Estatística e Modelos Probabilísticos - COE241 – 2018.2

Projeto de curso

Professora: Rosa Maria Meri Leão

Júlia Togashi de Miranda DRE: 117058609

Código: *https://github.com/jutogashi/COE241*

Resumo

Este se trata do relatório do projeto do curso Estatística e modelos probabilísticos. Nele trabalhamos com um conjunto de dados médicos reais (disponibilizados pelo Professor Claudio Gil Soares de Araujo, da CLINIMEX) que continham informações de idade, peso, carga máxima em determinado teste além da VO2 máxima (medida para condição cardiovascular).

O trabalho pode ser dividido em três grandes partes: num primeiro momento fazemos uma análise básica a respeito dos dados presentes no data set; depois buscamos através de diferentes métodos encontrar uma distribuição de variável aleatória que melhor se adeque a distribuição de nossos dados; por fim procurar relações entre as variáveis, podendo assim realizar previsões usando a inferência bayesiana.

O código e os gráficos desse relatório estão disponíveis em *https://github.com/jutogashi/COE241*. Foi utilizado o ambiente Jupyter em python 3.7, assim como extensamente os recursos das bibliotecas pandas, numpy, scipy, matplotlib e seaborn.

Estudo de Caso

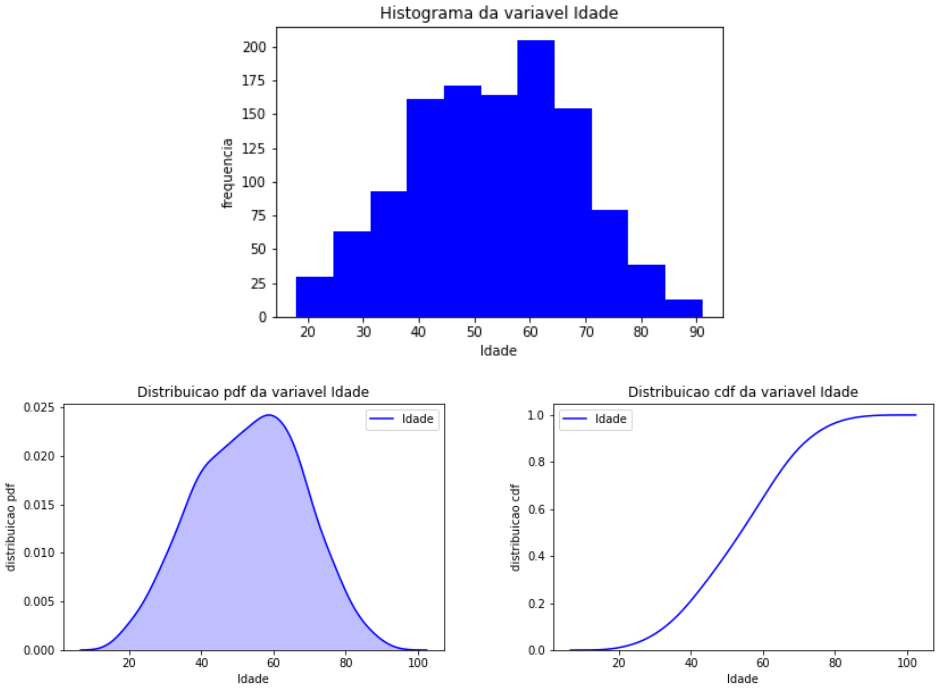
1- Histograma e Função Distribuição Empírica:

Temos que o histograma, a distribuição pdf e cdf empíricas de cada variável foram construídas com o auxílio da biblioteca seaborn.distplot, onde para essas últimas duas, o método de estimativa da função é a Estimativa de Densidade de Kernel.

O histograma representa a frequência na quantidade de amostras (não foi normalizado para mostra densidade, já que já temos isso representado na pdf empírica). O número de bins foi estimado a partir da formula m=1+3.3 log10(n), onde temos que m é igual ao número de intervalos e n é igual ao número de amostras. Como n é semelhante para todos as variáveis, temos que para todos os histogramas foi tomado m=11. (Temos que essa formula é derivada de uma distribuição binomial e implicitamente assume que temos uma distribuição próxima a normal para ser adequada, o que foi analisado mais à frente no trabalho).

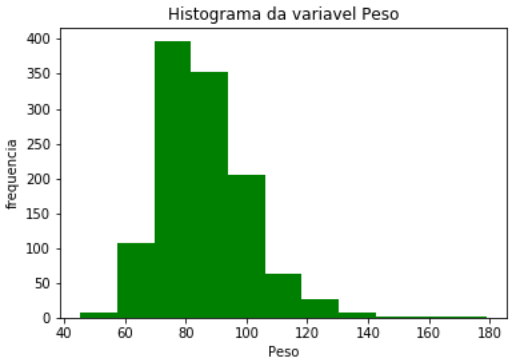
❖ Idade

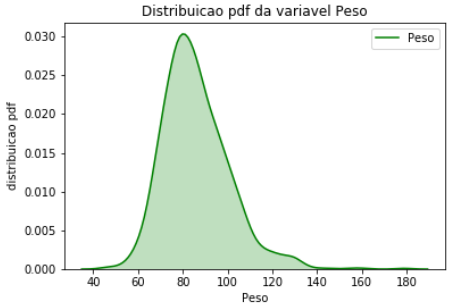
Já percebemos pela análise do histograma e da pdf que a idade possui uma variância considerável (dados coletados para pessoas de todas faixas de idade da vida adulta – amostras bem diversas). Não há nenhuma idade predominante no data set, sendo que majoritariamente os pacientes possuem entre 40 e 70 anos.

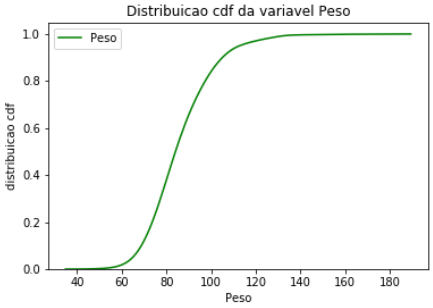


❖ Peso

Podemos concluir pela análise do histograma que possivelmente temos um número considerável de outliers com um peso consideravelmente maior que o padrão para o data set, que gira em torno de 70 a 100 kg (uma faixa padrão e peso para homens adultos). Isso acaba por uma análise gráfica de que os dados estão mais concentrados do que no caso da idade.

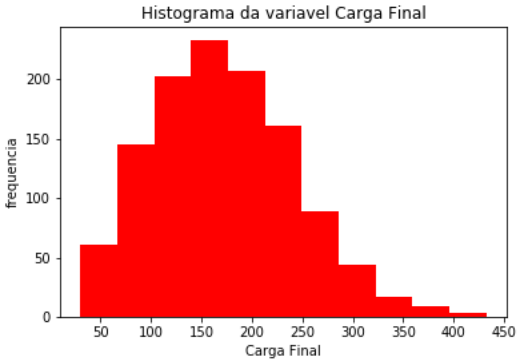


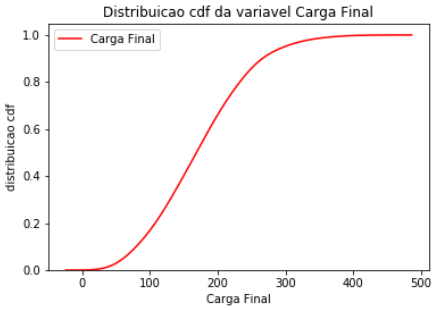


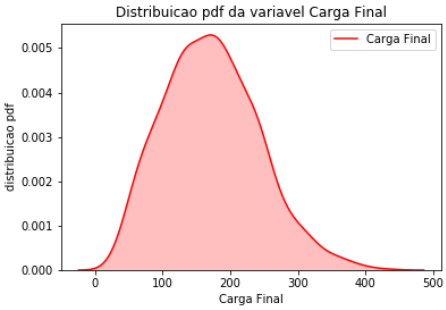


❖ Carga Final

Temos novamente o histograma e a pdf indicando que para a carga final temos um numero considerável de outliers superiores ao padrão (cerca de 70 a 250 watts), que provavelmente indicam pessoas do grupo amostral que possuem condicionamento físico muito superior (atletas por exemplo). Um fato a se atentar também é que já podemos observar que a distribuição da carga é muito semelhante à de O2 máximo, o que já nos pode ser indício de uma dependência entre essas variáveis, o que será analisado mais a fundo no tópico 6 do trabalho.

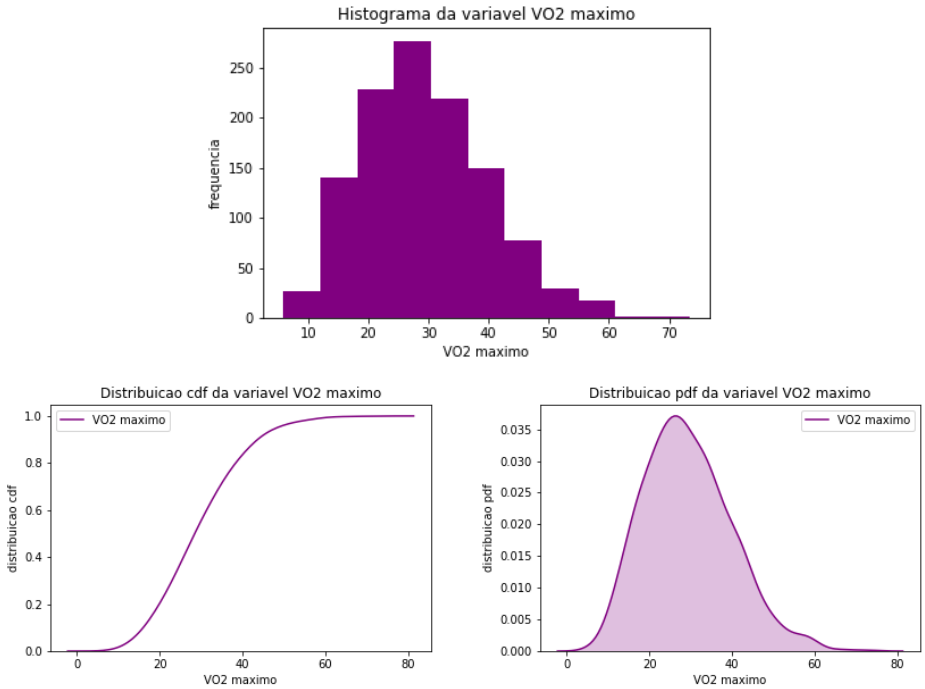






❖ VO2 Máximo

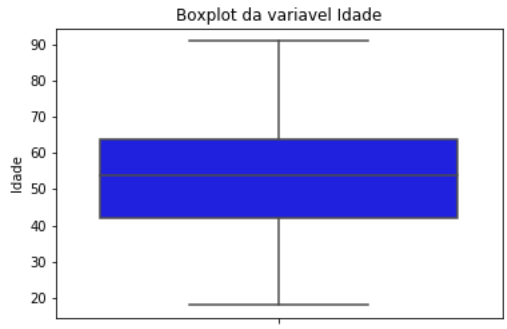
Temos que as considerações a respeito da taxa de VO2 máximo são muito semelhantes as para carga final, sendo que os pacientes estão majoritariamente concentrados em ter VO2 máximo entre 15 e 40 ml/(Kg.min).



2- Média, Variância e Boxplot

Para cálculo de média, desvio padrão e variância foi usado a biblioteca pandas que possui ferramentas para análise de dados em arquivos csv. Já para o boxplot foi usada novamente a função da biblioteca seaborn.

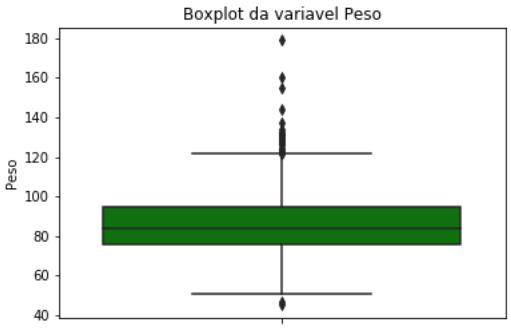
❖ Idade




```
Variável:  Idade
Media:  53.29095563139932
Desvio Padrão:  14.746296966880656
Variância:  217.45327423543367
```

Como já esperado pelo histograma, temos que a idade não possui outliers, ou seja, não há valores acima do upper_quartile+1.5*IQR ou abaixo do lower_quartile-1.5*IQR. Podemos perceber pela média e variância que as amostras realmente estão bem distribuídas com uma predominância central.

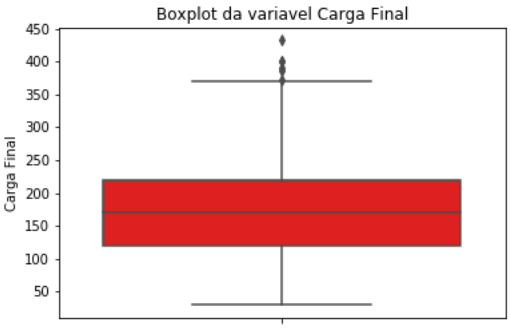
❖ Peso




```
Variável:  Peso
Media:  85.92577645051195
Desvio Padrão:  14.799113384059629
Variância:  219.0137569542528
```

Podemos perceber pela análise do boxplot que o peso possui tanto outliers inferiores quanto superiores. Contudo os superiores são muito mais numerosos e com uma discrepância muito significante ao conjunto de dados (mais que upper_quartile+3*IQR), o que realmente pode acabar afetando a análise de dados e a comparação com distribuições da literatura. Temos que a média é um valor esperado para um homem médio, e a variância não muito grande nas amostras.

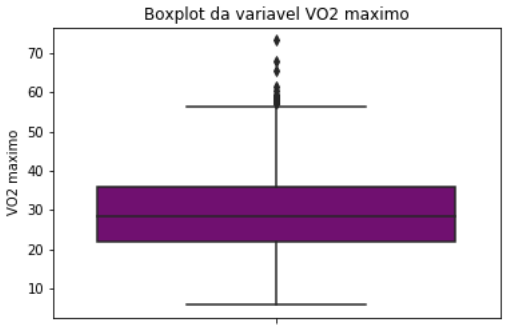
❖ Carga Final




```
Variável:  Carga Final
Media:  172.27150170648466
Desvio Padrão:  70.093123662472
Variância:  4913.0459847625925
```

O bloxplot da carga final nos indica que realmente a amostra possui somente outliers superiores, contudo esses não parecem ser muito numerosos ou tão discrepantes quanto para o peso, o que não deve acabar impactando negativamente a análise dos dados. Pela média e variância infere-se novamente que as amostras parecem bem distribuídas no data set.

❖ **VO2 máximo**




```
Variável:  VO2 maximo
Média:  29.39472792315316
Desvio Padrão:  10.49724989342601
Variância:  110.1922553250324
```

Assim como para a variável carga máxima, o boxplot de VO2 máximo nos indica que só possuímos outliers superiores na amostra. Contudo, mais semelhante ao peso, o numero de outliers parece considerável, e mais discrepante em relação ao upper whisker. Além disso, considerando que um valor normal para homens é de 35, temos que a média dos pacientes está abaixo desse valor. Isso pode ser um indicativo médico por exemplo para a qualidade de vida das regiões próximas à clínica, onde possivelmente seus frequentadores moram.

3- Parametrizando distribuições

Nessa etapa do trabalho o objetivo era comparar as distribuições empíricas das variáveis com distribuições conhecidas na literatura. Para isso devia-se utilizar o método da máxima verossimilhança (Maximum likelihood estimation – MLE) a fim de estimar os parâmetros das distribuições exponencial, gaussiana, lognormal e weibull.

O MLE é um método de estimar parâmetros de um modelo estatístico dado observações, achando os parâmetros que maximizam a likelihood de fazer as observações dado os parâmetros. Se temos um conjunto de amostras com valores observados temos que a pdf conjunta dessas amostras será o produtório da função densidade dado os parâmetros, sendo que esse é conhecido com função likelihood. Assim para maximiza-la, deveremos deriva-la e igualar a zero (e fazer a segunda derivada para conferir se de fato se trata de um máximo).

Agora temos as expressões analíticas encontradas para calcular os parâmetros de cada uma das distribuições da literatura.

❖ Exponencial

Temos a pdf para a variável aleatória exponencial:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ se } x \geq 0 \\ 0 \text{ se } x < 0 \end{cases}$$

Teremos a expressão da likelihood:

$$L(\lambda;x_1,\ldots,x_n) = \lambda^n \exp\left(-\lambda \sum_{j=1}^{n} x_j\right)$$

$$l(\lambda;x_1,\ldots,x_n) = n\ln(\lambda) - \lambda \sum_{j=1}^{n} x_j$$

Assim obtemos o parâmetro λ será estimado por:

$$\hat{\lambda}_n = \frac{n}{\sum_{j=1}^{n} x_j}$$

- Gaussiana

Temos a pdf para a variável aleatória gaussiana:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in (-\infty,\infty).$$

Teremos a expressão da likelihood:

$$L(\mu,\sigma^2;x_1,\ldots,x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

$$l(\mu,\sigma^2;x_1,\ldots,x_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

Assim obtemos os parâmetros μ e σ serão estimados por:

$$\hat{\mu}_n = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\hat{\mu})^2$$

- Lognormal

Temos a pdf para a variável aleatória lognormal:

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right], & \text{se } x > 0 \\ 0 \text{ caso contrário} \end{cases}$$

Teremos a expressão da likelihood:

$$\begin{aligned} \ell(\mu, \sigma \mid x_1, x_2, \ldots, x_n) &= -\sum_k \ln x_k + \ell_N(\mu, \sigma \mid \ln x_1, \ln x_2, \ldots, \ln x_n) \\ &= \text{constant} + \ell_N(\mu, \sigma \mid \ln x_1, \ln x_2, \ldots, \ln x_n). \end{aligned}$$

Assim obtemos os parâmetros μ e σ serão estimados por:

$$\widehat{\mu} = \frac{\sum_k \ln x_k}{n}, \qquad \widehat{\sigma}^2 = \frac{\sum_k (\ln x_k - \widehat{\mu})^2}{n}.$$

- Weibull

Temos a pdf para a variável aleatória weibull:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

Assim obtemos o parâmetro k será estimado por uma equação:

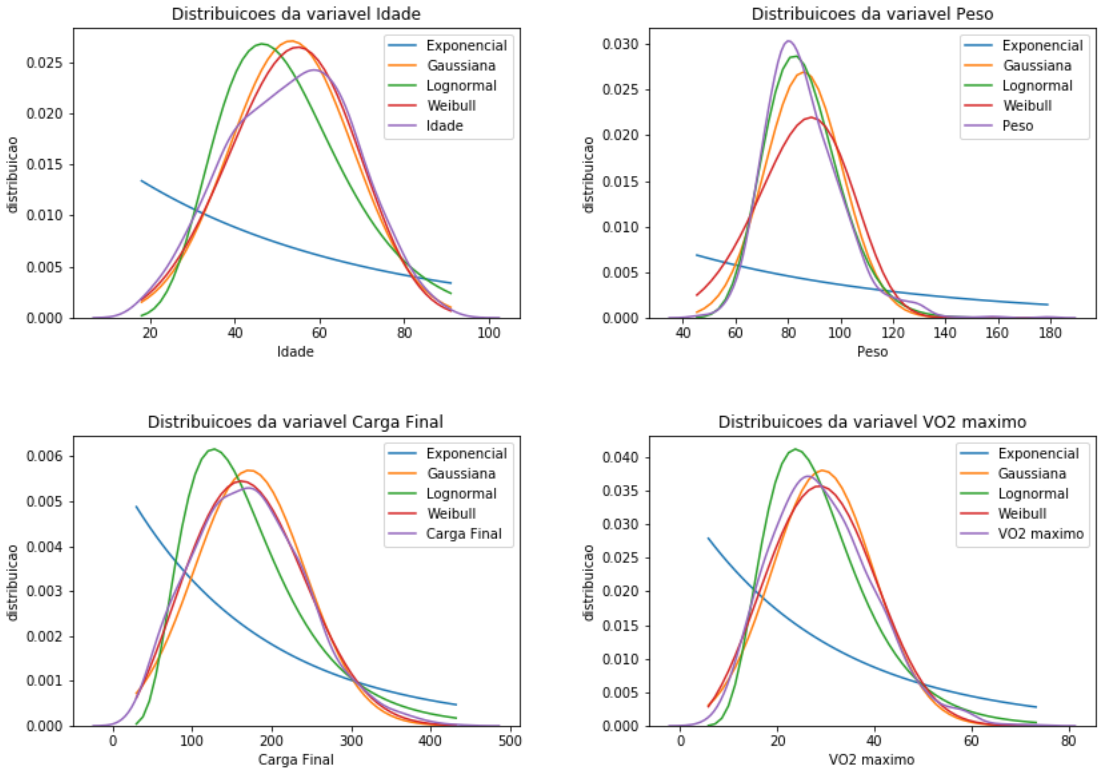
$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

E o parâmetro λ será estimado por:

$$\widehat{\lambda}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Para exponencial, gaussiana e lognormal, os parâmetros foram estimados utilizando diretamente as formulas encontradas acima, e a função distribuição foi plotada manualmente, utilizando as pdf de cada variável. Já para weibull que a solução de um de seus parâmetros é uma equação, foi utilizado a função do scipy para estimar MLE, e por isso foi plotada também usando a biblioteca para gera pdf do scipy.stats.

Temos então as distribuições encontradas para as quatro variáveis: idade, peso, carga fina e VO2 máximo, comparadas cada uma com essas quatro distribuições da literatura:



Primeiramente já podemos concluir que a função exponencial não representa bem nenhuma das variáveis, tendo forma completamente discrepante da buscada em todas as situações.

Já gaussiana, lognormal e weibull tem forma semelhante à de todas as distribuições, sendo que:

- Para idade: Percebemos que as que mais se adequam são a gaussiana e a weibull, não sendo possível somente por um teste visual inferir qual melhor se adequa.
- Para peso: Aparentemente a distribuição que melhor se adequa a distribuição empírica do peso é a lognormal, sendo que a gaussiana ainda apresenta relativa semelhança.
- Para carga final: Igualmente como para idade, temos que a gaussiana e a weibull possuem boa compatibilidade, mas nesse caso, tomando somente o critério da forma, a weibull parece ser mais compatível com a distribuição empírica.
- Para VO2 máximo: Mais uma vez gaussiana e weibull são suficientemente parecidas com a distribuição empírica, não sendo possível dizer qual é mais compatível apenas visualmente.

Temos também os parâmetros encontradas para cada variável e para cada distribuição:

Idade
Lambda da exponencial: 0.018764910258257682

Mu da gaussiana: 53.29095563139932
Sigma ao quadrado da gaussiana: 217.45327423543367

Mu da lognormal: 3.932509819486875
Sigma ao quadrado da lognormal: 0.0936331955793438

Constante da Weibull: 4.089481828645864
loc da Weibull: 0
Scale da Weibull: 58.78289005707875

Peso
Lambda da exponencial: 0.011637951279683105

Mu da gaussiana: 85.92577645051195
Sigma ao quadrado da gaussiana: 219.0137569542528

Mu da lognormal: 4.439451920143028
Sigma ao quadrado da lognormal: 0.027586997105752877

Constante da Weibull: 5.408013188534343
loc da Weibull: 0
Scale da Weibull: 92.24080850317551

Carga Final
Lambda da exponencial: 0.005804790636258545

Mu da gaussiana: 172.27150170648466
Sigma ao quadrado da gaussiana: 4913.0459847625925

Mu da lognormal: 5.0546544058509895
Sigma ao quadrado da lognormal: 0.2103368574854832

Constante da Weibull: 2.6469810001574725
loc da Weibull: 0
Scale da Weibull: 194.0388415799269

VO2 maximo

Lambda da exponencial: 0.03401970593551017

Mu da gaussiana: 29.39472792315316
Sigma ao quadrado da gaussiana: 110.1922553250324

Mu da lognormal: 3.3132400746591215
Sigma ao quadrado da lognormal: 0.14364411960908474

Constante da Weibull: 2.9978221690896216
loc da Weibull: 0
Scale da Weibull: 32.9274599599628

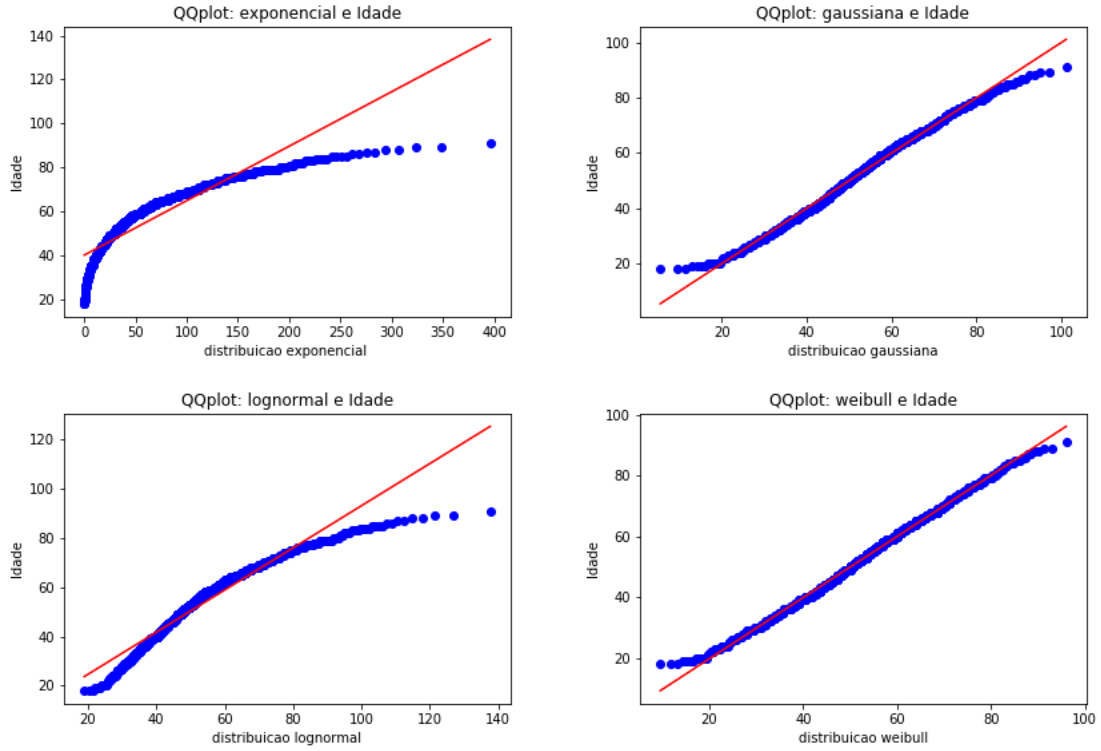
4- Gráfico QQplot ou ProbabilityPlot

O qqplot é um recurso gráfico para determinar se dois data sets tem uma distribuição comum, nele os pontos representam as frações de pontos da distribuição menores que um dado valor, para ambas as distribuições. As duas distribuições são consideradas equivalentes se o QQplot resultante se assemelha a uma reta de 45º em relação aos eixos.

Obs: Tanto para o item 4 quanto para o 5 foram usadas funções do scipy (stats.probplot e stats.kstest, respectivamente.). Nesse, muitas das pdf possuem notações diferentes, como a exponencial, que é escrita como f(x) = exp(-x), e possui um parâmetro denominado scale que é equivalente a 1/ λ. Sabendo disso, para facilitar a implementação e evitar erros, foi escolhido recalcular os parâmetros usando MLE para todas as distribuições usando a função que havia sido usada para calcular a Weibull. Podemos ainda converter os parâmetros do scipy olhando em sua documentação e obter os valores encontrados no item anterior.

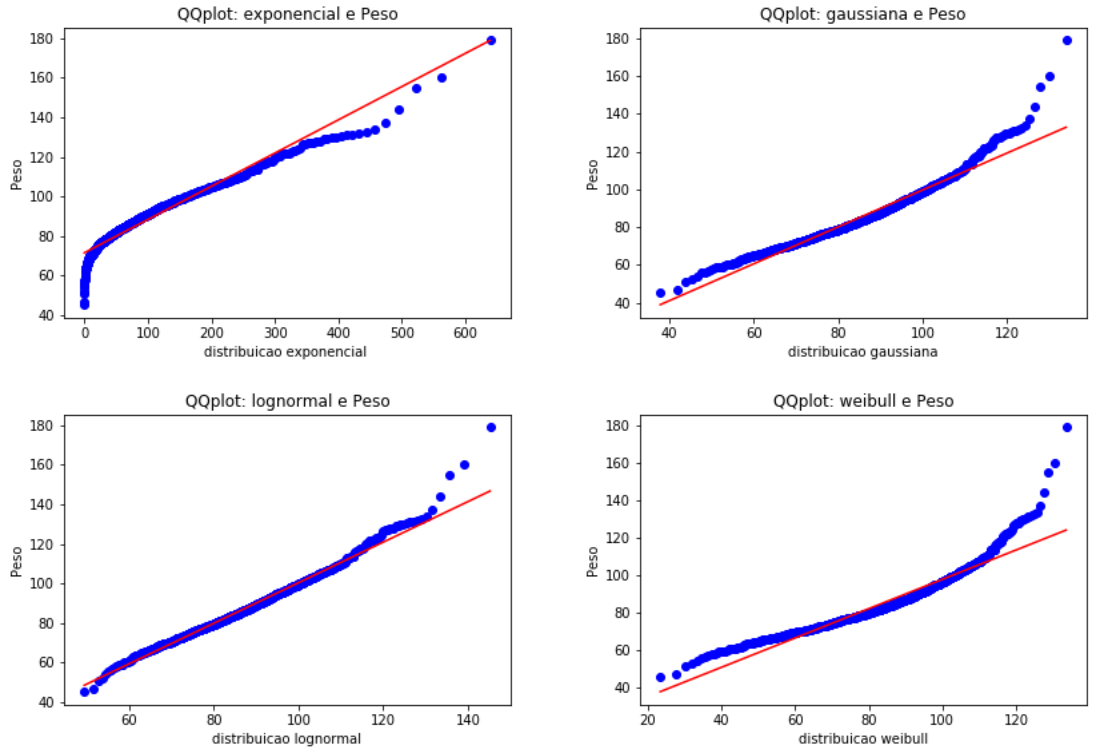
❖ Idade

Temos que para idade o qqplot reafirma as análises feitas na questão 3 e mostra que a distribuição empírica não é compatível com as distribuições exponencial e lognormal. Tanto para a gaussiana quanto para weibull as distribuições parecem compatíveis, sendo que elas variam da reta para valores nas margens superior e inferior. Contudo apesar de semelhante, a weibull parece ter compatibilidade um pouco maior, principalmente na margem superior dos dados.



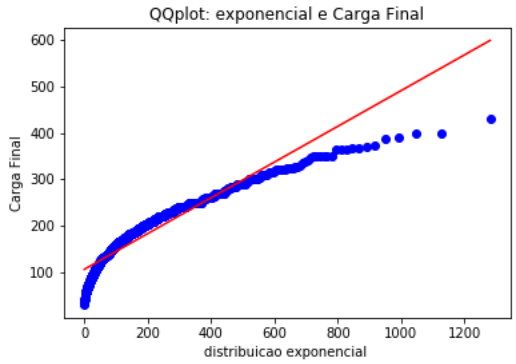
❖ Peso

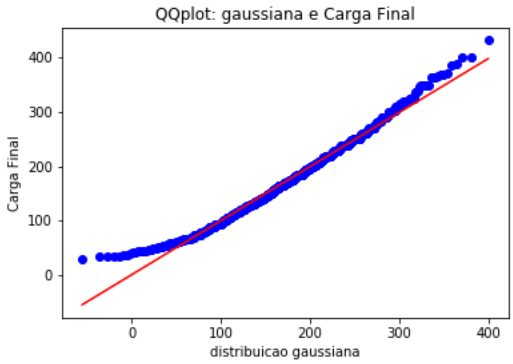
Nenhuma distribuição parece se adequar perfeitamente no QQplot, contudo como já é sabido pelo boxplot que o data set no quesito idade possui outliers significativos para cima, desconsiderando o resultado a partir de 140 kg (outliers), a função mais compatível parece ser a lognormal, como esperado.

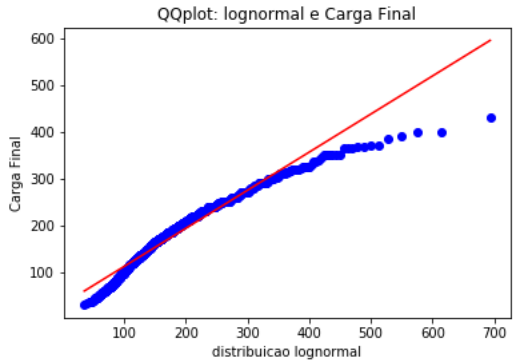


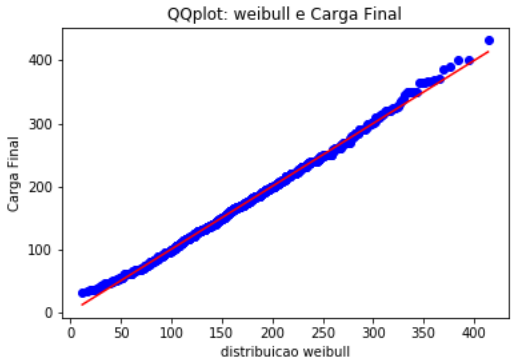
❖ Carga Final

Para o resultado do QQplot, no caso da carga final está muito claro que a função de maior compatibilidade é a weibull, formando quase perfeitamente a reta de 45º com o eixo, o que acaba sendo um indicativo para a dúvida estabelecida no item 3, onde a gaussiana ainda parecia ser uma distribuição válida para representação.



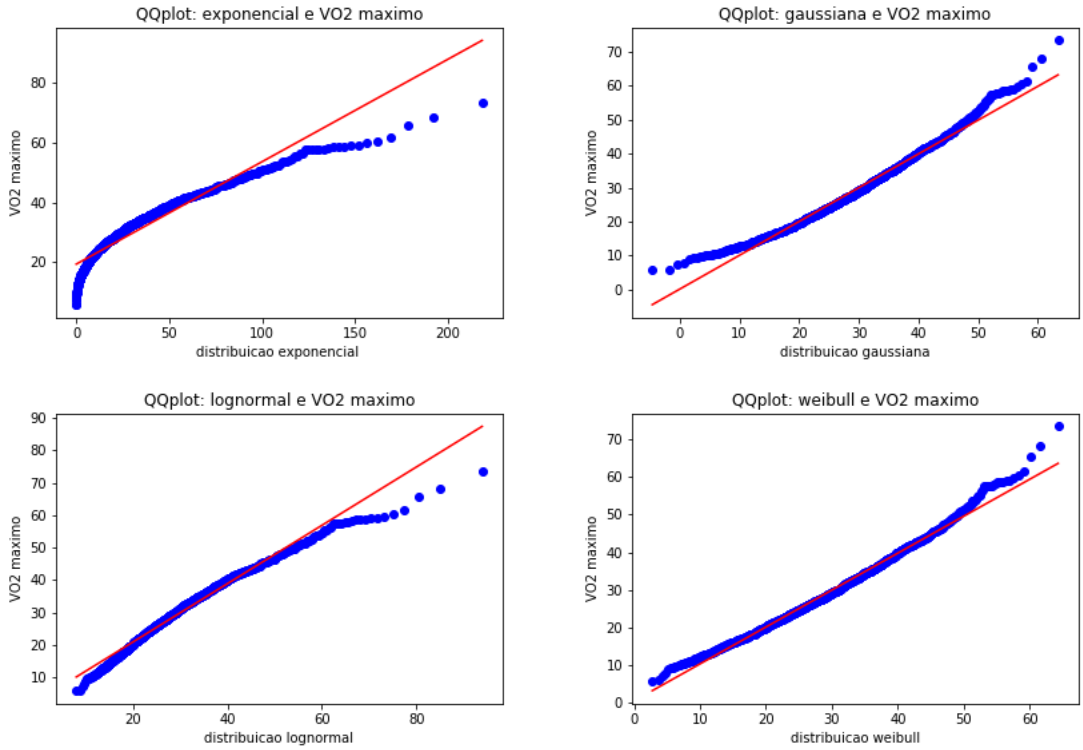






- **VO2 máximo**

Novamente, assim como ocorreu para o peso, os outliers parecem atrapalhar o resultado para os valores superiores dos dados. Novamente, se desconsiderarmos o gráfico a partir de aproximadamente 60, onde começam os outliers, percebemos que a distribuição mais compatível é a weibull, o que retira novamente a dúvida da analise somente da forma das distribuições no item 3, onde parecíamos ter a gaussiana e a weibull como distribuições igualmente válidas.



5- Teste de Hipótese

Nessa etapa do trabalho devemos utilizar o teste Kolmogorov–Smirnov (KS), para validar ou descartar a hipótese de que a nossa distribuição empírica e uma dada da literatura são compatíveis. Como já comentado no item anterior, foi utilizado uma função do scipy.stats para obter os resultados, no qual era retornado tanto o valor D do teste KS, quanto do p-valor.

O algoritmo de KS funciona sobre o seguinte princípio: primeiramente ordenamos os valores amostrais de forma crescente. Calculamos a função distribuição empírica F(x). A estatística Komolgorov-Smirnov é definida como:

$$D_n = \sup_x \left| \hat{F}_n(x) - F_0(x) \right|$$

A hipótese H0 é rejeitada para um dado nível de significância α se: D > dα. Já o p-valor é o menor valor de α (nível de significância do teste) para o qual a hipótese H0 é rejeitada.

Temos que uma hipótese (que as funções possuem distribuições compatíveis), caso o valor de D seja suficientemente pequeno, ou o de p-valor suficientemente grande. Adotamos para análise α=0.05

De antemão podemos dizer que, como esperado, a hipótese para exponencial foi rejeitada para todos os casos. Ela está tão distante que, mesmo se considerasse um α =0.10, ela continuaria não passando para qualquer uma das variáveis aleatórias.

❖ Idade

Tomando α=0.05, podemos dizer que não podemos rejeitar a hipótese para gaussiana e para Weibull, o que já era esperado pelas análises anteriores. Contudo, mais uma vez, a Weibull possui resultados para o teste KS inferiores a gaussiana, e um p-valor superior, o que nos leva a concluir que realmente é a melhor representação para a variável idade.

```
-Exponencial
D =  0.372755615059967
p_value =  0.0

-Gaussiana
D =  0.04408368872194113
p_value =  0.02039175142102323

-Lognormal
D =  0.084730460447627
p_value =  9.073029882955552e-08

-Weibull
D =  0.033037815723893305
p_value =  0.15145170476000502
```

❖ Peso

Temos para o peso que, dado um α=0.05, a única variável aleatória cuja a hipótese não é rejeitada, tanto no critério do KS quanto no do p-valor é a lognormal. Levando em consideração as análises do item 3 e 4, concluímos que essa realmente é a variável que melhor representa nossa distribuição empírica.

```
-Exponencial
D =  0.4954410013455397
p_value =  0.0

-Gaussiana
D =  0.06661818817785059
p_value =  5.758423507362644e-05

-Lognormal
D =  0.032285259002662436
p_value =  0.17003957723543306

-Weibull
D =  0.1032173331741221
p_value =  2.5226265520927882e-11
```

❖ Carga Final

Podemos dizer que para o teste de hipótese, gaussiana e Weibull podem ser consideradas válidas. Todavia, a Weibull possui resultados expressivamente superiores, principalmente no quesito do p-valor. Isso somado ao teste visual da forma e ao QQplot, nos leva a concluir que a Weibull é a função que melhor se enquadra a variável carga final.

```
-Exponencial
D =  0.28651634266099946
p_value =  0.0

-Gaussiana
D =  0.039233911356943985
p_value =  0.052776560691338625

-Lognormal
D =  0.08035970386976421
p_value =  4.962162909460943e-07

-Weibull
D =  0.02457022560625388
p_value =  0.47886304960046483
```

❖ VO2 máximo

A variável VO2 máximo é a que possui mais hipóteses que não podem ser rejeitadas pelo teste de Kolmogorov–Smirnov, a gaussiana, lognormal e Weibull. A Weibull, contudo, possui resultado de D um pouco inferior as outras, e é a única que possui p-valor acima de 0.05, isso aliado aos resultados anteriores nos leva a concluir que é a melhor representação. Importante atentar para o fato de que tanto o VO2 máximo quanto a carga final são melhores representas pela mesma variável aleatória, o que pode nos ser outro indicativo que talvez possuam alguma relação.

```
-Exponencial
D =  0.3348896789424037
p_value =  0.0

-Lognormal
D =  0.04056142112151151
p_value =  0.041130681450034956

-Gaussiana
D =  0.044531849851028094
p_value =  0.018572422090608276

-Weibull
D =  0.03674655683834849
p_value =  0.08234531803549583
```

6- Análise de dependência entre as variáveis, modelo de regressão

O objetivo desse item do projeto é fazer uma análise sobre possíveis relações entre as variáveis idade, peso, carga final e a variável VO2 máximo, para no item 7, fazermos predições a respeito dos resultados. Possivelmente sendo os resultados de maior interesse prático médico, foi utilizado a biblioteca do seaborn para fazer o scatter plot e o numpy para a regressão linear.

Obs: A função de regressão linear do numpy nos retorna coeficientes mesmo quando as variáveis não parecem ter qualquer relação linear. Os resultados dos coeficientes retornados serão apresentados nesse trabalho, contudo será discutido se esses possuem sentido.

A técnica utilizada na regressão linear do numpy é o de mínimos quadrados. Nesse, é buscada uma linha que minimize a soma dos quadrados das diferenças entre o valor estimado e os dados observados (residuais).

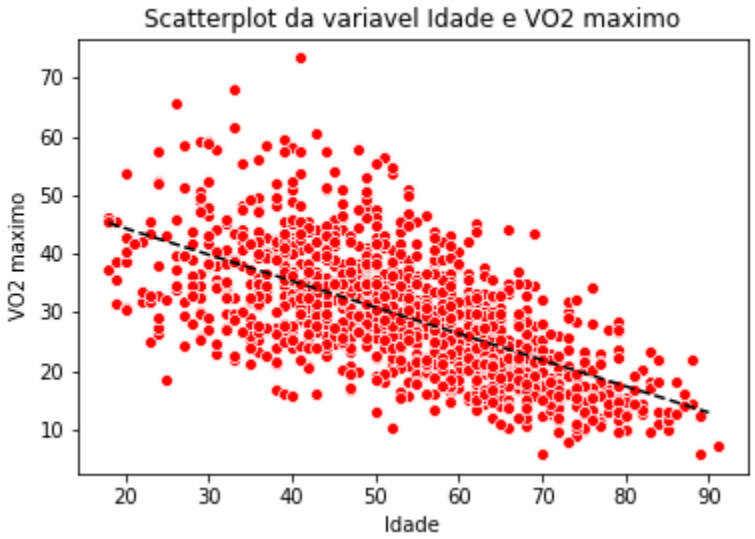
Temos também que o coeficiente de correlação entre as variáveis foi calculado manualmente pela formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Onde n é o número de amostras; r o coeficiente de correlação; x e y as variáveis.

❖ Idade e VO2 máximo

Dado a visualização gráfica do scatter plot e o coeficiente de correlação entre as variáveis podemos dizer que há uma maior tendência de pessoas mais jovens terem uma taxa de VO2 superior. Esse resultado nos parece intuitivo, visto que pessoas com menor idade tendem a ser mais ativas, praticar mais esportes, logo terem uma condição aeróbica superior. Contudo, podemos perceber que isso se traduz somente em uma tendência, não percebemos uma relação linear explicita entre as duas variáveis (principalmente dado o coeficiente, que não se encontra tão próximo de -1), visto justamente que fatores como sedentarismo e atividade física vão muito além da faixa etária. Assim também concluímos que não há muito sentido em falar sobre o modelo de regressão linear nesse caso.



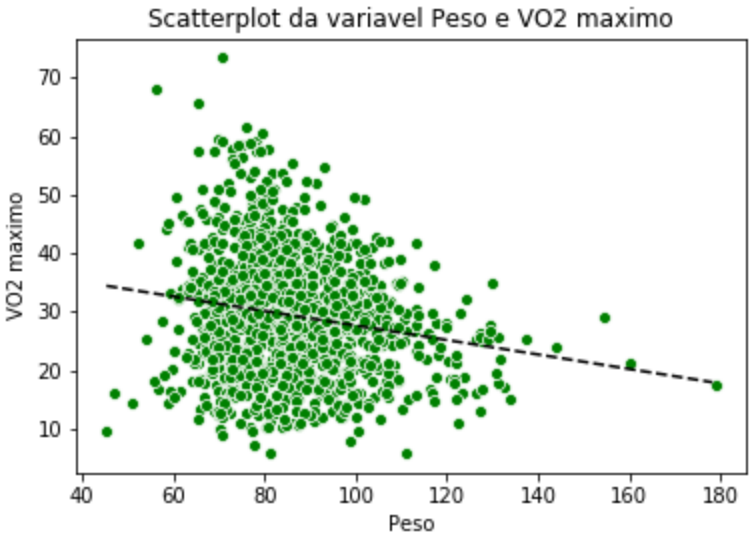

```
Coeficiente de correlação entre Idade e VO2 máximo:  -0.6300720192503418

Regressão linear: [-0.44852097 53.2968391]
```

❖ Peso e VO2 máximo

Dado o gráfico e o coeficiente de correlação muito próximo de 0, concluímos que não há relação entre peso e VO2 máximo. Uma conclusão que podemos tirar desse resultado é que talvez a medida somente do peso seja muito genérica, não contribuindo tanto para uma análise médica. Temos que a forma física de um paciente, que de fato deve impactar em sua condição aeróbica, não parece ser definida bem somente pelo peso. A questão por exemplo da altura, o índice de gordura ou a massa magra (muscular) de um paciente pode implicar em pesos semelhantes, contudo significados físicos completamente distintos.

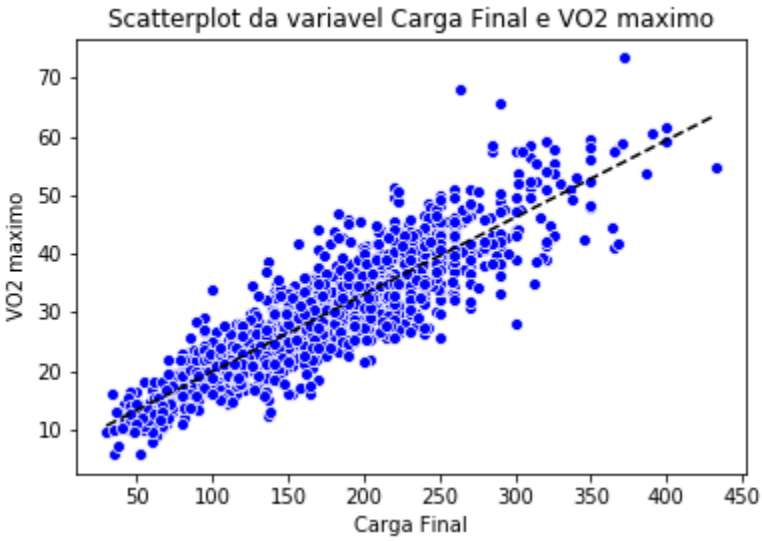
Novamente não há nenhum sentido em falarmos de um modelo de regressão linear nessa situação.




```
Coeficiente de correlação entre Peso e VO2 máximo:  -0.17440061829630799
Regressão linear: [-0.12370517 40.02419091]
```

- **Carga Final e VO2 máximo**

Percebe-se que pelo coeficiente de correlação muito próximo de um que carga final e VO2 máximo possuem uma relação linear, o que traz sentido ao resultado da regressão linear. É natural pensarmos também que pacientes que suportaram uma carga maior em determinado teste físico possuam uma capacidade aeróbica superior. Esse resultado médico é importante pois nos permite inferir VO2 pela carga, ou seja, possivelmente prever grupos de risco por exemplo através de outros exames clínicos.




```
Coeficiente de correlação entre Carga Final e VO2 máximo:0.878325609405961
Regressão linear: [0.13153934 6.73424783]
```

7- Inferência Bayesiana

Dado que o resultado do item anterior nos mostra que a variável com maior relação linear com VO2 máximo é a carga final, realizaremos a inferência bayesiana para esse par de fatores. Foi escolhido 10 intervalos da carga final para a tabela.

Temos primeiramente 2 tabelas, uma que tem como data VO2 máximo ser menor que 35, e outra que seja maior ou igual a 35.

O cálculo do numerador de bayes e feito multiplicado a prior pela likelihood, e o cálculo da posterior é feito dividindo o numerador de bayes pelo total do numerador de bayes (soma para todos os intervalos). Assim o resultado que obtemos para a posterior indica qual a probabilidade, dado um paciente ter VO2 maior ou menor que 35, de ele pertencer a um intervalo de carga final.

- **Dado: VO2 máximo < 35**

Hypothesis	Prior	Likelihood	Bayes Numerator	Posterior
(30.0, 70.2)	0.071672	1.0	0.071672	**0.099291**
(70.2, 110.4)	0.132253	1.0	0.132253	**0.183215**
(110.4, 150.6)	0.208191	0.987705	0.205631	**0.284870**
(150.6, 190.8)	0.209044	0.848980	0.177474	**0.245863**
(190.8, 231.0)	0.182594	0.588785	0.107509	**0.148936**
(231.0, 271.2)	0.119454	0.2	0.023891	**0.033097**
(271.2, 311.4)	0.041809	0.061224	0.002560	**0.003546**
(311.4, 351.6)	0.024744	0.034483	0.000853	**0.001182**
(351.6, 391.8)	0.007679	0.0	0.0	**0.0**
(391.8, 432.0)	0.001706	0.0	0.0	**0.0**
TOTAL	**1**	**No Sum**	**0.721843**	**1**

Podemos concluir pela tabela que dado que VO2 é menor que 35, é impossível que o paciente tenha carga final entre 351.6 e 432. Também percebemos que 100% dos indivíduos com carga final entre 30 e 110.4, possuem VO2 menor que 35, contudo como esses representam uma porcentagem menor do total, isso acaba diminuindo a posterior para esse grupo. Por fim, concluímos que a maior probabilidade é que o paciente esteja situado na faixa de 110.4 e 231.

- **Dado: VO2 máximo > 35**

Hypothesis	Prior	Likelihood	Bayes Numerator	Posterior
(30.0, 70.2)	0.071672	0.0	0.0	**0.0**
(70.2, 110.4)	0.132253	0.0	0.0	**0.0**
(110.4, 150.6)	0.208191	0.012295	0.002560	**0.009231**
(150.6, 190.8)	0.209044	0.151020	0.031570	**0.113846**
(190.8, 231.0)	0.182594	0.411215	0.075085	**0.270769**
(231.0, 271.2)	0.119454	0.8	0.095563	**0.344615**
(271.2, 311.4)	0.041809	0.938776	0.039249	**0.141538**
(311.4, 351.6)	0.024744	0.965517	0.023891	**0.086154**
(351.6, 391.8)	0.007679	1.0	0.007679	**0.027692**
(391.8, 432.0)	0.001706	1.0	0.001706	**0.006154**
TOTAL	**1**	**No Sum**	**0.277303**	**1**

Podemos concluir pela tabela que dado que VO2 é maior que 35, é impossível que o paciente tenha carga final entre 30 e 110.4. Também percebemos que 100% dos indivíduos com carga final entre 351.6 e 432, possuem VO2 maior que 35, contudo como esses representam uma porcentagem menor do total, isso acaba diminuindo a posterior para esse grupo. Por fim, concluímos que a maior probabilidade é que o paciente esteja situado na faixa de 190.8 e 311.4.

- **Previsão**

Por fim, queremos realizar uma previsão de melhora no VO2 máximo do paciente. Ou seja, qual a probabilidade de que dado paciente que tenha uma medida de VO2 abaixo da média, melhore-a a ponto de ficar acima da média. (P [V O2 máximo ≥ 35| V O2 máximo < 35])

Temos que a tabela é atualizada utilizando a prior, as likelihoods, e o numerador de bayes do caso de VO2<35 e a posterior do caso VO2<35. O cálculo da previsão então é feito multiplicando a posterior <35 pela likelihood >35. Temos o resultado final para cada intervalo de carga final e a probabilidade da previsão de mlhora sendo igual à soma desses.

Dado o resultado de 0.13, podemos concluir que a probabilidade de ocorrer uma melhora é baixa, principalmente pela forte relação linear entre carga e VO2, ou seja, caso de fato ocorresse uma melhora, o mais provável seria que tivéssemos uma melhora também na carga final, o que acabaria resultando nesse paciente passar a pertencer a outro intervalo.

Hypothesis	Prior	Likelihood1	Bayes Numerator	Posterior 1	Likelihood2	Prediction
(30.0, 70.2)	0.071672	1.0	0.071672	0.099291	0.0	**0.0**
(70.2, 110.4)	0.132253	1.0	0.132253	0.183215	0.0	**0.0**
(110.4, 150.6)	0.208191	0.987705	0.205631	0.284870	0.012295	**0.003502**
(150.6, 190.8)	0.209044	0.848980	0.177474	0.245863	0.151020	**0.037130**
(190.8, 231.0)	0.182594	0.588785	0.107509	0.148936	0.411215	**0.061245**
(231.0, 271.2)	0.119454	0.2	0.023891	0.033097	0.8	**0.026478**
(271.2, 311.4)	0.041809	0.061224	0.002560	0.003546	0.938776	**0.003329**
(311.4, 351.6)	0.024744	0.034483	0.000853	0.001182	0.965517	**0.001141**
(351.6, 391.8)	0.007679	0.0	0.0	0.0	*	**0.0**
(391.8, 432.0)	0.001706	0.0	0.0	0.0	*	**0.0**
TOTAL	**1**	**No Sum**	**0.721843**	**1**	**No Sum**	**0.132825**

Conclusão

Nesse projeto pode-se entrar em contato prático com diversos métodos e conceitos apresentados na aula de Estatística e modelos probabilísticos, assim como as diversas bibliotecas com funções implementadas para estudos nesse tema.

Em um primeiro momento pode-se inferir sobre o conjunto de dados disponíveis, analisando como estavam distribuídos e o que isso implicava sobre o conjunto de pacientes na prática.

Já nos itens 3, 4 e 5, chegou-se à conclusão de quais distribuições da literatura melhor adequavam-se as distribuições empíricas. Obtemos para a idade a distribuição Weibull; para peso a lognormal e novamente para carga final e VO2 máximo a Weibull. Todavia, esses resultados não querem dizer que não há outras distribuições na literatura que se adequem ainda melhor ao nosso conjunto de dados, contudo foi considerado que para todas as variáveis obtemos resultados suficientemente bons.

Por fim pode-se observar também que variáveis como peso e VO2 máximo não possuem relação linear, enquanto carga final e VO2 possuem forte relação. Ainda trabalhado esses dados na área de inferência, obteve-se resultados que provavelmente representam uma das maiores contribuições da computação e estatística na área de saúde, permitindo sinalizar grupos de risco por exemplo para uma característica dado testes para outra distinta.

Referencias

[1] http://www.land.ufrj.br/~classes/est-prob-2018/

[2] https://docs.scipy.org/doc/scipy-0.15.1/reference/index.html

[3] https://seaborn.pydata.org/index.html

[4] http://www.portalaction.com.br/inferencia-0

[5] https://medium.com/data-hackers/uma-introdu%C3%A7%C3%A3o-simples-ao-pandas-1e15eea37fa1

[6] https://matplotlib.org/index.html

[7] http://www.numpy.org/