11.1.1: Empirical Frequency (Section 10.2.2)

Let's walk through how we computed empirical frequency of rain in San Francisco.

First we load the data:

```
# we will remove the STATION and NAME variables
# since they are identical for all rows
SFrain <- read_csv("data/SanFranciscoRain/1329219.csv") %>%
  dplyr::select(-STATION, -NAME)

glimpse(SFrain)
```

```
## Observations: 365
## Variables: 2
## $ DATE <date> 2017-01-01, 2017-01-02, 2017-01-03, 2017-01…
## $ PRCP <dbl> 0.05, 0.10, 0.40, 0.89, 0.01, 0.00, 0.82, 1.…
```

We see that the data frame contains a variable called `PRCP` which denotes the amount of rain each day. Let's create a new variable called `rainToday` that denotes whether the amount of precipitation was above zero:

```
SFrain <-
  SFrain %>%
  mutate(rainToday = as.integer(PRCP > 0))

glimpse(SFrain)
```

```
## Observations: 365
## Variables: 3
## $ DATE      <date> 2017-01-01, 2017-01-02, 2017-01-03, 20…
## $ PRCP      <dbl> 0.05, 0.10, 0.40, 0.89, 0.01, 0.00, 0.8…
## $ rainToday <int> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, …
```

Now we will summarize the data to compute the probability of rain:

```
pRainInSF <-
  SFrain %>%
  summarize(
    pRainInSF = mean(rainToday)
  ) %>%
  pull()

pRainInSF
```

```
## [1] 0.2
```

This page titled 11.1.1: Empirical Frequency (Section 10.2.2) is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Russell A. Poldrack via source content that was edited to the style and standards of the LibreTexts platform.