073742 + \binom{10}{1}(0.2)^1(0.8)^9 \approx 0.38.$

```
# dbinom(x,n,p) calculates P(X=x) for Bin(n,p)
dbinom(0,10,0.2)

## [1] 0.1073742
```

```
# pbinom(x,n,p) calculates P(X<=x) for Bin(n,p)
pbinom(1,10,0.2)

## [1] 0.3758096
```

Appendix: Poisson Distribution

The Poisson distribution is used in 2nd year courses, so is given here for reference.

Definition (Poisson Distribution)

The Poisson distribution models a context in which we have

- events occuring in an interval;
- the average number of events occuring in an interval is λ (rate).

If $X =$ number of events in the interval, then $X \sim Po(\lambda)$ and

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots \text{ and } \lambda > 0.$$

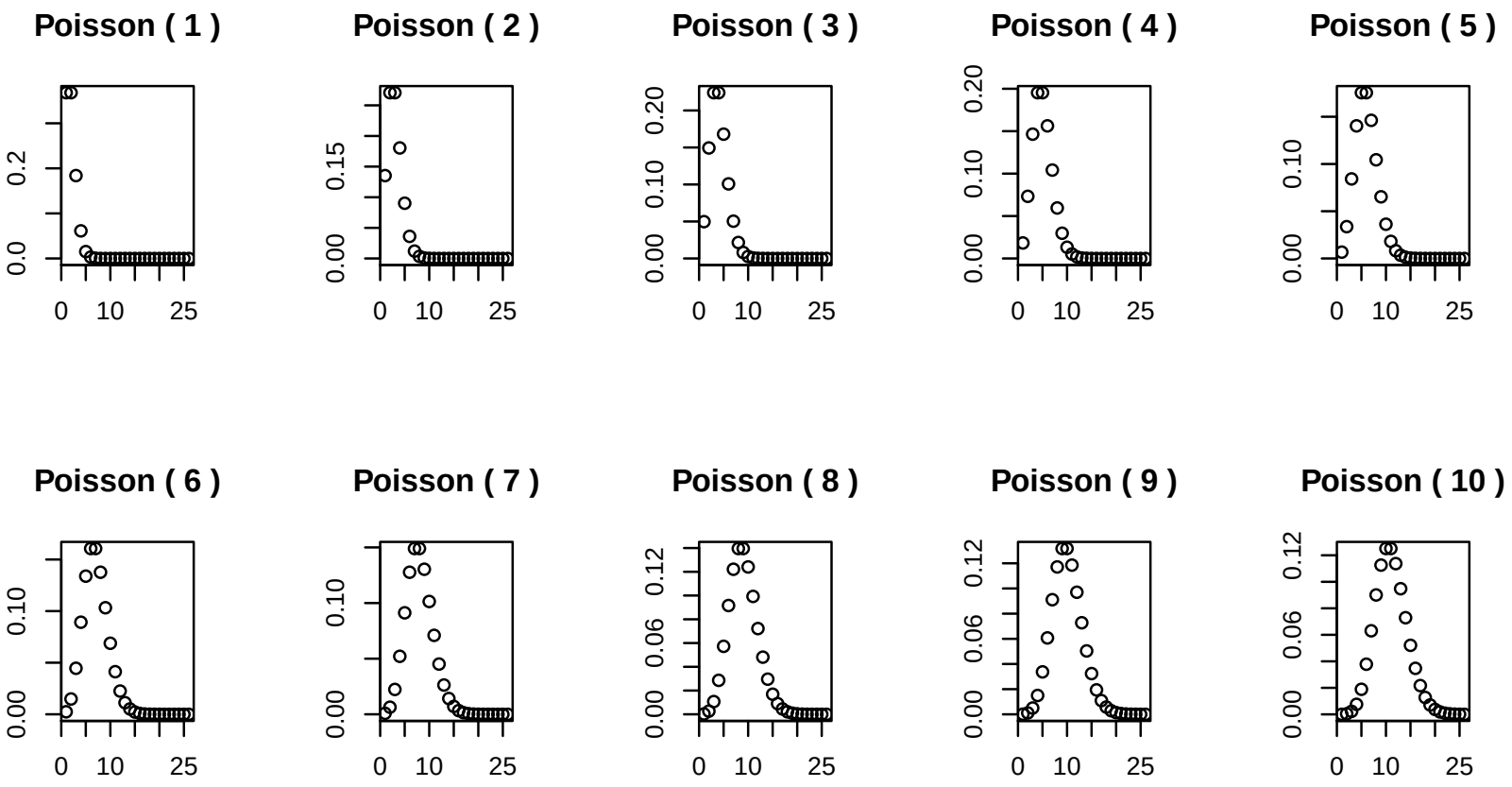
Notes:

- λ is pronounced 'lambda'.
- Poisson is pronounced 'pwasonn'.
- The Poisson distribution has 1 parameter: λ.
- The Poisson Distribution models rare events, where λ is small. For example, $\lambda =$ the average number of lime trees exhibiting calcium deficiency in an orchard in a year.
- It can be shown (by algebra) that the Poisson distribution has mean $E(X) = Var(X) = \lambda$.

Extension: For large n and small p, $X \sim Bin(n, p)$ can be approximated by $Y \sim Po(\lambda = np)$.

Examples of Poisson Distributions with changing λ

$X \sim P_0(\lambda)$, for different $\lambda = 1, 0.2, \ldots 10.$



Example

Magnesium deficiency

While magnesium deficiency occurs in most districts in New South Wales, calcium deficiency is rarely seen in citrus orchards. A deficient range is below 1.6 percent of dry leaf matter and satisfactory range is 3-5.5.

For a small size orchard, let X represent the number of trees with calcium deficiency in a year, where $\lambda = 2$. For ordering an expensive fertilizer, what is the chance that more than 1 tree will be calcium deficient? ▸ NSW agriculture

1. Identify the model:
   $X =$ the number of trees that are calcium deficient $\sim Po(\lambda)$,
   where $\lambda =$ average number of deficient trees in a year $= 2$.
2. Calculate probabilities:
   $P(\text{X=0}) = \frac{2^0 e^{-2}}{0!} = e^{-2} = 0.1353353$
   $P(\text{X} = 1) = \frac{2^1 e^{-2}}{1!} = 0.2706706$
   $P(X > 1) = 1 - P(X = 0) - P(X = 1) = 0.5939941$

```r
# dpois(x,l) calculates P(X=x) for Po(l)

dpois(0,2)

## [1] 0.1353353
```

```r
# ppois(x,l) calculates P(X<=x) for Po(l)

1-ppois(1,2)

## [1] 0.5939942
```