wk1_notes: Foundation for Data Science

D. ODay

10/11/2021

Reading Notes

Will need the following packages for these notes * cobmbinat *gtools

```
library('combinat')
```

```
##
## Attaching package: 'combinat'

## The following object is masked from 'package:utils':
##
##     combn
```

```
library('gtools')
```

Permuations

We can calculate a permutation doing the following. Remember is simply the factorial of the number of objects.

```
perms = combinat::permn(c("A", "B", "C"))

head(perms)
```

```
## [[1]]
## [1] "A" "B" "C"
##
## [[2]]
## [1] "A" "C" "B"
##
## [[3]]
## [1] "C" "A" "B"
##
## [[4]]
## [1] "C" "B" "A"
##
## [[5]]
## [1] "B" "C" "A"
##
## [[6]]
## [1] "B" "A" "C"
```

Binomial Coefficient

A binomial coefficient is represented by $\binom{n}{x}$. Or in other words, "n choose x". The full formula is

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

Definition: It gives the number of ways that x objects can be chosen from a population of n objects

Intuition: Will use '5 choose 3' as an example. The numerator gives us the total permutations. When we have picked three, there are 2 ways to order the remaining choices. Thus, 5! is over counting what we want (3!), by 2!. Therefore, we divide by 2! which is the (n-x)! part. Then we also divide by 3! since we don't care about the order of the groups (x! in the generic equation)

```r
#Generate people labeled 1 to 5
committees = gtools::combinations(n=5, r=3)

#Should get choose(5,3) = factorial(5)/(factorial(3) * factorial(2)) = 10
committees
```

```
##       [,1] [,2] [,3]
##  [1,]    1    2    3
##  [2,]    1    2    4
##  [3,]    1    2    5
##  [4,]    1    3    4
##  [5,]    1    3    5
##  [6,]    1    4    5
##  [7,]    2    3    4
##  [8,]    2    3    5
##  [9,]    2    4    5
## [10,]    3    4    5
```

Sampling Table

When Order Matters with replacement:

$$n^k$$

Basically, every single time we will have n options and we are picking k times. Since order matters,we aren't over counting.

When Order Matters Without Replacement:

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

Same as a regular binomial coefficient except without *k!*. This is because *k!* was used to divide out the over counting of results from differences of order. But we want them now since order matters.

When Order Doesn't Matter With Replacement:

$$\binom{n+k-1}{k}$$

This box is known as the 'Bose-Einstein' result

When Order Doesn't Matter Without Replacement:

$$\binom{n}{k}$$

The basic binomial coefficient counts the number of ways to select k objects from a population of n objects where we are not replacing the objects and the order is not important

Lecture Notes

Intro to Probability

What is Statistics?

Statistics: Statistics is the science of using data effectively to gain new knowledge.

Population: Those individuals or objects from which we want to acquire information or draw a conclusion. A **subset** of it is called a sample

In probability we make assumptions about the population, then we ask about the nature of a sample. We want to say, with some degree of confidence, whether the whole population has this characteristic or not.

Sample Spaces and Events

Probability studies randomness and uncertainty by giving these concepts a mathematical foundation. Gives us a framework to quantify uncertainty.

Terminology

Experiment: any action or process that generates observations **Sample Space:** denoted with S, is the set of all possible outcomes, of an experiment. **Event:** any possible outcome or combination of outcomes, of an experiment **Cardinality:** the number of outcomes it contains. *|S|* represents the cardinality of the sample space.

Examples

Experiment 1: Flip a coin once

- S = 0,1
- |S| = 2

Experiment 2: Flip a coin twice

- S = 00, 01, 10, 11
- |S| = 4

Experiment 3: Flip a coin until we get a tail

- S = 1, 01, 001, 0001,...
- |S| = infinity

Axioms of Probability

What is Probability?

The goal is to assign some number, P(A), called the probability of the event A, which will give a precise measure to the chance that A will occur. In statistics, we draw a sample from a population, and give an estimate. You'll be able to understand statistics more thoroughly and deeply if you first understand probabilities

- Start with an experiment that generates outcomes
- Organize all of the outcomes into a sample space, S
- Let A be some even contained in S. That is, A is some collection of outcomes from the experiment

Axioms of Probability

Axiom 1: For any event A,

$$0 \leq P(A) \leq 1$$

Axiom 2:

$$P(S) = 1$$

Axiom 3: If A_1, A_2, ..., A_n are a collection of n mutually exclusive events (i.e. the intersection of any two is the empty set), then

$$P(\cup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k)$$

Consequences of Axioms

- $A \cap A^c = \emptyset$ and $A \cup A^c = S$ so, $1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$ which implies $P(A^c) = 1 - P(A)$
- if $A \cap B = \emptyset$, then $P(A \cap B) = 0 = P(\emptyset)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

These are useful as they help us calculate many probabilities

Example 2 notes

If we have P(A), P(B), P(C), $P(A \cap B)$, $P(A \cap C)$, $P(B \cap C)$, $P(A \cap B \cap C)$. Then the probability of at least one of these events occuring is:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

We do this because we must at subtract the intersection because $P(A \cap B \cap C)$ was counted in three times with P(A), P(B), P(C) and then counted out three times with $P(A \cap B)$, $P(A \cap c)$, $P(B \cap C)$. Thus, we need to add it back in.

Counting: Permutations and Combinations

Counting

If a sample space S has N single events, and if each of these events is equally likely to occur, then we need only count the number of events to find the probability (let A be an event in S, $P(E_k) = 1/N$, $S = \{E_1, E_2, \ldots, E_n\}$ and k=1,2,...,N), then

$$P(A) = \frac{number of simple events in A}{N}$$

Example Experiment: Roll a six-sided die twice
|S| = 36
$S = \{(i,j) | i,j \in \{1,2,3,4,5,6\}\}$
Each of the outcomes is equally likely • Let A be the even of rolling a 1 on the first roll
$P(A) = P(\{11, 12, 13, 14, 15, 16\}) = \frac{1}{6}$

Permutations

Any ordered sequence of k objects taken from a set of n distinct objects is called a permutation of size k. Notation: $P_{k,n}$

Example

Organization has 60 people, 1 person selected as president, another as vice president and another as treasurer. What is the cardinality of the sample space (i.e. how many ways can this be done?)?
Answer: $P_{3,60} = 60 \cdot 59 \cdot 58 = \frac{60!}{57!}$

Combinations

Given same number of people, we just want to choose 3 people. Order doesn't matter (no roles). Notation $C_{k,n}$.
Answer: $|S_c| = \frac{60!}{57!3!} = \binom{60}{3}$

Example Continued

Suppose we have the same 60 people, 35 are female and 25 are male. We need to select a committee of 11 people.
• How many ways can such a committee be formed?
The number of committees of 11 = $\binom{60}{11}$
• What is the probability that a randomly selected committee will contain at least 5 men and at least 5 women? (Assume each committee is equally likely)

$$P(\text{at least 5 men and at least 5 women on the committee}) =$$
$$P(\text{5 men and 6 women}) + P(\text{6 men and 5 women}) =$$
$$\frac{\binom{25}{5}\binom{35}{6}}{\binom{60}{11}} + \frac{\binom{25}{6}\binom{35}{5}}{\binom{60}{11}}$$

We had to add the probabilities together because there were two different possible combinations for the answer

Another Example

A city has bought 20 buses. Shortly after being put into service, some of them develop cracks in the frame. the buses are inspected and 8 have visible crack.

• How many ways can the city select a sample of 5 for thorough inspection? (Assume each bus is equally likely to be chosen)

$$\binom{20}{5} = |S|$$

• If 5 buses are chosen at random, find the probability that exactly 4 have cracks.

$$P(4 \text{ with cracks}) = \frac{\binom{8}{4}\binom{12}{1}}{\binom{20}{5}}$$

• If 5 buses are chosen at random, find the probability that at least 4 have cracks

$$P(\text{at least 4 have cracks}) = \frac{\binom{8}{4}\binom{12}{1}}{\binom{20}{5}} + \frac{\binom{8}{5}}{\binom{20}{5}}$$

Homework Notes

Learning probability simulation within R

#Simulating a Die Simulating a regular die with 5 trials ran.

```
die = 1:6
sample(die, size = 5, replace = TRUE)
```

```
## [1] 1 3 6 1 5
```

Simulating an unfair die

```
die = 1:6
result = sample(die, size = 500, replace = TRUE, prob=c(.1, .1, .5, .1,.1,.1))
table(result)
```

```
## result
##   1   2   3   4   5   6
##  34  53 275  44  49  45
```

```
table(result) / 500
```

```
## result
##     1     2     3     4     5     6
## 0.068 0.106 0.550 0.088 0.098 0.090
```