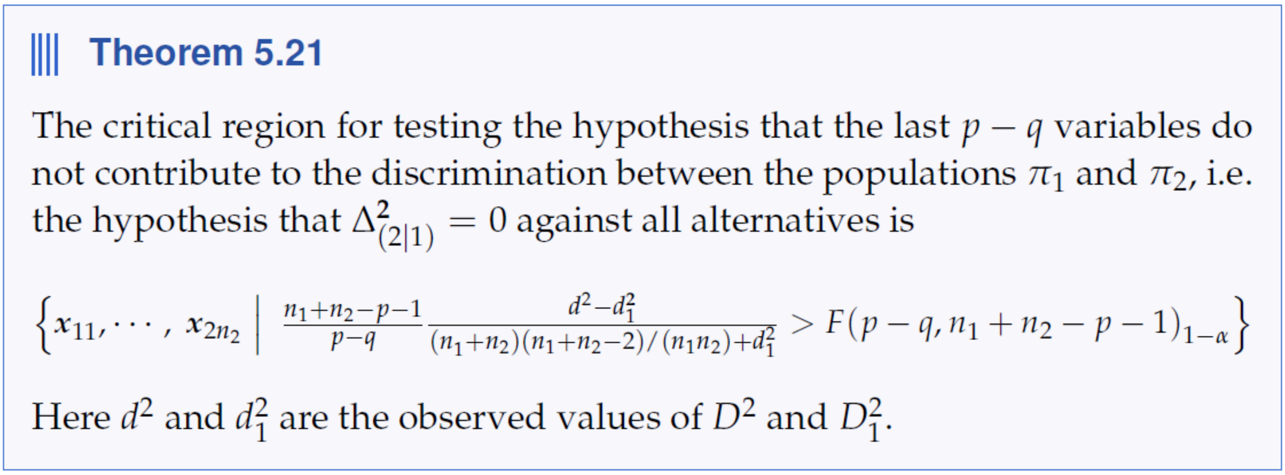
Solution for Exam 2011 Q. 3.3 – 3.6

ANYM, 20190924

Q 3.3

We use the following theorem, page 358



Theorem 5.21

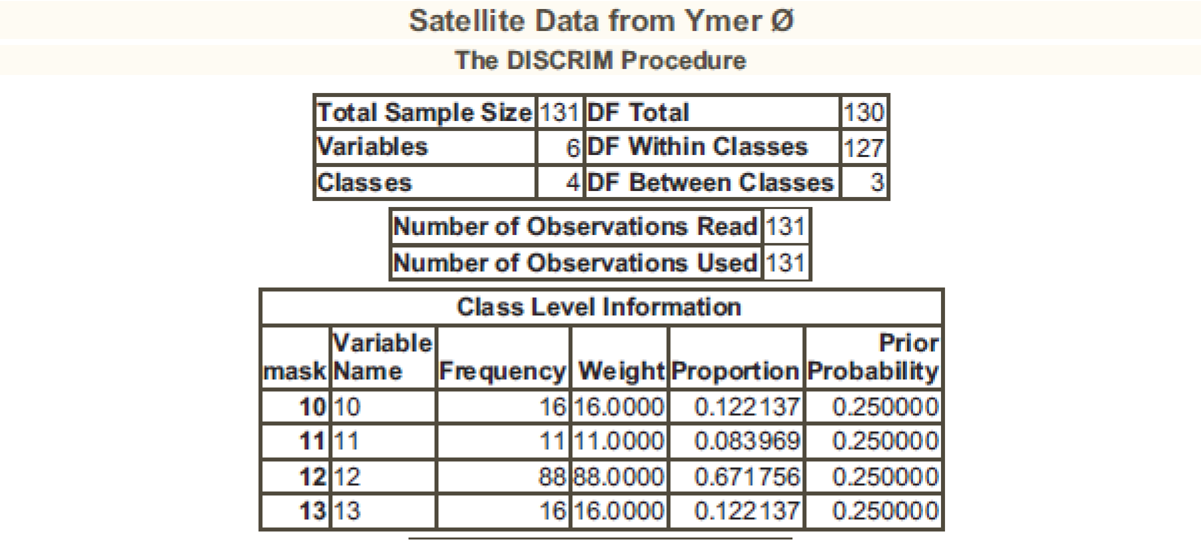
The critical region for testing the hypothesis that the last $p-q$ variables do not contribute to the discrimination between the populations π_1 and π_2, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

$$\left\{x_{11},\cdots,\, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p-q}\,\frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} > F(p-q, n_1+n_2-p-1)_{1-\alpha}\right\}$$

Here d^2 and d_1^2 are the observed values of D^2 and D_1^2.

If we have equal dispersion/covariance matrix and we have equal priors, the Mahalanobis' distance is equal to the squared generalized distance.

From the SAS enclosure B we have:



Satellite Data from Ymer Ø

The DISCRIM Procedure

Total Sample Size	131	DF Total	130
Variables	6	DF Within Classes	127
Classes	4	DF Between Classes	3

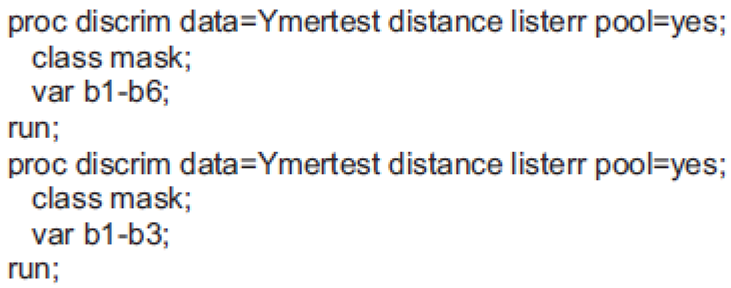
Number of Observations Read	131
Number of Observations Used	131

Class Level Information

mask	Variable Name	Frequency	Weight	Proportion	Prior Probability
10	10	16	16.0000	0.122137	0.250000
11	11	11	11.0000	0.083969	0.250000
12	12	88	88.0000	0.671756	0.250000
13	13	16	16.0000	0.122137	0.250000

We see unit 10: n1 = 16, unit 13: n2 = 16, variables: p=6, q=3.

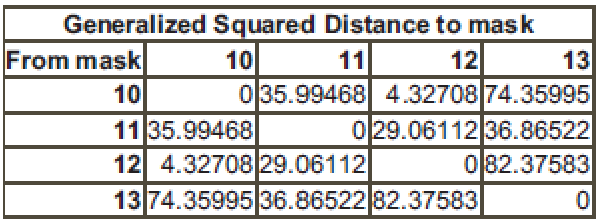
We now only need d and d1. We see from the sas-code:



```
proc discrim data=Ymertest distance listerr pool=yes;
  class mask;
  var b1-b6;
run;
proc discrim data=Ymertest distance listerr pool=yes;
  class mask;
  var b1-b3;
run;
```

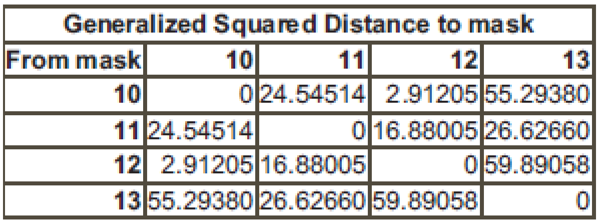
That these has been calculated.

Using all variables:



Generalized Squared Distance to mask				
From mask	**10**	**11**	**12**	**13**
10	0	35.99468	4.32708	74.35995
11	35.99468	0	29.06112	36.86522
12	4.32708	29.06112	0	82.37583
13	74.35995	36.86522	82.37583	0

Using only b1-b3:



Generalized Squared Distance to mask				
From mask	**10**	**11**	**12**	**13**
10	0	24.54514	2.91205	55.29380
11	24.54514	0	16.88005	26.62660
12	2.91205	16.88005	0	59.89058
13	55.29380	26.62660	59.89058	0

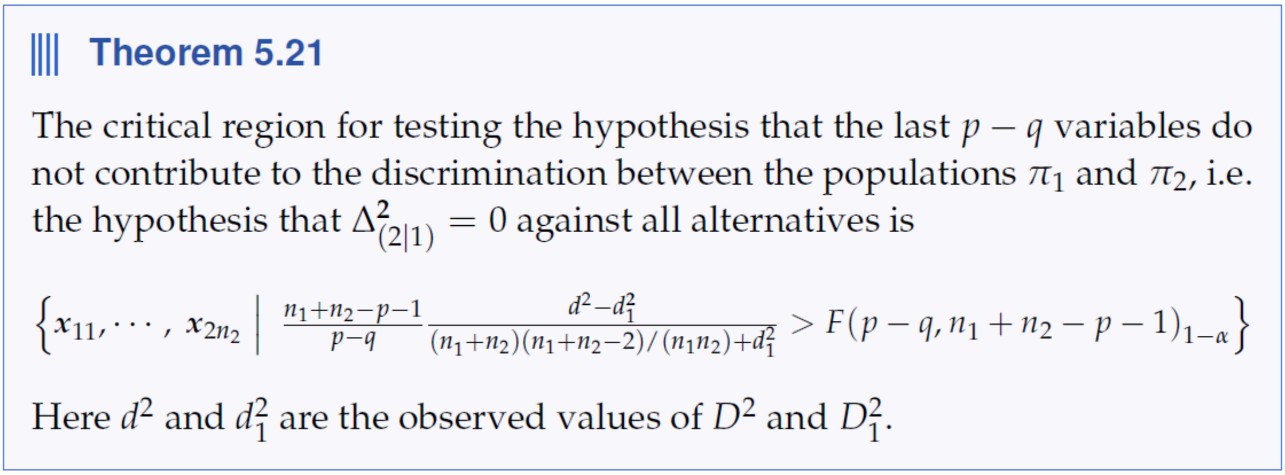
We can now insert:

$$\begin{aligned}\frac{n_1+n_2-p-1}{p-q}\frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} &= \frac{16+16-6-1}{6-3}\frac{74.36-55.29}{(16+16)(16+16-2)/(16\cdot 16)+55.29} \\ &= \frac{25}{3}\frac{74.36-55.29}{32\cdot 30/256+55.29}\end{aligned}$$

And see the correct answer is 1.

Q 3.4

We again refer to page 358



Theorem 5.21

The critical region for testing the hypothesis that the last $p-q$ variables do not contribute to the discrimination between the populations π_1 and π_2, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

$$\left\{ x_{11}, \cdots, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p-q} \frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1 n_2)+d_1^2} > F(p-q, n_1+n_2-p-1)_{1-\alpha} \right\}$$

Here d^2 and d_1^2 are the observed values of D^2 and D_1^2.

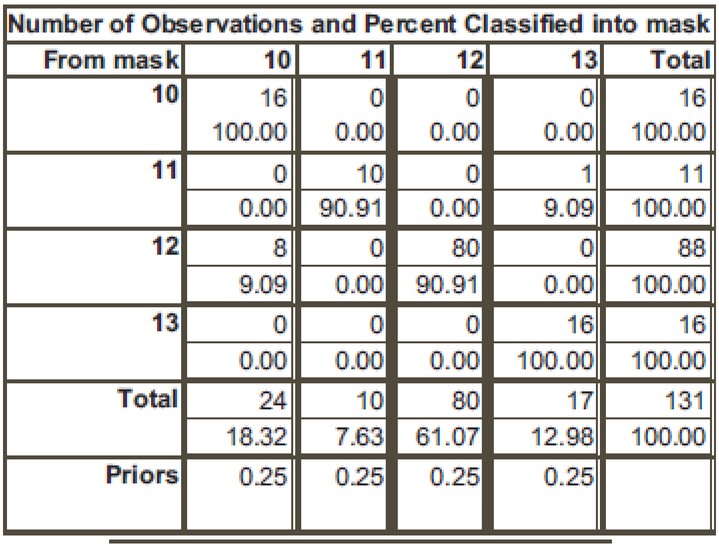
We simply insert in the test statistic:

$$F(p-q, n_1+n_2-p-1) = F(6-3, 16+16-6-1) = F(3, 25)$$

The correct answer is 3.

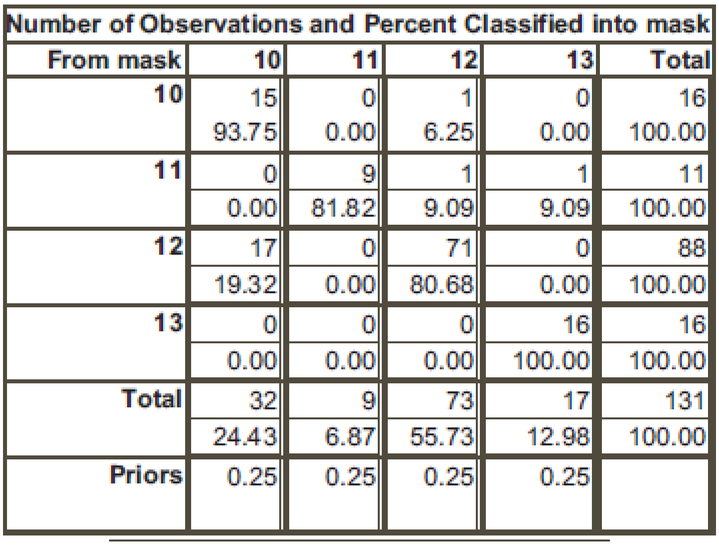
Q 3.5

We inspect the confusion matrix from using all the variables and only the visible light. Using all variables:



Number of Observations and Percent Classified into mask					
From mask	10	11	12	13	Total
10	16	0	0	0	16
	100.00	0.00	0.00	0.00	100.00
11	0	10	0	1	11
	0.00	90.91	0.00	9.09	100.00
12	8	0	80	0	88
	9.09	0.00	90.91	0.00	100.00
13	0	0	0	16	16
	0.00	0.00	0.00	100.00	100.00
Total	24	10	80	17	131
	18.32	7.63	61.07	12.98	100.00
Priors	0.25	0.25	0.25	0.25	

Omitting the infrared channels:

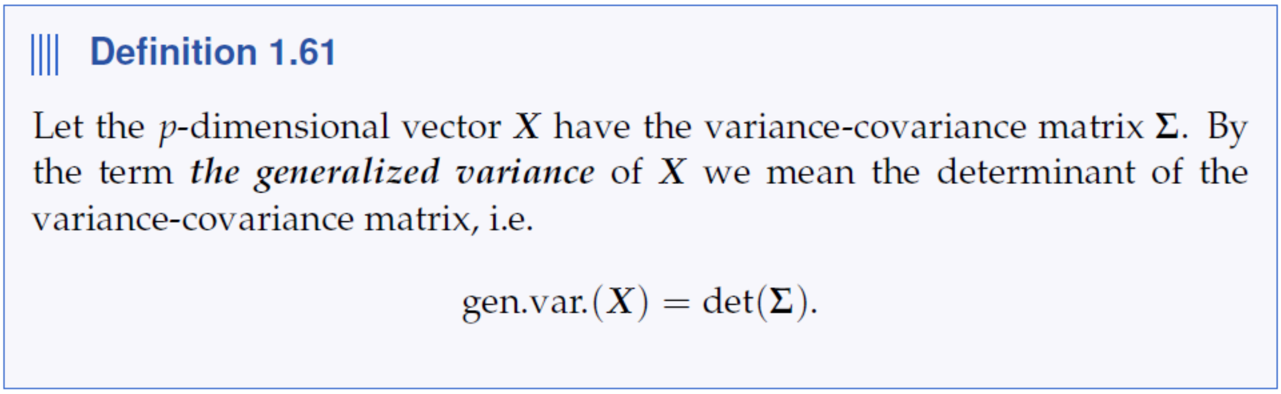


Number of Observations and Percent Classified into mask					
From mask	10	11	12	13	Total
10	15	0	1	0	16
	93.75	0.00	6.25	0.00	100.00
11	0	9	1	1	11
	0.00	81.82	9.09	9.09	100.00
12	17	0	71	0	88
	19.32	0.00	80.68	0.00	100.00
13	0	0	0	16	16
	0.00	0.00	0.00	100.00	100.00
Total	32	9	73	17	131
	24.43	6.87	55.73	12.98	100.00
Priors	0.25	0.25	0.25	0.25	

We count the off-diagonal classifications, which is 9 in the first case and 20 in the latter. That means an increase of 11, answer 4.

Q 3.6

We first need to know what generalized variance is (page 58).



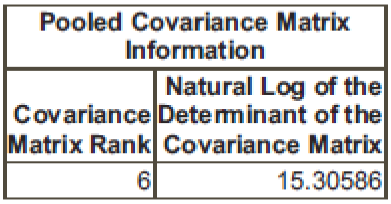
Definition 1.61

Let the p-dimensional vector $\boldsymbol{X}$ have the variance-covariance matrix $\boldsymbol{\Sigma}$. By the term ***the generalized variance*** of $\boldsymbol{X}$ we mean the determinant of the variance-covariance matrix, i.e.

$$\text{gen.var.}(\boldsymbol{X}) = \det(\boldsymbol{\Sigma}).$$

In the univariate case we have the concept of variance, which we generalize to the dispersion matrix in the multivariate case. The generalized variance is a way to boil down all the numbers in the dispersion matrix to one number. It is e.g. used when testing equivalence of dispersion matrices.

We look at the SAS enclosure and find:



Pooled Covariance Matrix Information	
Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
6	15.30586

I.e. the logarithm of the generalized variance. To get the gen.var. we thus need to take the exponent: $e^{15.31}$, i.e. answer 2.