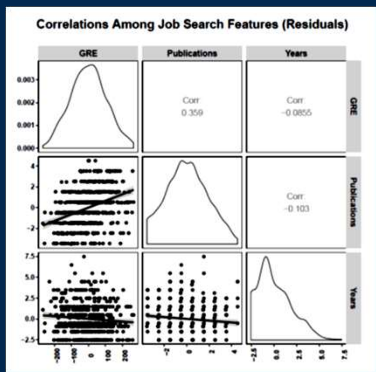
Discriminant Analysis

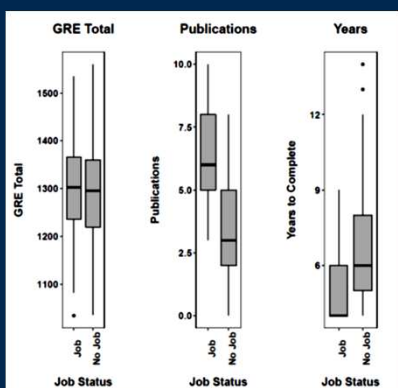
Today . . .

- Another example
- Classification details
- Alternatives when assumptions fail

In this hypothetical example, data from 500 graduate students seeking jobs were examined. Available for each student were three predictors: GRE(V+Q), Years to Finish the Degree, and Number of Publications. The outcome measure was categorical: “Got a job” versus “Did not get a job.”



The data are not as clean as the iris data. The measures are less continuous and years has a lower bound (4 years) that creates a skewed and truncated distribution.



Years and publications individually separate the groups. GRE appears to make no difference.

```
Job_MANOVA <- manova(as.matrix(Job[, 1:3]) ~ Job$job)
summary(Job_MANOVA)

##            Df Pillai approx F num Df den Df Pr(>F)
## Job$job     1   0.41      115      3    496 <2e-16 ***
## Residuals 498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary.aov(Job_MANOVA)

##  Response gre :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## Job$job       1    5693    5693    0.53   0.47
## Residuals   498 5362834   10769
##
##  Response pubs :
##              Df Sum Sq Mean Sq F value Pr(>F)
## Job$job       1    927     927     266 <2e-16 ***
## Residuals   498   1733       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response years :
##              Df Sum Sq Mean Sq F value Pr(>F)
## Job$job       1    263   262.6    70.9  4e-16 ***
## Residuals   498   1844     3.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA confirms that years to finish and publications individually separate the groups.

```
Job_CDA$coeffs.raw

##            Can1
## gre   -0.003253
## pubs   0.513837
## years -0.198684

Job_CDA$coeffs.std

##          Can1
## gre   -0.3375
## pubs   0.9584
## years -0.3824

Job_CDA$structure

##           Can1
## gre    0.05083
## pubs   0.92155
## years -0.55108
```

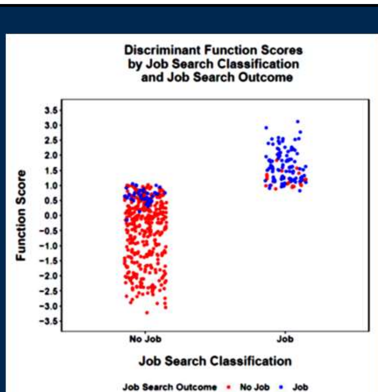
```
Canonical Discriminant Analysis for as.factor(job):

  CanRsq Eigenvalue Difference Percent Cumulative
1   0.41      0.696                100        100

Test of H0: The canonical correlations in the
current row and all that follow are zero

  LR test stat approx F num Df den Df Pr(> F)
1         0.59      347      1    498 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The lone possible discriminant function provides significant discrimination. Years and publications are unique contributors. The number of publications is the most important variable. The message: publish often and finish quickly. Note the standardized coefficient for GRE.

```
Job_CDA$coeffs.raw

##            Can1
## gre   -0.003253
## pubs   0.513837
## years -0.198684

Job_CDA$coeffs.std

##          Can1
## gre   -0.3375
## pubs   0.9584
## years -0.3824

Job_CDA$structure

##           Can1
## gre    0.05083
## pubs   0.92155
## years -0.55108
```

One problem we confront in assessing these coefficients is that we have no information about their standard errors. Those are important if we want to make claims about coefficients being different from 0.

Later we will consider bootstrapping as a means to get that information.



The separation of the groups on the discriminant function is clear.

Classification based on the discriminant analysis can be expected to be fairly accurate.

```r
table(Original = Job$job_num, Predicted = predict(Job_LDA)$class)

##         Predicted
## Original   1   2
##        1 344  20
##        2  41  95

Proportion_of_Correct_Classification <- sum(diag(table(Original = Job$job_num,
    Predicted = predict(Job_LDA)$class)))/sum(table(Original = Job$job_num,
    Predicted = predict(Job_LDA)$class))
Proportion_of_Correct_Classification

## [1] 0.878
```

Simple classification of all cases.

```r
Job_Jack <- lda(job_num ~ gre + pubs + years, data = Job, CV = TRUE)
table(Original = Job$job_num, Predicted = Job_Jack$class)

##         Predicted
## Original   1   2
##        1 339  25
##        2  42  94

Proportion_of_Correct_Classification <- sum(diag(table(Original = Job$job_num,
    Predicted = Job_Jack$class)))/sum(table(Original = Job$job_num,
    Predicted = Job_Jack$class))
Proportion_of_Correct_Classification

## [1] 0.866
```

Jackknife classification.

```r
training_sample <- sample(1:500, 250)
Job_Train <- lda(job_num ~ gre + pubs + years, data = Job, CV = FALSE,
    subset = training_sample)
Job_Predict <- predict(Job_Train, newdata = Job[-training_sample,
    ])
Job_Original <- as.data.frame(Job[-training_sample, 6])
Job_Cross <- cbind(Job_Original, Job_Predict$class)
names(Job_Cross) <- c("Original", "Predicted")
table(Original = Job_Cross$Original, Predicted = Job_Cross$Predicted)
##         Predicted
## Original   1   2
##        1 166  18
##        2  16  50

Proportion_of_Correct_Classification <- sum(diag(table(Original = Job_Cross$Original,
    Predicted = Job_Cross$Predicted)))/sum(table(Original = Job_Cross$Original,
    Predicted = Job_Cross$Predicted))
Proportion_of_Correct_Classification

## [1] 0.864
```

Split-sample classification.

Classification accuracy here will depend on the particular random split. Bootstrapping will help us see the variability of these estimates.

Under the assumption of multivariate normality, the probability of a particular profile of scores (X) given a particular group centroid and pooled covariance matrix can be estimated:

$$p(X|Group_g) \propto \frac{1}{\sqrt{|C_w|(2\pi)^{\frac{k}{2}}}} e^{-\frac{1}{2}(X-\bar{X}_g)'(C_w^{-1})(X-\bar{X}_g)}$$

The pooled within-group covariance matrix is used if the homogeneity assumption holds.

The posterior probabilities take prior probability of group membership into account:

$$p(Group_g|X) = \frac{p(Group_g)p(X|Group_g)}{\sum_{i=1}^{G} p(Group_i)p(X|Group_i)}$$

In the iris data, the prior probabilities were equal (.33) but for the current data, the prior probability of getting a job is smaller (.272) than the prior probability of not getting a job (.728; from the sample proportions). We assume the sample is representative.

The case is assigned to the group with the highest posterior probability. The posterior probabilities are particularly useful because they indicate our confidence in classification.

	LD1	1	2	
1	-0.2686	0.9077	0.092347	1
2	-2.0386	0.9963	0.003695	1
3	0.9581	0.4975	0.502488	2
4	0.4677	0.7125	0.287474	1
5	0.1521	0.8173	0.182719	1
6	-0.9081	0.9702	0.029837	1

Is the classification better than would be expected by chance? We first need to know the correct classification that would occur by chance:

		Expected		
		No Job	Job	All
Actual	No Job	$\frac{(q_1N)^2}{N}$		q_1N
	Job		$\frac{(q_2N)^2}{N}$	q_2N
	All	q_1N	q_2N	N

Note that we assume the expected marginals to be the same as the actual marginals.

Is the classification better than would be expected by chance? We first need to know the correct classification that would occur by chance:

		Expected		
		No Job	Job	All
Actual	No Job	$\frac{(q_1N)^2}{N}$		364
	Job		$\frac{(q_2N)^2}{N}$	136
	All	364	136	500

Note that we assume the expected marginals to be the same as the actual marginals.

Is the classification better than would be expected by chance? We first need to know the correct classification that would occur by chance:

		Expected		
		No Job	Job	All
Actual	No Job	264.99	99.01	364
	Job	99.01	36.99	136
	All	364	136	500

(364 x 364)/500

The total number of correct classifications that would occur by chance (302, 60.4%) can be tested against the actual number of correct classifications given the discriminant analysis model (439, 87.8%, for the simple prediction of all cases).

A *t*-test can be calculated:

$$t = \frac{439 - 302}{\sqrt{500(.604)(1 - .604)}} = 12.53$$

where the denominator is the standard error of the number of correct classifications by chance (the null hypothesis).

The difference between chance expected and actual classification can be tested with a chi-square as well.

$$\chi^2 = \sum_{i=1}^{C} \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}}$$

$$\chi^2 = \frac{(439-302)^2}{302} + \frac{(61-198)^2}{198} = 156.94$$

Because this is a single degree of freedom test, $t^2 = \chi^2$.

Klecka's tau(τ) is sometimes calculated for classification results:

$$\tau = \frac{n_o - \sum_{i=1}^{G} p_i n_i}{N - \sum_{i=1}^{G} p_i n_i}$$

n_o is the number of correct classifications, n_i is the number of cases in group i, p_i is the proportion of the total sample expected to be in group i, G is the number of groups, and N is the total sample size.

$$\tau = \frac{439-302}{500-302} = .69$$

This is interpreted as the proportional improvement in classification over random assignment to groups. Values can range from 0 to 1. This is very similar to Cohen's kappa.

Is an obtained τ value different from 0? The sampling distribution of τ is unknown and so a convenient formula for the standard error is not available. The bootstrap procedure will help here as well.

- Several additional classification indices can be useful. In the following, an “event” is “getting a job”:
  - Precision: "What percentage of predicted events are correct?"
  - Recall: "What percentage of events were correctly predicted?"
  - F1: Harmonic mean of precision and recall.
  - Prevalence: "What is the proportion of actual events in the sample?"
  - Detection Rate: "What proportion of the entire sample are correctly predicted events?"

		Actual		
		Absent	*Present*	*Marginal*
Prediction	Absent	d	c	Row 1 = d+c
	Present	b	a	Row 2 = b+a
	Marginal	Column 1 = d+b	Column 2 = c+a	N=a+b+c+d

$$Recall = \frac{a}{a+c} \qquad Precision = \frac{a}{a+b}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Prevalence = \frac{a+c}{a+b+c+d} \qquad Detection\ \ Rate = \frac{a}{a+b+c+d}$$

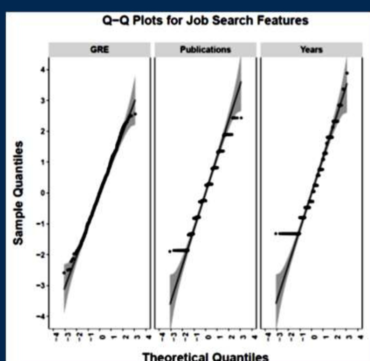
```
Confusion Matrix and Statistics

          Reference
Prediction No Job Job
    No Job    339  42
    Job        25  94

               Accuracy : 0.866
                 95% CI : (0.833, 0.895)
    No Information Rate : 0.728
    P-Value [Acc > NIR] : 8.02e-14

                  Kappa : 0.648
 Mcnemar's Test P-Value : 0.0506

            Sensitivity : 0.691
            Specificity : 0.931
         Pos Pred Value : 0.790
         Neg Pred Value : 0.890
              Precision : 0.790
                 Recall : 0.691
                     F1 : 0.737
             Prevalence : 0.272
         Detection Rate : 0.188
   Detection Prevalence : 0.238
      Balanced Accuracy : 0.811

       'Positive' Class : Job
```

The discriminant analysis produced predictions that were correct quite often (.79) and in particular correctly identified getting a job fairly well (.69), despite getting a job being a relatively rare event (.27).

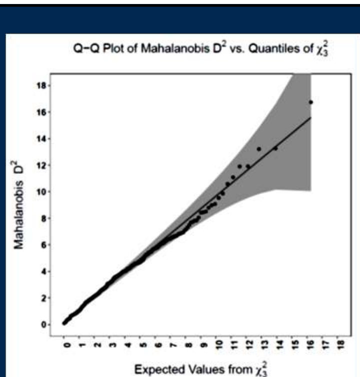
```
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Job[, 1:3]
## Chi-Sq (approx.) = 105.7, df = 6, p-value < 2.2e-16

boxM(Job[, 1:3], Job$job)$cov

## $Job
##              gre     pubs     years
## gre   10455.78 56.3826 -25.2563
## pubs     56.38  2.2065   0.5251
## years   -25.26  0.5251   1.4056
##
## $`No Job`
##              gre     pubs     years
## gre   10885.13 74.3135 -14.0374
## pubs     74.31  3.9523  -0.7029
## years   -14.04 -0.7029   4.5581

boxM(Job[, 1:3], Job$job)$pooled

##              gre    pubs    years
## gre   10768.74 69.453 -17.079
## pubs     69.45  3.479  -0.370
## years   -17.08 -0.370   3.704
```

If the assumptions fail seriously then classification and inferences can be affected.

The covariance matrices are not homogeneous.



GRE appears to be normally distributed. Publications and years, however, look a little flaky (technical statistical term).



Collectively there are no outliers.

```
mvn(Job[, 7:9], mvnTest = "mardia")

## $multivariateNormality
##              Test         Statistic               p value Result
## 1 Mardia Skewness  86.730301008051 2.37724296170123e-14     NO
## 2 Mardia Kurtosis -1.6503616863684   0.0988689824798465    YES
## 3             MVN             <NA>                 <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   gre_R      0.9962  0.2749      YES
## 2 Shapiro-Wilk  pubs_R      0.9683  <0.001      NO
## 3 Shapiro-Wilk  years_R     0.9349  <0.001      NO
```

The significance tests for normality are surely being influenced by the very large sample size.

If we have doubts about the legitimacy of significance tests or classification based on normality, we can use methods that don't rely on that assumption.

When the homogeneity of covariance matrices assumption fails, a pooled within-groups covariance matrix is not appropriate. An alternative form of analysis known as quadratic discriminant analysis can be used. This uses the separate group variance- covariance matrices in the classification process.

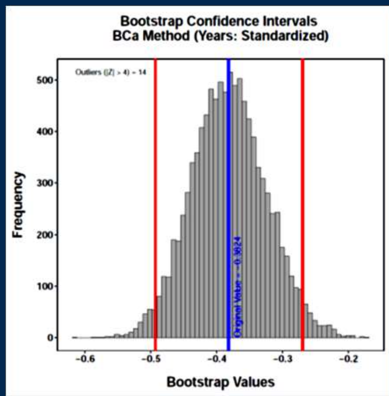
```
table(Original = Job$job_num, Predicted = Job_QDA_P$class)

##         Predicted
## Original   1   2
##        1 333  31
##        2  28 108
```

The jackknife classification is nearly identical to the original analysis.

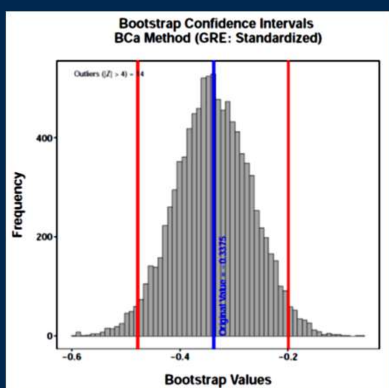
If the assumptions underlying the discriminant analysis (homogeneous covariance matrices, multivariate normality) are not viable, the bootstrapping approach can be taken.

In bootstrapping, we assume that whatever population the sample came from, the sample is representative of that population. Therefore we can sample randomly from the sample, with replacement, to get multiple samples of the same size on which we can repeat the analyses.

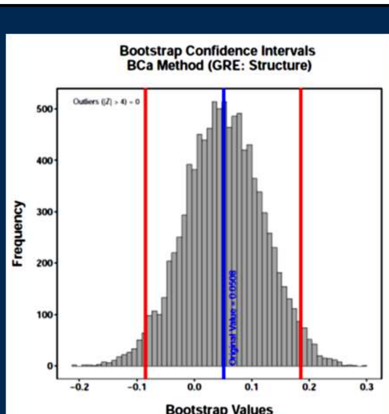
The resulting empirical sampling distributions can be examined to assist inferences.



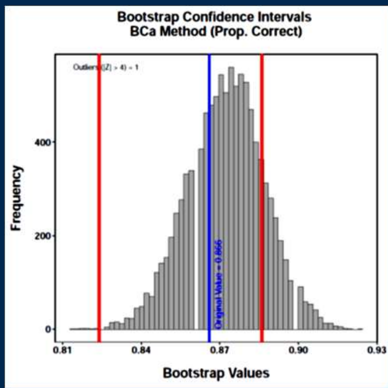
The 95% confidence intervals for the standardized discriminant coefficient for years excludes 0.



Similarly, the bootstrap estimates for the standardized discriminant coefficient for GRE are less than 0.

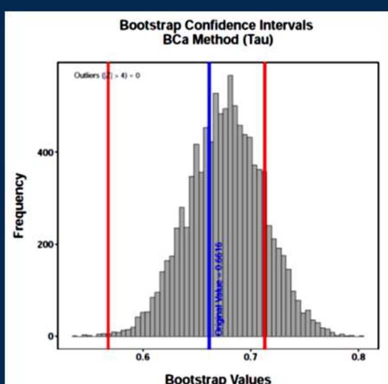


But, the 95% confidence interval for the GRE structure coefficient includes 0.



Classification is accurate and between about 82% and 88% with 95% confidence.

Note the shift in the confidence interval relative to the distribution.



The proportional gain in classification over chance (τ) is moderate. Bootstrapping is particularly useful here because the sampling distribution of τ is unknown.

Bootstrapping is helpful if multivariate normality is out of the question, homogeneity of covariance matrices is not tenable, or sampling distributions are unknown.

But, other “problems” require a different approach. What if the groups are ordered in some way that we would like to incorporate into the statistical model? What if we have particular interest in comparisons among some of the groups? What if we would like to include interactions or polynomial terms with the predictors?

Next time . . .

Logistic regression