Check-in 7

Jack Rellamas

Remember, **follow the instructions below and use R Markdown to create a pdf document with your code and answers to the following questions on Gradescope.** You may find a template file by clicking "Code" in the top right corner of this page.

Load the flips data using the following code:

```r
library(tidyverse)
flips <- read_csv("https://math167r-s24.github.io/static/flips.csv")
head(flips)
```

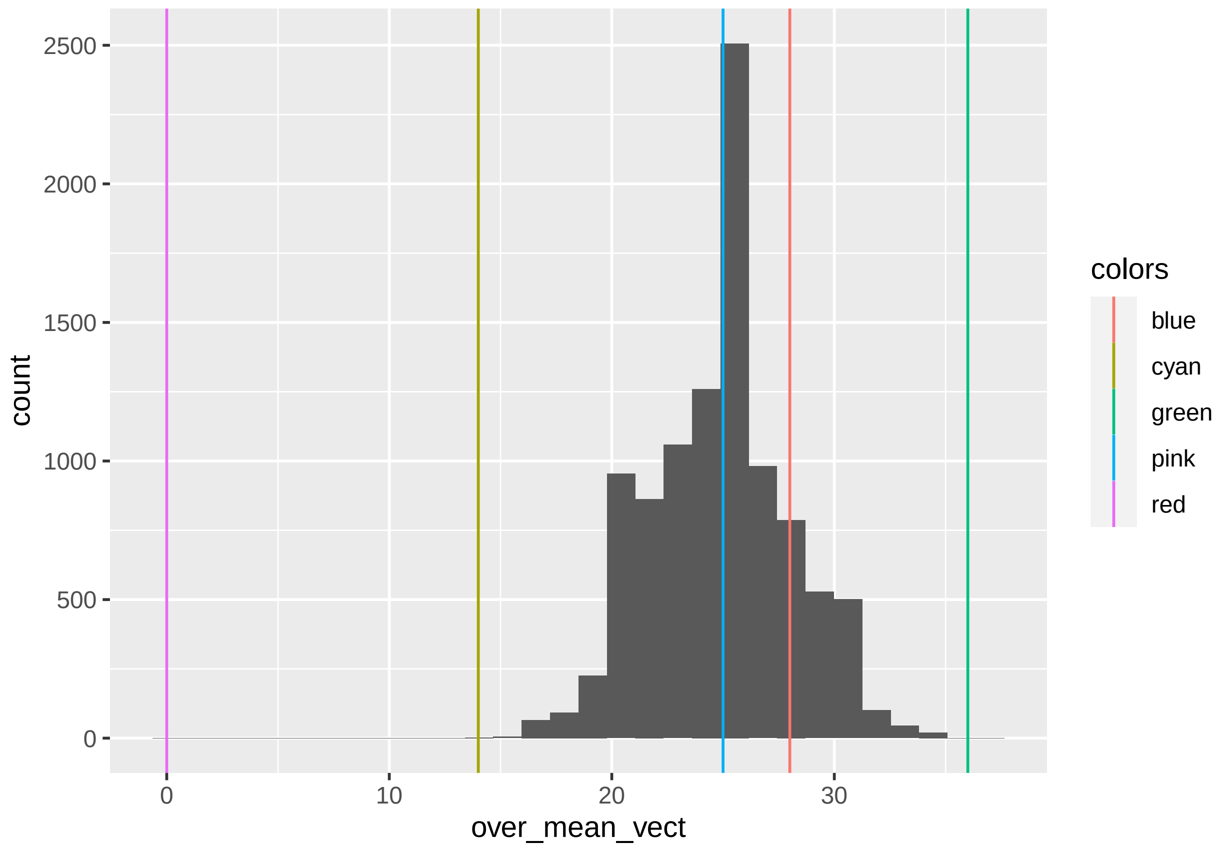
```
## # A tibble: 6 x 5
##   A     B     C     D     E
##   <chr> <chr> <chr> <chr> <chr>
## 1 T     H     H     T     H
## 2 H     H     H     H     H
## 3 T     H     T     H     H
## 4 H     T     H     H     H
## 5 H     T     T     H     T
## 6 T     H     T     T     H
```

1. Design your own hypothesis test to try to identify the sequence of real flips. Design your own test statistic and simulate the null distribution of your test statistic. Can you identify which sequence is the real one?

```r
#TODO then see num times a rand sample has a greater streak length > streak length avg

#TODO: count the number of streaks in each RS


set.seed(94088)

trials <- 200
experiments <- 10000

# table of flips of experiments
flip_table <- data.frame(replicate(experiments, sample(x = c("H", "T"), size = trials, replace = T)))

# get list of lists with streak lengths of random samples
streak_list <- list()
for (i in colnames(flip_table)) {
  temp <- rle(flip_table[[i]])
  streak_list[[length(streak_list) + 1]] <- temp$lengths
}
```

```r
streak_list_sample <- list()
for (i in colnames(flips)) {
  temp <- rle(flips[[i]])
  streak_list_sample[[length(streak_list_sample) + 1]] <- temp$lengths
}

# get sample mean streak length
mean_streak <- mean(sapply(streak_list, mean))

# count num times an experiment has a streak > mean_streak
over_mean_vect <- c()
for (i in streak_list) {
  count_over_mean <- sum(i > mean_streak)
  over_mean_vect <- c(over_mean_vect, count_over_mean)
}
over_mean_vect_sample <- c()
for (i in streak_list_sample) {
  count_over_mean <- sum(i > mean_streak)
  over_mean_vect_sample <- c(over_mean_vect_sample, count_over_mean)
}

# histogram of number of times an experiment had a streak > mean_streak
colors <- c("red", "green", "blue", "pink", "cyan")

ggplot(data = data.frame(over_mean_vect), aes(x = over_mean_vect)) +
  geom_histogram() +
  geom_vline(data = data.frame(over_mean_vect_sample),
             aes(xintercept = over_mean_vect_sample, color = colors))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# calculate p-values
count_over_avg_streak_mean <- mean(over_mean_vect)
z_scores <- sapply(over_mean_vect_sample,
                   function (x) (x - count_over_avg_streak_mean) / var(over_mean_vect))
p_vals <- sapply(z_scores, function (x) pnorm(q = x, lower.tail = T))
# p_vals
p_vals <- c()
for (x in z_scores) {
  if (x >= 0) {
    p_vals <- c(p_vals, pnorm(q = x, lower.tail = F))
  } else {
    p_vals <- c(p_vals, pnorm(q = x, lower.tail = T))
  }
}
p_vals
```

```
## [1] 0.004354111 0.120485383 0.371063392 0.494927128 0.125663634
```

From the computed 'p_vals' vector, we can only eliminate the random with 0 streaks over the population mean streak. However, from visual inspection, we can say that the samples with a counts = 28 and 25, the cyan and red vertical lines are most likely to be the real sequence.