Probability

02.09.2022

Click here while you wait

Bulletin

- Reminder:
  - lab 04 is due Friday Feb 11
  - double check: link your group members on gradescope
- Mid-course survey on sakai under quizzes. Due Friday Feb 11.
- AE grades

Today

By the end of today you will

- have a working understanding of the terms **probability**, **sample space**, **event**, **population** and **sample**.
- compute **probabilities** of **events** from data
- create a contingency table using `pivot_wider()` and `kable()`
- use a contingency table to explore the relationship between two categorical variables.

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.
  - We use capital letters, e.g. A to denote events.
  - For any event A and its complement, A^C, $\text{Pr(A)} + \text{Pr}(A^C) = 1$.
- 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}, each item brackets is a distinct **outcome** from the questionnaire.
  - The probability of the entire sample space is 1.

Introduction

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

```
sta199 <- read_csv("https://sta199-sp22-003.netlify.app/class_data/sta199-sp22-form.csv")
```

For this Application Exercise, we will look at our newly collected data.

Data includes

- `year`: Year in school
- `animal`: Whether you prefer cats or dogs
- `tv`: Favorite TV genre
- `major`: probable major (statistical science or not)

Exercise 1

Give two examples of an event from the data set.

Let A be the event that a student's favorite animal is a dog.

Let A be the event that a student is pursuing a statsci major.

Let A be the event that a student is a sophomore *or* senior.

Exercise 2

Let's take a look at favorite TV genre. Note that we have categorized genres so that each person can only have one favorite genre.

- What is the sample space for favorite TV genre? You can use code to identify the sample space.

```
sta199 %>%
  count(tv) %>%
  select(tv)
```

```
## # A tibble: 7 x 1
##   tv
##   <chr>
## 1 Action/Adventure
## 2 Anime
## 3 Comedy
## 4 Drama
## 5 News
## 6 Sports
## 7 Thriller
```

Exercise 3

- Let's make a table that includes the TV genre, the number of people who prefer each, and the associated probabilities.

```
# code here
sta199 %>%
  count(tv) %>%
  # summarize(tv, n, prob = n / sum(n))
  mutate(prob = n / sum(n))
```

```
## # A tibble: 7 x 3
##   tv                   n   prob
##   <chr>            <int>  <dbl>
## 1 Action/Adventure     4 0.0909
## 2 Anime                8 0.182
## 3 Comedy               7 0.159
## 4 Drama               13 0.295
## 5 News                 1 0.0227
## 6 Sports               7 0.159
## 7 Thriller             4 0.0909
```

```
  # summarize(sum(prob))
```

Exercise 4

How large is the sample space of any individual's response? Can we check this in `R`?

```
# code here
sta199 %>%
  count(year, animal, tv, major) %>%
  nrow()
```

```
## [1] 24
```

The sample space of our data is 24 different outcomes.

Exercise 5

- What is the probability a randomly selected STA 199 student favors cats?

```
# code here
sta199 %>%
  count(animal) %>%
  mutate(prob = n / nrow(sta199))
```

```
## # A tibble: 2 x 3
##   animal     n  prob
##   <chr>  <int> <dbl>
## 1 Cats       5 0.114
## 2 Dogs      39 0.886
```

- What is the probability a randomly selected STA 199 student is not a senior and prefers dogs?

Let A be the event that someone is not a senior and prefers dogs.

```
# code here
sta199 %>%
  mutate(A =  ((year != "Senior") & (animal == "Dogs"))
  ) %>%
  summarize(prob = mean(A))
```

```
## # A tibble: 1 x 1
##    prob
##   <dbl>
## 1 0.841
```

- What is the probability a randomly selected STA 199 student is a first year and a statistics major?

Let B be the event someone is a first-year and a stat major.

```
# code here
sta199 %>%
  mutate(B =  ((year == "First-year") & (major == "Statistical Science"))
  ) %>%
  summarize(prob = mean(B))
```

```
## # A tibble: 1 x 1
##    prob
##   <dbl>
## 1 0.114
```

Exercise 6

Now let's make at table looking at the relationship between year and favorite tv.

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

```
## # A tibble: 16 x 3
##    year       tv                   n
##    <chr>      <chr>            <int>
##  1 First-year Action/Adventure     2
##  2 First-year Anime                5
##  3 First-year Comedy               6
##  4 First-year Drama                7
##  5 First-year Sports               5
##  6 First-year Thriller             3
##  7 Junior     Anime                2
##  8 Junior     Comedy               1
##  9 Junior     Sports               1
## 10 Senior     Drama                1
## 11 Senior     News                 1
## 12 Senior     Thriller             1
## 13 Sophomore  Action/Adventure     2
## 14 Sophomore  Anime                1
## 15 Sophomore  Drama                5
## 16 Sophomore  Sports               1
```

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 tv show, 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, tv) %>%
  pivot_wider(id_cols = c(year, tv),#how we identify unique obs
              names_from = tv, #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	Action/Adventure	Anime	Comedy	Drama	Sports	Thriller	News
First-year	2	5	6	7	5	3	0
Junior	0	2	1	0	1	0	0
Senior	0	0	0	1	0	1	1
Sophomore	2	1	0	5	1	0	0

- How many students in STA 199 are juniors and like dramas?

Exercise 7

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?

```
# code here
9/44
```

```
## [1] 0.2045455
```

```
count(sta199, year) %>%
  filter(year == "Sophomore") %>%
  pull(n) / nrow(sta199)
```

```
## [1] 0.2045455
```

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

```
# code here
sta199 %>%
  count(major) %>%
  mutate(prob = n / sum(n))
```

```
## # A tibble: 2 x 3
##   major                       n  prob
##   <chr>                   <int> <dbl>
## 1 Not Statistical Science    38 0.864
## 2 Statistical Science         6 0.136
```

- What is the probability that a randomly selected STA 199 student is a sophomore **or** a statistics major?

```
# code here
total_obs = nrow(sta199)

sta199 %>%
  mutate( A =  (major == "Statistical Science" | year == "Sophomore")) %>%
  summarize(mean(A))
```

```
## # A tibble: 1 x 1
##   `mean(A)`
##       <dbl>
## 1     0.341
```

```
sta199 %>%
  filter( (major == "Statistical Science" | year == "Sophomore")) %>%
   nrow() / total_obs
```

```
## [1] 0.3409091
```

- What is the probability that a randomly selected STA 199 student is a sophomore **and** and a statistics major?

```
# code here
sta199 %>%
  filter((major == "Statistical Science" & year == "Sophomore")) %>%
  nrow() / nrow(sta199)
```

```
## [1] 0
```

More definitions

Population: the entire group you want to learn about. Often, it's useful to think the population is "truth"

Sample: Your sample of the population from which you draw inference.

Resources

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