Week 9: Discrete Probability Distributions

Data Analysis for Psychology in R 1

Marju Kaps

Department of Psychology
The University of Edinburgh

Course Overview

Exploratory Data Analysis	Research design and data
	Describing categorical data
	Describing continuous data
	Describing relationships
	Functions
Probability	Probability theory
	Probability rules
	Random variables (discrete)
	Random variables (continuous)
	Sampling

Foundations of inference	Confidence intervals
	Hypothesis testing (p-values)
	Hypothesis testing (critical values)
	Hypothesis testing and confidence intervals
	Errors, power, effect size, assumptions
Common hypothesis tests	One sample t-test
	Independent samples t-test
	Paired samples t-test
	Chi-square tests
	Correlation

Learning Objectives

1. Understand concept of a random variable
2. Understand the process of assigning probabilities to all outcomes
3. Apply the understanding of discrete probability distributions to the example of the binomial distribution
4. Understand the difference between a probability mass function (PMF) and a cumulative probability function (CDF)

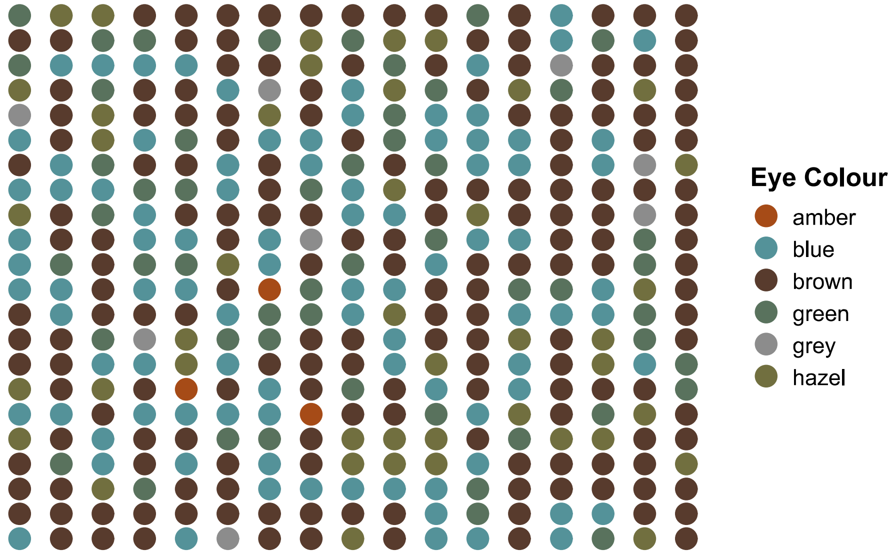
Probability as it relates to Psychology...

- Recall our definition of a **random experiment:**
  - It could (theoretically) be infinitely repeated under the same conditions
  - The outcome is uncertain
- When we conduct a random experiment, we are sampling simple events from a *sample space* to get an outcome
- We can't be 100% certain which outcome will occur each time the experiment is repeated
- An outcome's probability provides us with information that can be used to make decisions about data when we're faced with randomness

Probability as it relates to Psychology...

- *Sample Space:* all student eye colours
- *Simple Event:* the eye colour of an individual student
- *Random Experiment:* Randomly selecting a student and checking their eye colour

DapR Student Eye Colours



Random variables

- A **random variable** is a set of values that quantify the outcome of the random experiment
  - Allows you to map the outcomes of a random experiment to numbers
  - Usually denoted with a capital letter

Random Experiment: Checking eye colour

Random Variable:

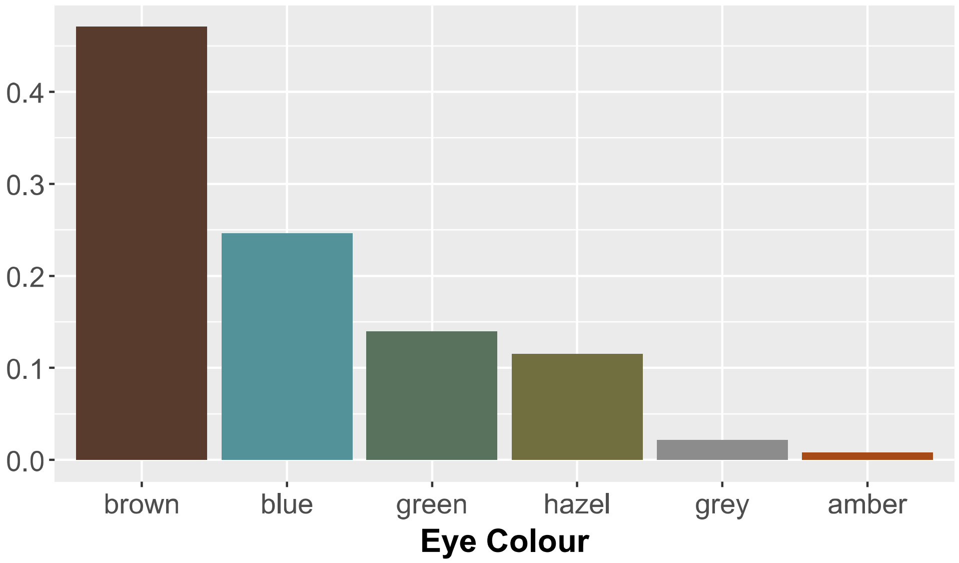
$$X = \begin{cases} 1\ if\ amber \\ 2\ if\ blue \\ 3\ if\ brown \\ 4\ if\ green \\ 5\ if\ grey \\ 6\ if\ hazel \end{cases}$$

- A **discrete random variable** can assume only a finite number of different values
  - e.g. outcome of a coin toss; number of children in a family
- A **continuous random variable** is arbitrarily precise, and thus can take all values in some range
  - e.g. height, age, distance

> **Test your understanding:** What kind of variable is eye colour?

Probability distributions

- A probability distribution maps the values of a random variable to the probability of it occurring



Probability Mass Function

- A **probability mass function** gives the probability that a **discrete random variable** exactly equals a specific value:

$$f(x) = P(X = x)$$

- In the case of our eye colour example:

$$f(hazel) = P(X = hazel)$$

Probability Mass Function

$$f(x) = P(X = x)$$

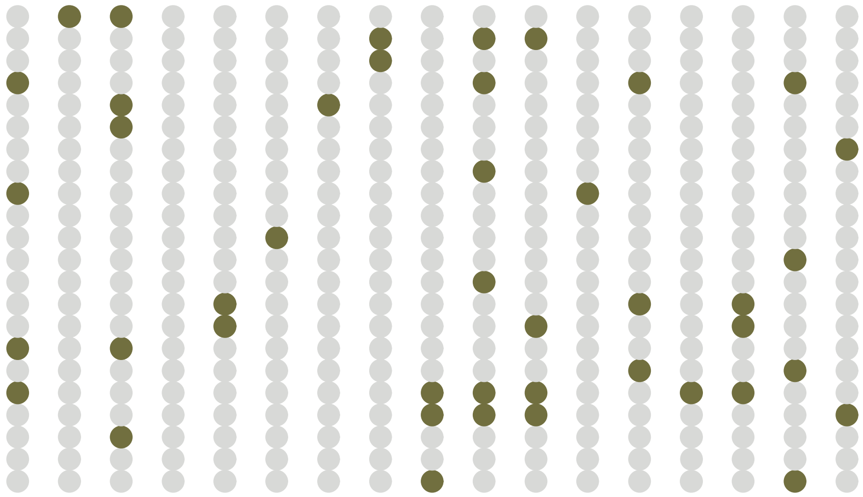
- Some observations (remember probability rules from last week):
  - If you have a random experiment with N possible outcomes, then:

    $$\sum_{i=1}^{N}(f(x_i)) = 1$$

  - For any subset A of the sample space:

    $$P(A) = \sum_{i \in A}(f(x_i))$$

$$P(hazel) = \sum_{i \in hazel}(f(h_i))$$



$$P(hazel) = P(h_1) + P(h_2)... + P(h_{43}) = \frac{43}{374}$$

Discrete random variables: An example

- **Simple Experiment:** Rolling two 6-sided dice
- **Discrete random variable:** The sum of the two upward facing sides
- **Assumptions:**
  1. Dice are fair (numbers between 1 and 6 all equally likely)
  2. The outcome of each dice is *independent* of the outcome of the other

Discrete random variables: An example

Sample space, S:

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

- We can represent S as a frequency distribution.
- **Frequency distribution:** Mapping the values of the random variable with how often they occur

Outcome	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1

- Probabilities are just frequency over total possible outcomes:

$$P(x) = \frac{ways\ x\ can\ happen}{total\ possible\ outcomes}$$

> **Test Your Understanding:** What is the probability of the dice summing to 7?

Discrete random variables: An example

- First, we need to **sum the frequencies** to the get total number of possible outcomes:

```
sum(table_data$Frequency)
```

```
## [1] 36
```

- Next, we **divide the frequency of each outcome by the total frequency**:

$$P(X = 2) = \frac{1}{36} = .03$$

$$P(X = 3) = \frac{2}{36} = .06$$

$$\vdots$$

$$P(X = 12) = \frac{1}{36} = .03$$

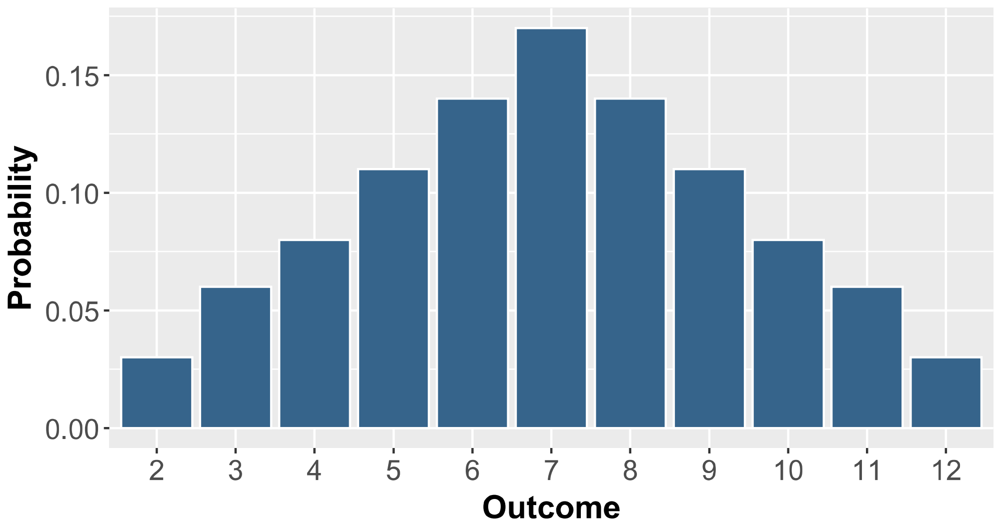
- This gives us a **discrete probability distribution:**

Outcome	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Probability	0.03	0.06	0.08	0.11	0.14	0.17	0.14	0.11	0.08	0.06	0.03

Probability mass function

- You can plot a discrete probability distribution using a bar plot:

Outcome	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Probability	0.03	0.06	0.08	0.11	0.14	0.17	0.14	0.11	0.08	0.06	0.03



Questions?

Binomial Distributions

- A common type of discrete probability distribution is the **binomial distribution**
- Properties:
  - There are only two possible outcomes, one reflecting `success` and one reflecting `failure`
  - The number of observations (*n*) is fixed
  - Each observation is independent of each other
  - The probability of success (*p*) is the same for each observation
- We are interested in the number of successes (*k*) given a fixed number of trials (*n*)

> **Test your understanding:** Identify `success` and n in the following examples:
>
> - The number of tails in a sequence of 5 coin tosses
> - The incidence of a disease in a sample of 100 participants

Binomial Probability Mass Function

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

- k = number of `successes`
- n = total trials
- p = probability of `success`
- $q = 1 - p$, i.e. probability of `failure`
- $\binom{n}{k}$ = n choose k, or the number of ways to select k `successes` from n observations (aka a *combination*).

Binomial PMF - Worked Example

$$P(X=3) = \binom{n}{k} p^k q^{n-k}$$

- **Example:**
  - Random Experiment - Participants were asked to guess which hand a coin is in 5 times.
  - We want to calculate the probability of the participant selecting the correct hand 3 times of the 5
- This looks overwhelming, but let's break it down into it's separate parts.

> Step 1 - Identify n, p, q, and k and plug them into the equation
>
> - n = 5
> - p = 0.5
> - q = 0.5
> - k = 3

Binomial PMF - Worked Example

$$P(X=3) = \binom{5}{3} \times 0.5^3 \times 0.5^{5-3}$$

> Step 2 - $\binom{5}{3}$

- Reflects the number of ways we could get 3 `successes` from 5 trials
- This could happen in multiple ways
- We could calculate this by hand, but it's much easier to use the formula for $\binom{n}{k}$:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Y	Y	Y	N	N
Y	Y	N	Y	N
Y	Y	N	N	Y
Y	N	Y	Y	N
Y	N	Y	N	Y
Y	N	N	Y	Y
N	Y	Y	Y	N
N	Y	Y	N	Y
N	Y	N	Y	Y
N	N	Y	Y	Y

Binomial PMF - Worked Example Step 2

$$\binom{5}{3} = \frac{5!}{3!(5-3)!}$$

$$5! = 5 * 4 * 3 * 2 * 1 = 120$$

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{120}{6 \times 2} = 10$$

- There are 10 ways to get 3 `successes` from 5 trials

Binomial PMF - Worked Example Steps 3 & 4

$$P(X = 3) = 10 \times 0.5^3 \times 0.5^{5-3}$$

> Step 3 - p^k

- $0.5^3 = 0.125$

> Step 4 - q^{n-k}

- $0.5^{5-3} = 0.5^2 = 0.25$

Binomial PMF - Worked Example Step 5

> Step 5 - Put it all together

$$P(X = 3) = 10 \times 0.125 \times 0.25 = 0.3125$$

- Congratulations! We've worked out the probability of one possible outcome ($X = 3$) of our random experiment! ... but we still have 5 more.

k	$P(X = k)$
0	?
1	?
2	?
3	.3125
4	?
5	?

Binomial PMF in R

- Luckily, you can use the `dbinom` function in R to calculate these things for you:

```
dbinom(x, size, prob)
```

- Where:
  - `x` $= k$
  - `size` $= n$
  - `prob` $= p$

```
dbinom(3, 5, 0.5)
```

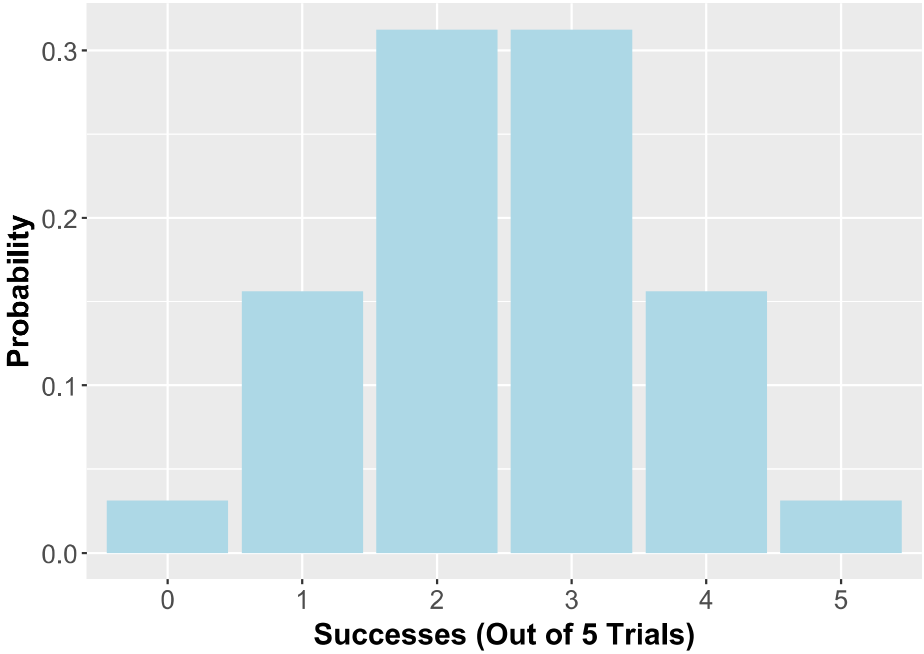
```
## [1] 0.3125
```

Questions?

Visualising binomial probability distribution

- We can pass these values to `ggplot` to produce a bar plot that shows the binomial probability distribution for this random experiment:

k	Pk
0	0.03
1	0.16
2	0.31
3	0.31
4	0.16
5	0.03

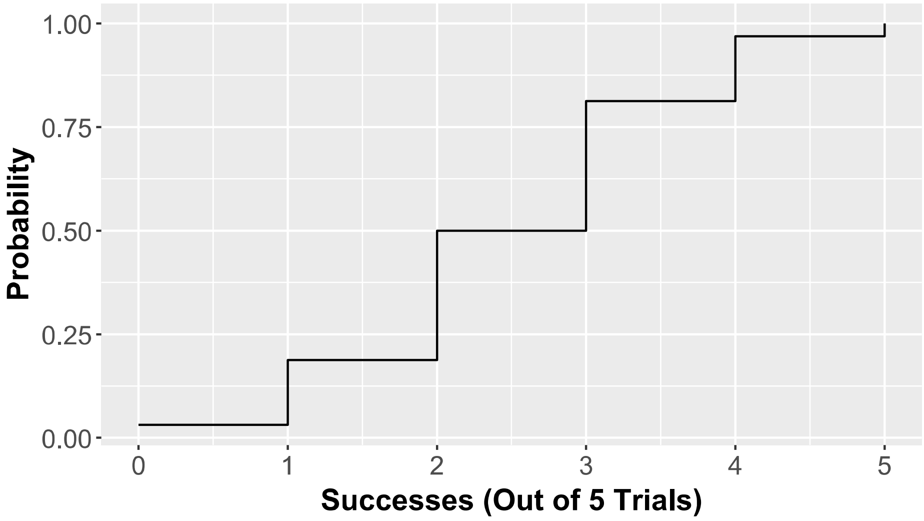


Cumulative probability

- We've been looking at the **probability mass function** to investigate the total probability of a discrete outcome.
- The **Cumulative distribution function** allows us to see the total probability of all values before or after a given point.
- With a binomial distribution, the cumulative probability function simply sums the probabilities of the individual outcomes.
- In R, we can use `pbinom` to get cumulutive probabilities:

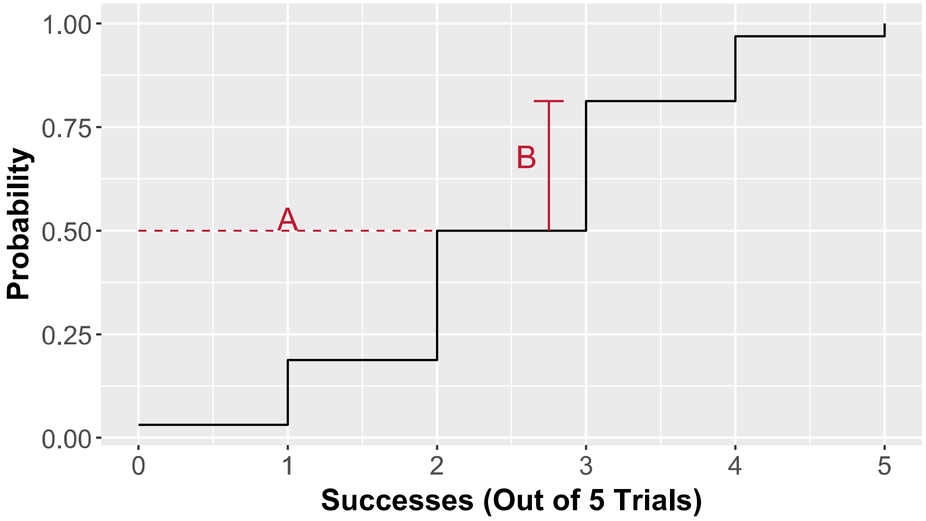
```
round(pbinom(0:5, 5, 0.5), 2)
```

```
## [1] 0.03 0.19 0.50 0.81 0.97 1.00
```



Interpreting CDF

- **A** reflects the probability of selecting the correct hand 0, 1, or 2 out of five trials
    - In this example, 50%
- **B** reflects the individual probability of selecting the correct hand 3 out of 5 trials
    - The difference between the probability of selecting the correct hand 0, 1, 2, or 3 trials and the probability of selecting the correct hand 0, 1, or 2 trials
    - $0.81 - 0.5 = 0.3125$



Questions?

Summary of today

We discussed:

- Random variables and random experiments
- Assigning probabilities to outcomes and defining a probability distribution
- Probability mass functions vs. cumulative distribution functions
- The binomial distribution for assigning probabilities to sets of outcomes

- Tomorrow, I'll present a live R session focused on computing and plotting discrete probability distributions
- Next week, we will talk about continuous probability distributions

This week

Tasks

- Attend both lectures
- Attend your lab and work together on the lab tasks
- Complete the weekly quiz
  - Opened Monday at 9am
  - Closes Sunday at 5pm

Support

- **Office hours**: for one-to-one support on course materials or assessments
  (see LEARN > Course information > Course contacts)
- **Piazza**: help each other on this peer-to-peer discussion forum
- **Student Adviser**: for general support while you are at university
  (find your student adviser on MyEd/Euclid)