MATH 3070 Lab Project 11

Prachi Aswani

November 14, 2024

Contents

Remember: I expect to see commentary either in the text, in the code with comments created using **#**, *or (preferably) both!* ***Failing to do so may result in lost points!***

Because randomization is used in this assignment, I set the seed here, in addition to beginning each code block. ***Do not change the seed!***

```
set.seed(6222016)
```

Problem 1 (Verzani problem 7.1)

Simulate 1000 rolls of a pair of dice, and compute the sum of each pair. Which is more common, a roll of 7 or 8?

```
rolls <- replicate(1000, sum(sample(1:6, 2, replace = TRUE)))

# Count occurrences of 7 and 8
count_7 <- sum(rolls == 7)
count_8 <- sum(rolls == 8)

# Output the counts
count_7
```

```
## [1] 175
```

```
count_8
```

```
## [1] 133
```

```
# Roll of 7 is more common
```

Problem 2 (Verzani problem 7.2)

For the ***rivers*** *data set, take 1000 random samples of size 10. Compare the mean of the sample means computed from these samples, with the sample mean of the data in* ***rivers****.*

```
# Load the rivers dataset
data("rivers")

# Generate a matrix where each row is a sample of size 10 from the rivers dataset
samples <- replicate(1000, sample(rivers, 10, replace = TRUE))

# Calculate the mean of each sample by taking the row means
sample_means <- colMeans(samples)

# Calculate the overall mean of the rivers data
overall_mean <- mean(rivers)

# Compare means
mean_of_sample_means <- mean(sample_means)

# Output results
mean_of_sample_means
```

```
## [1] 592.1487
```

```
overall_mean
```

```
## [1] 591.1844
```

```
# Interpretation
# In this comparison, the mean of the sample means (approximately 592.15)
# is very close to the overall mean of the rivers dataset (approximately 591.18).
# This similarity aligns with the Central Limit Theorem, suggesting that
# the average of sample means should approach the population mean as the
# number of samples increases, even when each sample size is relatively small.
```

Problem 3

The data set ***Melanoma*** *(**MASS**) includes data for 205 Danish patients with malignant melanoma. The variable* ***time*** *describes survival time in days, and* ***sex*** *describes the sex of the patient. Does survival time differ between the sexes?*

1. *Find* $\bar{x}_{men} - \bar{x}_{women}$*, the mean difference in survival time (****time****) between men and women (****sex****) in* ***Melanoma****.*

```
# Load the Melanoma dataset
library(MASS)
data("Melanoma")

# Calculate mean survival times
mean_men <- mean(Melanoma$time[Melanoma$sex == 1])
```

```r
mean_women <- mean(Melanoma$time[Melanoma$sex == 0])

# Mean difference
mean_difference <- mean_men - mean_women
mean_difference
```

```
## [1] -336.934
```

```r
# The mean survival time difference between men and women is approximately
# -336.93 days, indicating that, on average, men have a shorter
# survival time than women in this dataset.
```

2. *Investigate whether the difference you observed in part 1 is significant, using procedures explored in the lecture. There are two groups in this investigation: male (`Melanoma$sex == 0`) and female (`Melanoma$sex == 1`). Randomly reassign the data in the `time` variable to the two groups, and compute the mean difference. Repeat 2000 times (this needs to be done relatively quickly; if it takes over a few minutes, I will dock points), and determine how frequently the difference in the mean survival time between men and women (that is, $\bar{x}_{men} - \bar{x}_{women}$ in the simulated data is less than the same difference observed in the actual data. Does this analysis suggest the difference is due to "noise", or due to an actual difference in survival time between men and women?*

```r
# Permutation test
set.seed(6222016)
perm_diffs <- replicate(2000, {
  permuted_times <- sample(Melanoma$time)
  perm_men <- mean(permuted_times[Melanoma$sex == 1])
  perm_women <- mean(permuted_times[Melanoma$sex == 0])
  perm_men - perm_women
})

# Calculate p-value
p_value <- mean(perm_diffs <= mean_difference)
p_value
```

```
## [1] 0.0185
```

```r
# The permutation test, which involves randomly reassigning survival times
# to the two groups 2000 times, yields a p-value of 0.0185.
# This low p-value (typically, a p-value below 0.05 is considered
# significant) suggests that the observed difference in survival
# times is unlikely to be due to random noise. Instead, it points
# to a statistically significant difference in survival times
# between men and women in this dataset.
```