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

O Salmão é um peixe cujo mercado tem uma importância significante para economia, como podemos ver na Tabela 1. A fronteira do Canadá com o Alasca é uma importante área de pesca de salmão, exerce forte influência na escassez ou abundância deste peixe (em um próximo ciclo reprodutivo) no local de reprodução. Existem basicamente dois tipos de Salmão nessa região: um que nasce no Alasca e outro que nasce no Canadá. Devido a proximidade, um salmão nascido no Alasca, pode acabar sendo pescado no mar por um pescador do Canadá e vice versa. Os salmões nascem em água doce mas migram para o mar e retornam, posteriormente, para o local onde nasceram, para fins de reprodução. Por isso, caso grande parte da população de salmão nascida em um local específico for pescado, ocorre uma diminuição na quantidade de salmãos que conseguirão se reproduzir neste local, gerando escassez destes peixes.

Os pescadores do Alasca eram conhecidos por interceptar grandes quantidades de salmão canadense, deixando os canadenses com menos oportunidade de interceptar salmão originário do Alasca. Este fato gerou alguns conflitos entre Estados Unidos e Canadá, tanto que em 1985 estes países fizeram um tratado para pesca de salmão do Oceano Pacífico (*Pacific Salmon Treaty*), proibindo a pesca de salmão do tipo que nasce no Canadá por pescadores norte americanos e do tipo que nasce no Alasca por pescadores canadenses. Portanto, para facilitar o seguimento do tratado, é imprescindível conseguir diferenciar os tipos de salmão por sua região de origem. Mais informações sobre esse conflito na referência 1 da bibliografia.

Neste relatório, pretende-se criar uma regra de classificação, visando descomplicar a identificação da origem dos salmões pescados. O banco de dados utilizado contém duas variáveis intituladas como DGAD e DGM (diâmetro da guelra (em mm) na fase de água doce e diâmetro da guelra (em mm) na fase no mar respectivamente) medidas em 50 salmões provenientes do Alasca e em 50 salmões provenientes do Canadá, assim como o gênero de cada peixe. Neste relatório, nos referiremos a origem dos salmõesindistintamente como Região, localidade e/ou grupo.

Todas as análises foram realizadas com o suporte dos softwares *R* versão 3.4.2 e *RStudio* versão 1.1.383.

Foi considerado um nível de significância de 5% para tomada de decisões quanto aos resultados dos testes estatísticos aqui apresentados.

Tabela 1: Informações sobre pesca de Salmão no ano de 2015 para Alasca (veja referência 2) e para Canadá (veja referência 3)

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

2. Análise Descritiva

A Tabela 2, apresenta algumas medidas resumo para as variáveis DGAD e DGM separadas por região (Alasca e Canadá). É possível notar que as média amostral para a variável DGAD é maior no Canadá, enquanto para a variável DGM ocorre o oposto. Podemos notar também que os desvios padrão para a variável DGM é consideravelmente diferente entre o Canadá e o Alasca, já para a variável DGAD essa diferença é bem menor.

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 variáveis separadas por Região (Alasca e Canadá), podemos observar que os peixes do Canadá tendem a ter um diâmetro (em mm) da guelra durante a fase de água doce maior que os peixes do Alasca e durante a fase no Mar tendem a ter um diâmetro menor (em mm). Se olharmos separadamente os dois grupos, vemos que existe uma correlação levemente positiva de valor 0,25 entre os peixes do Canadá, e levemente negativa de valor -0,31 para os peixes do Alasca. Vale ressaltar que se considerarmos ambos os Grupos, parecer haver uma correlação negativa entre as variáveis.

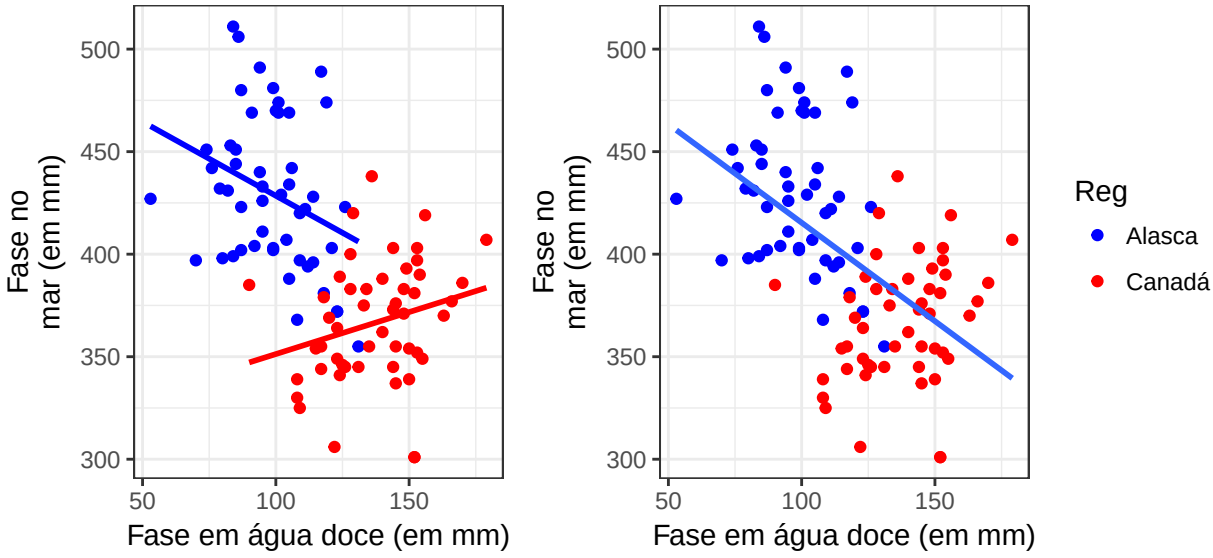

Figura 1: Gráfico de dispersão entre os diâmetros da guelra de salmões em água doce e no mar

Na Figura 2 notamos que o diâmetro da guelra é menor na água doce para os peixes de ambos os grupos, o que constatamos

também na distribuição de densidade na Figura 3 e é reforçado nos valores das medidas centrais, como a da média e mediana, na Tabela 2. Estes resultados são esperados devido a natureza dos dados, pois as medidas em água doce foram obtidas na fase inicial 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. Durante a fase no mar, ocorre o contrário. Para ambas as variáveis (DGAD, DGM) é possível notar uma sobreposição em boa parte da distribuições apresentadas na Figura 3. As distribuições parecem ser levemente assimétricas, mais evidente para o diâmetro da guelra no mar dos peixes do Canadá, e menos evidente para diâmetro da guelra na água doce dos salmões do Alasca.

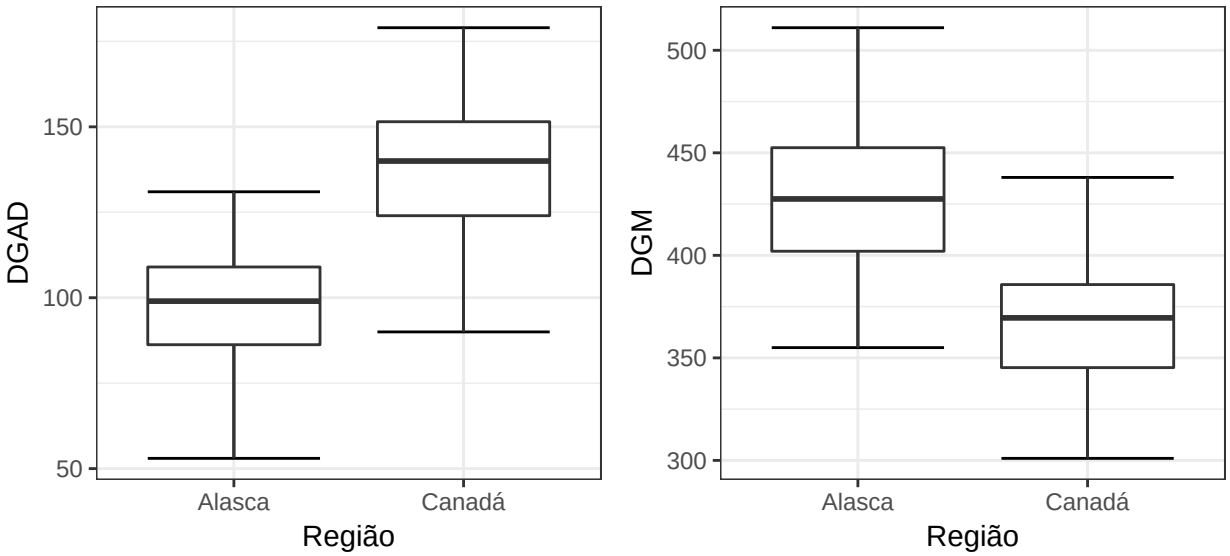

Figura 2: Boxplots por grupo

Podemos observar na Figura 4 que para cada variável 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 peixes do Alasca, os pontos se comportaram de maneira razoavelmente aleatória em torno da linha de referência, não apresentando sinais evidentes de tendência. Os outros três gráficos, apresentam uma certa sistematização no comportamento dos dados em torno da linha de referência, evidenciando uma certa tendência, embora pareça ser menos contundente, e configurando um ponto contra a suposição de normalidade dos dados> Porém não seria irrazoável supor normalidade para as variáveis neste caso dada a fraca intensidade desta tendência.

Na Figura 5, podemos ver que a normalidade bivariada não parece ser uma suposição razoável 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 irrazoável. Contudo, a

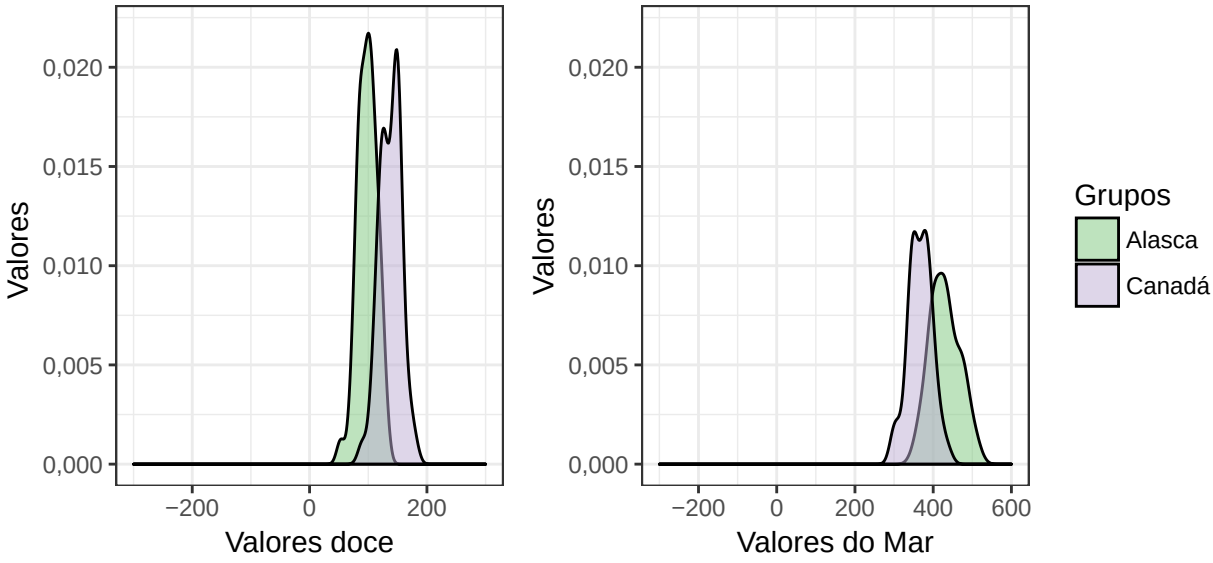

Figura 3: Distribuição estimada por grupo

técnica de análise discriminante de Fisher foi desenvolvida sem supor normalidade dos dados. Assim, a suposição de normalidade avaliadas com base nos gráficos 4 e 5 não é relevante para a aplicação da técnica de análise discriminante de Fisher.

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

Também foi realizado um 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 fêmeas, indicando que não parece ser razoável a suposição de igualdade das matrizes de covariâncias entre os tipos de salmão femea. O teste de box, portanto, indicou que a suposição de homocedasticidade neste caso não parece ser razoável quando se considera os grupos de salmão do sexo feminino originários do Alasca e do Canadá.

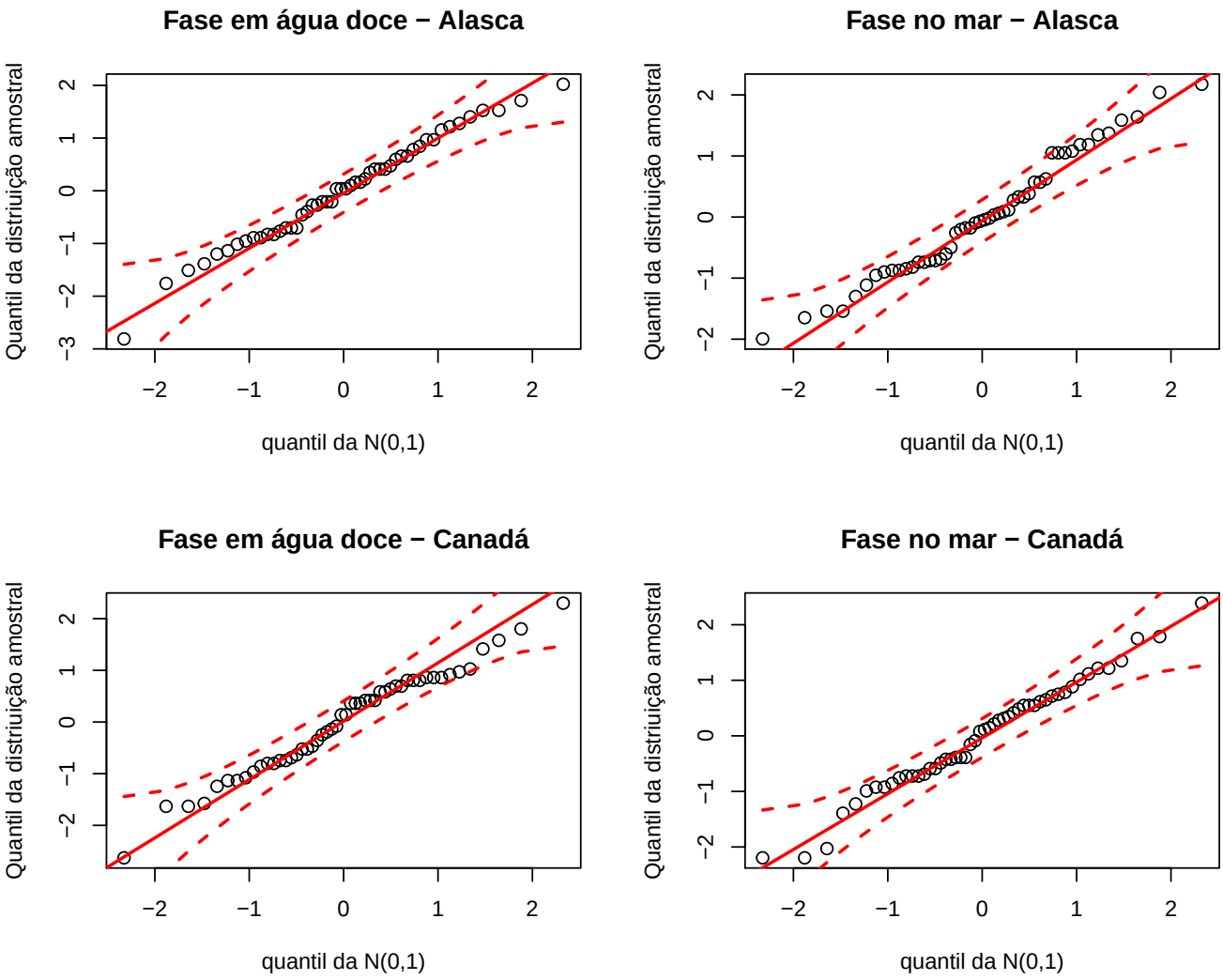

Figura 4: Quantil-quantil para cada Grupo

3. Análise Inferencial

Não encontramos nenhuma informação que nos dê direcionamento direto à definição de probabilidades a priori de um salmão ser proveniente de uma ou outra localidade (Alasca ou Canadá), 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 orientado utilizar probabilidades diferentes para cada localidade, utilizamos os dados sobre toneladas de salmão comercial pescados e seus respectivos valores monetários gerados 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 $”). Observamos 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 e 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.

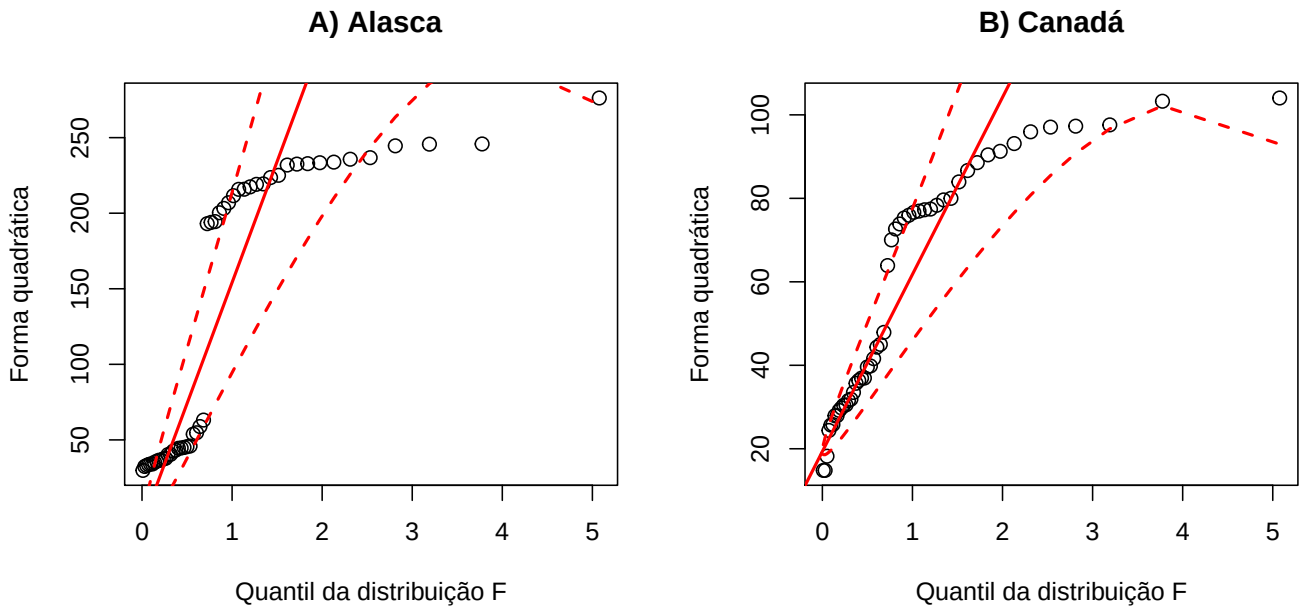

Figura 5: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis; A) Alasca, B) Canadá

Acreditamos que considerar a probabilidade a 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 1 e da relação levantada entre a probabilidade do salmão pertencer a uma determinada localidade e o local da pesca, uma vez que não se tem 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 a priori destacadas acima, obtivemos uma regra de classificação a qual estão apresentados os seguintes resultados.

Tabela 3: Resultados da classificação da amostra teste

	Alasca	Canadá
Alasca	23	2
Canadá	1	24

Ao observar a Tabela 3 (Resultados da classificação da amostra teste), observamos a qualidade da regra de classificação e obtemos uma taxa de erro aparente TEA = 6 %, valor bem próximo ao da taxa ótima de erro TOE = 5,263 % ao qual leva em consideração a validade da suposição de igualdade das matrizes de covariância relativas aos dois tipos de salmão, ou seja, mesmo com as observações indicativas à fuga da suposição mencionadas, a regra de classificação mostrou uma taxa de erro bem próxima à taxa ótima, indicando uma boa performance da regra proposta.

A Tabela 4 (Medidas resumo para os valores função discriminante aplicada na amostra teste, por grupo:) apresenta medidas resumo para os valores da função discriminante e a Figura 6 (Boxplots da função discriminante aplicada à amostra teste, por grupo) apresenta os boxplots para estes mesmos valores. Notamos que o valor máximo para o grupo Alasca é bem próximo ao valor da mediana para o grupo Canadá. O mesmo acontece para o mínimo do grupo Canadá em comparação com a mediana do grupo Alasca, o desvio padrão dos valores dos dois grupos é bastante similar (diferença de 0,19). Para ambos os grupos, os valores da média e mediana se apresentam com valores bastante próximos e os boxplots parecem ser simétricos (ou pouco assimétricos).

A Figura 7 (Densidade estimada da função discriminante aplicada à amostra teste, por grupo)) apresenta as densidades estimadas da função discriminante para os grupos. À luz do comportamento apresentado pelos boxplots temos uma interseção entre estas densidades, o que demonstra que a regra de decisão obtida não consegue isolar totalmente as distribuições e consequentemente, diferenciar totalmente os tipos de salmão. Porém isso é intrínseco ao banco de dados, já que as variáveis propriamente não tem um comportamento totalmente distinto entre os tipos de salmão.

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

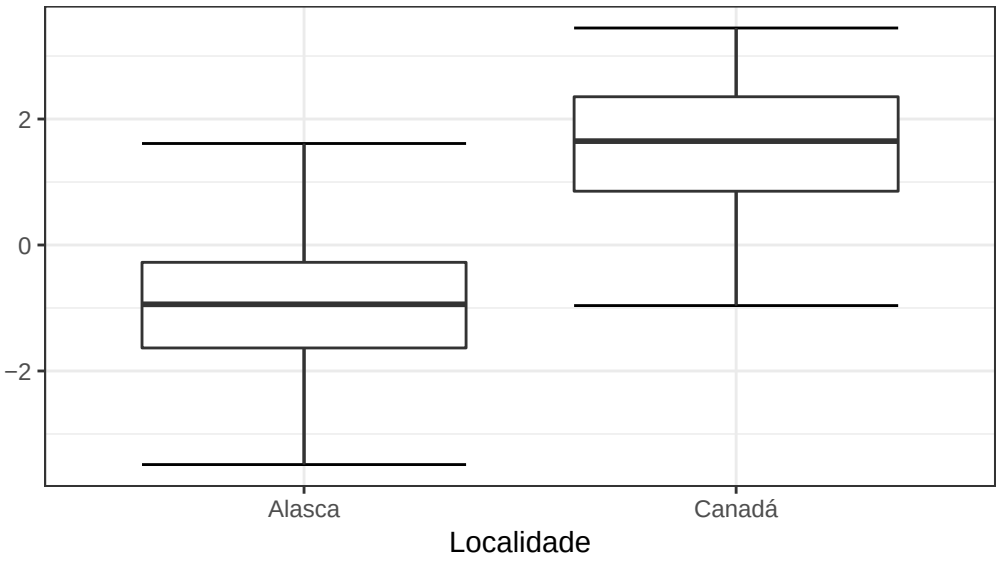

Figura 6: Boxplots da função discriminante aplicada à amostra teste, por grupo

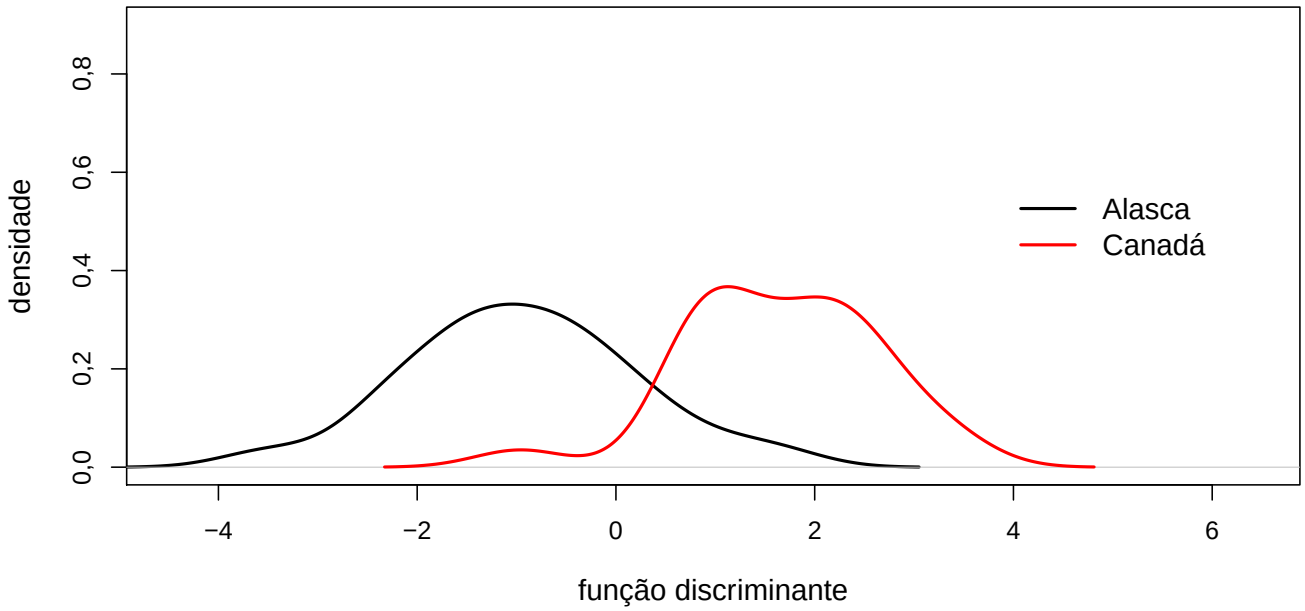

Figura 7: Densidade estimada da função discriminante aplicada à amostra teste, por grupo

Análise considerando a variável gênero

4. Análise Descritiva

Na Tabela 5, podemos observar as medidas resumo de cada uma das variaveis separadas por gênero.

Tabela 5: Medidas Resumo das variáveis por região, considerando a variável gênero

	Gênero	n	Media	Variancia	Desvio Padrao	CV(%)	Minimo	Mediana	Maximo
1	Fêmea	52	118,058	777,114	27,877	23,613	53	118,5	179
2	Macho	48	117,771	580,734	24,098	20,462	79	117	170
3	Fêmea	52	396,327	1808,773	42,53	10,731	301	397,5	481
4	Macho	48	400,104	2533,457	50,333	12,58	306	388,5	511

A partir da Figura 8 e 9, consta o gráfico de dispersão entre as variaveis separadas por Região (Alasca e Canadá). Para o gênero Fêmea, podemos observar que os peixes 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 de valor 0,22 entre os indivíduos do Canadá, e levemente negativa de valor -0,35 para os índividuos do Alasca. Vale ressaltar que se considerarmos ambos os Grupos, parecer haver uma correlação levemente negativa entre as variaveis. O mesmo é observado para o grupo dos machos,se olharmos separadamente os dois grupos, vemos que existe uma correlação levemente positiva de valor 0,27 entre os indivíduos do Canadá, e

levemente negativa de valor -0,36 para os índividuos do Alasca.

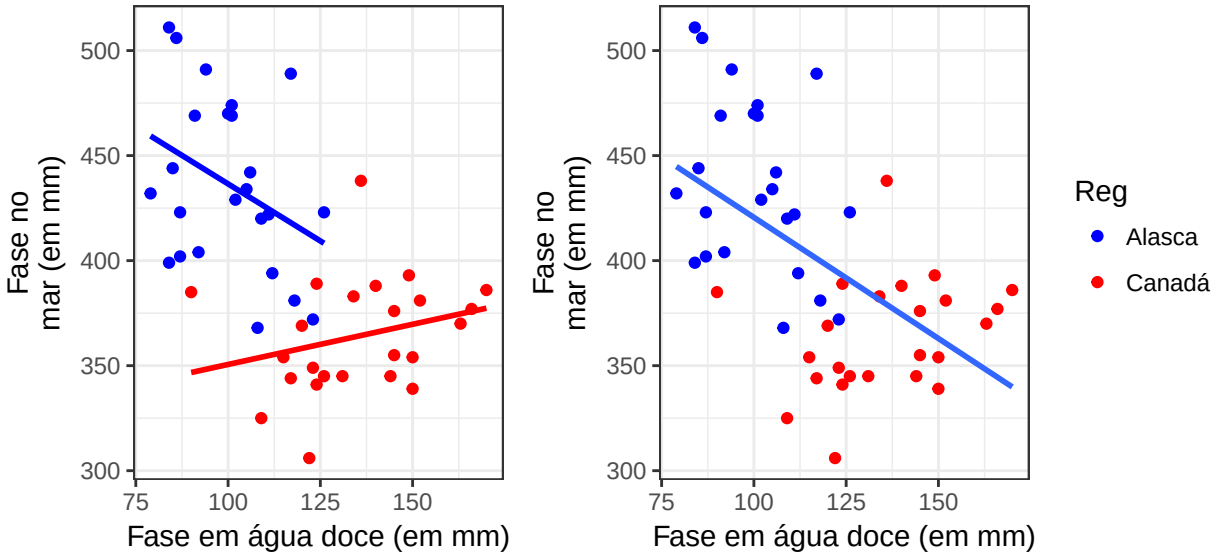

Figura 8: Gráfico de dispersão entre os diâmetros da guelra de salmões em água doce e no mar - Macho

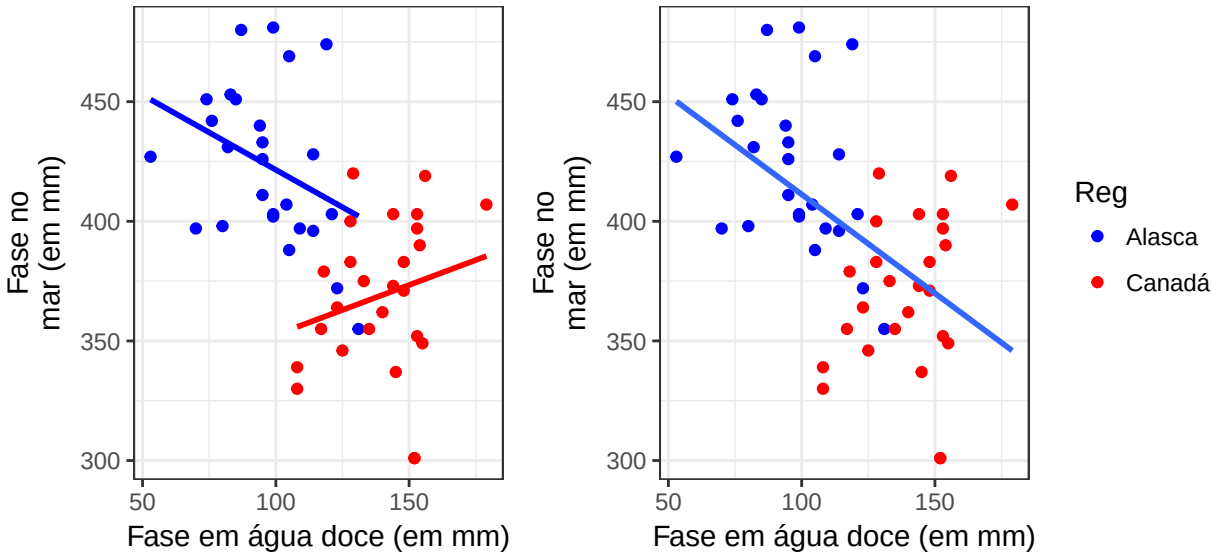

Figura 9: Gráfico de dispersão entre os diâmetros da guelra de salmões em água doce e no mar- Fêmea

A partir dos Box-plot da Figura 10, podemos ver para o gênero macho que o diâmetro das guelras do Canadá na fase em agua doce consideravelmente maior que o Grupo Alasca, já para a Fase no mar o contrario ocorre.

Além disso podemos ver que o diâmetro das guelras para o salmão fêmea do Canadá é consideravelmente maior do que os do Alasca, já durante a fase no mar, o contrario ocorre e para o Canadá o valor máximo quase atinge a mediana do grupo Alasca.

Portanto é possível notar a partir nas densidades da Figura 11 e 12, que existe uma sobreposição em boa parte da distribuições apresentadas, o que já foi notado nos boxplots anteriormente.

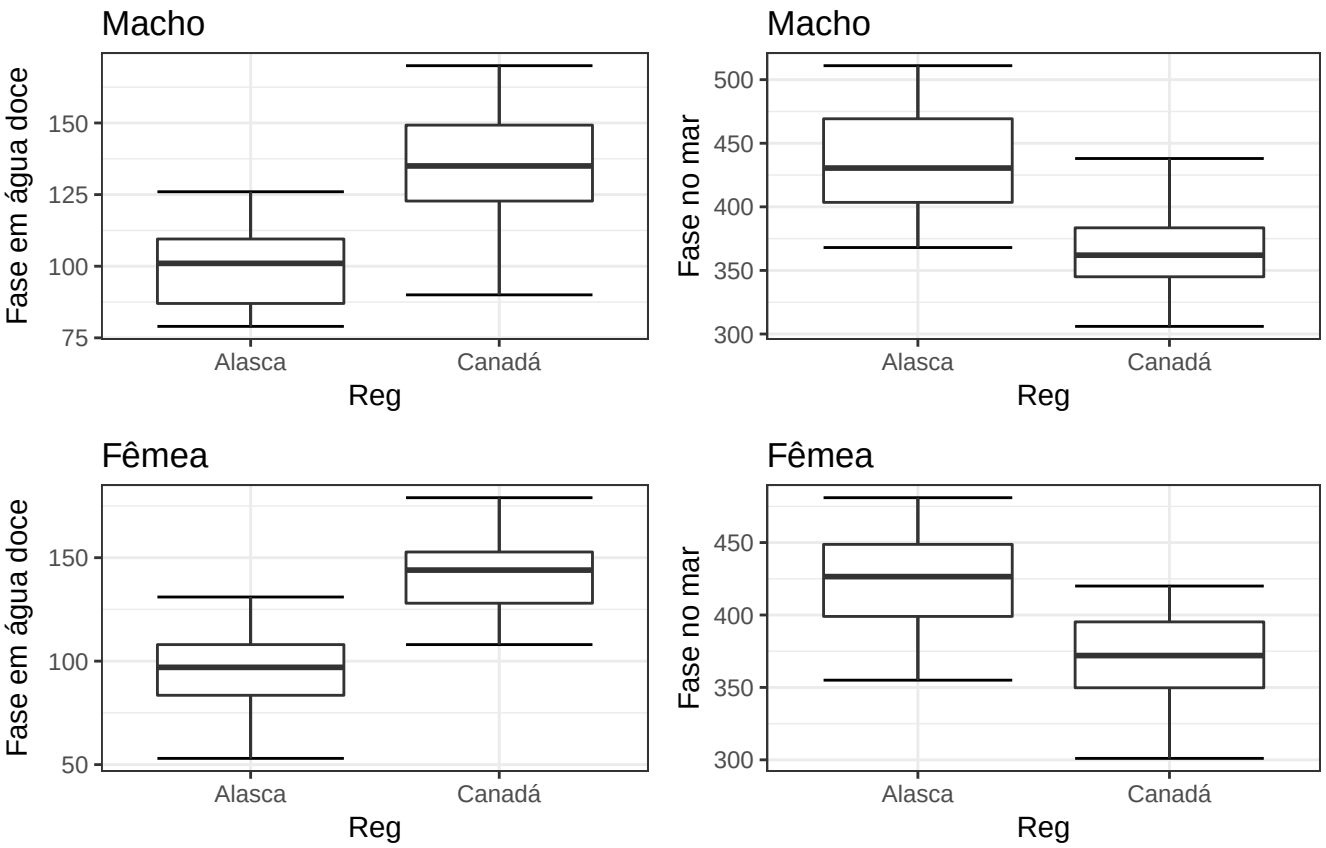

Figura 10: Boxplots dos Grupos por Gênero

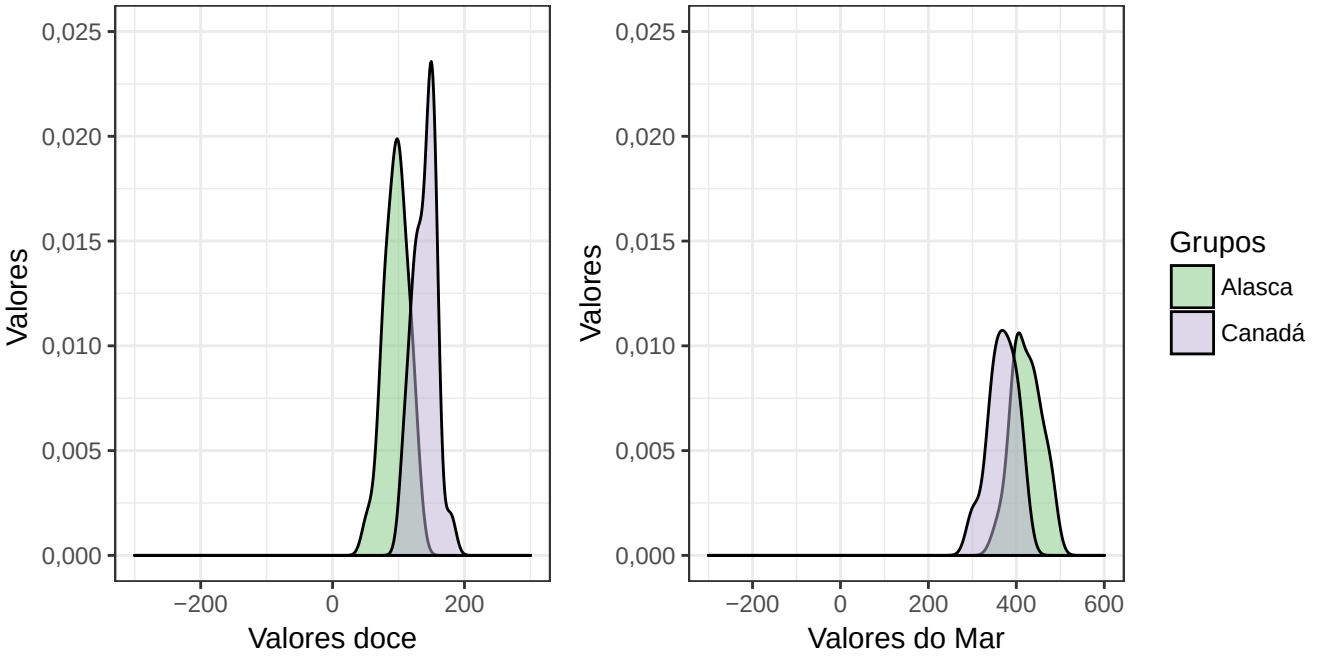

Figura 11: Densidades dos grupos do gênero - Fêmea

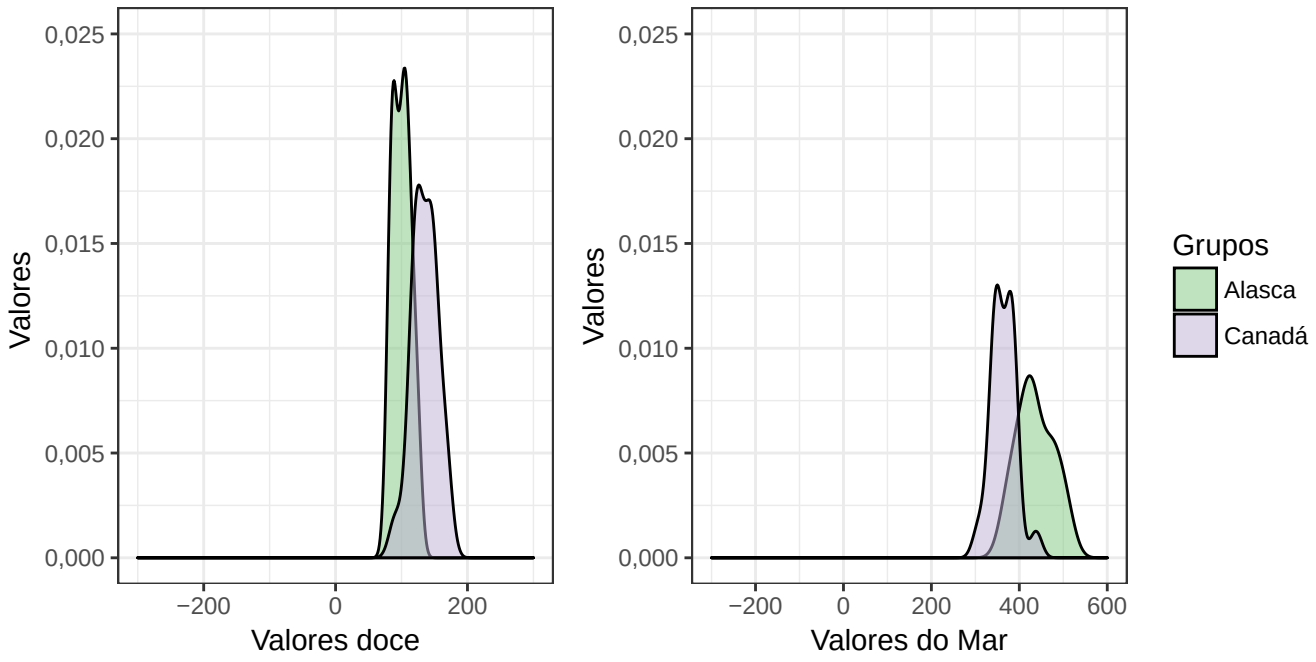

Figura 12: Densidades dos grupos do gênero - Macho

5. Análise Inferencial

Outras duas regras, baseadas na metodologia de Análise discriminante de Fisher, foram geradas a partir do banco de dados. Porém agora os grupos (Alasca e Canadá) foram divididos por sexo fêmea e macho. Aqui também foi suposto a presença de homocedasticidade entre as populações, com custos de classificação errada iguais e com as mesmas prioris definidas anteriormente, ou seja, as probabilidades a priori de um salmão pertencer ao Grupo Alasca foi definida como 0,6 e pertencer ao grupo Canadá foi 0,4 para ambos os sexos.

Como já foi discutido anteriormente, o teste de Box indicou para ambos os sexos que as matrizes de covariância são diferentes, de modo que a suposição de homocedasticidade não parece ser razoável para ambos os casos.

Os resultados referentes às regras de classificação para salmões fêmea e macho são apresentados na tabela 6.

Tabela 6: Resultados da classificação da amostra teste

		Classificado	
	Observado	Alasca	Canadá
Fêmea	Alasca	12	1
	Canadá	0	13
Macho	Alasca	12	0
	Canadá	1	11

Tabela 7: TOE e TEA para cada gênero

	Gênero	
	Feminino	Masculino
TEA	0,038 %	0,042 %
TOE	0,102 %	0,111 %

Ao observar a Tabela 6, observamos a qualidade da regra de classificação para os salmões fêmea. De acordo com a tabela 7, obtivemos uma taxa de erro aparente TEA = 0,038 %, valor bem abaixo ao da taxa ótima de erro TOE = 0,102 % ao qual leva em consideração a validade das suposições de igualdade das matrizes de covariância relativas aos dois tipos de salmão, ou seja, mesmo com as observações indicativas à fuga da suposição mencionadas, a regra de classificação mostrou uma taxa de erro ainda menor que a taxa ótima, indicando uma boa performance da regra proposta para o banco de dados em questão.

Para os machos, observamos a qualidade da regra de classificação para os salmões macho. Obtivemos uma taxa de erro aparente TEA = 0,042 %, valor bem abaixo ao da taxa ótima de erro TOE = 0,111 % ao qual leva em consideração a validade das suposição de igualdade das matrizes de covariância relativas aos dois tipos de salmão, ou seja, mesmo com as observações indicativas a fuga da suposição mencionadas, a regra de classificação mostrou uma taxa de erro ainda menor que a taxa ótima, indicando uma boa performance da regra proposta para o banco de dados em questão.

Observamos também que as regras de classificação para os salmões fêmea e macho apresentam valores de TEA bastante similares, já que ambos tiveram apenas um erro de classificação e a quantidade de peixes de cada sexo na amostra onde as regras foram testadas é praticamente a mesma.

Na tabela 8 podemos verificar as medidas resumos dos grupos divididos por macho e fêmea, onde são apresentadas os valores da função discriminante aplicada na amostra teste. Além disso é possível observar os box-plots para cada sexo na figura 13 (boxplot fêmea).

Para salmões fêmea, observamos que o valor mínimo referente ao grupo Canadá é menor que o máximo referente ao grupo Alasca, porém, pode-se observar por meio dos box-plots correspondentes que a menos de alguns valores discrepante, os box-plots parecem ser bastante distintos, uma vez que, pode-se perceber uma separação clara entre valores presentes no grupo Alasca, se comparados com os valores presentes no grupo Canadá. Este comportamento também é observado, embora seja menos evidente, na figura 14 (função discriminante fêmea) ao qual apresenta as densidades estimadas para os valores da função discriminante para os grupos. Nesta figura observa-se uma área de interseção pequena entre as densidades, o que favorece a distinção da origem dos almões fêmea por meio das variáveis DGAD e DGM.

O comportamento referente aos valores da função discriminante para salmões macho é análogo ao apresentado para salmões fêmea. Os box-plots correspondentes, a menos de dois pontos, parecem ser bastante distintos, assim como existe pouca área de interseção entre as densidades estimadas para os valores da função discriminante para os grupos Alasca e Canadá (Figura

14 ((funçao discriminante macho)), o que favorece a distinção da origem dos salmões macho por meio das variáveis DGAD e DGM.

Tabela 8: Medidas resumo para os valores função discriminante aplicada na amostra teste considerando a variável gênero

Gênero	Localidade	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Fêmea	Alasca	-1,18	1,11	1,23	-2,39	-1,44	1,49	13
	Canadá	1,80	0,79	0,62	-0,71	1,80	3,26	13
Macho	Alasca	-1,50	0,87	0,76	-3,23	-1,68	0,26	12
	Canadá	1,53	0,95	0,90	-0,80	-1,81	2,58	12

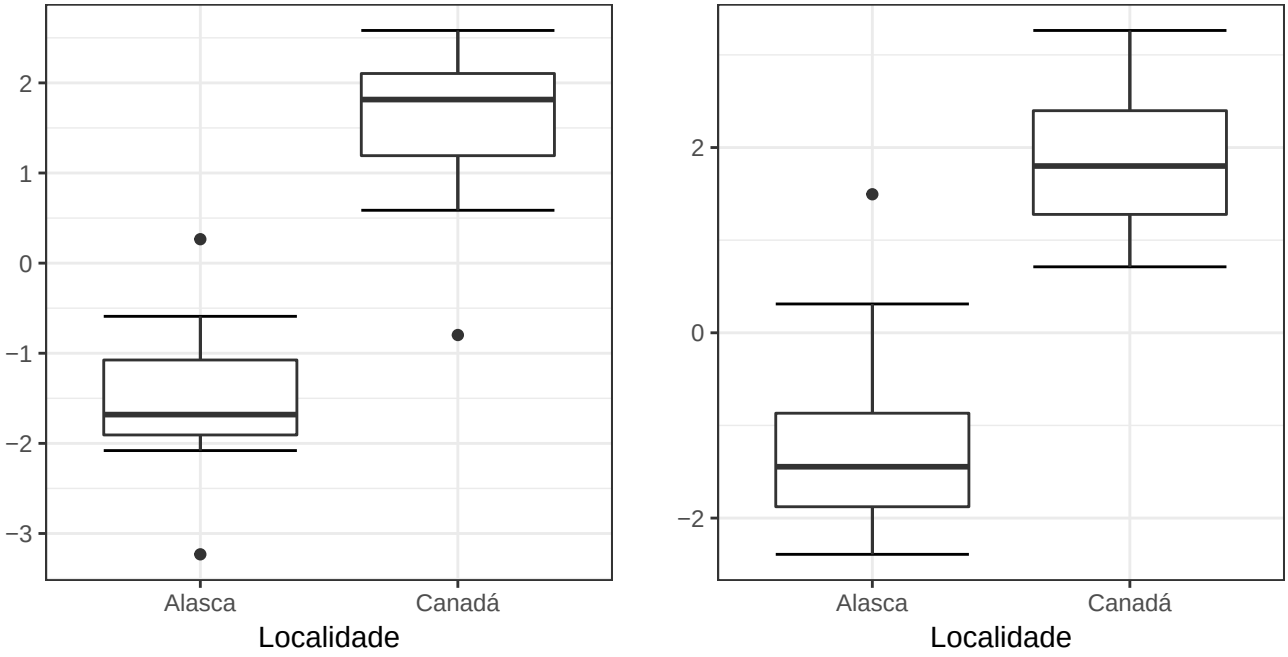

Figura 13: Boxplots da função discriminante aplicada à amostra teste, por grupo

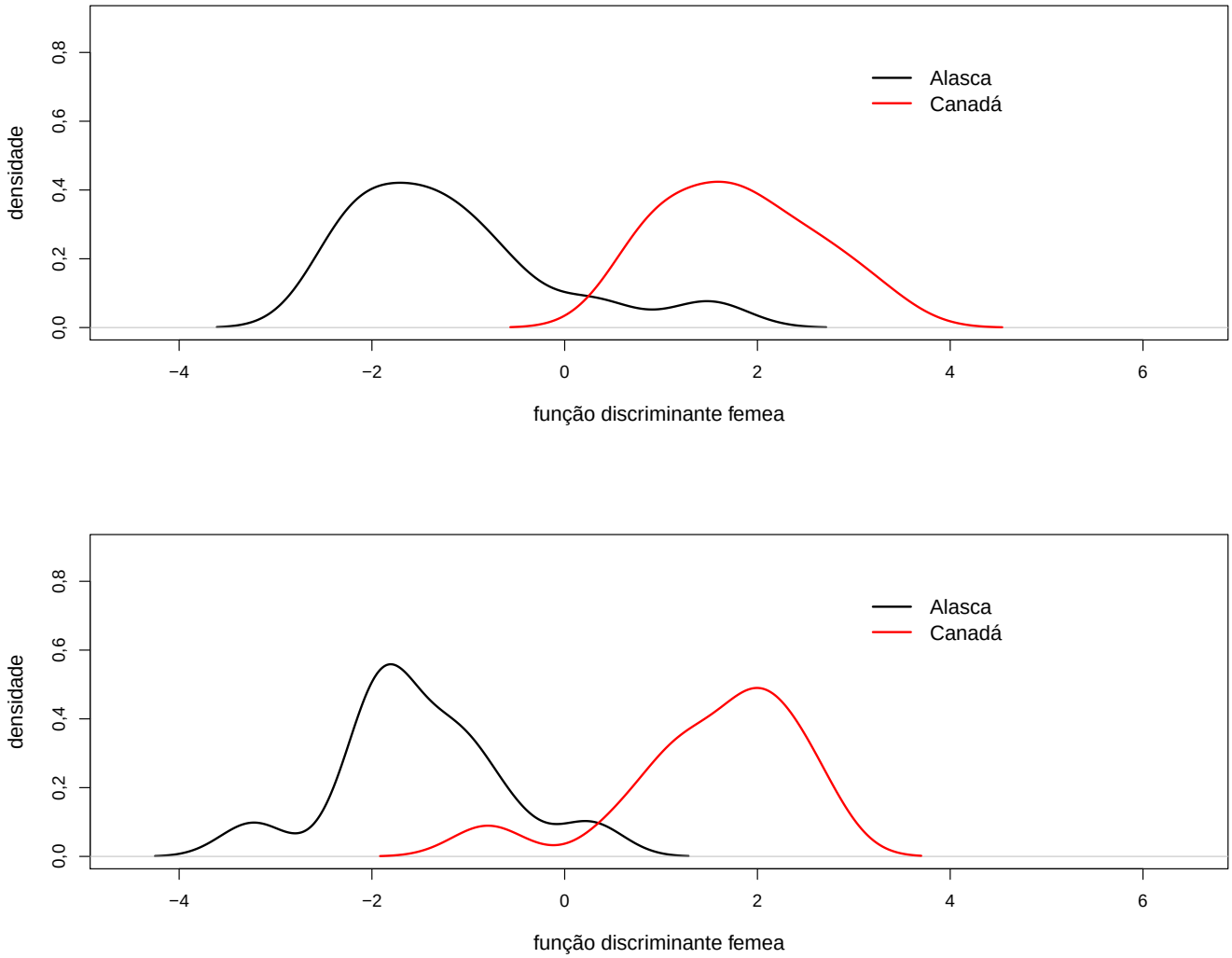

Figura 14: Densidade estimada da função discriminante aplicada à amostra teste, por grupo

6.Conclusões

As análises apresentadas aqui indicaram que as regras de classificação de salmões quanto à sua origem (Alasca ou Canadá) obtidas por meio do método de análise discriminante de Fisher foram razoavelmente boas, mesmo sendo a suposição de homocedasticidade irrazoável para todos os conjuntos analisados (Salmões de ambos os sexos, Salmões macho e Salmões fêmea).

Concluimos que todas as regras de classificação obtidas neste presente estudo poderiam ser sugeridas, embora as regras obtidas separando-se os salmões por sexo tenham sido um pouco mais efetivas, errando a classificação de 2 salmões em contrapartida aos 3 erros de classificação obtidos pela regra que não leva em consideração os sexos dos peixes para a mesma quantidade de peixes analisados. Porém deve-se avaliar sempre o contexto no qual a regra será aplicada, o custo, a utilidade e a influência que esta regra de classificação pode exercer.

7. Bibliografia

1. THE PACIFIC SALMON TREATY: A BRIEF TRUCE IN THE CANADA/U.S.A. PACIFIC SALMON WAR.
2. site:"http://www.adfg.alaska.gov/index.cfm?adfg=commercialbyfisherysalmon.salmon_combined_historical"
3. site:"http://www.pac.dfo-mpo.gc.ca/stats/comm/summ-somm/annsumm-sommann/2015/ANNUAL15_USER_three_party_groups-eng.htm"
4. Azevedo, C. L. N. (2017). Notas de aula sobre análise multivariada de dados note= "http://www.ime.unicamp.br/~cnaber/Material_AM_2S_2017.htm"
5. Johnson, R. A. e Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. 6 a edição, Upper Saddle River, NJ: Pearson Prentice Hall.