Homework #03: The Joy of Probability

due February 24 11:59 PM

Dav King

2/17/2022

Load Packages and Data

```
library(tidyverse)
library(fivethirtyeight)
library(viridis)
library(knitr)
```

Exercise 1

```
bob_ross %>%
  count(tree) %>%
  mutate(tree_prob = n/sum(n))
```

```
## # A tibble: 2 x 3
##    tree     n tree_prob
##   <int> <int>     <dbl>
## 1     0    42     0.104
## 2     1   361     0.896
```

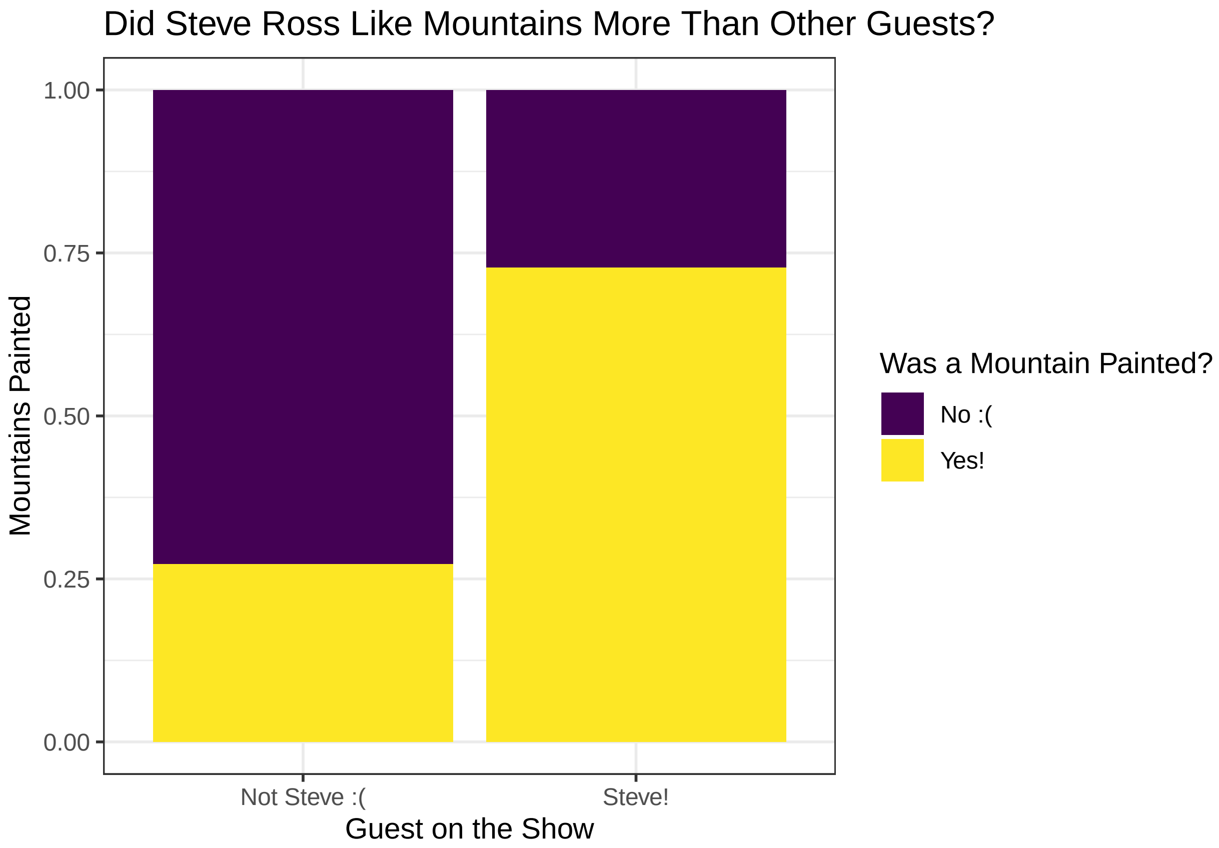
There are 403 episodes, in 361 of which a tree was painted. The probability that a randomly selected episode featured a tree is 0.896.

Exercise 2

```
bob_ross %>%
  filter(guest == 1) %>%
  summarize(prob_steve = mean(steve_ross))
```

```
## # A tibble: 1 x 1
##   prob_steve
##        <dbl>
## 1        0.5
```

```
bob_ross %>%
  filter(guest == 1) %>%
  mutate(steve = if_else(steve_ross == 1, "Steve!", "Not Steve :(")) %>%
  mutate(mount = if_else(mountain == 1, "Yes!", "No :(")) %>%
  ggplot(aes(x = steve, fill = factor(mount))) +
  geom_bar(position = "fill") +
  labs(title = "Did Steve Ross Like Mountains More Than Other Guests?",
       x = "Guest on the Show", y = "Mountains Painted",
       fill = "Was a Mountain Painted?") +
  scale_fill_viridis(discrete = T) +
  theme_bw()
```



The conditional probability of Steve Ross being the guest on the show is 0.5. He did indeed like to paint mountains more than other guests - or at least, he did so with far greater frequency.

Exercise 3

```
ross_paintings <- bob_ross %>%
  filter(guest == 0)
```

Exercise 4

```
ross_paintings %>%
  filter(cirrus == 1 & cumulus == 1) %>%
  select(episode, cirrus, cumulus)
```

```
## # A tibble: 1 x 3
##   episode cirrus cumulus
##   <chr>    <int>   <int>
## 1 S02E09       1       1
```

They are not disjoint - though there is only one instance in which the two clouds were painted together, there was still such an instance; thus, they cannot be mutually exclusive.

Exercise 5

```
M <- ross_paintings %>%
  filter(cabin == 1) %>%
  count(cabin)

X <- ross_paintings %>%
  filter(cabin == 1) %>%
  filter(lake == 1) %>%
  count(lake)

set.seed(2182022) # don't change the seed
num_lakes = rbinom(100000, M$n, prob = 0.5)
cabin_lakes = data.frame(num_lakes)

cabin_lakes %>%
  mutate(prob_lakes = if_else(num_lakes <= X$n, 1, 0)) %>%
  summarize(answer = mean(prob_lakes))
```

```
##    answer
## 1 0.00999
```

The probability that Bob Ross painted `X` or fewer lakes, given that he flipped a fair coin to decide whether or not to paint one every time, is 0.00999. 68 paintings feature a cabin, and 24 of those also feature a lake.

Exercise 6

```
A <- ross_paintings %>%
  summarize(prob_mountain = mean(mountain), prob_river = mean(river))

B <- ross_paintings %>%
  filter(river == 1) %>%
  summarize(mountain_given_river = mean(mountain))

(B$mountain_given_river * A$prob_river)/A$prob_mountain
```

```
## [1] 0.3221477
```

```
A$prob_mountain
```

```
## [1] 0.3910761
```

```
B$mountain_given_river
```

```
## [1] 0.3870968
```

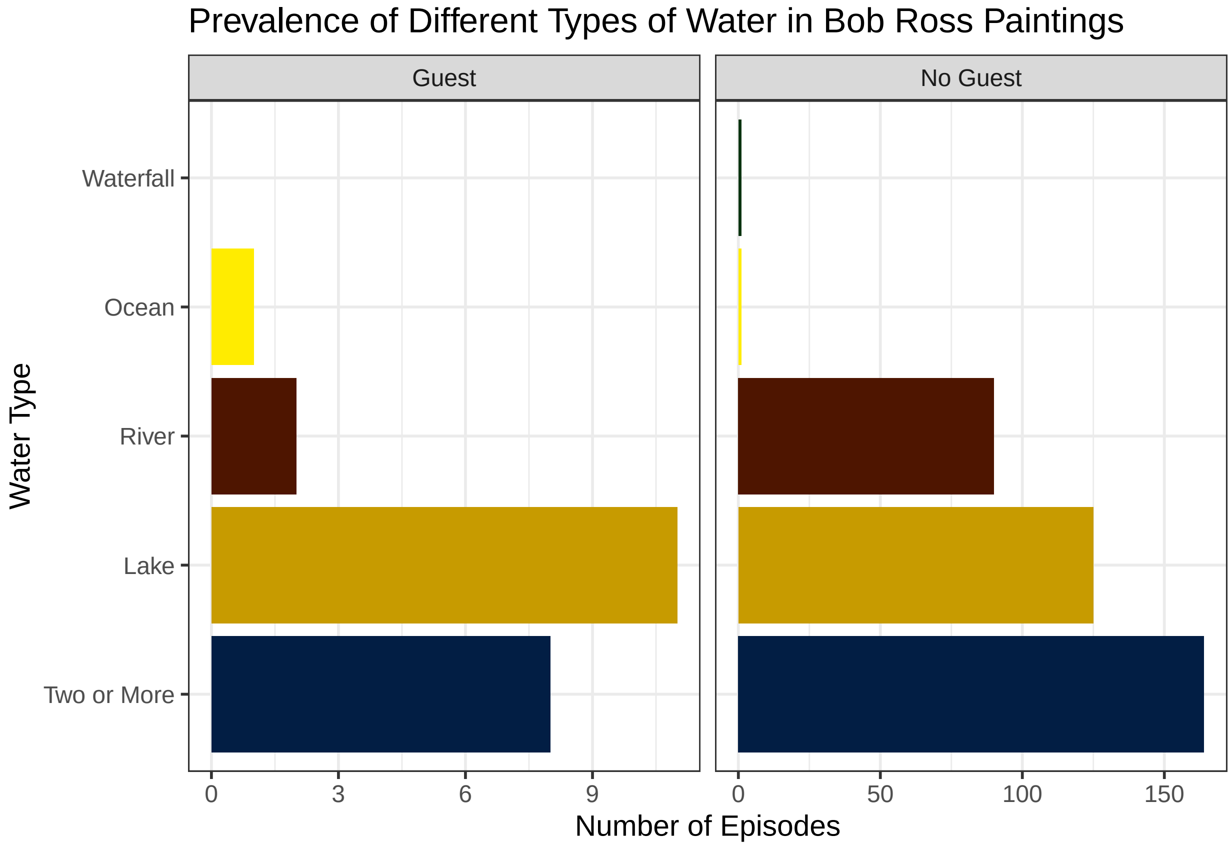
Bob Ross does not paint mountains independent of whether or not he paints rivers. The probability that he paints a mountain is 0.391. If they were independent, we would expect the probability of a mountain to be painted given that a river was painted to also be 0.391. However, in paintings where a river was painted, this probability is 0.387. Though this difference is slight, it is still present; thus, we can safely say that these two events are not independent.

Exercise 7

Does the relative prevalence of different types of water in this show's paintings vary according to the presence of guests on the show and/or an episode's place within a season?

```
ross3 <- bob_ross %>%
  mutate(water_type = case_when(
    waves == 1 & ocean == 0 & waterfall == 0 & river == 0 &
      lake == 0 ~ "Waves",
    waves == 0 & ocean == 1 & waterfall == 0 & river == 0 &
      lake == 0 ~ "Ocean",
    waves == 0 & ocean == 0 & waterfall == 1 & river == 0 &
      lake == 0 ~ "Waterfall",
    waves == 0 & ocean == 0 & waterfall == 0 & river == 1 &
      lake == 0 ~ "River",
    waves == 0 & ocean == 0 & waterfall == 0 & river == 0 &
      lake == 1 ~ "Lake",
    T ~ "Two or More"
  )) %>%
  mutate(nom_guest = if_else(guest == 0, "No Guest", "Guest"))

ross_palette <- c("River" = "#4E1500", "Ocean" = "#FFEC00",
                  "Waterfall" = "#0A3410", "Two or More" = "#021E44",
                  "Lake" = "#C79B00")

ross3 %>%
  group_by(water_type, nom_guest) %>%
  summarize(water_presence = n()) %>%
  ggplot(aes(x = reorder(water_type, -water_presence),
                  y = water_presence, fill = water_type)) +
  geom_col() +
  labs(title = "Prevalence of Different Types of Water in Bob Ross Paintings",
       x = "Water Type", y = "Number of Episodes") +
  theme_bw() +
  coord_flip() +
  scale_fill_manual(values = ross_palette) +
  guides(fill = "none") +
  facet_wrap(~ nom_guest, scales = "free_x")
```

```
## ‘summarise()‘ has grouped output by ’water_type’. You can override using the ‘.groups‘ argument.
```



```
pop <- count(ross3)

pWater <- function(water){
  ross3 %>%
    filter(water_type == water) %>%
    count(water_type) %>%
    mutate(p = n/pop)
}

pWaterfall <- pWater("Waterfall")
pOcean <- pWater("Ocean")
```

```r
pRiver <- pWater("River")
pLake <- pWater("Lake")
pTwo <- pWater("Two or More")

marginal_p_water <- c(pLake$p$n, pOcean$p$n, pRiver$p$n,
                      pTwo$p$n, pWaterfall$p$n)

ross4 <- ross3 %>%
  count(water_type, nom_guest) %>%
  pivot_wider(names_from = nom_guest, values_from = n)

ross4$marginal_p_water <- marginal_p_water
kable(ross4)
```

water_type	Guest	No Guest	marginal_p_water
Lake	11	125	0.3374690
Ocean	1	1	0.0049628
River	2	90	0.2282878
Two or More	8	164	0.4267990
Waterfall	NA	1	0.0024814

This graph and table suggest that there is likely a difference in the prevalence of each type of water depending on the presence of a guest, but does not necessarily statistically confirm it - this, however, will.

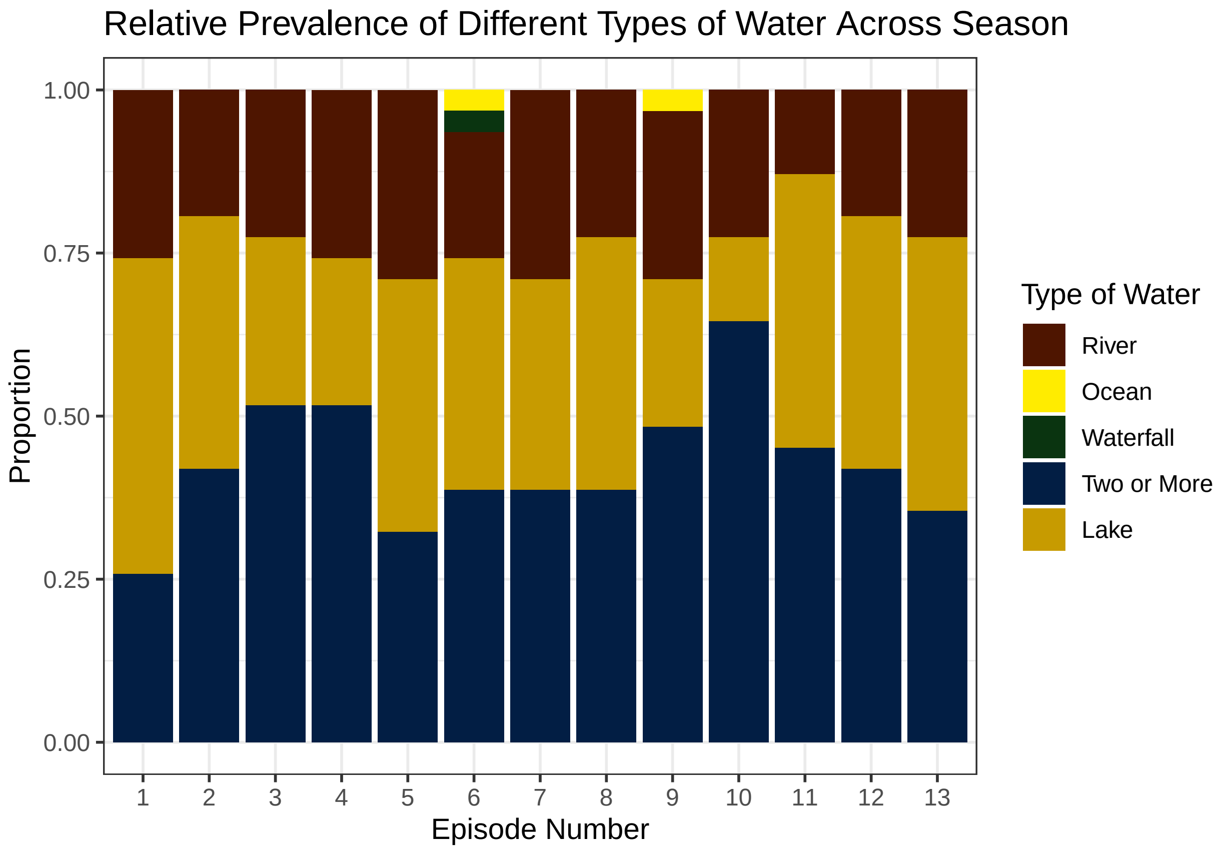
```r
given <- function(g, water){
  ross3 %>%
    filter(nom_guest == g) %>%
    count(water_type) %>%
    mutate(p = n/sum(n)) %>%
    filter(water_type == water)
}

p_guest_ocean <- given("Guest", "Ocean")
p_guest_river <- given("Guest", "River")
p_guest_lake <- given("Guest", "Lake")
p_guest_two <- given("Guest", "Two or More")

guest_probs <- c("X", p_guest_ocean$p, p_guest_river$p,
                 p_guest_lake$p, p_guest_two$p)

p_no_guest_wf <- given("No Guest", "Waterfall")
p_no_guest_ocean <- given("No Guest", "Ocean")
p_no_guest_river <- given("No Guest", "River")
p_no_guest_lake <- given("No Guest", "Lake")
p_no_guest_two <- given("No Guest", "Two or More")

no_guest_probs <- c(p_no_guest_wf$p, p_no_guest_ocean$p, p_no_guest_river$p,
                    p_no_guest_lake$p, p_no_guest_two$p)

A <- "X"
B <- p_guest_ocean$p == p_no_guest_ocean$p
```

```r
C <- p_guest_river$p == p_no_guest_river$p
D <- p_guest_lake$p == p_no_guest_lake$p
E <- p_guest_two$p == p_no_guest_two$p

across_conditions <- c(A, B, C, D, E)

A1 <- "X"
B1 <- p_guest_ocean$p == pOcean$p$n
C1 <- p_guest_river$p == pRiver$p$n
D1 <- p_guest_lake$p == pLake$p$n
E1 <- p_guest_two$p == pTwo$p$n

guest_vs_baseline <- c(A1, B1, C1, D1, E1)

A2 <- p_no_guest_wf$p == pWaterfall$p$n
B2 <- p_no_guest_ocean$p == pOcean$p$n
C2 <- p_no_guest_river$p == pRiver$p$n
D2 <- p_no_guest_lake$p == pLake$p$n
E2 <- p_no_guest_two$p == pTwo$p$n

no_guest_vs_baseline <- c(A2, B2, C2, D2, E2)

cProb_table <- data.frame(ross4$water_type, marginal_p_water, guest_probs,
                          no_guest_probs, guest_vs_baseline,
                          no_guest_vs_baseline, across_conditions)
cProb_table %>%
  rename(
    "Water Type" = ross4.water_type,
    "Marginal Probability" = marginal_p_water,
    "Probability Given Guest" = guest_probs,
    "Probability Given No Guest" = no_guest_probs,
    "Guest vs Baseline" = guest_vs_baseline,
    "No Guest vs Baseline" = no_guest_vs_baseline,
    "Guest vs No Guest" = across_conditions) %>%
  kable()
```

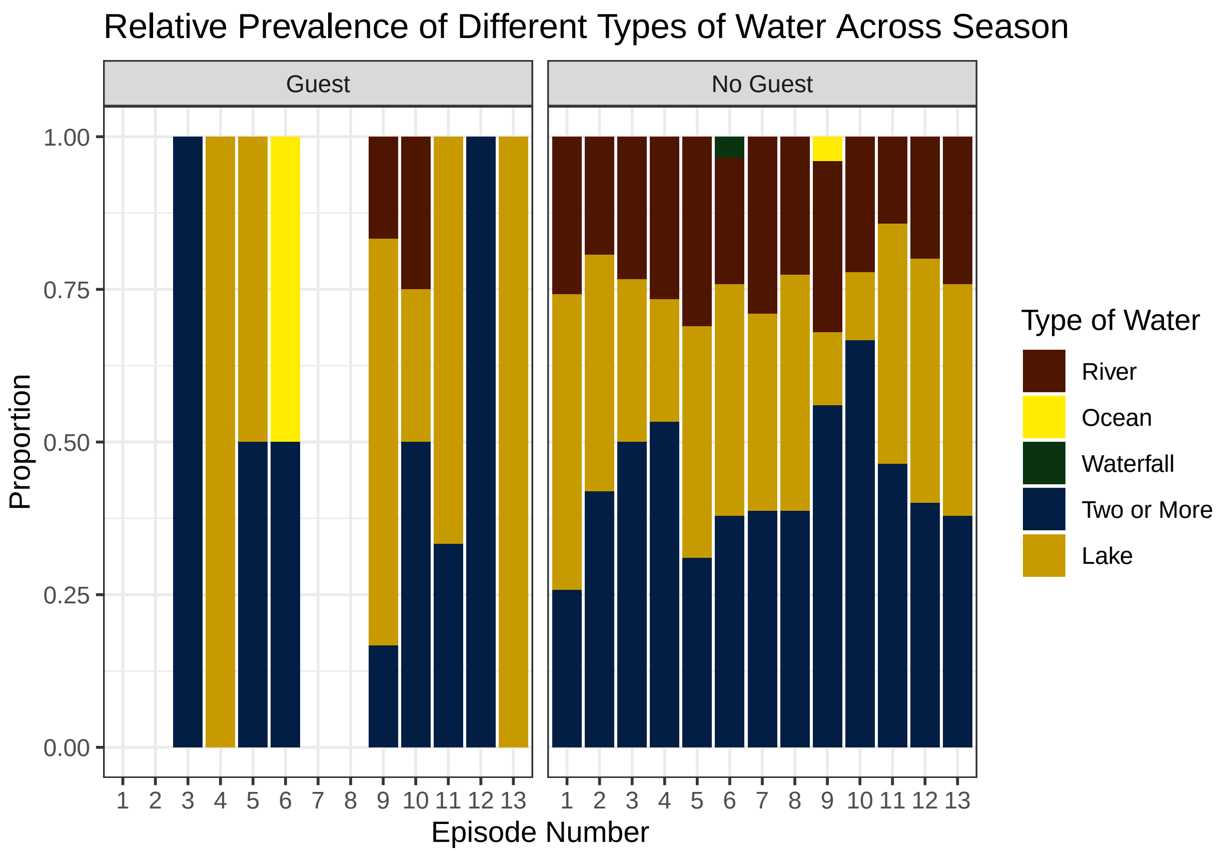
Water Type	Marginal Probability	Probability Given Guest	Probability Given No Guest	Guest vs Baseline	No Guest vs Baseline	Guest vs No Guest
Lake	0.3374690	X	0.0026247	X	FALSE	X
Ocean	0.0049628	0.0454545454545455	0.0026247	FALSE	FALSE	FALSE
River	0.2282878	0.0909090909090909	0.2362205	FALSE	FALSE	FALSE
Two or More	0.4267990	0.5	0.3280840	FALSE	FALSE	FALSE
Waterfall	0.0024814	0.363636363636364	0.4304462	FALSE	FALSE	FALSE

As we can see, in none of these conditions (for any type of water, guest or no guest as compared both to the total dataset and to each other) are the marginal probabilities equal to one another; thus, these two events are not independent of one another.

```r
ross3 %>%
  group_by(water_type, nom_guest) %>%
  mutate(water_presence = n()) %>%
  ggplot(aes(x = factor(episode_num),
             fill = reorder(water_type, water_presence))) +
  geom_bar(position = "fill") +
  labs(title = "Relative Prevalence of Different Types of Water Across Season",
       x = "Episode Number", y = "Proportion", fill = "Type of Water") +
  theme_bw() +
  scale_fill_manual(values = ross_palette)
```



```
ross3 %>%
  group_by(water_type, nom_guest) %>%
  mutate(water_presence = n()) %>%
  ggplot(aes(x = factor(episode_num),
             fill = reorder(water_type, water_presence))) +
  geom_bar(position = "fill") +
  labs(title = "Relative Prevalence of Different Types of Water Across Season",
       x = "Episode Number", y = "Proportion", fill = "Type of Water") +
  theme_bw() +
  scale_fill_manual(values = ross_palette) +
  facet_wrap(~ nom_guest)
```



These once again suggest both that A) there is a difference in the relative prevalence of each type of water across the season and B) that this varies according to whether or not there was a guest on the show; but again, let us turn to statistics:

```
timeline <- ross3 %>%
  #For simplicity's sake, we will group the episodes into three chunks
  mutate(chunked = case_when(
    episode_num <= 4 ~ "Early",
    episode_num <= 8 ~ "Middle",
    T ~ "Late"
  )) %>%
  mutate(chunked = factor(chunked, levels = c("Early", "Middle", "Late"),
                          ordered = T)) %>%
  arrange(chunked)

time_prob <- function(time){
  timeline %>%
    filter(chunked == time) %>%
    count(water_type) %>%
    mutate(p = n/sum(n))
}

early_prob <- time_prob("Early")
middle_prob <- time_prob("Middle")
late_prob <- time_prob("Late")
```

```r
early_prob
```

```
## # A tibble: 3 x 3
##   water_type      n     p
##   <chr>       <int> <dbl>
## 1 Lake           42 0.339
## 2 River          29 0.234
## 3 Two or More    53 0.427
```

```r
middle_prob
```

```
## # A tibble: 5 x 3
##   water_type      n       p
##   <chr>       <int>   <dbl>
## 1 Lake           45 0.363
## 2 Ocean           1 0.00806
## 3 River          31 0.25
## 4 Two or More    46 0.371
## 5 Waterfall       1 0.00806
```

```r
late_prob
```

```
## # A tibble: 4 x 3
##   water_type      n       p
##   <chr>       <int>   <dbl>
## 1 Lake           49 0.316
## 2 Ocean           1 0.00645
## 3 River          32 0.206
## 4 Two or More    73 0.471
```

```r
#In the interest of time, I won't do the logic calculations - we can all see
#that none of those numbers are the same across groups
```

```r
split_time_prob <- function(time, g){
  timeline %>%
    filter(chunked == time & guest == g) %>%
    count(water_type) %>%
    mutate(p = n/sum(n))
}

early_guest <- split_time_prob("Early", 1)
middle_guest <- split_time_prob("Middle", 1)
late_guest <- split_time_prob("Late", 1)

early_no_guest <- split_time_prob("Early", 0)
middle_no_guest <- split_time_prob("Middle", 0)
late_no_guest <- split_time_prob("Late", 0)

early_prob
```

```
## # A tibble: 3 x 3
##   water_type      n     p
##   <chr>       <int> <dbl>
## 1 Lake           42 0.339
## 2 River          29 0.234
## 3 Two or More    53 0.427
```

```
early_guest
```

```
## # A tibble: 2 x 3
##   water_type      n     p
##   <chr>       <int> <dbl>
## 1 Lake            1   0.5
## 2 Two or More     1   0.5
```

```
early_no_guest
```

```
## # A tibble: 3 x 3
##   water_type      n     p
##   <chr>       <int> <dbl>
## 1 Lake           41 0.336
## 2 River          29 0.238
## 3 Two or More    52 0.426
```

```
middle_prob
```

```
## # A tibble: 5 x 3
##   water_type      n       p
##   <chr>       <int>   <dbl>
## 1 Lake           45 0.363
## 2 Ocean           1 0.00806
## 3 River          31 0.25
## 4 Two or More    46 0.371
## 5 Waterfall       1 0.00806
```

```
middle_guest
```

```
## # A tibble: 3 x 3
##   water_type      n     p
##   <chr>       <int> <dbl>
## 1 Lake            1  0.25
## 2 Ocean           1  0.25
## 3 Two or More     2  0.5
```

```
middle_no_guest
```

```
## # A tibble: 4 x 3
##   water_type      n       p
##   <chr>       <int>   <dbl>
## 1 Lake           44 0.367
## 2 River          31 0.258
## 3 Two or More    44 0.367
## 4 Waterfall       1 0.00833
```

```
late_prob
```

```
## # A tibble: 4 x 3
##   water_type      n       p
##   <chr>       <int>   <dbl>
## 1 Lake           49 0.316
## 2 Ocean           1 0.00645
## 3 River          32 0.206
## 4 Two or More    73 0.471
```

```
late_guest
```

```
## # A tibble: 3 x 3
##   water_type      n     p
##   <chr>       <int> <dbl>
## 1 Lake            9 0.562
## 2 River           2 0.125
## 3 Two or More     5 0.312
```

```
late_no_guest
```

```
## # A tibble: 4 x 3
##   water_type      n       p
##   <chr>       <int>   <dbl>
## 1 Lake           40 0.288
## 2 Ocean           1 0.00719
## 3 River          30 0.216
## 4 Two or More    68 0.489
```

The only value held in common (and thus even remotely plausible to be independent) anywhere in this calculation is a 50% chance of two or more types of water in both the early and middle chunks of the season on shows that had guests - clearly, these values cannot be considered independent of one another.