AE 10: Probability

Dav King

2/10/2022

Learning goals

- Introduce probabilities and how we can use them to understand categorical data
- Create a contingency table using `pivot_wider()` and `kable()`
- Use a contingency table to explore the relationship between two categorical variables.

Introduction

```
library(tidyverse)
library(knitr)
```

```
sta199 <- read_csv("sta199-fa21-year-major.csv")
```

For this Application Exercise, we will look at the year in school and majors for students taking STA 199 in Fall 2021. The data set includes the following variables:

- `section`: STA 199 section
- `year`: Year in school
- `major_category`: Major / academic interest.
  - For the purposes of this AE, we'll call this the student's "major".

Definitions

- The **probability** of an event tells us how likely an event is to occur, and it can take values from 0 to 1, inclusive. It can be viewed as
  - the proportion of times the event would occur if it could be observed an infinite number of times.
  - our degree of belief an event will happen.
- An **event** is the basic element to which probability is applied, e.g. the result of an observation or experiment.
  - Example: **A** is the event a student in STA 199 is a sophomore.
- A **sample space** is the set of all possible outcomes. Each outcome in the sample space is **disjoint** or **mutually exclusive** meaning they can't occur simultaneously.
  - Example: The sample space for year is {First-year, Sophomore, Junior, Senior}

Exercise 1

Let's take a look at the majors. Note that we have categorized majors so that each student can only be in one major category.

- What is the sample space for major? You can use code to identify the sample space.

```
unique(sta199$major_category)
```

```
## [1] "other"              "pubpol only"        "stats only"
## [4] "compsci only"       "undecided"          "stat + other major"
## [7] "econ only"
```

- Let's make a table that includes the majors, the number of students in each, and the associated probabilities.

```
sta199 %>%
  count(major_category) %>%
  mutate(prop_major = n/sum(n))
```

```
## # A tibble: 7 x 3
##   major_category         n prop_major
##   <chr>              <int>      <dbl>
## 1 compsci only          40     0.162
## 2 econ only             15     0.0607
## 3 other                 98     0.397
## 4 pubpol only           38     0.154
## 5 stat + other major    36     0.146
## 6 stats only            10     0.0405
## 7 undecided             10     0.0405
```

- What is the probability a randomly selected STA 199 student is a "pubpol only" major?

```
sta199 %>%
  count(major_category) %>%
  mutate(prop_major = n/sum(n)) %>%
  filter(major_category == "pubpol only")
```

```
## # A tibble: 1 x 3
##   major_category     n prop_major
##   <chr>          <int>      <dbl>
## 1 pubpol only       38      0.154
```

- What is the probability a randomly selected STA 199 student is studying statistics?

```
sta199 %>%
  count(major_category) %>%
  mutate(prop_major = n/sum(n)) %>%
  filter(major_category == "stat + other major" |
           major_category == "stats only")
```

```
## # A tibble: 2 x 3
##   major_category         n prop_major
##   <chr>              <int>      <dbl>
## 1 stat + other major    36     0.146
## 2 stats only            10     0.0405
```

- What is the probability a randomly selected STA 199 student is not a “pubpol only” major?

```
sta199 %>%
  count(major_category) %>%
  mutate(prop_major = n/sum(n)) %>%
  filter(major_category != "pubpol only") %>%
  mutate(answer = sum(prop_major)) %>%
  summarize(answer) %>%
  slice(1)
```

```
## # A tibble: 1 x 1
##   answer
##    <dbl>
## 1  0.846
```

Exercise 2

Now let’s make at table looking at the relationship between year and major.

```
sta199 %>%
  count(year, major_category)
```

```
## # A tibble: 23 x 3
##    year       major_category         n
##    <chr>      <chr>              <int>
##  1 First-year compsci only           8
##  2 First-year econ only              6
##  3 First-year other                 39
##  4 First-year pubpol only           22
##  5 First-year stat + other major    26
##  6 First-year stats only             7
##  7 First-year undecided              5
##  8 Junior     compsci only           7
##  9 Junior     econ only              3
## 10 Junior     other                 12
## # ... with 13 more rows
```

We’ll reformat the data into a **contingency table**, a table frequently used to study the association between two categorical variables. In this contingency table, each row will represent a year, each column will represent a major, and each cell is the number of students have a particular combination of year and major.

To make the contingency table, we will use a new function in `dplry` called `pivot_wider()`. It will take the data frame produced by `count()` that is current in a “long” format and reshape it to be in a “wide” format.

We will also use the `kable()` function in the `knitr` package to neatly format our new table.

```
sta199 %>%
  count(year, major_category) %>%
  pivot_wider(id_cols = c(year, major_category),#how we identify unique obs
              names_from = major_category, #how we will name the columns
              values_from = n, #values used for each cell
              values_fill = 0) %>% #how to fill cells with 0 observations
  kable() # neatly display the results
```

year	compsci only	econ only	other	pubpol only	stat + other major	stats only	undecided
First-year	8	6	39	22	26	7	5
Junior	7	3	12	4	1	0	0
Senior	2	0	5	1	1	0	0
Sophomore	23	6	42	11	8	3	5

- How many students in STA 199 are first-years and in the "econ only" majors category. 6
- How many students in STA 199 are in the "other" major category? 98

Exercise 3

For each of the following exercises:

(1) Calculate the probability using the contingency table above.

(2) Then write code to check your answer using the `sta199` data frame and `dplyr` functions.

- What is the probability a randomly selected STA 199 student is a sophomore? .397

```
sta199 %>%
  count(year) %>%
  mutate(prop_year = n/sum(n))
```

```
## # A tibble: 4 x 3
##   year            n prop_year
##   <chr>       <int>     <dbl>
## 1 First-year    113    0.457
## 2 Junior         27    0.109
## 3 Senior          9    0.0364
## 4 Sophomore      98    0.397
```

- What is the probability that a randomly selected STA 199 student is a "compsci only" major? .162

```
sta199 %>%
  count(major_category) %>%
  mutate(prop_major = n/sum(n))
```

```
## # A tibble: 7 x 3
##   major_category          n prop_major
##   <chr>               <int>      <dbl>
## 1 compsci only           40     0.162
## 2 econ only              15     0.0607
## 3 other                  98     0.397
## 4 pubpol only            38     0.154
## 5 stat + other major     36     0.146
## 6 stats only             10     0.0405
## 7 undecided              10     0.0405
```

- What is the probability that a randomly selected STA 199 student is a sophomore **or** a "compsci only" major? .466

```
sta199 %>%
  mutate(soph_cs = case_when(
    year == "Sophomore" | major_category == "compsci only" &
                                    year != "Sophomore" ~ 1,
    T ~ 0
  )) %>%
  count(soph_cs) %>%
  mutate(answer = n/sum(n))
```

```
## # A tibble: 2 x 3
##   soph_cs     n answer
##     <dbl> <int>  <dbl>
## 1       0   132  0.534
## 2       1   115  0.466
```

- What is the probability that a randomly selected STA 199 student is a sophomore **and** a "compsci only" major? .093

```
sta199 %>%
  count(year, major_category) %>%
  mutate(prop_maj_year = n/sum(n)) %>%
  filter(year == "Sophomore" & major_category == "compsci only")
```

```
## # A tibble: 1 x 4
##   year      major_category     n prop_maj_year
##   <chr>     <chr>          <int>         <dbl>
## 1 Sophomore compsci only      23        0.0931
```

Resources

- Notes on `pivot_wider` and `pivot_longer`
  - Click here for slides
  - Click here for video