Module 2 Lab

This lab explores the error rate perils of simultaneous inference, and shows how two "correction" methods work.

Bonferroni's Correction

First we examine coverage rates of t-based confidence intervals for the mean. To do this, we need to:

1. Generate the random variables
2. Calculate the confidence interval for the mean
3. Assess if the interval covers the true mean
4. Repeat the first three steps many times
5. Calculate the proportion of iterations in which the interval covers the true parameter value

The following code carries out these steps. We should see very close to 95% coverage.

```
CI <- function(){
  X <- rnorm(50, 0, 1)          # Generate 50 N(0,1) random variables
  CI <- t.test(X)$conf.int      # Extract confidence interval from t.test()
  CI[1] < 0 & 0 < CI[2]         # Assess if CI covers true mean (TRUE/FALSE)
}
reps <- replicate(10000, CI())  # repeat many times
mean(reps)                      # Calculate proportion of CI's that contain true mean
```

```
## [1] 0.9496
```

Okay, that worked like it is supposed to; this part we already knew. What about simultaneous confidence intervals? Let's repeat the above steps, but with a few data sets. We will give each data set its own mean to avoid confusion, and help us follow the process. The question we are exploring: Does a family of 95% confidence intervals cover their respective true parameters 95% of the time?

```
CI_simul <- function(){

  # Generate 50 N(0,1) random variables 3 times
  X1 <- rnorm(50, 0, 1)
  X2 <- rnorm(50, 5, 1)
  X3 <- rnorm(50, 10, 1)

  # Extract each confidence interval from their individual t.test()
  CI1 <- t.test(X1)$conf.int
  CI2 <- t.test(X2)$conf.int
  CI3 <- t.test(X3)$conf.int

  # Assess if all three CI's cover their true means (TRUE/FALSE)
  CI1[1] < 0 & 0 < CI1[2]  &
```

```
  CI2[1] < 5 & 5 < CI2[2]  &
  CI3[1] < 10 & 10 < CI3[2]

}
reps_simul <- replicate(10000, CI_simul())  # repeat many times
# Calculate proportion of iterations where all three CI's contain their true mean
mean(reps_simul)
```

```
## [1] 0.8632
```

The family of intervals has far less than 95% coverage! In fact, very close to $(1-\alpha)^3$ coverage, the result we expect if the three data sets (and confidence intervals) are independent. In this situation, we need a simultaneous inference correction. We need a correction because we want to be able to say "Our family of intervals collectively cover their true parameters 95% of the time."

The Bonferroni correction increases the individual confidence intervals to $1-\alpha/k$, where k is the number of intervals, to achieve a family-wise coverage of 95%.

```
CI_simul_bonf <- function(){

  # Generate 50 N(0,1) random variables 3 times
  X1 <- rnorm(50, 0, 1)
  X2 <- rnorm(50, 5, 1)
  X3 <- rnorm(50, 10, 1)

  # Increase each CI to "1 - 0.05/k", where k = 3 groups
  k <- 3
  CI1 <- t.test(X1, conf.level = 1 - 0.05/k)$conf.int
  CI2 <- t.test(X2, conf.level = 1 - 0.05/k)$conf.int
  CI3 <- t.test(X3, conf.level = 1 - 0.05/k)$conf.int

  # Assess if all three CI's cover their true means (TRUE/FALSE)
  CI1[1] < 0 & 0 < CI1[2]  &
  CI2[1] < 5 & 5 < CI2[2]  &
  CI3[1] < 10 & 10 < CI3[2]

}
reps_simul_bonf <- replicate(10000, CI_simul_bonf())  # repeat many times
# Calculate proportion of iterations where all three CI's contain their true mean
mean(reps_simul_bonf)
```

```
## [1] 0.9491
```

We are back at 95%!! The Bonferroni correction works in a wide range of situations, but it is our most conservative choice for controlling the family-wise error rates. Another option for multiple comparison correction is "Tukey's honest significant difference," or Tukey's HSD for short.

Tukey's HSD

For this portion of the lab we use the data from the Module 1 homework with the heights of baseball, soccer, and basketball players. With whatever method you prefer, load in `Sports_Heights.csv`; it is included in the Module 2, Lab folder.

```
Heights <- read.csv("Sport_Heights.csv", row.names = 1)
```

The data contains the height of players from the three different sports, and we want to answer the question: Do any two of the sports have the same mean height? The naive option is to calculate individual t-based confidence intervals for each of the three pairwise differences.

```
# Create individual height vectors by sport
soccer <- subset(Heights, Sport == "soccer", "Height")$Height
baseball <- subset(Heights, Sport == "baseball", "Height")$Height
basketball <- subset(Heights, Sport == "basketball", "Height")$Height

# Individual confidence intervals
t.test(basketball, baseball, var.equal = FALSE)$conf.int
```

```
## [1] -3.0781678  0.9524436
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(soccer, baseball, var.equal = FALSE)$conf.int
```

```
## [1] -2.012451 -0.110158
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(soccer, basketball, var.equal = FALSE)$conf.int
```

```
## [1] -1.998016  2.001131
## attr(,"conf.level")
## [1] 0.95
```

However, from the previous example, we know this set of intervals will not have a 95% family-wise coverage rate. Therefore, we apply Tukey's correction, and compare.

`TukeyHSD` requires that its data come from a fitted model, so the first line of code below fits a model with the function `aov()`, which stands for "Analysis of Variance." Passing the `aov()` fit of `Height` against `Sport` tells `TukeyHSD` that `Height` is grouped by `Sport`.

```
fit <- aov(Height ~ Sport, data = Heights) # aov = Analysis of Variance
summary(fit) # May look familiar!
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Sport         2   37.6   18.80   1.005  0.369
## Residuals   147 2751.2   18.72
```

```
TukeyHSD(fit) # Notice wider intervals
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Height ~ Sport, data = Heights)
```

```
##
## $Sport
##                              diff       lwr       upr     p adj
## basketball-baseball -1.062862085 -3.111477 0.9857525 0.4383693
## soccer-baseball     -1.061304543 -3.109919 0.9873100 0.4394202
## soccer-basketball    0.001557541 -2.047057 2.0501721 0.9999982
```

The first line fits the model and creates the Anova table returned by the second line. This table should look familiar from the Module 1 homework! The first line reminds you of the procedure being used, and the family-wise confidence level for the intervals calculated. Confidence intervals for the differences are provided in the table, in the columns labeled "`lwr`" and "`upr`." Notice that, as expected, they are all wider than their respective t-based non-corrected intervals. With these wider intervals, the family-wise coverage rate will be 95% in repeated experiments. Also notice that the individual t-based confidence interval for the difference between mean soccer and baseball heights does not contain zero. However, the corrected interval does contain zero, meaning that we would reach different conclusions depending on the method used.

Food for thought–how would you use a simulation to explore the Tukey HSD corrected confidence interval coverage rates?

See Chapter 6 in "The Statistical Sleuth" for more details on Bonferroni, Tukey HSD, and other simultaneous inference correction procedures.