Sheep Example Discriminant Analysis

April 29, 2020

0.1 Example 12.6

Five diagnostic tests were carried out on two groups of sheep:

1. those suffering with Scrapie (a mild disease);
2. those suffering from ovine CJD (a more serious disease).

The measurements of the two tests were:

[7]:
```
# 5 tests for Scrapie group

srp.t1 <- c(11, 33, 20, 18, 22)
srp.t2 <- c(18, 27, 28, 26, 23)
srp.t3 <- c(15, 31, 27, 18, 22)
srp.t4 <- c(18, 21, 23, 18, 16)
srp.t5 <- c(15, 17, 19, 9, 10)
```

[8]:
```
# The same 5 diagnostic tests for the ovine CJD group

cjd.t1 <- c(18, 31, 14, 25, 36)
cjd.t2 <- c(17, 24, 16, 24, 28)
cjd.t3 <- c(20, 31, 17, 31, 24)
cjd.t4 <- c(18, 26, 20, 26, 26)
cjd.t5 <- c(18, 20, 17, 18, 29)
```

[11]:
```
scrapie <- data.frame(srp.t1, srp.t2, srp.t3, srp.t4, srp.t5)
cjd <- data.frame(cjd.t1, cjd.t2, cjd.t3, cjd.t4, cjd.t5)
```

0.1.1 Scrapie data: 5 observations

[12]:
```
scrapie
```

A data.frame: 5 × 5

srp.t1 <dbl>	srp.t2 <dbl>	srp.t3 <dbl>	srp.t4 <dbl>	srp.t5 <dbl>
11	18	15	18	15
33	27	31	21	17
20	28	27	23	19
18	26	18	18	9
22	23	22	16	10

0.1.2 Ovine CJD data: 5 observations

[13]:
```
cjd
```

A data.frame: 5 × 5

cjd.t1 <dbl>	cjd.t2 <dbl>	cjd.t3 <dbl>	cjd.t4 <dbl>	cjd.t5 <dbl>
18	17	20	18	18
31	24	31	26	20
14	16	17	20	17
25	24	31	26	18
36	28	24	26	29

0.1.3 Need the sample sizes of the two samples

[21]:
```
no.scrapie <- dim(scrapie)[1]
no.cjd <- dim(cjd)[1]

c(no.scrapie, no.cjd)
```

1. 5 2. 5

0.1.4 Our pooled estimate of the covariance matrix: $\hat{\Sigma} = \mathbf{S}_U$

$\mathbf{S}_U$ is the unbiased estimate of Σ

$$\mathbf{S}_U = \frac{(n_1 - 1)\mathbf{S}_{1U} + (n_2 - 1)\mathbf{S}_{2U}}{n_1 + n_2 - 2}$$

Where $\mathbf{S}_{1U}$ is the sample covariance matrix for sample 1.

[46]:
```
s.pooled <- (no.scrapie - 1) * var(scrapie) + (no.cjd - 1) * var(cjd)

s.pooled <- s.pooled / (no.scrapie + no.cjd - 2)

s.pooled
```

A matrix: 5 × 5 of type dbl

	srp.t1	srp.t2	srp.t3	srp.t4	srp.t5
srp.t1	72.700	33.025	41.65	18.675	22.300
srp.t2	33.025	21.250	21.30	12.725	11.925
srp.t3	41.650	21.300	41.30	16.350	9.850
srp.t4	18.675	12.725	16.35	11.450	10.200
srp.t5	22.300	11.925	9.85	10.200	21.650

[51]:
```
# calculate the inverse of the sample covariance matrix
s.pooled.inv <- solve(s.pooled)

s.pooled.inv
```

A matrix: 5 × 5 of type dbl

	srp.t1	srp.t2	srp.t3	srp.t4	srp.t5
srp.t1	0.09627223	-0.12756676	-0.07536557	0.1508991	-0.06570241
srp.t2	-0.12756676	0.32560846	0.06744833	-0.3102937	0.06755138
srp.t3	-0.07536557	0.06744833	0.12750144	-0.2041784	0.07866334
srp.t4	0.15089911	-0.31029369	-0.20417838	0.6566855	-0.20100846
srp.t5	-0.06570241	0.06755138	0.07866334	-0.2010085	0.13556886

Now getting the sample means

Our sample means will be vectors

The length of the sample mean vectors will be 5 - due to the 5 tests

[52]:
```
# MARGIN = 2 means we're getting column-wise means

sample.mean.scrapie <- apply(scrapie, MARGIN = 2, mean)
sample.mean.cjd <- apply(cjd, MARGIN = 2, mean)
```

[53]:
```
# little bit of R syntax to turn these two objects into vectors

names(sample.mean.scrapie) <- NULL
names(sample.mean.cjd) <- NULL
```

0.2 We can now calculate the $\widehat{\mathbf{L}}$ part of the Sample Linear Discriminant Function, $\widehat{\mathbf{L}}\mathbf{x}$

$\mathbf{L} = \Sigma^{-1}(\mu_1 - \mu_2)$

$\widehat{\mathbf{L}} = \mathbf{S}_U^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

[56]:
```
# in R, %*% performs a dot product
L.hat <- s.pooled.inv %*% (sample.mean.scrapie - sample.mean.cjd)
L.hat
```

A matrix: 5 × 1 of type dbl

srp.t1	-0.7491324
srp.t2	2.0307983
srp.t3	0.5350933
srp.t4	-2.3422912
srp.t5	0.2175097

0.2.1 Our estimate for α, the squared Mahalanobis distance, is given by:

$\alpha = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$

$\alpha = (\mu_1 - \mu_2)^T \mathbf{L} = \mathbf{L}^T(\mu_1 - \mu_2)$

$\hat{\alpha} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_U^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

$\hat{\alpha} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \widehat{\mathbf{L}} = \widehat{\mathbf{L}}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

[69]:
```
alpha.hat <- t(sample.mean.scrapie - sample.mean.cjd) %*% s.pooled.inv %*%␣
↪(sample.mean.scrapie - sample.mean.cjd)
```

```
alpha.hat <- alpha.hat[1,1]

alpha.hat
```

15.1835214211121

0.2.2 Probability that misclassified

Total Probability that disease is misclassified $= \Phi\left(-\frac{\sqrt{\hat{\alpha}}}{2}\right)$

[72]:
```
pnorm(-0.5 * sqrt(alpha.hat), lower.tail = TRUE)
```

0.0256894240288836