Handout #12

Biting-Fly Data: Comparing Species

Problem 6.28 "Two species of biting flies (genus *Leptoconops*) are so similar morphologically, that for many years they were thought to be the same. Biological differences such as sex ratios of emerging flies and biting habits were found to exist." Using the taxonomic data in Table 6.15 examine whether or not there is evidence for differences in the two species *L. carteri* and *L. torrens.* Measurements were taken on $n = 70$ specimens, 35 from each species. We will examine $p = 5$ of the seven variables reported in Table 6.15: wing length (x_1), wing width (x_2), third palp length (x_3), third palp width (x_4), and fourth palp length (x_5).

After assessing the appropriateness of the multivariate normal assumption, test H_0: $\underline{\mu}_c = \underline{\mu}_t$ versus H_a: $\underline{\mu}_c \neq \underline{\mu}_t$, where $\underline{\mu}_c$ and $\underline{\mu}_t$ are 5×1 mean vectors for species *L. carteri* and species *L. torrens*, respectively. If H_0 is rejected, examine evidence for differences in species, variable-by-variable.

Outliers & Normality. Examining the scatterplot matrix (separately for each species), one *L. carteri* observation (case 36) seems to be unusual, tending to have a small value for each of the 5 variables, but particularly so for wing width. This case was omitted from all subsequent analysis. With this point omitted, the Q-Q plot to assess multivariate normality for the data on the five variables, for each species separately, showed no evidence of departure from normality – though assessment of the univariate normality of each variable was not examined here.

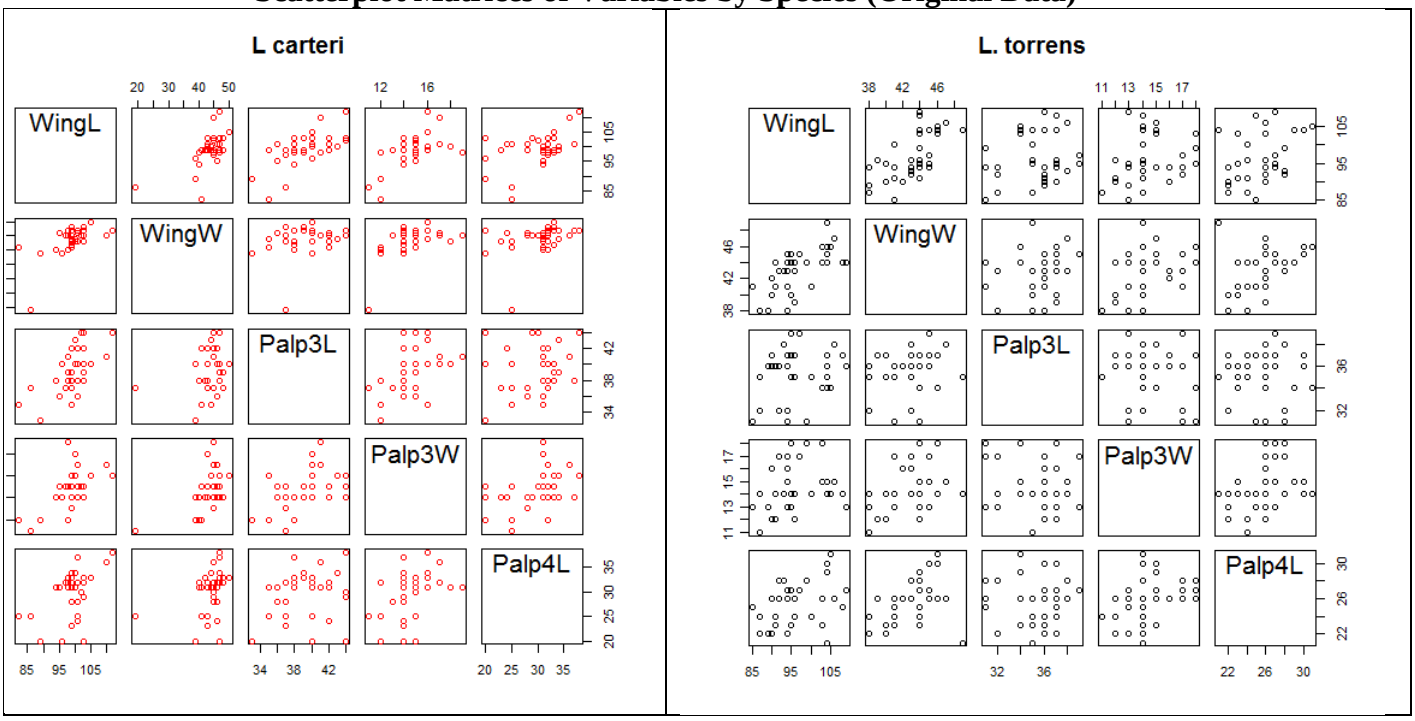
Multivariate Two-Sample T^2 Test. Assuming that the population variance-covariance matrices for 5 variables are the same across the two species, the ***R*** commands `manova` and `summary` are used to carry out the test of H_0 (see attached output). The value of the resulting test statistic, a scaled version of Hotellings T^2, was $F^* = 12.34$. Under the null hypothesis, this test statistic is an observation from an F distribution with $p = 5$ numerator and $n_1 + n_2 - p - 1 = 63$ denominator degrees of freedom. The p-value is 2.35×10^{-8}, providing strong evidence against the null hypothesis, and supporting the conclusion that the mean vector for these taxonomic variables differs by species.

Univariate Analyses. Using `summary.aov` on the output of `manova` produces the 5 separate univariate analyses. (It only makes sense to examine these analyses when the multivariate test leads to rejection of the null hypothesis.) To control for the overall error rate among these five tests, each difference will be judged significant at $\alpha = .05$ only if the reported p-value is less than $\alpha/5 = .01$. By this criterion, there is no evidence for differences between species *L. torrens* and *L. carteri* in mean wing length, mean wing width, or mean third palp width. However there is strong evidence for differences between species in both mean third palp length and mean fourth palp length.

Simultaneous Confidence Intervals. Simultaneous 95% confidence intervals for the difference in species (*L. carteri* minus *L. torrens*) were formed for each of the five taxonomic variables using the Bonferroni correction and the pooled estimate of the common variance-covariance matrix. The Bonferroni critical value is $t_{n_1+n_2-2}(\frac{\alpha}{2p}) = t_{67}(.005) = 2.65$, and the pooled variance estimates are 33.81, 7.48, 6.45, 2.88, and 13.59 respectively for the 5 variables. The confidence intervals are reported after the `manova` output.

Only the confidence intervals for third palp length and fourth palp length do not include zero, this is suggestive that the observed overall significant difference in species mean vectors, is due to the species having different population mean third palp length and fourth palp length – entirely consistent (as it must be) with the reported univariate analyses above.

Scatterplot Matrices of Variables by Species (Original Data)



Covariance Matrices by Species (after removal of outlier in L. carteri data)

L carteri (*n* = 34)

	WingL	WingW	Palp3L	Palp3W	Palp4L
WingL	26.685383	7.704100	9.164884	3.784314	11.191622
WingW	7.704100	7.468806	2.026738	1.447415	5.352941
Palp3L	9.164884	2.026738	8.122103	1.971480	2.336007
Palp3W	3.784314	1.447415	1.971480	2.367201	2.641711
Palp4L	11.191622	5.352941	2.336007	2.641711	21.159537

L torrens (*n* = 35)

	WingL	WingW	Palp3L	Palp3W	Palp4L
WingL	40.726050	11.716807	2.3252101	2.1991597	6.263025
WingW	11.716807	7.492437	1.8268908	1.8394958	3.261345
Palp3L	2.325210	1.826891	4.8285714	-0.7848739	0.612605
Palp3W	2.199160	1.839496	-0.7848739	3.3747899	1.696639
Palp4L	6.263025	3.261345	0.6126050	1.6966387	6.240336

Pooled Estimate of *S*

	WingL	WingW	Palp3L	Palp3W	Palp4L
WingL	33.810498	9.740399	5.694005	2.979907	8.690543
WingW	9.740399	7.480798	1.925323	1.646382	4.291534
Palp3L	5.694005	1.925323	6.450759	0.572733	1.461445
Palp3W	2.979907	1.646382	0.572733	2.878515	2.162122
Palp4L	8.690543	4.291534	1.461445	2.162122	13.588599

Hotellings T^2 Analysis (Manova output)

```
> #
> # Using built-in MANOVA code
> #
> # Note: it is important that the two species be coded 0 and 1 here
> # Flysn is the modified dataset with the outlier deleted
> #
> fit<-manova(cbind(WingL,WingW,Palp3L,Palp3W,Palp4L) ~ Species, data=Flysn)
> #
> # Produce the F statistic version of Hotellings T^2
> summary(fit, test="Wilks")
          Df   Wilks approx F num Df den Df    Pr(>F)
Species    1 0.50522   12.339      5     63 2.348e-08 ***
Residuals 67
---


> # Produce Univariate Analyses
> summary.aov(fit)
 Response WingL :
            Df  Sum Sq Mean Sq F value Pr(>F)
Species      1  185.33  185.33  5.4816 0.0222 *
Residuals   67 2265.30   33.81
---
 Response WingW :
            Df Sum Sq Mean Sq F value  Pr(>F)
Species      1  41.77  41.772  5.5839 0.02104 *
Residuals   67 501.21   7.481
---
 Response Palp3L :
            Df Sum Sq Mean Sq F value    Pr(>F)
Species      1 277.45 277.451  43.011 9.275e-09 ***
Residuals   67 432.20   6.451
---
 Response Palp3W :
            Df  Sum Sq Mean Sq F value Pr(>F)
Species      1   1.082  1.0815  0.3757  0.542
Residuals   67 192.861  2.8785

 Response Palp4L :
            Df Sum Sq Mean Sq F value    Pr(>F)
Species      1 352.11  352.11  25.913 3.109e-06 ***
Residuals   67 910.44   13.59
---
```

Variable	Simultaneous 95% Confidence Interval
Wing length	3.28±3.71 = (-0.43,6.99)
Wing width	1.56±1.75 = (-0.19,3.30)
Third palp length*	4.01±1.62 = (2.39,5.63)
Third palp width	0.25±1.08 = (-0.83,1.33)
Fourth palp length*	4.52±2.35 = (2.17,6.87)

(Courtesy of Dr. Roy St. Laurent)