Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
DEPARTAMENTO DE ESTATÍSTICA

TRABALHO - PARTE 2
RELATÓRIO - QUESTÃO 1

ELIANE RAMOS DE SIQUEIRA RA:155233
GUILHERME PAZIAN RA:160323
HENRIQUE CAPATTO RA:146406
MURILO SALGADO RAZOLI RA:150987

Disciplina: **ME731 - Análise Multivariada**
Professor: **Caio Lucidius Naberezny Azevedo**

Campinas - SP
24 de Novembro de 2017

1. Introdução

Os salmões nascem em água doce mas migram para o mar retornando, posteriormente, para o local onde nasceram, para fins de reprodução, por isso, caso grande parte da população de salmão nascidos em um local específico for pescado, ocorre uma diminuição na quantidade de salmãos que conseguirão se reproduzir neste local, gerando excasses destes peixes. A fronteira do Canadá com o Alasca é uma importante área de pesca de salmão, e o mercado deste peixe tem significante importância na economia (veja tabela 1 "tabela com as quantidades de salmão EUA/CANADA e $") assim como exerce forte influência na excasses ou abundancia deste peixe (num próximo ciclo reprodutivo) no local de reprodução. Existem basicamente dois tipos de Salmão nessa região, uma que nasce no Alasca e outra que nasce no Canadá, pela proximidade, um salmão nascido no Alasca, pode acabar sendo pescado no mar por um pescador do Canada e vice versa. Os dados são desconhecidos, mas os pescadores do Alasca eram conhecidos por interceptar grandes quantidades de salmão Canadense, e os pescadores Canadenses tinham menos oportunidade de interceptar salmão originário do Alaska. Este fato gerou alguns conflitos entre Estados Unidos e Canada, tanto que em 1985 estes países fizeram um tratado para pesca de salmão do Oceano Pacífico (Pacific Salmon Treaty'), ao qual proibe a pesca de salmão do tipo que nasce no Canadá por pescadores Norte Americanos e do tipo que nasce no Alasca por pescadores Canadenses. A fim de seguir o tratado é impressindível conseguir diferenciar os tipos de salmão originário do Alasca e do Canadá.

Veja mais sobre esse conflito em (THE PACIFIC SALMON TREATY: A BRIEF TRUCE IN THE CANADA/U.S.A. PACIFIC SALMON WAR (COLOCAR NAS REFERÊNCIAS)).

Com o presente trabalho, pretende-se criar uma regra de classificação, visando poder identificar mais facilmente a origem de salmões pescados, utilizando-se um banco de dados contendo duas variáveis intituladas aqui como DGAD e DGM (diâmetro da guelra (em mm) na fase de água doce e na fase no mar respectivamente) medidas em 50 salmões provenientes do Alasca e em 50 salmões provenientes do Canadá, assim como o sexo destes peixes. A origem dos salmões (Alasca ou Canadá) é identificada por uma variável aqui utilizada indistintamente como Região, localidade e/ou grupo.

Todas as análises serão realizadas com o suporte dos softwares *R* versão 3.4.2 e *RStudio* versão 1.1.383. Foi-se considerado um nível de significância de 5% para a tomar decisões quanto aos testes estatísticos aqui apresentados.

Tabela 1: Informações sobre pesca de Salmão no ano de 2015 para Alasca (veja $(link)[http://www.adfg.alaska.gov/index.cfm?adfg=commercialbyfisherysalmon.salmon_combined_historical]$ (((colocar nas referencias)))) e para Canadá (veja $(link)[http://www.pac.dfo-mpo.gc.ca/stats/comm/summ-somm/annsumm-sommann/2015/ANNUAL15_USER_three_party_groups-eng.htm]$ (((colocar nas referencias))))

Origem	Toneladas	Mil Dólares (EUA)
Alaska	120280	494783
Canadá	6534	14168

2. Análise Descritiva

A tabela 1, apresenta mostra algumas medidas resumo para as vaiáveis DGAD e DGM separadas por região (Alasca e Canadá).

Tabela 2: Medidas Resumo das variáveis por região

	Região	n	Media	Variancia	Desvio Padrao	CV(%)	Minimo	Mediana	Maximo
DGAD	Alasca	50	98,38	260,608	16,143	16,409	53	99	131
	Canadá	50	137,46	326,09	18,058	13,137	90	140	179
DGM	Alasca	50	429,66	1399,086	37,404	8,706	355	427,5	511
	Canadá	50	366,62	893,261	29,887	8,152	301	369,5	438

A partir da Figura 1, que consta o gráfico de dispersão entre as variaveis separadas por Região (Alasca e Canadá), podemos observar que os indíviduos do Canáda tendem a ter um diâmetro (em mm) da guelra durante a fase de água doce maior que os indivíduos do Alasca e durante a fase no Mar tendem a ter um diametro menor (em mm). Se olharmos separadamente os dois grupos, vemos que existe uma correlação levemente positiva entre os índividuos do Canadá, e levemente negativa para os índividuos do Alasca. Vale ressaltar que se considerarmos ambos os Grupos, parecer haver uma correlação negativa entre as variaveis.

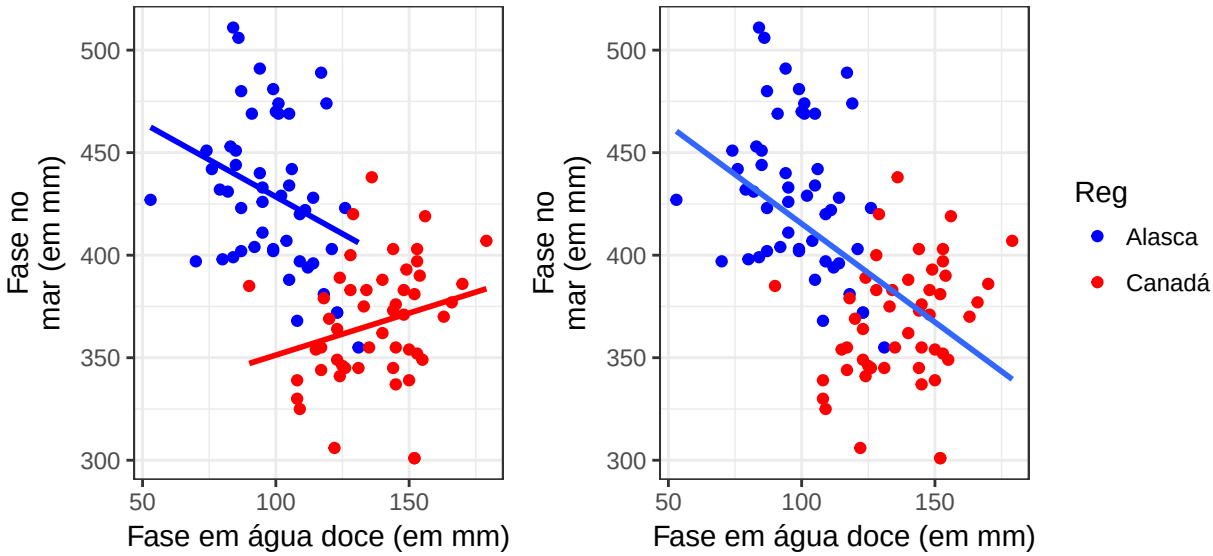

Figura 1: Figura 1: Gráfico de dispersão entre os diâmetros da guelra de salmões em água doce e no mar

Na Figura 2 notamos que o diâmetro da guelra é menor na água doce para os índividuos de ambos os grupos, constatamos isso também na distribuição de densidade na Figura X, e reforçado nos valores das medidas centrais, como a da média e mediana, na Tabela X. Estes resultados são esperados devido a natureza dos dados, pois as medidas em água doce foram obtidas na fase inícial da vida dos peixes e as medições obtidas no mar aconteceram na maturidade destes salmões.

Podemos observar também nos box-plot, que o diâmetro das guelras para o salmão do Canadá é consideravelmente maior

do que os do Alasca, já durante a fase no mar, o contrário ocorre. Para ambas as variáveis (DGAD, DGM) é possível notar uma sobreposição em boa parte da distribuições apresentadas Figura 3. As distribuições parecem ser levemente assimétricas, mais evidente para o diâmetro da guelra no mar de indivíduos do Canadá, e menos evidente para diâmetro da guelra na água doce de indivíduos do Alasca.

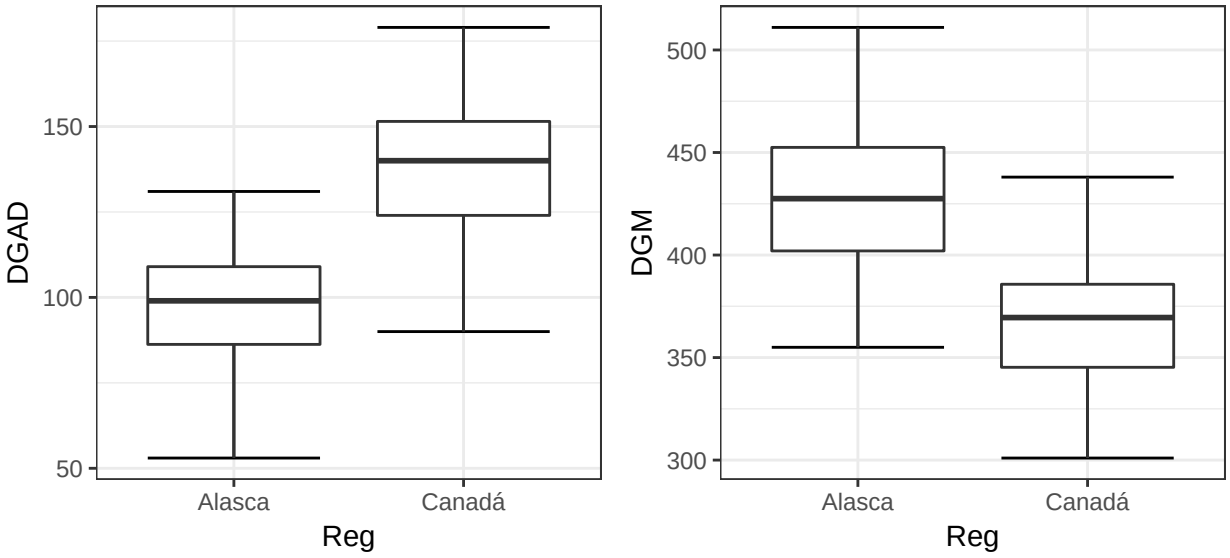

Figura 2: Boxplots por grupo

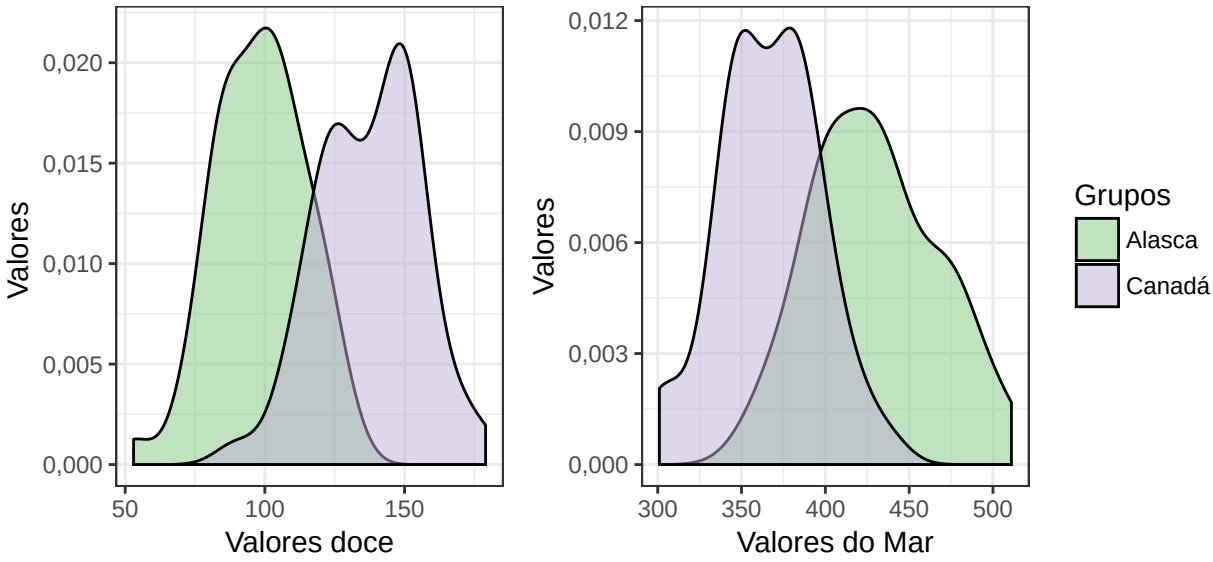

Figura 3: Figura 3: Distribuição estimada por grupo

```
## NULL
```

Podemos observar na Figura 4 que para cada variavel DGAD e DGM foi realizado um gráfico de quantil-quantil para cada Grupo (Canadá e Alasca), vemos que para o diâmetro da guelra na fase em água doce de indíviduos do Alasca, os pontos se comportaram de maneira razoávelmente aleatória em torno da linha de referência, não apresentando sinais evidentes de tendência,

já os outros três gráficos, apresentam uma certa sitematização no comportamento dos dados em torno da linha de referência, evidenciando uma certa tendência, embora pareça ser uma tendencia amena, é um ponto contra a suposição de normalidade dos dados, porém não seria irrazuável supor normalidade para as variáveis neste caso dada a fraca intensidade desta tendência.

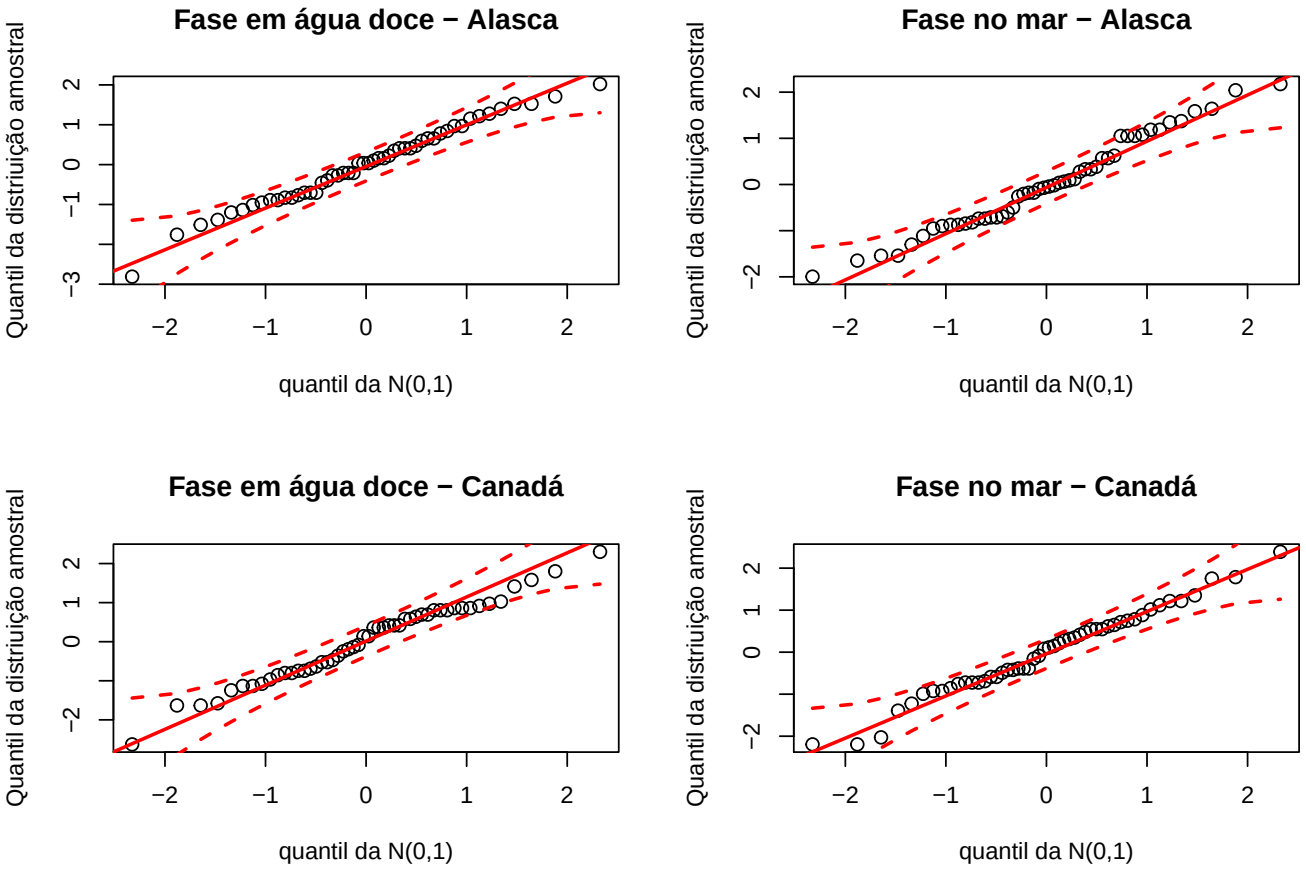

Figura 4: Figura 4: Quantil-quantil para cada Grupo

No gráfico abaixo podemos ver que a normalidade bivariada não parece ser uma suposição razoavel para nenhum dos grupos, porque além da sistematização em torno da linha de referência, existem muitos pontos fora das bandas de confiança. Dadas as observações, vemos que a suposição de normalidade multivariada para ambas as regiões parece ser irrazuável. Contudo, a técnica de análise discriminante de Fisher foi desenvolvida sem supor normalidade dos dados, portanto, a suposição de normalidade avaliadas com base nos gráficos 4 e 5 não são relevantes para a aplicação da técnica de análise discriminante de Fisher.

Foi realizado o teste de Box para igualdade de matrizes de covariâncias dos dados dos dois tipos de salmão (alasca e Canadá), ao qual resultou num p-valor = 0,013, indicando que existe diferença estatisticamente significante entre as matrizes de covariâncias dos tipos de salmão, indicando que não parece ser razoável a suposição de igualdade das matrizes de covariâncias entre os tipos de salmão. Este resultado têm grande relevância, uma vez que a técnica de análise discriminante de Fisher supõe homocedásticidade multivariada, ou seja, igualdade das matrizes de covariâncias para os grupos. O teste de box, portanto, indicou que a suposição de homocedásticidade neste caso não parece ser razoável quando se considera os grupos de salmão originários do

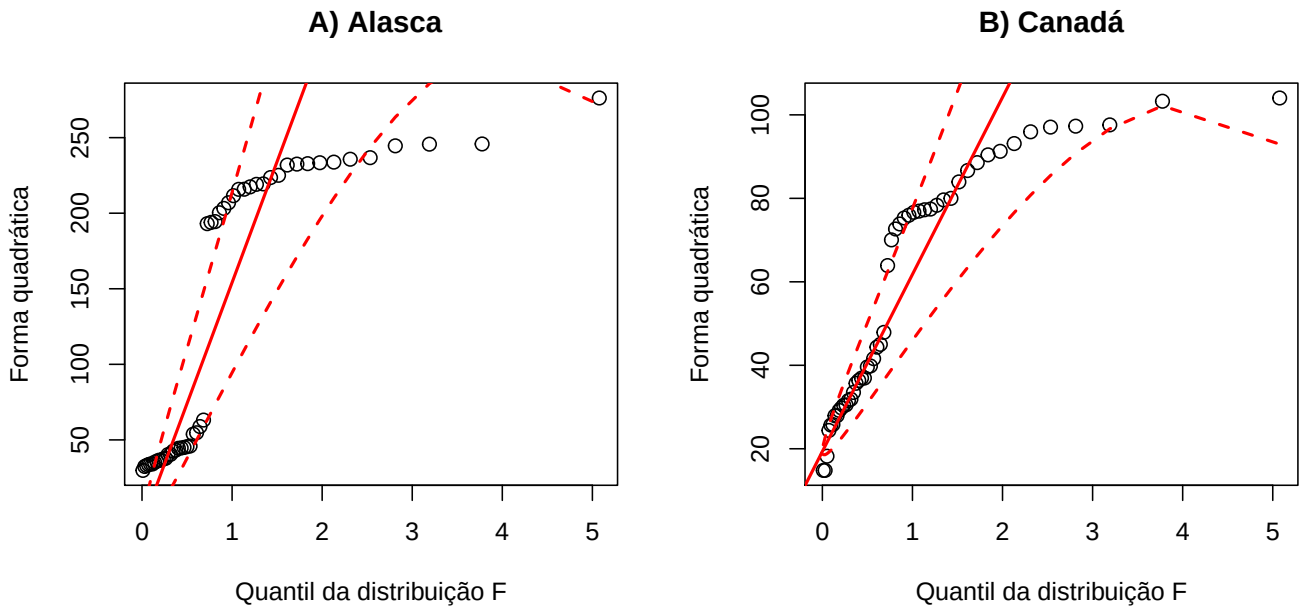

Figura 5: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis; A) Alasca, B) Canadá

Alasca e do Canadá.

teste de igualdade de matrizes de covariância considerando os sexos

Foi realizado o teste de Box para igualdade de matrizes de covariâncias dos dados dos dois tipos de salmão macho, ao qual resultou num p-valor 0,013, indicando que existe diferença estatisticamente significante entre as matrizes de covariâncias dos tipos de salmão machos, indicando que não parece ser razoável a suposição de igualdade das matrizes de covariâncias entre os tipos de salmão macho. Este resultado têm grande relevância, uma vez que a técnica de análise discriminante de Fisher supõe homocedásticidade multivariada, ou seja, igualdade das matrizes de covariâncias para os grupos. O teste de Box, portanto, indicou que a suposição de homocedásticidade neste caso não parece ser razoável quando se considera os grupos de salmão do sexo masculino originários do Alasca e do Canadá.

Foi realizado o teste de Box para igualdade de matrizes de covariâncias dos dados dos dois tipos de salmão fêmea, ao qual resultou num p-valor 0,013, indicando que existe diferença estatisticamente significante entre as matrizes de covariâncias dos tipos de salmão femeas, indicando que não parece ser razoável a suposição de igualdade das matrizes de covariâncias entre os tipos de salmão femea. Este resultado têm grande relevância, uma vez que a técnica de análise discriminante de Fisher supõe homocedásticidade multivariada, ou seja, igualdade das matrizes de covariâncias para os grupos. O teste de box, portanto, indicou que a suposição de homocedásticidade neste caso não parece ser razoável quando se considera os grupos de salmão do sexo feminino originários do Alasca e do Canadá.

3. Análise Inferencial

Não encontramos nenhuma informação que nos dê direcionamento direto à definição de probabilidades à priori de um salmão ser proveniente de uma ou outra localidade (Alasca ou Canada), uma vez que essa probabilidade está muito relacionada ao local onde o salmão foi pescado. Por não ter informações suficientes iríamos supor probabilidades iguais para cada localidade, porém como foi orientação utilizar probabilidades diferentes para cada localidade, utilizamos os dados sobre toneladas de salmão comercial pescados e o respectivo valor monetário gerado no ano de 2015 (dado mais atual) a partir dessa pesca para as duas localidades, estes valores são apresentados na tabela 01 ("tabela com as quantidades de salmão Alasca/CANADA e $"). Observe que o volume de pesca de salmão para o ano de 2015 é muito maior no Alasca em comparação com o Canadá, isso nos leva a acreditar que a população de salmão do Alasca é maior que a população de salmão do Canadá, o que nos leva a conjecturar que a probabilidade de um salmão ser originário do Alasca é maior do que um salmão ser originário do Canadá. Acreditamos que considerar a probabilidade à priori de um salmão ser originário do Alasca como sendo 0,6 e ser originário do Canadá como sendo 0,4 parece ser razoável diante dos dados da tabela 01 e da relação levantada entre a probabilidade do salmão pertencer a uma determinada localidade e o local da pesca, uma vez que não se têm informações mais precisas quanto às populações de salmão e seus respectivos comportamentos migratórios destes de ambas as localidades.

Com base na metodologia de Análise discriminante de Fisher, considerando custos iguais de classificação errada e probabilidades à priori destacadas acima, obtivemos uma regra de classificação à qual estão apresentados os resultados abaixo.

Tabela 3: Resultados da classificação da amostra teste

	Alasca	Canadá
Alasca	23	2
Canadá	1	24

Ao observar a tabela XX (Resultados da classificação da amostra teste), observamos a qualidade da regra de classificação, obtemos uma taxa de erro aparente TEA = 6 %, valor bem próximo ao da taxa ótima de erro TOE = 5,26 % ao qual leva em consideração a validade das suposições de normalidade multivariada e igualdade das matrizes de covariância relativas aos dois tipos de salmão, ou seja, mesmo com as observações indicativas à fuga das suposições mencionadas anteriormente, a regra de classificação mostrou uma taxa de erro bem próxima à taxa ótima, indicando uma boa performance da regra proposta. TEA = 6 %

TOE = 5,2634733 %

Tabela 4: Medidas resumo para os valores função discriminante aplicada na amostra teste, por grupo:

	Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
1	Alasca	-0,63	1,15	1,33	-3,16	-0,62	1,94	25
2	Canadá	1,97	0,96	0,93	-0,64	1,97	3,77	25

4.Conclusões

5. Bibliografia

- Azevedo, C. L. N. (2017). Notas de aula sobre análise multivariada de dados http://www.ime.unicamp.br/~cnaber/Material_AM_2S_2017.htm
- Johnson, R. A. & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. 6 a edição, Upper Saddle River, NJ: Pearson Prentice Hall.

0.1 GENERO

```
## Confusion Matrix and Statistics
##
##              gf.pred
## gf.data.teste Alasca Canadá
##        Alasca     12      1
##        Canadá      0     13
##
##                Accuracy : 0,9615
##                  95% CI : (0,8036, 0,999)
##     No Information Rate : 0,5385
##     P-Value [Acc > NIR] : 2,383e-06
##
##                   Kappa : 0,9231
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 1,0000
##             Specificity : 0,9286
##          Pos Pred Value : 0,9231
##          Neg Pred Value : 1,0000
##              Prevalence : 0,4615
##          Detection Rate : 0,4615
##    Detection Prevalence : 0,5000
##       Balanced Accuracy : 0,9643
##
```

```
##        'Positive' Class : Alasca
##
## Confusion Matrix and Statistics
##
##              gm.pred
## gm.data.teste Alasca Canadá
##        Alasca     12      0
##        Canadá      1     11
##
##                Accuracy : 0,9583
##                  95% CI : (0,7888, 0,9989)
##     No Information Rate : 0,5417
##     P-Value [Acc > NIR] : 8,672e-06
##
##                   Kappa : 0,9167
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0,9231
##             Specificity : 1,0000
##          Pos Pred Value : 1,0000
##          Neg Pred Value : 0,9167
##              Prevalence : 0,5417
##          Detection Rate : 0,5000
##    Detection Prevalence : 0,5000
##       Balanced Accuracy : 0,9615
##
##        'Positive' Class : Alasca
##
```

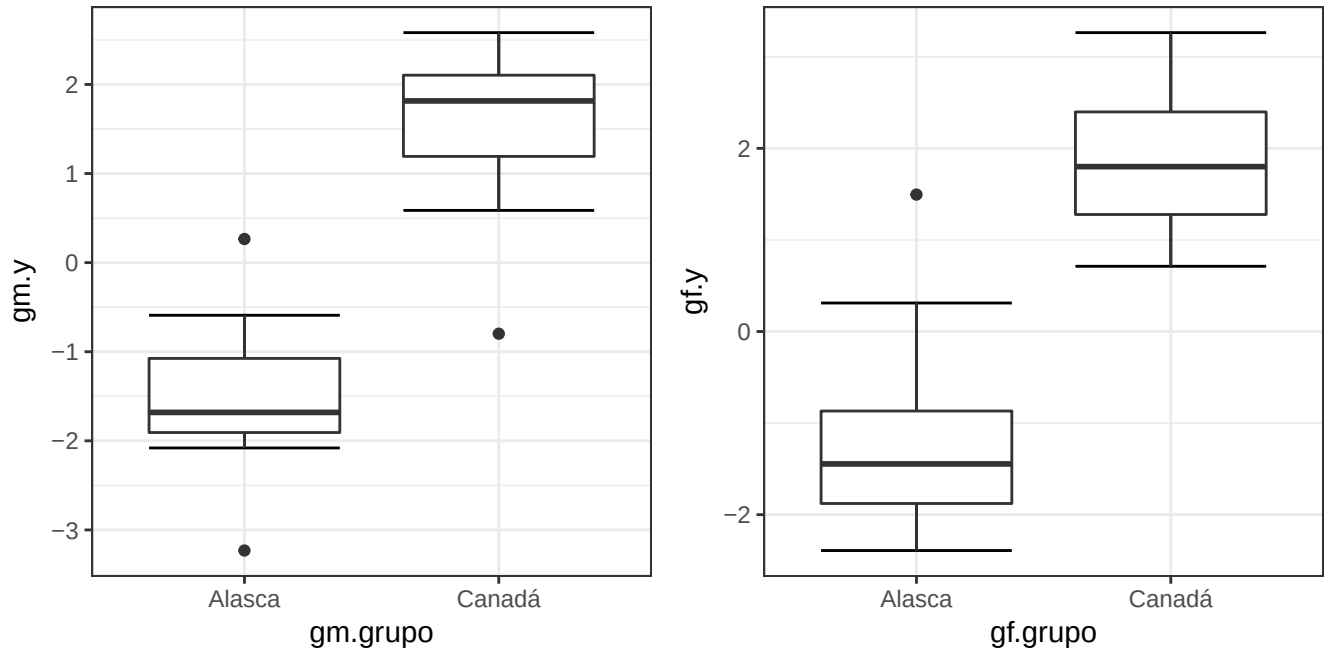

Figura 6: Boxplots da função discriminante aplicada à amostra teste, por grupo

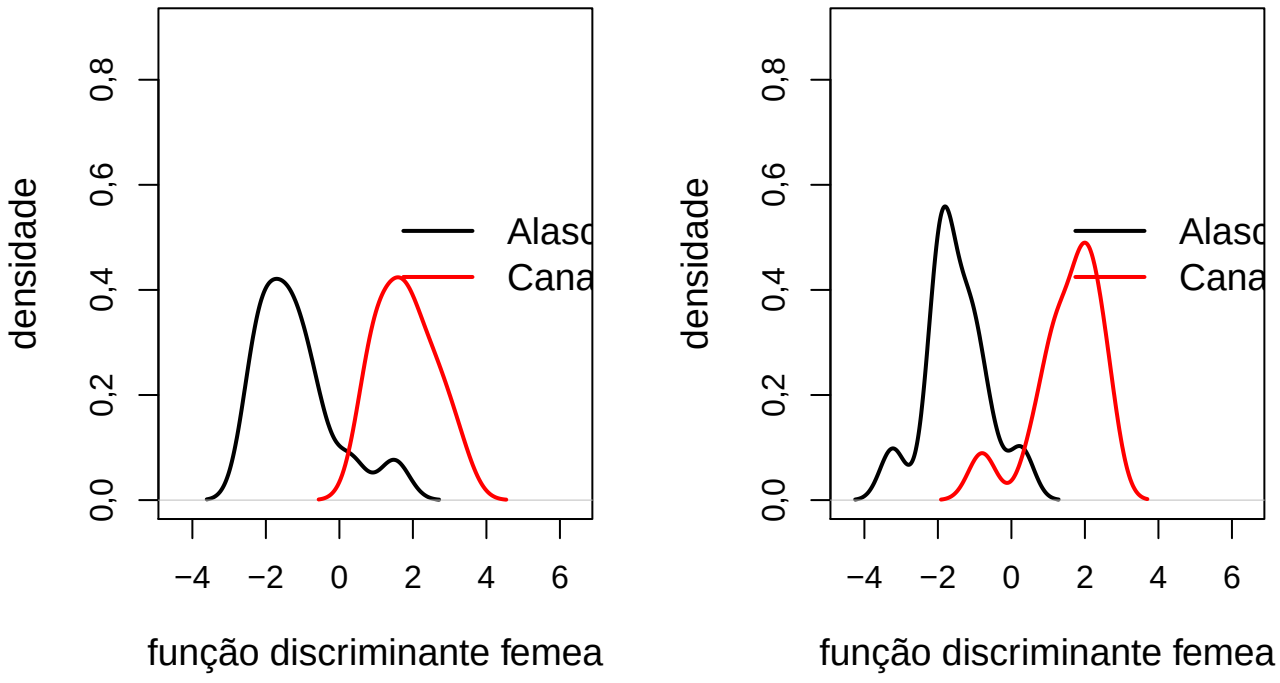

Figura 7: Densidade estimada da função discriminante aplicada à amostra teste, por grupo