STAT636 - Homework 3

Daniel Osorio - dcosorioh@tamu.edu
Department of Veterinary Integrative Biosciences
Texas A&M University

1. Consider the Auto data. These represent a two-factor experiment on cars. The two factors are (i) the number of cylinders (4 and 6 were considered) and (ii) origin (three origins were considered). We have 4 cars under each of the $2 \times 3 = 6$ factor combinations. For each car, we have measurements of three weight variables: X_1 = displacement, X_2 = horsepower, and X_3 = acceleration. So, in terms of a two-way MANOVA model, $g = 2$, $b = 3$, and $n = 4$.

   (a) Test for a location effect, a variety effect, and a location-variety interaction at $\alpha = 0.05$. Do this using the manova function in R. Overall, what do you conclude about these data?

   (b) Construct the two-way MANOVA table by computing SSPFAC 1, SSPFAC 2, SSPINT, SSPRES, and SSPCOR. Provide R code that matches the Wilks' statistics computed by manova. Note that your p-values (computed according to the notes) will not match those of manova, because the distributional results we have learned for two-way MANOVA are large-sample approximations. That said, how do your p-values compare to those of manova?

2. Conduct a simulation study to investigate the coverage probabilities of different confidence interval types with multivariate data. Let the sample size be $n = 30$, the number of variables $p = 5$, the number of simulations $B = 10000$, and

$$\mu' = (0, 0, 0, 0, 0)$$

and

$$\Sigma = \begin{pmatrix} 1.0 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.6 & 1.0 & 0.6 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1.0 & 0.6 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1.0 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 1.0 \end{pmatrix}$$

For each of B times, simulate a dataset of size n from the $N_p(\mu, \Sigma)$ distribution, compute 95% confidence intervals of types one-at-a-time, T^2 simultaneous, and Bonferroni simultaneous, and record whether each interval contains its corresponding population mean component value. Report a 3×5 matrix of estimated coverage probabilities. The rows of your matrix should correspond to the 3 different interval types, and the columns should correspond to the p mean components; be sure to clearly indicate which row goes with which interval type. Comment on the performance of the different interval types.

```
> simmulationFunction <- function(n, p, rho, alpha, B = 1000) {
+   mu <- rep(0, p)
+   Sigma <- matrix(rho, nrow = p, ncol = p); diag(Sigma) <- 1
+   F_crit <- (n - 1) * p * qf(1 - alpha, p, n - p) / (n - p)
+   bon_crit <- qt(1 - alpha / (2 * p), n - 1)
+   cov_95 <- cov_T2 <- cov_bon <- matrix(NA, nrow = B, ncol = p)
+   for(i in seq_len(B)) {
```

```
+       X <- MASS::mvrnorm(n, mu, Sigma)
+       x_bar <- colMeans(X)
+       S <- var(X)
+       for(k in 1:p) {
+         ci_95 <- quantile(X[,k],c((alpha/2),(1-(alpha/2))))
+         ci_T2 <- x_bar[k] + c(-1, 1) * sqrt(F_crit * S[k, k] / n)
+         ci_bon <- x_bar[k] + c(-1, 1) * bon_crit * sqrt(S[k, k] / n)
+
+         cov_95[i, k] <- (ci_95[1] <= mu[k]) & (mu[k] <= ci_95[2])
+         cov_T2[i, k] <- (ci_T2[1] <= mu[k]) & (mu[k] <= ci_T2[2])
+         cov_bon[i, k] <- (ci_bon[1] <= mu[k]) & (mu[k] <= ci_bon[2])
+       }
+     }
+     out <- matrix(data = NA, nrow = 3, ncol = p)
+     rownames(out) <- c("95%", "T2", "Bonferroni")
+     out[1,] <- apply(cov_95, 2, mean)
+     out[2,] <- apply(cov_T2, 2, mean)
+     out[3,] <- apply(cov_bon, 2, mean)
+     return(out)
+ }
> simmulationFunction(n = 30, p = 5, alpha = 0.05, rho = 0.6, B = 10000)

             [,1]   [,2]   [,3]   [,4]   [,5]
95%        1.0000 1.0000 1.0000 1.0000 1.0000
T2         0.9993 0.9993 0.9992 0.9996 0.9992
Bonferroni 0.9890 0.9883 0.9886 0.9905 0.9899
```