Lecture 2
Introduction to probability theory

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Outline

- Terminology
- Bernoulli versus Binomial
- Random Variables
- Combinatorics
- Binomial distribution / probabilities
- Birthday Problem analogy

Terminology

For our purposes we use the following terminology:

- An *Experiment* is any process that generates a set of data
- e.g. flipping a coin five times
- *Trials* is the number of repetitions of an experiment
- e.g. if we repeat the above experiment one hundred times, we perform 100 trials

Terminology cont.

- An *outcome* is the set of possible results for an experiment
- e.g. for the experiment of flipping a single coin, the outcomes are {H, T}
- An *event* is a set of outcomes of an experiment to which a probability is assigned
- e.g. consider the possible outcomes to be the result of randomly selecting a single card from a set of 52 playing cards, an event could be getting a *red* or a *black* card
- $P(CARD = red) = 0.5$

Terminology cont.

- *Permutations* are the set of all possible arrangements
- e.g. all possible events from flipping a coin three times

1	H	H	H
2	H	H	T
3	H	T	H
4	H	T	T
5	T	H	H
6	T	H	T
7	T	T	H
8	T	T	T

- In the context of permutations, each and every row is unique
- There are eight possible permutations

Terminology cont.

- *Combinations* is a selection of items such that their order does not matter
- e.g. possible events from flipping a coin three times

				Unique rows
1	H	H	H	First
2	H	H	T	Second
3	H	T	H	Second
4	H	T	T	Third
5	T	H	H	Second
6	T	H	T	Third
7	T	T	H	Third
8	T	T	T	Four

- e.g. rows 2, 3, and 4 are unique, or the same, since they have two heads and one tail
- There are four possible combinations

Bernoulli versus Binomial

A Bernoulli random variable has two possible outcomes: 0 or 1. A binomial distribution is the sum of independent and identically distributed Bernoulli random variables.

- Representing a single coin is a Bernoulli random variable
- Representing the result of flipping multiple coins is a binomial experiment
- $\therefore$ all Bernoulli distributions are binomial distributions, but *not* all binomial distributions are Bernoulli distributions

Random Variables (RV)

- A RV is much like any sort of variable, be it mathematical, or in programming, but:
  - its possible values correspond to outcomes, e.g. H, T
  - the precise value is unknown, since it depends on randomness
- RV distinguish between a model and its measurements
- RV are usually represented by capital letters
- RV are described by their probability distributions

Bernoulli RV

- Single coin toss is an example of a Bernoulli RV
  $P(X = H) = p$
  $P(X = T) = 1 - p$
- p is a parameter associated with this Bernoulli RV and also defines the associated Bernoulli distribution

Binomial RV

Suppose a fair coin is tossed 3 times

- There are eight equally likely outcomes:
  TTT, TTH, THT, HTT, THH, HTH, HHT and HHH
- Three coins is an example of a binomial RV
- If we are interested in the number of heads, the possible counts are 0, 1, 2 and 3
- $\therefore$ we use counts (product of combinations) to summarize events for a binomial RV

Head count	0	1	2	3
Combinations	1	3	3	1

- But how can we calculate count in general?

Combinations

What if we toss four coins? Or toss a single coin four times; same thing

Head count	0	1	2	3	4
Combinations	1	4	6	4	1

- How do we compute the number of combinations?
- From combinatorics we use "n choose k"

  also known as C_k^n and $\binom{n}{k}$; in this case $\binom{4}{\text{head count}}$

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

where

$$n! = n \times (n-1) \times (n-2) \cdots \times 1$$

e.g.

$$4! = 4 \times 3 \times 2 \times 1$$

How to do “n choose k” using *R*

$\binom{4}{1} = choose(4, 1)$
$\binom{4}{1} = 4$

So the combinations returned are

Head count	0	1	2	3	4
Combinations	1	4	6	4	1

are found using
$\{1, 4, 6, 4, 1\}$= *choose*(4, 0:4)

Binomial distribution

A Binomial experiment requires the following conditions

- n independent events
- Each event has the same probability p of success
- We are interested in the successes from n trials

The probability of k successes from n trials is a Binomial distribution with probabilities:

$$P(k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

In the case of a fair coin, the above simplifies to:

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

Binomial distribution cont.

In the case of a fair coin, the probability of success for k events is

$$P(k) = \binom{n}{k} \times p^n$$

Consider the combinations and probabilities for a pair of fair coins:

- possible events are: TT, TH, HT, HH
- probability of each of the four events is $1/2 \times 1/2$

Head count	Combinations	Event prob.[1]	Probability
0	1	1/4	1/4
1	2	1/4	1/2
2	1	1/4	1/4

1 Event prob. is the probability for each and every event

Binomial distribution cont.

Instead of using

$$P(k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

an easier way in *R* is

dbinom(0:2, 2, 0.5)

- *dbinom* obtains the density value for a binomial distribution
- the density value is the probability for a particular outcome, say zero successes
- 0:2 generates the outcomes of interest, in this case all possible outcomes: 0, 1, 2
- the second argument, the 2, is the number of trials, or coins in this case
- the last argument, 0.5 is the probability of success, or p

R Binomial Distribution functions

To get this help in, say *R*Studio, type

?*dbinom*

R: The Binomial Distribution ▾ Find in Topic

Binomial {stats} R Documentation

The Binomial Distribution

Description

Density, distribution function, quantile function and random generation for the binomial distribution with parameters `size` and `prob`.

This is conventionally interpreted as the number of 'successes' in `size` trials.

Usage

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

Binomial problem

Imagine we have three hard drives and each has a 10% probability of failing after one year.

What is the probability of having 0, 1, 2 and 3 failures after a year?

$$P(k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

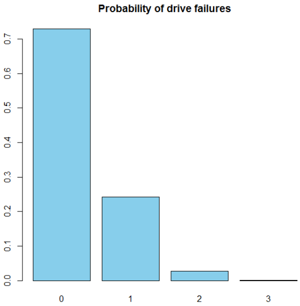
Consider

- k being the outcomes of interest, hence 0:3
- n being the trials or number of drives, hence 3
- say success is the failure of a hard drive
  $\therefore$ p = 0.1

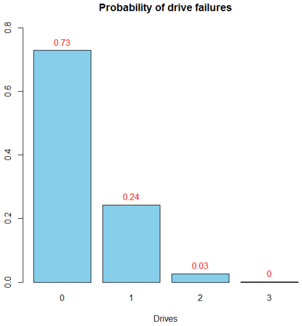
Or using *R* via *dbinom*(k, n, p)
giving the respective probabilities of 0.729, 0.243, 0.027 and 0.001

Binomial problem cont.

```r
d <- dbinom(0:3, 3, 0.1)
barplot(d, names.arg = 0:3, col = 'skyblue',
        main = 'Probability of drive failures')
```




```r
d <- dbinom(0:3, 3, 0.1)
res <- barplot(d, names.arg = 0:3, col = 'skyblue',
               main = 'Probability of drive failures',
               ylim = c(0, 0.8), xlab = 'Drives')
text(res, d, round(d, 2), pos = 3, col = 'red')
```

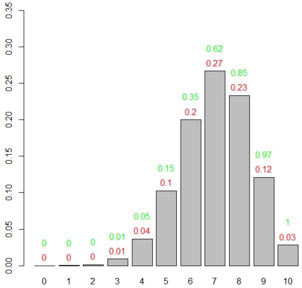


Binomial problem cont.

```r
# Simulate flipping a coin ten times,
# where probability of success is 0.7
outcomes <- 0:10
p <- 0.7

# Individual outcome probabilities
d <- dbinom(outcomes, 10, p)
res <- barplot(d, names.arg = outcomes, ylim = c(0, 0.35))
text(res, d, round(d, 2), pos = 3, col = 'red')

# Cumulative probabilities
prob <- pbinom(outcomes, 10, p)
text(res, d + 0.02, round(prob, 2), pos = 3, col = 'green')
```



The Birthday Problem

Imagine this scenario:

- A gathering of people
- How many people in the gathering are needed so
  - there is a probability of at least 50%
  - in finding at least two people with the same birthday?

The Birthday Problem analogy

Consider a simplifying analogy

- Imagine a world where there is only 6 days in a year
- $\therefore$ a die models possible birthdays
- Limit the gathering to no more than five people
- $\therefore$ each individual is modeled by a die

The Birthday Problem analogy cont.

```r
# Birthday problem analogy
die <- 1:6        # Analogous to birthdays; just six possible birthdays
rolls <- 5        # Number people at the gathering
trials <- 100000  # Number of replicates in order to provide accuracy

d <- replicate(trials,
               {
                 # Randomly decide birthdays for each person
                 res <- sample(die, size = rolls,
                               replace = TRUE)
                 # Determine if any birthdays matched someone else's
                 length(unique(res)) != rolls
               })

res <- table(d) # determine number of matches
res[2] / trials # proportion of matches
```

The Birthday Problem analogy cont.