UFRJ- Universidade Federal do Rio de Janeiro
Estatística e Modelos Probabilísticos
Trabalho Final

Profª Rosa M. M Leão
Felipe Schreiber Fenandes DRE: 116206990
https://github.com/FelipeSchreiber/Estat-stica-TrabalhoFinal/tree/master/TrabalhoFinal

Motivação

Esse trabalho consistiu em, a partir do dataset disponibilizado pelo Professor Claudio Gil Soares de Araujo, obter dados estatísticos de cada uma das variáveis (idade, peso, carga final, VO2 máximo), verificar se há alguma correlação/dependência entre as mesmas e por fim determinar um modelo para estimar o VO2 máximo a partir das demais variáveis.

Para tal, foi usado Jupyter notebook(incluso na plataforma Anaconda para o python2.7), além das bibliotecas padrões da linguagem, tais como: “pandas” para a leitura do arquivo no formato csv e obtenção de alguns dados estatísticos , “math” para calcular o espaçamento dos bins, “numpy” para obter os bins dos histogramas e para obter a função cumulativa de probabilidade, “matplotlib” e “seaborn” pra plotar os gráficos necessários.

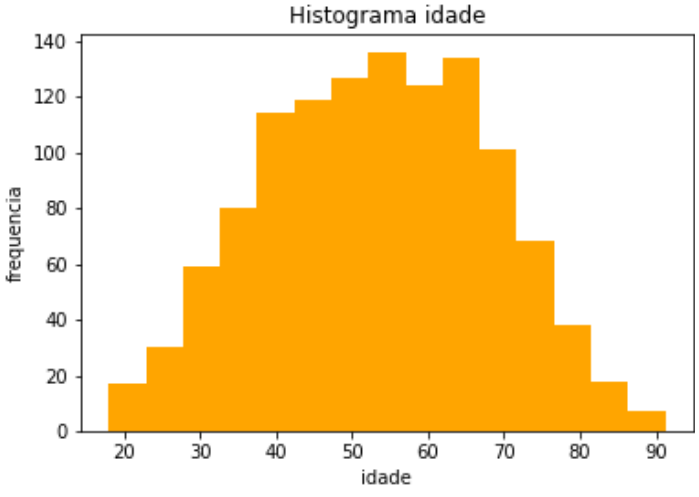
Histograma e Função Distribuição Empírica

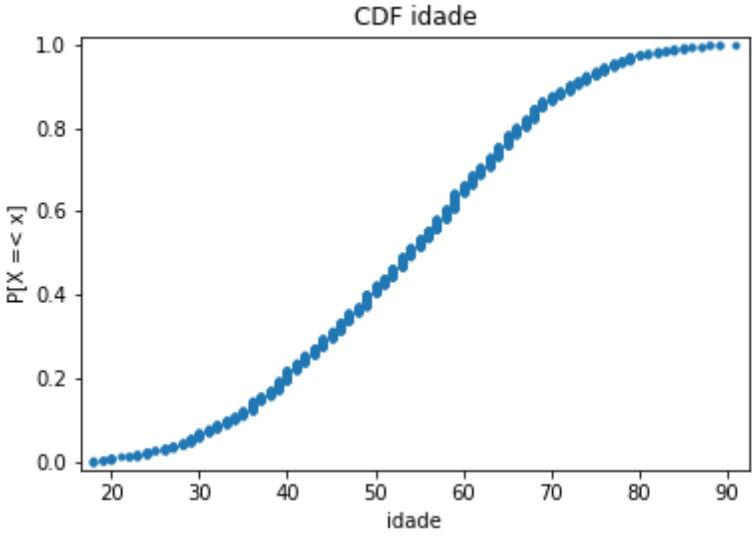
Aqui plotamos os histogramas correspondentes, onde o eixo “x” representa a respectiva variável e o eixo “y” a frequência com que determinado intervalo aparece no dataset. Os bins foram obtidos da seguinte forma: primeiro calculamos o espaçamento entre eles utilizando a fórmula binwidth = 3.49 * σ * n^(-1/3), onde “n” é a quantidade de amostras da variável - que é igual para todas as variáveis - e “σ” o desvio padrão da mesma. Então foi usada a função do “numpy” np.arange(min(X), max(X) + binwidth, binwidth) para obter os intervalos, sendo X a variável em questão.

Já a função distribuição acumulativa empírica (ECDF): No eixo “x” ordenamos os dados usando função np.sort(), também do “numpy”. No eixo “y” pegamos o intervalo que vai de 1 até n+1 e o dividimos por “n” a fim de obter um intervalo entre 0-1 que representa a contribuição individual de cada amostra para a ECDF – cada amostra contribui 1/n para a CDF, portanto o eixo “y” foi dividido com esse espaçamento.

Dessa forma, o primeiro ponto possui abscissa corresponde ao menor valor do conjunto de dados e ordenada 1/n. O segundo ponto tem abscissa correspondente ao segundo menor valor da amostra e ordenada 2/n, assim por diante incrementando a probabilidade acumulada em 1/n até que o último ponto possua abscissa correspondente ao maior valor do dataset e ordenada 1. Uma vez obtidos as coordenadas “x” e “y” de cada ponto, efetuamos a plotagem.

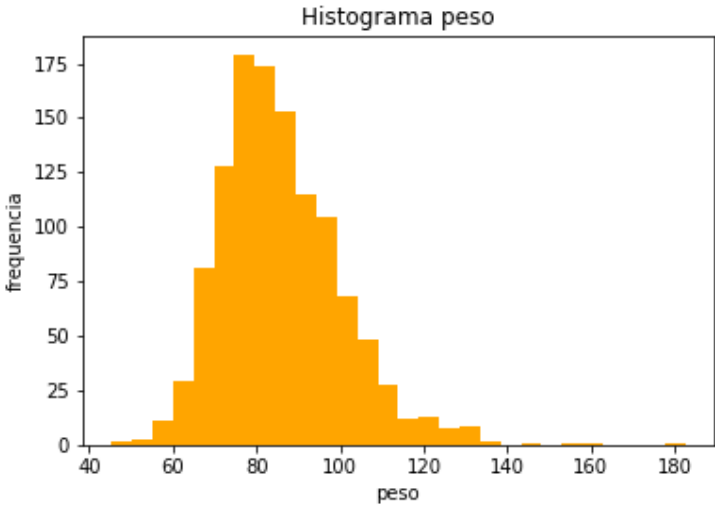
1- Idade

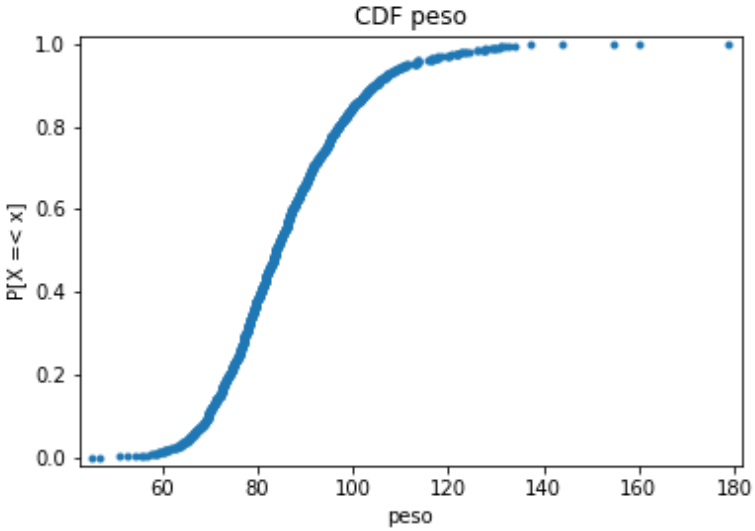




A partir da CDF vemos que cerca de 20% das pessoas estão abaixo dos 40 anos, enquanto que menos 20% aproximadamente estão acima dos 70 anos (Pela CDF a probabilidade de alguém ter menos do que 70 anos é perto de 0.8, fazendo 1 – 0.8 encontramos a probabilidade de 0.2 estar acima dessa idade). Assim, concluímos que a variância desse conjunto de dados é considerável.

2- Peso



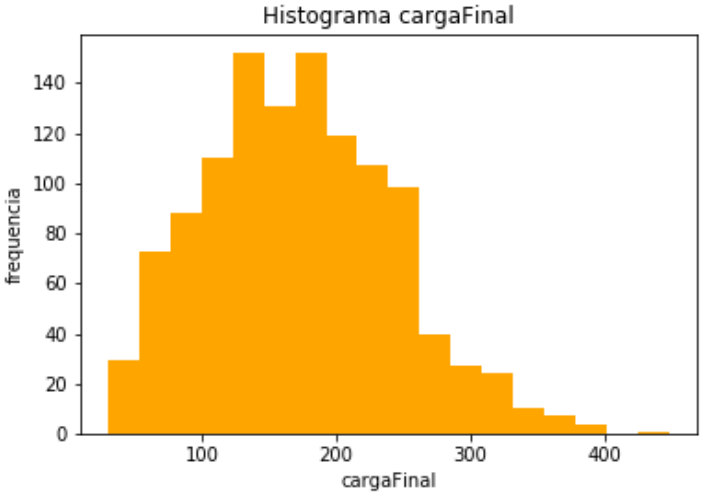


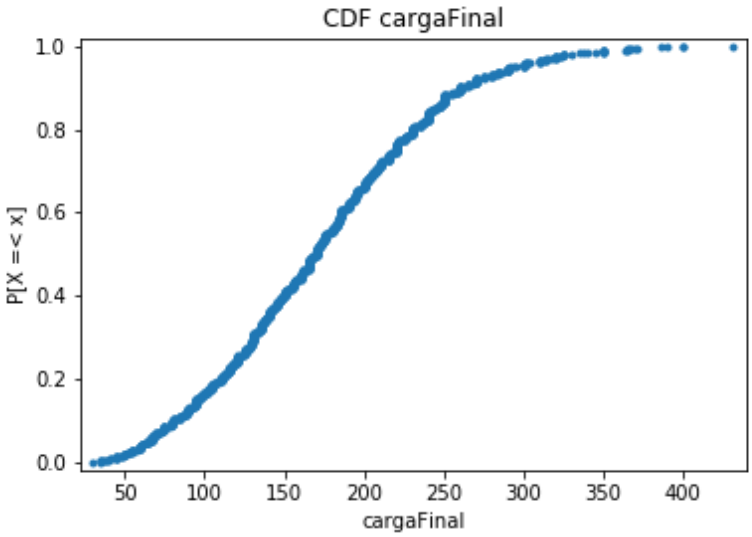
Pela CDF vemos que aproximadamente 40% dos pacientes possuem menos do que 80 kg, ao passo que apenas 20% estão acima dos 100 kg (Novamente, analisando a CDF, temos que a probabilidade de um paciente ter menos de 100 kg é de 0.8, assim aplicamos o raciocínio análogo feito para a idade)

Podemos presumir que a variância nesse caso é menor, pois analisando o histograma vemos uma frequência alta e estreita em torno dos 80 kg. Isso também pode ser constatado pela CDF: entre 80 kg e 100 kg, se traçarmos uma reta ligando os pontos, notamos uma inclinação próxima de 90 graus devido ao crescimento abrupto e alta concentração de pacientes nessa faixa.

Outro ponto observável é a presença de outliers: No gráfico da CDF, há pontos muito afastados da curva de distribuição, representando pacientes com peso muito acima, destoando dos demais. Isso pode ser um indicativo de paciente que possuem hábitos diferentes dos demais.

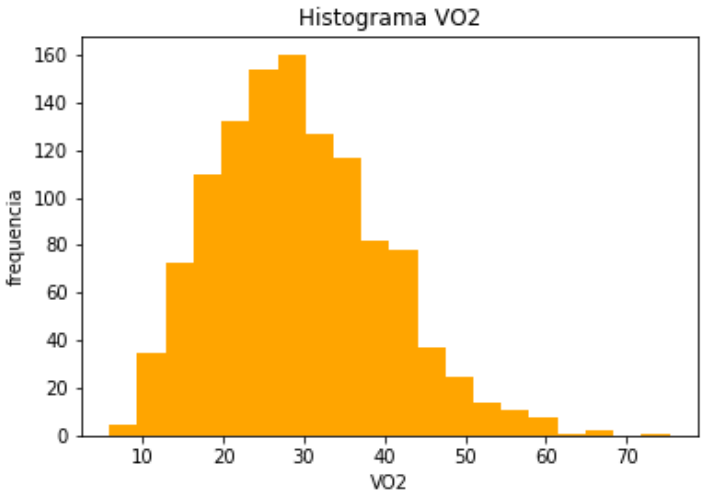
3- Carga Final

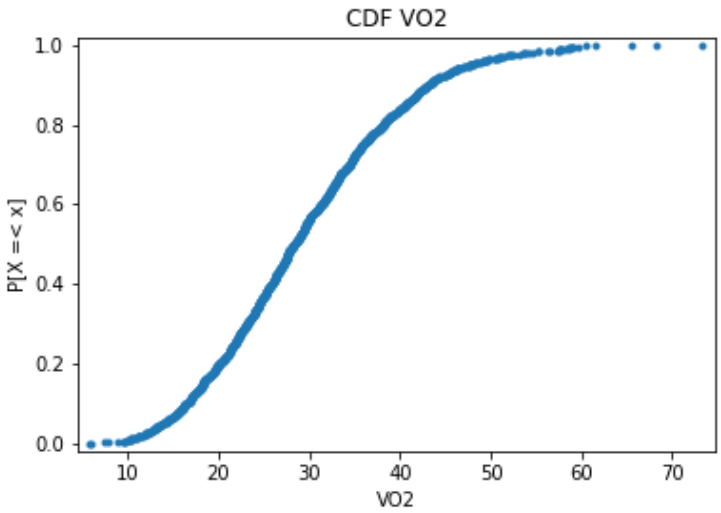




Observando a CDF, estima-se que entorno de 60% possui carga final abaixo de 200 watts e 40% abaixo de 150 watts. Além disso, tal como observado na distribuição do peso, temos alguns pontos afastados da curva, podendo indicar pessoas com capacidade física muito acima do restante (outliers).

4-VO2 Máximo



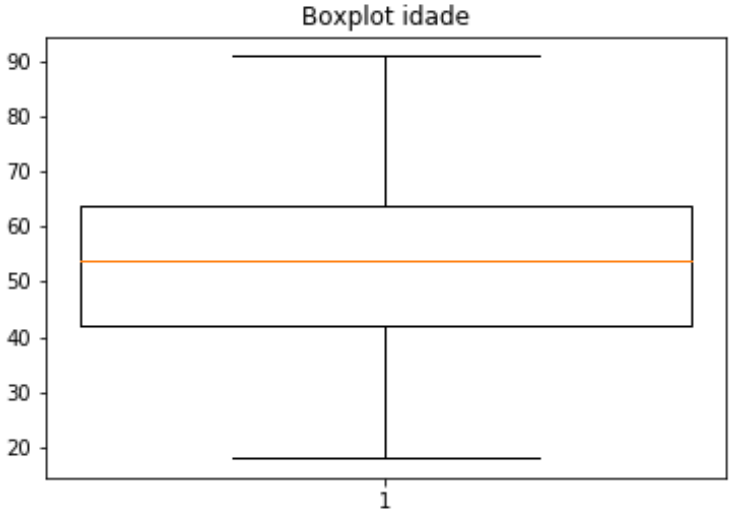


Notou-se que cerca de 60% dos pacientes tem VO2 máximo abaixo de 30 ml/(Kg.min), de acordo com os gráficos acima. Novamente, a presença de outliers nesse caso pode indicar pacientes mais adeptos a atividades aeróbicas, pois segundo o artigo "**Efeitos do Estado e Especificidade do Treinamento Aeróbio na Relação %VO2max versus %FCmax Durante o Ciclismo**", cujos autores são Fabrizio Caputo, Camila Coelho Greco, Benedito Sérgio Denadai (retirado de http://www.scielo.br/pdf/abc/v84n1/23000.pdf em 7/12/2018), a taxa de VO2 máximo é maior em pessoas adeptas a esse tipo de exercício físico (ciclistas, triatletas, corredores, etc.).

Média, Desvio Padrão, Variância e Boxplot

Utilizando a biblioteca "pandas", obtemos a variância, desvio padrão e média dos dados. Já para o boxplot, foi utilizada a função plt.boxplot() da biblioteca "matplotlib". Ainda, utilizando a função np.percentile() da biblioteca "numpy" obtemos a mediana e os valores Q1 e Q3, que correspondem ao lower quartile e upper quartile respectivamente. Assim obtemos outras medidas, tais como: IQR (Inter Quartile Range) fazendo Q3 – Q1 ; limite superior fazendo Q3 + 1.5 * IQR e limite inferior fazendo Q1 – 1.5 * IQR

1- Idade

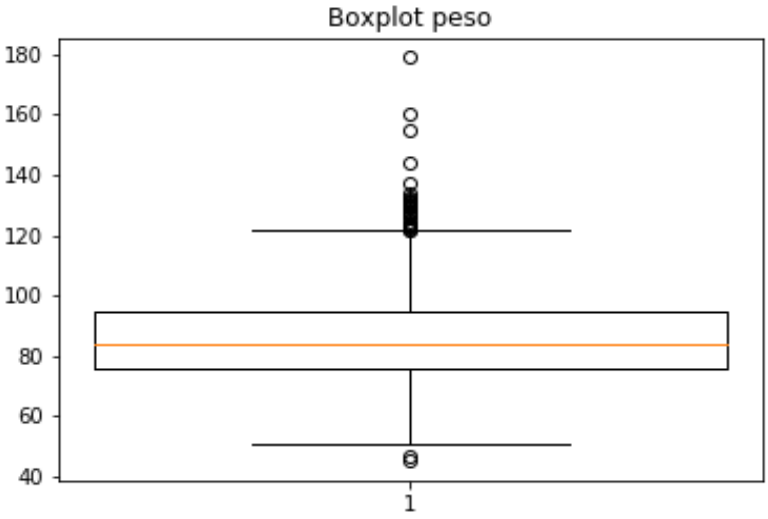


Media: 53.29095563139932
Desvio Padrao: 14.746296966880655
Variancia: 217.45327423543358
mediana: 54.0
Q1: 42.0 ,Q3: 64.0
IQR: 22.0
limite inferior: 9.0
limite superior: 97.0
Idade min: 18
Idade max: 91

Diretamente do boxplot, vemos que o limite superior foi ajustado para 91, já que não há nenhum dado acima desse valor, embora o calculado seja 97. Da mesma forma o limite inferior foi ajustado para 18, já que não há nesse dataset nenhum valor abaixo disso, embora o calculado seja 9.

Observa-se ainda que não há outliers para a variável idade, conforme era esperado analisando a ECDF da mesma (não possui nenhum ponto destoante dos demais). Pelo IQR, vemos que a variabilidade das amostras é razoável, o que indica que os dados não estão nem muito concentrados entorno da média, nem muito afastados da mesma.

2-Peso

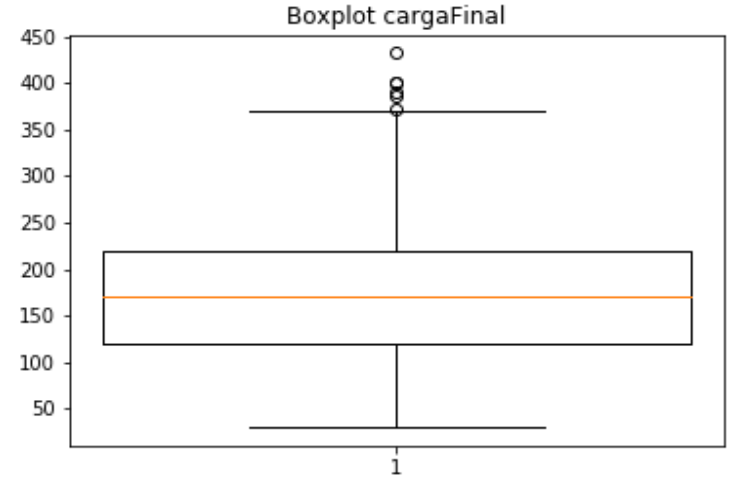


Media: 85.92577645051215
Desvio Padrao: 14.799113384059632
Variancia: 219.01375695425293
mediana: 83.7
Q1: 76.1 ,Q3: 94.45
IQR: 18.35
limite inferior: 48.575
limite superior: 121.975
Peso min: 45.3
Peso max: 178.9

Aqui percebemos a presença de outliers tanto superiores quanto inferiores. Os outliers superiores já eram esperados pela análise feita observando-se a CDF, no entanto o fato de haver, também, outliers inferiores só foi possível determinar com maior precisão analisando o boxplot.

Numa pesquisa feita com a população de Brasília, retirada de http://www.scielo.br/pdf/rbepid/v16n1/1415-790X-rbepid-16-01-0157.pdf em 7/12/2018, apontou que o peso médio masculino era 78,5 kg. A pesquisa, que durou cerca de 13 anos, trabalhou com a faixa etária entre 20 – 91 anos, cuja média amostral é de 44 anos. Dada a semelhança entre os dados coletados tanto pelo estudo descrito, quanto pelo professor Claudio (faixa etária idêntica, ambos se tratam da população masculina e diferença pequena entre os resultados obtidos), futuramente pode-se fazer uma análise comparativa de outras características dos habitantes (hábitos alimentares, prática esportiva, poder financeiro etc.) e portanto determinar aquelas que levaram a obtermos resultados tão próximos.

Outra possibilidade é: suponha que as populações do Rio de Janeiro e de Brasília possuíam a 13 anos atrás a mesma distribuição para o peso. No contexto da luta contra a obesidade, queremos determinar se uma política governamental adotada em Brasília foi efetiva. Para tal, podemos fazer um teste de hipótese com relação à média e o teste de Kolmogorov-Smirnov para aceitar/rejeitar a hipótese de que houve de fato redução da média e se a distribuição variou.

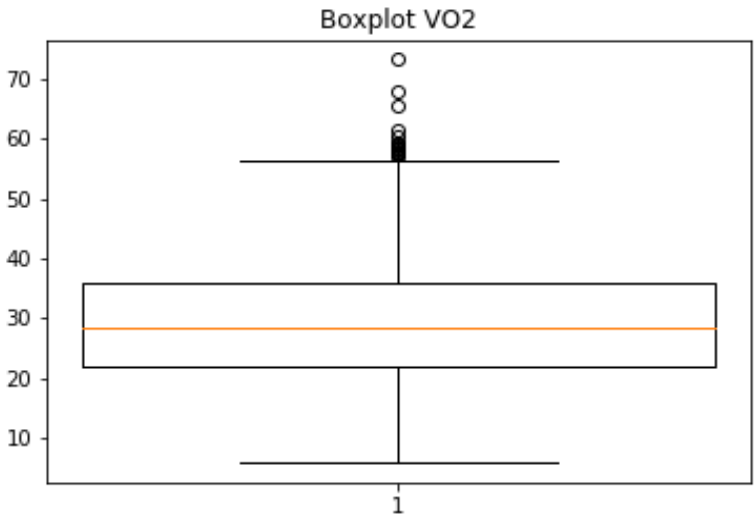


3-Carga Final

Media: 172.27150170648466
Desvio Padrao: 70.09312366247202
Variancia: 4913.045984762596
mediana: 170.0
Q1: 120.0 ,Q3: 220.0
IQR: 100.0
limite inferior: -30.0
limite superior: 370.0
Carga Final min: 30.0
Carga Final max: 432.0

Pelo boxplot podemos ver que o limite superior se manteve igual ao calculado, no entanto o limite inferior foi ajustado para 30 watts, já que na amostra não nenhum valor abaixo desse, embora o calculado seja -30 pela fórmula. Aqui há portanto apenas outliers superiores - o que já era de se esperar observando a CDF - mas não tão discrepantes quanto aqueles observados para a variável peso.

4-VO2 Máximo



Media: 29.394727923153184)
Desvio Padrao: 10.497249893426014
Variancia: 110.19225532503248
mediana: 28.32665964175
Q1: 21.7974228575 ,Q3: 35.853793263325
IQR: 14.056370405825003
limite inferior: 0.7128672487624961
limite superior: 56.938348872062505
VO2 min: 5.85
VO2 max: 73.33

Novamente, percebemos que o limite inferior foi ajustado para 5.85 ml/(kg.min), já que na amostra não há nenhum valor abaixo desse, embora o calculado pela fórmula seja 0.71 ml/(kg.min). Já o limite superior se manteve igual ao calculado pela fórmula. Assim, pudemos observar a presença de apenas outliers superiores que, assim como a carga final, possui discrepância menor que a variável peso.

Observações gerais: Se levarmos em conta apenas a “altura da caixa” do boxplot de cada uma das variáveis, então aquela que teria a menor variabilidade dos dados é o peso. Por outro lado, se considerarmos o desvio padrão, seria a taxa de VO2 máximo, o que nos leva a concluir que em certas ocasiões o boxplot pode ser tendencioso. Uma análise da variabilidade dos dados usando boxplot deve ser feita comparando-se dados da mesma natureza com populações distintas.

Parametrizando distribuições

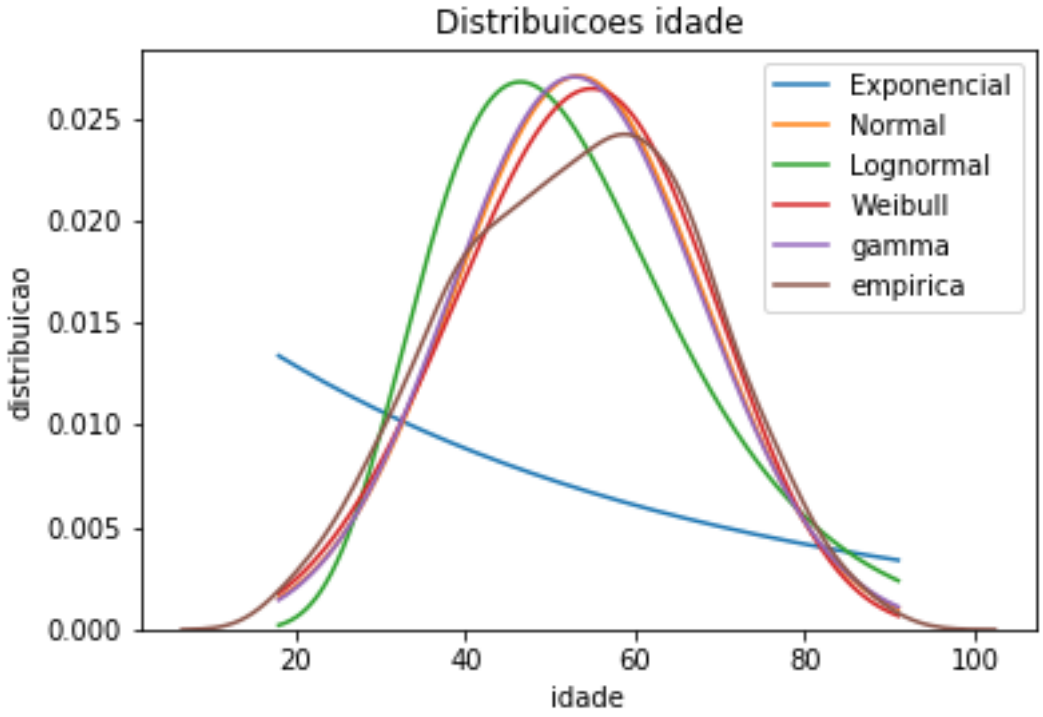
Utilizando o método do Maximum Likelihood Estimator (MLE) estimamos os parâmetros das distribuições lognormal, gaussiana, Weibull e da exponencial para cada uma das variáveis. A expressão analítica para os parâmetros da Weibull, conforme veremos a seguir, mostrou-se muito complexa. Portanto os mesmos foram obtidos por meio da função **scipy.stats.weibull_min() do python.**

O modelo MLE consiste em, dada as observações - isto é, os dados coletados – encontrar os parâmetros que maximizam a likelihood, parametrizando assim a distribuição mais verossímil para aquela população. A função likelihood nada mais é do que a pdf conjunta das amostras e, se supormos a independência mútua, a mesma será igual ao seguinte produtório:

$$L(\theta_1, \theta_2, \ldots, \theta_k | x_1, x_2, \ldots, x_k) = \prod_{i=1}^{n} f(x_i | \theta_1, \theta_2, \ldots, \theta_k)$$

θ’s: parâmetros; x’s: amostras

Uma vez obtidas as distribuições da literatura parametrizadas, comparou-se com a distribuição empírica obtida anteriormente. Dessa forma podemos verificar qual que mais se assemelha e assim utilizá-la para previsões futuras.



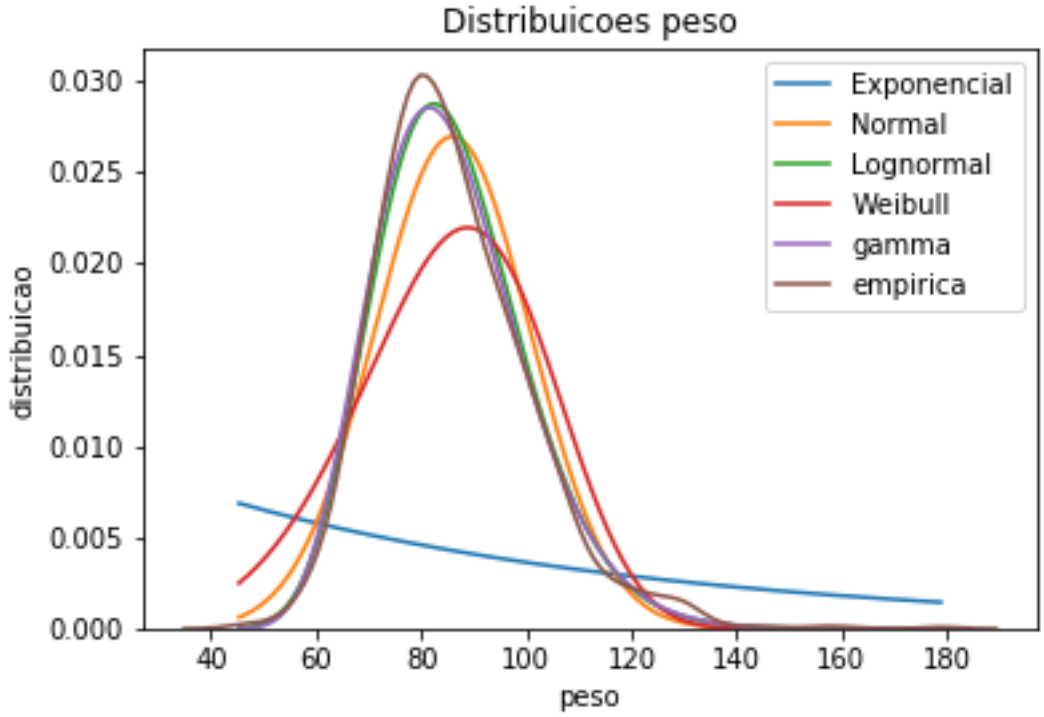

```
Valor do  λ  da Exponencial:  0.0187649102583

Valor de  μ  (média) da Normal:  53.2909556314
Valor de  σ  (desvio padrão) da Normal:  14.7462969669

Valor de  μ  (scale) da lognormal:  3.932509819486875
Valor de  σ ^2 (shape) da lognormal:  0.0936331955793438

Valor de  β  (scale) da weibull:  58.78289005707875
Valor de  α  (shape) da weibull:  4.089481828645864
Valor de  γ  (location) da weibull:  0

Valor de  β   da gamma:  0.43453226781141874
Valor de  α   da gamma:  1155.4445044051527
Valor do location da gamma:  -448.81311781571185
```



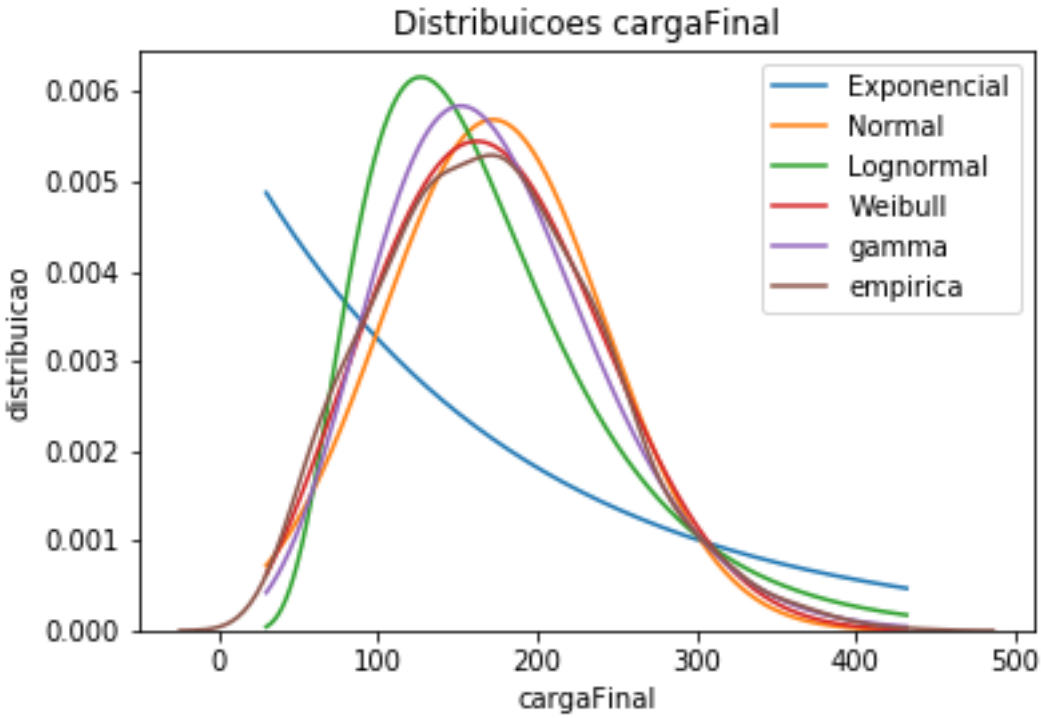

```
Valor do  λ  da Exponencial:  0.0116379512797

Valor de  μ  (média) da Normal:  85.9257764505
Valor de  σ  (desvio padrão) da Normal:  14.7991133841

Valor de  μ  (scale) da lognormal:  4.439451920143028
Valor de  σ ^2 (shape) da lognormal:  0.027586997105752877

Valor de  β  (scale) da weibull:  92.24080850317551
Valor de  α  (shape) da weibull:  5.408013188534343
Valor de  γ  (location) da weibull:  0

Valor de  β   da gamma:  4.141585952256193
Valor de  α   da gamma:  12.226068852907186
Valor do location da gamma:  35.290455713407525
```



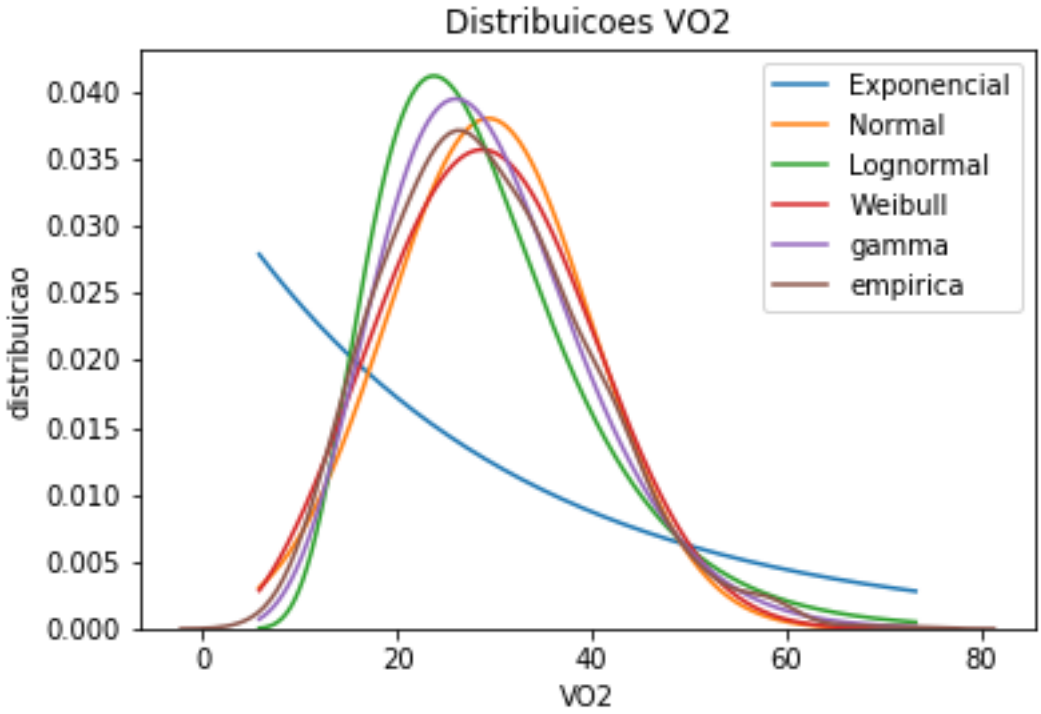
Valor do λ da Exponencial: 0.00580479063626

Valor de μ (média) da Normal: 172.271501706
Valor de σ (desvio padrão) da Normal: 70.0931236625

Valor de μ (scale) da lognormal: 5.0546544058509895
Valor de σ ^2 (shape) da lognormal: 0.2103368574854832

Valor de β (scale) da weibull: 194.0388415799269
Valor de α (shape) da weibull: 2.6469810001574725
Valor de γ (location) da weibull: 0

Valor de β da gamma: 20.172265942417688
Valor de α da gamma: 12.281388335183824
Valor do location da gamma: -75.47193343770181




```
Valor do  λ  da Exponencial:  0.0340197059355

Valor de  μ  (média) da Normal:  29.3947279232
Valor de  σ  (desvio padrão) da Normal:  10.4972498934

Valor de  μ  (scale) da lognormal:  3.3132400746591215
Valor de  σ ^2 (shape) da lognormal:  0.14364411960908474

Valor de  β  (scale) da weibull:  32.9274599599628
Valor de  α  (shape) da weibull:  2.9978221690896216
Valor de  γ  (location) da weibull:  0

Valor de  β   da gamma:  3.330202880351707
Valor de  α   da gamma:  10.048783233395653
Valor do location da gamma:  -4.069760606930359
```

Todos os gráficos foram traçados usando a biblioteca matplotlib.pyplot.plot(), exceto pela distribuição empírica, que foi feita utilizando seaborn.

Por que foi incluída a Gamma se já havia distribuições próximas da empírica para todas as variáveis?

Ao efetuarmos as parametrizações e plotarmos as respectivas funções de densidade de probabilidade pela primeira vez, constatou-se que nenhuma delas se adequava suficientemente bem para a distribuição empírica do peso. Portanto, comparando essa última com outras distribuições da literatura, optou-se pela Gamma já que possuía formato similar.

Contudo, ao efetuarmos o QQPlot para o peso, notamos que tanto a lognormal quanto a gamma eram idênticas, apresentando uma inconsistência com o observável no gráfico de parametrizações: a gamma e a lognormal possuíam uma discrepância considerável. Foi então que descobriu-se que a lognormal estava plotada erroneamente.

Posteriormente ajustes no código foram realizados, obtendo a plotagem correta para a lognormal apresentada acima. Com esse novo resultado a lognormal não só se adequava muito bem para a distribuição do peso, quanto ela e a Gamma estavam praticamente sobrepostas, reafirmando as observações feitas no QQPlot.

Porém, o trabalho não foi de todo em vão: verificou-se mais tarde que para a VO2 Máxima a distribuição que de longe melhor se ajustou foi a Gamma! Se não fosse esse pequeno erro inicial na plotagem da lognormal, provavelmente não teríamos inserido essa nova distribuição, muito menos chegado à tal resultado!!!

Os parâmetros da gamma que nos leva a obter a Likelihood máxima foram obtidos usando a função stats.gamma.fit(), da biblioteca "scipy". Exceto para a variável peso, cujo parâmetro location foi ajustado manualmente de forma a aproximar a distribuição gamma da empírica, todos as demais tiveram o location obtido pelo método fit() mencionado acima.

Observações:

-Para a idade percebemos que as mais próximas da empírica são a Weibull e a Normal, não sendo possível determinar visualmente qual a melhor. Além dessas, vemos que gamma praticamente acompanha toda a trajetória da gaussiana, sendo portanto outra possível candidata para representar a empírica.
-Para o peso, de todas a Lognormal é a que tem a kurtose mais perto da empírica. Além disso notamos que ela e a Gamma estão praticamente sobrepostas uma na outra, sendo ambas igualmente boas para representar os dados.

-Para a carga final, tal como para a idade, tanto a Weibull quanto a Normal e a Gamma se assemelham bastante da empírica. Porém, se analisarmos a kurtose - isto é, o alongamento vertical – conclui-se que a Weibull seja mais adequada. Já a lognormal possui uma kurtose ainda maior, levando ao aumento da discrepância em relação à empírica, sem mencionar o fato de estar mais afastada longitudinalmente à esquerda – isso se dá devido ao fato dela possuir skewness (coeficiente de assimetria) maior e positivo.

-Para VO2 Max, a Gaussiana, a Gamma e a Weibull são as que mais se aproximam da empírica (novamente, a lognormal é a mais afastada da empírica, tanto em relação ao eixo das ordenadas quanto das abscissas). No entanto, dada a tamanha proximidade das três curvas, não é possível inferir qual delas melhor se adéqua, deixando esse teste inconclusivo.

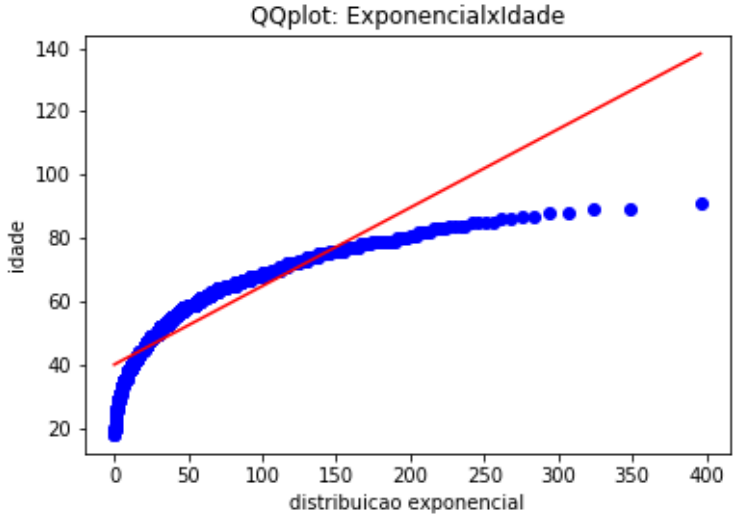
Em todos os gráficos, a distribuição exponencial foi a que apresentou maior discrepância em relação à empírica

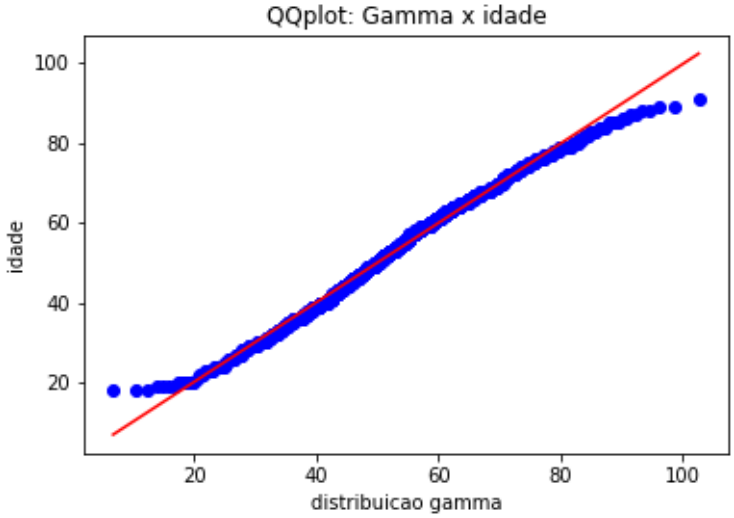
Gráfico QQPlot (ProbabilityPlot)

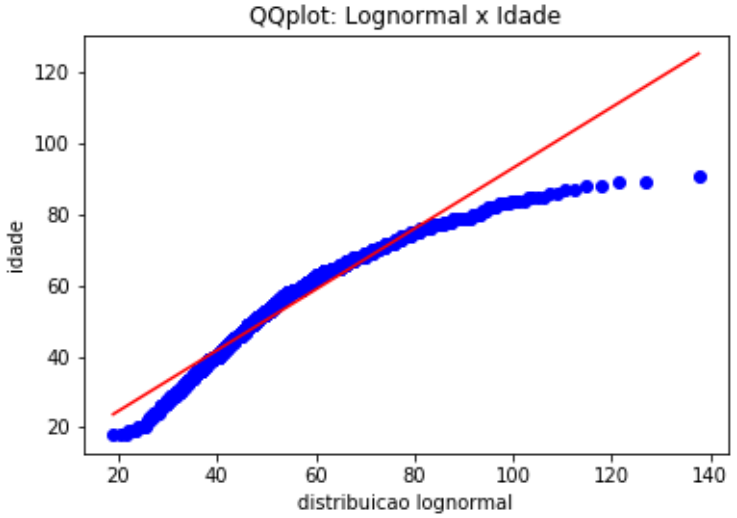
O quantile – quantile (q- q) plot é uma técnica gráfica para determinar se dois conjuntos de dados vem de populações com a mesma distribuição. Nesse caso, um data set será os dados coletados pelo professor, e o outro serão os dados obtidos pela distribuição de interesse.
O gráfico é feito da seguinte forma: primeiro, atribuímos um valor de quantile para cada amostra de um data set. Em seguida, pega-se a distribuição com a qual queremos comparar e adicionamos a mesma quantidade de quantiles que criamos para o data set. Se no data set tínhamos "N" amostras, então teremos "N" quantiles para ambos conjuntos de dados. Na prática, isso significa dividir os data sets em grupos de tamanhos iguais. Para a curva de distribuição, dividir em grupos de tamanho igual significa ter a mesma probabilidade de se observar um valor em cada grupo.
Por exemplo, se estamos comparando com a distribuição normal, os grupos mais nas bordas possuem probabilidade menor de ocorrência e para compensar esse fato eles precisam ser mais largos (abranger um intervalo maior de valores que a variável aleatória pode assumir). Já aqueles que estão mais próximos dá média da normal precisam ser mais estreitos devido à maior probabilidade de se observar um valor nessa faixa.

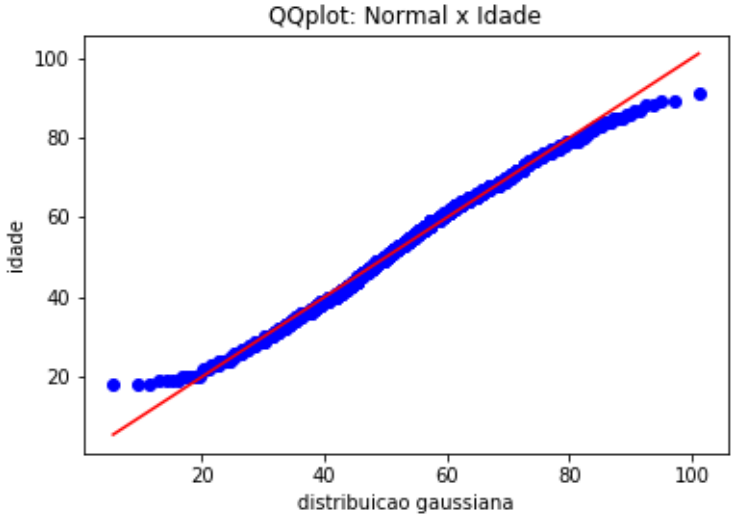
A plotagem do gráfico, por fim, começa pegando o menor quantile de ambos data set's e associar um com a ordenada e outro com a abscissa, continuando o processo até que todos os quantiles sejam considerados (supõe-se que eles já estejam ordenados). Assim, se um ponto possui coordenadas (1, 2) e 25% das amostras de um data set encontram-se abaixo de 1, então também 25% do outro data set encontram-se abaixo de 2. Portanto, em caso de ambos data set's possuírem a mesma distribuição, então os valores dos quantiles serão iguais, formando uma linha que faz 45° com o eixo das abscissas. Sendo assim, esse método se mostra bem eficaz para determinar a semelhança entre duas distribuições e se traçarmos a reta y = x a visualização fica bem simples.

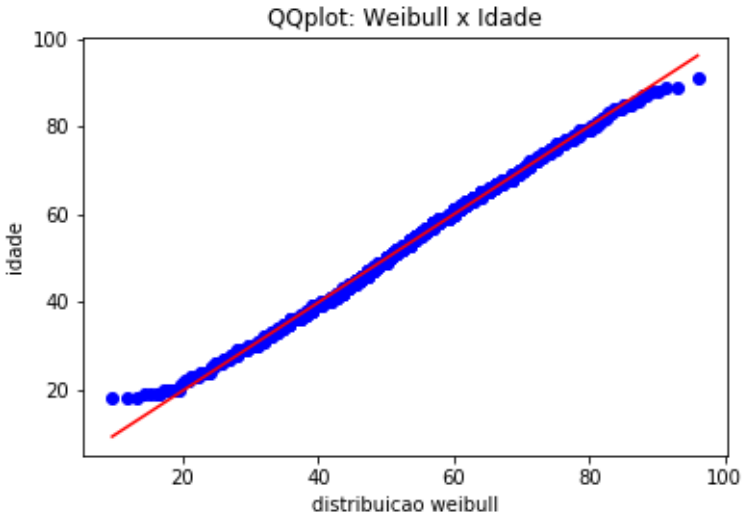
1-Idade





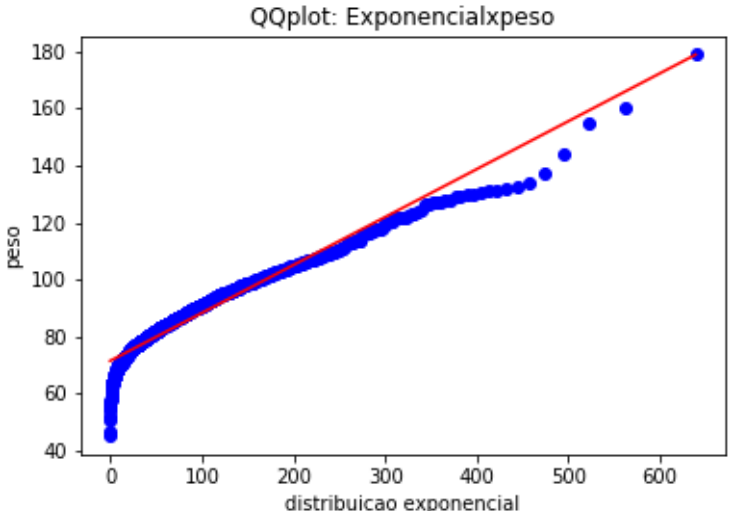


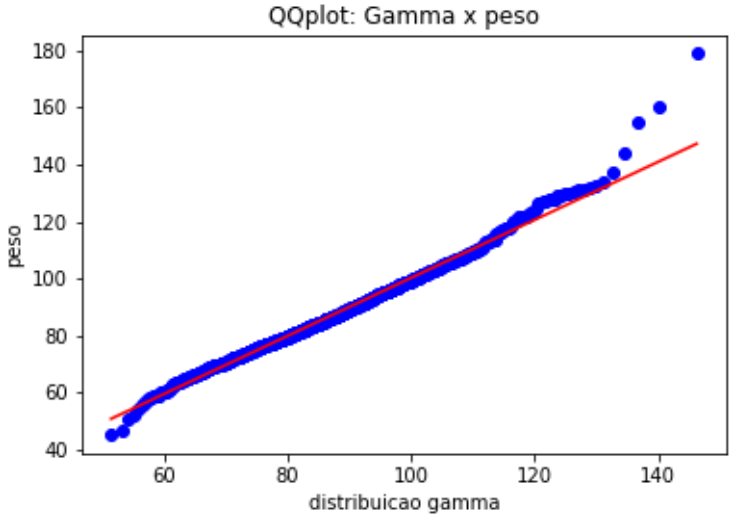


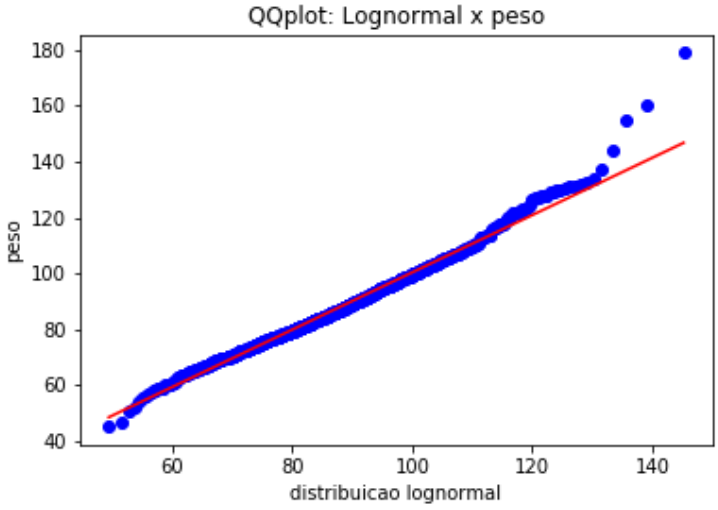


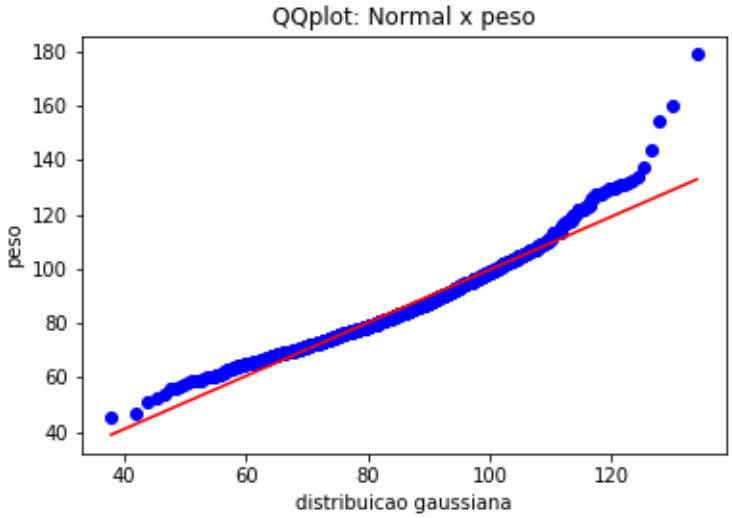
Numa primeira análise vemos que nem a lognormal nem a exponencial são compatíveis com a empírica. Nos resta analisar a Weibull, Normal e a Gamma. Tanto a Gamma quanto a Normal possuem desvios muito semelhantes em relação a reta y = x, conforme era de se esperar pois, na análise feita no item anterior, vimos que suas trajetórias se sobrepõe em grande parte. Porém, pelo fato da Weibull possuir um desvio menor ainda em relação a ambas, concluímos que ela é que melhor se adéqua.

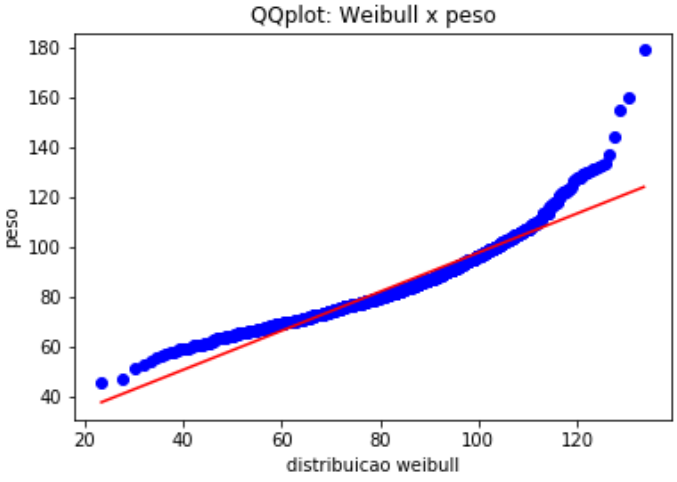
2-Peso









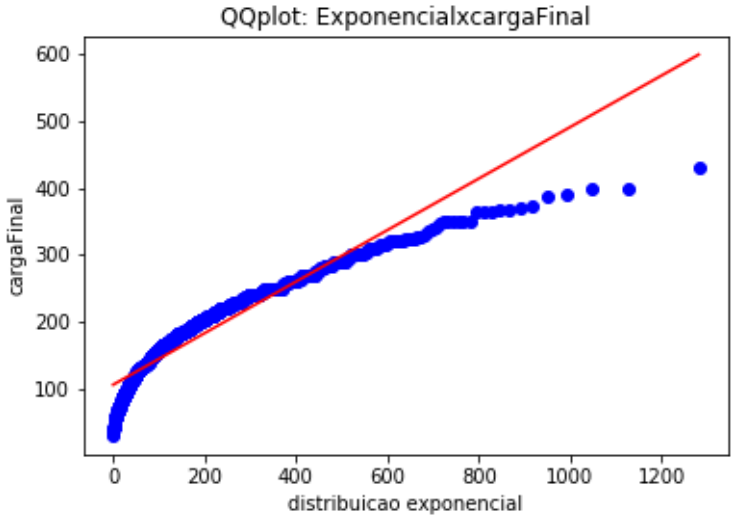


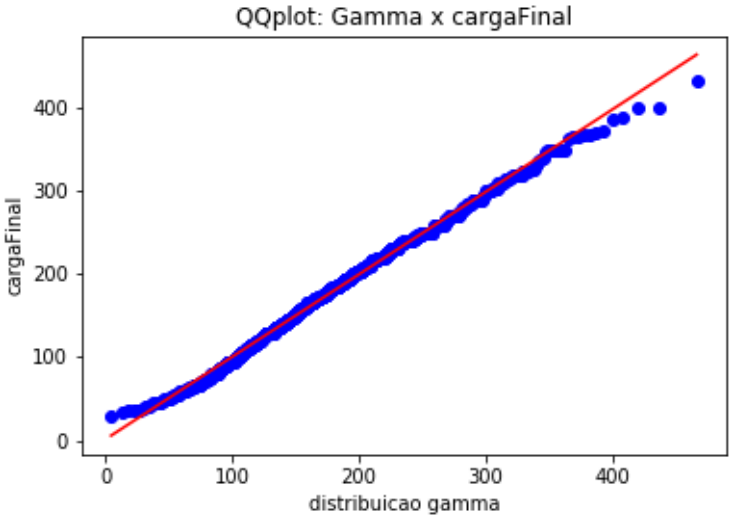
Nenhuma das distribuições parece se adequar tão bem à empírica, devido ao aumento significativo do desvio em relação à reta conforme aumenta o valor do peso.
Lembrando dos resultados obtidos no primeiro item, constatamos que alguns pontos se situavam bem distantes da ECDF, podendo indicar grande

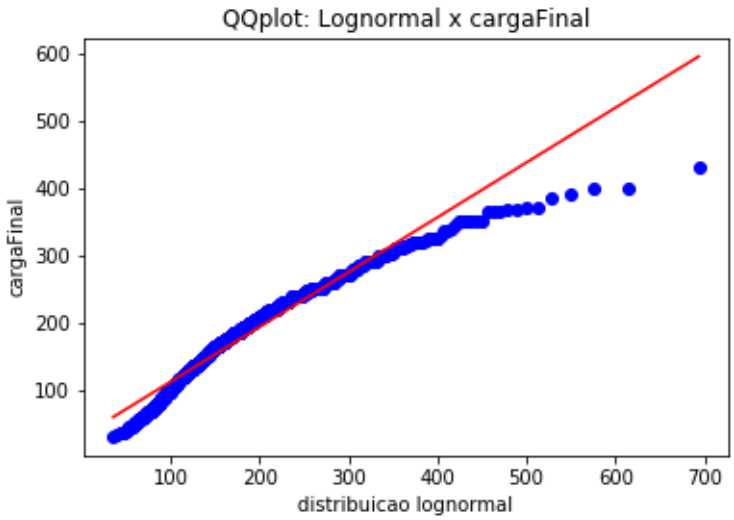
quantidade de outliers superiores, fato esse reforçado quando obtemos o Boxplot da variável peso. Assim, essa parece ser a principal causa dos desvios aqui observados: a presença significativa de indivíduos que destoam do restante.

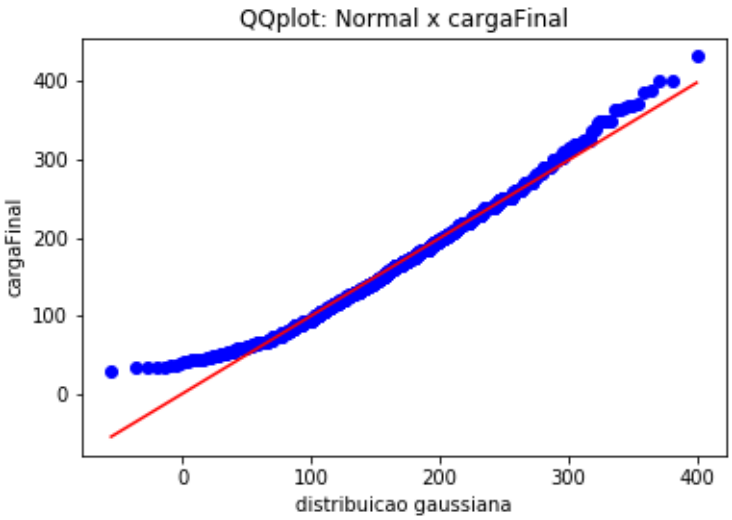
Contudo, se ignorarmos tais desvios - isto é, os valores acima do limite superior da Boxplot (121.975 kg) e os abaixo do limite inferior da mesma (48.575 kg) - vemos que tanto a Gamma quanto a Lognormal apresentam a melhor adequação à empírica. Não conseguimos determinar qual a melhor, dado que ambas possuem QQPlot idênticos (isso já era de se esperar, pois na parametrização constatamos que ambas tinham as suas respectivas distribuições praticamente sobrepostas).

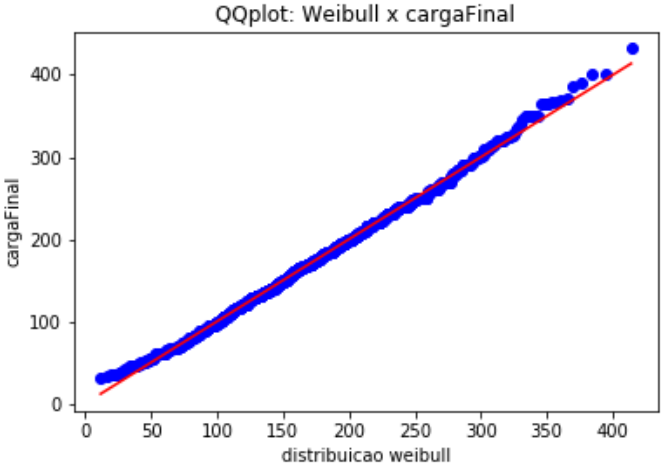
3-Carga Final







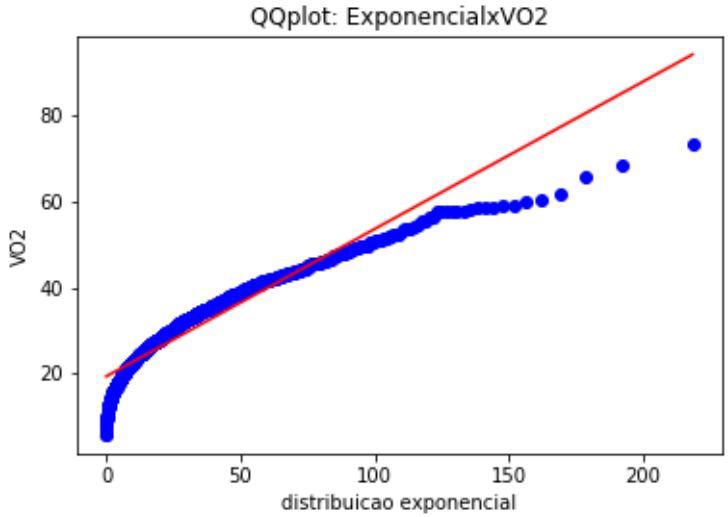


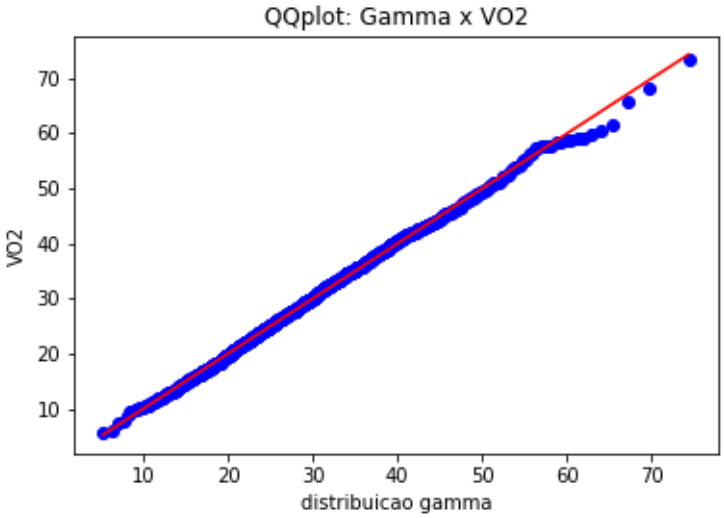


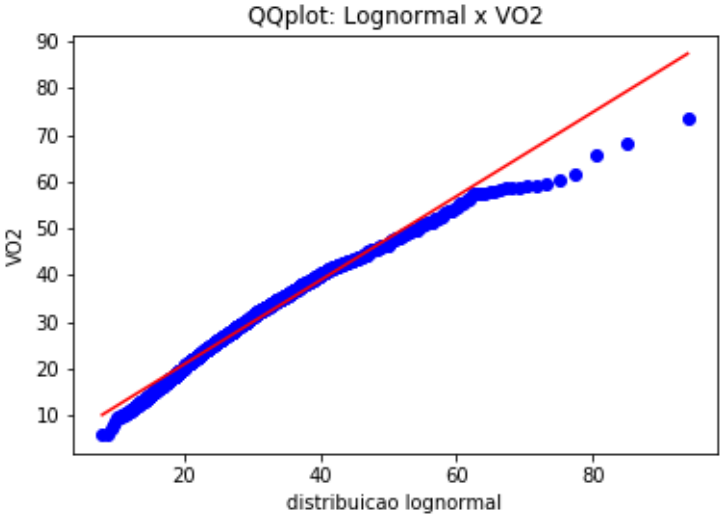
Para a carga final, o Boxplot nos diz que não há tantos outliers assim. Portanto os desvios aqui observados são devido ao quanto uma distribuição destoa de outra. Analisando o QQPlot, de fato a Normal, a Gamma e a Weibull se assemelham muito com a empírica, conforme pode ser constatado pelo fato da maioria dos pontos se situarem sobre a reta.

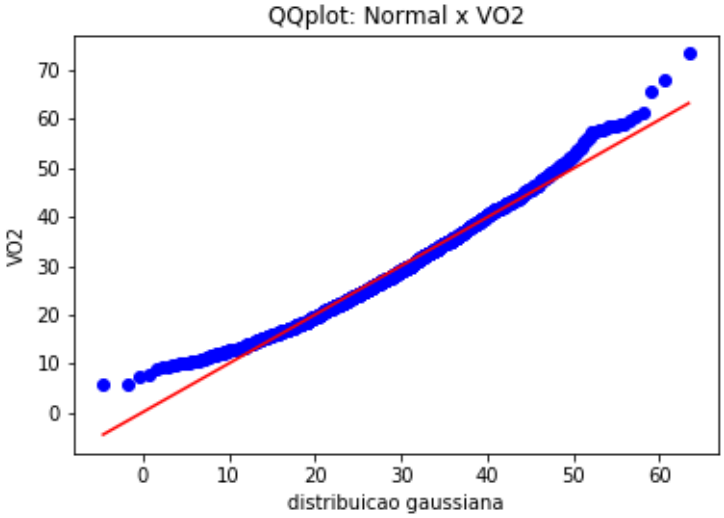
No entanto, a Weibull é a que melhor se adéqua, o que já era de se esperar com o argumento da kurtose levantado no item anterior.

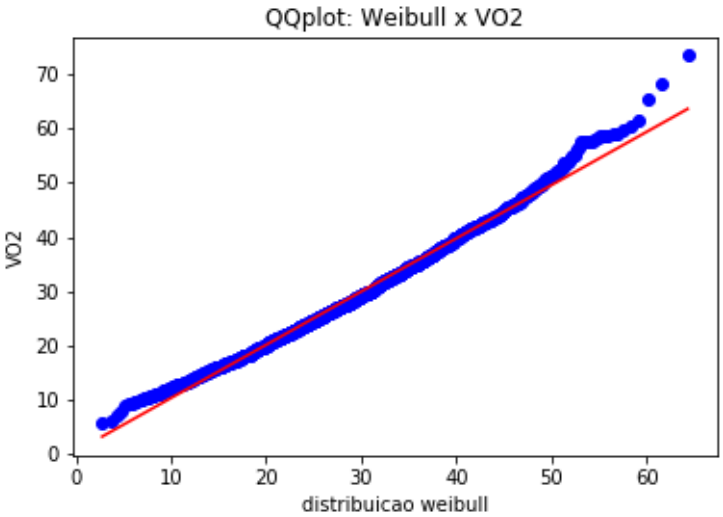
4-VO2 MÁX











Tal como para o peso, o Boxplot nos diz que há muitos outliers superiores no conjunto de dados coletados. Portanto, devemos levar isto em consideração ao analisarmos os desvios do QQPlot.

De fato, se observarmos os gráficos aqui apresentados, acima do limite superior do Boxplot (56.93 ml/(kg.min)) observamos que todas as curvas começam a apresentar desvios maiores em relação à reta. Para efeitos de comparação desconsideraremos, portanto, tais anomalias.

Se anteriormente na parametrização era impossível distinguir qual das três

distribuições - Gamma, Weibull ou Normal - representa melhor a empírica, aqui podemos determinar com maior clareza que é a Gamma. É também importante ressaltar que a tríade possui gráficos QQPlot bem parecidos e todas apresentam grande quantidade de pontos sobre a reta, reflexo esse da notável proximidade entre as parametrizações e das mesmas com a empírica.

Teste de Hipótese

Aqui utilizaremos o teste Kolmogrov-Smirnov (KS) para validar/descartar as distribuições parametrizadas obtidas para cada uma das variáveis. Esse teste toma como base a máxima diferença absoluta Dn entre a função distribuição acumulada e a distribuição empírica dos dados para uma amostra aleatória X1, X2,…,Xn:

$$D_n = \sup_x |F(x) - F_n(x)|$$ Onde F(x) é cdf assumida para os dados e Fn(x) a cdf empírica

Assim, quanto mais próximo o valor de Dn estiver de zero, maior a probabilidade de ambas distribuições serem compatíveis.

Então compara-se tal diferença com um valor crítico para um dado nível de significância α, isto é, a probabilidade de rejeitar a hipótese nula quando ela for verdadeira – configurando assim o erro do tipo 1. Nossa hipótese nula H0 será que as duas distribuições são idênticas e caso Dn > dα a rejeitaremos. O valor de dα pode ser facilmente obtido de acordo com a tabela abaixo. Adotando o nível de significância igual a 0,05 e sendo a quantidade de amostras maior do que 35, usaremos a fórmula para obter dα igual a 0.03972.

Para o teste KS foi feito da seguinte forma: primeiramente estimamos os parâmetros da distribuição com a qual queremos comparar por meio da função stats.Nome_da_Distribuição.fit(), da biblioteca scipy. Em seguida, chamamos a função stats.kstest() com os parâmetros estimados anteriormente e as nossas variáveis aleatórias.

A função kstest() retorna tanto o valor de Dn obtido, quanto o p-valor, isto é, o menor valor de α para o qual a hipótese H0 é rejeitada. Se por um lado o Dn é usado para aceitar/rejeitar a amostra, por outro o p-valor pode ser uma medida importante se quisermos distinguir duas distribuições aceitas. Isso porque o p-valor é definido como a probabilidade de se observar um valor da estatística de teste maior ou igual ao encontrado, de forma que quanto mais próximo de 1 for o p-value, melhor a distribuição parametrizada se adéqua a empírica.

Valores Críticos do Teste KS

Tamanho da amostra (N)	Nível de significância para $D_{crit} = máx\|F_{obs}(X) - F_{esp}(X)\|$				
	0,20	0,15	0,10	0,05	0,01
1	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
Mais de 35	$\frac{1,07}{\sqrt{N}}$	$\frac{1,14}{\sqrt{N}}$	$\frac{1,22}{\sqrt{N}}$	$\frac{1,36}{\sqrt{N}}$	$\frac{1,63}{\sqrt{N}}$

1-Idade

```
Exponencial
D =  0.372755615059967
p_value =  0.0

Gamma
D =  0.04751884106246462
p_value =  0.00972503090633925

Weibull
D =  0.033037815723893305
p_value =  0.15145170476000502

Normal
D =  0.04408368872194113
p_value =  0.02039175142102323

Lognormal
D =  0.084730460447627
p_value =  9.073029882955552e-08
```

Tomando os respectivos Dn e lembrando que para valores maiores do que 0,039 a hipótese H0 é rejeitada, a única distribuição que não pode ser descartada para idade é a Weibull. Isso reitera as observações feitas no Qqplot: Apesar da Weibull e a Normal possuírem certa semelhança, os menores desvios em relação à reta y=x são obtidos com a primeira. Importante notar também que a Normal apresenta um Dn e p-valor bastante próximo aos da Weibull.

2-Peso

```
Exponencial
D =  0.4954410013455397
p_value =  0.0

Gamma
D =  0.02845849451831739
p_value =  0.29399800878157434

Weibull
D =  0.1032173331741221
p_value =  2.5226265520927882e-11

Normal
D =  0.06661818817785059
p_value =  5.7584235073626644e-05

Lognormal
D =  0.032285259002662436
p_value =  0.17003957723543306
```

Para o peso a Gamma e a Lognormal são as únicas que não foram rejeitadas ao analisarmos o Dn. Uma das principais dúvidas até então era qual das duas representava melhor a empírica. Essa pergunta não pode ser respondida apenas analisando os gráficos das parametrizações, nem o Qqplot. No entanto, se tomarmos o p-valor como critério de desempate, constatamos por meio desse teste que a Gamma, por possuir um p-valor maior, é a que melhor se adéqua.

3-Carga Final

```
Exponencial
D =  0.28651634266099946
p_value =  0.0

Gamma
D =  0.03250279883942098
p_value =  0.16448849232302964

Weibull
D =  0.02457022560625388
p_value =  0.47886304960046483

Normal
D =  0.039233911356943985
p_value =  0.052776560691338625

Lognormal
D =  0.08035970386976421
p_value =  4.962162909460943e-07
```

Para a carga final, as distribuições não rejeitadas são a Weibull, a Gamma e a Normal, conforme era de se esperar devido à incrível proximidade das mesmas com a empírica no gráfico das

parametrizações. Contudo, se adotarmos o p-valor como critério de desempate, vemos que a Weibull é que tem o maior p-valor, e portanto é a mais adequada. Isso reforça o observado tanto pelo gráfico das parametrizações sob o argumento da curtose, quanto no Qqplot sob o argumento dos menores desvios em relação à reta y=x.

4-VO2 Máximo

```
Exponencial
D =  0.3348896789424037
p_value =  0.0

Gamma
D =  0.01825409900634445
p_value =  0.829669491626449

Weibull
D =  0.03674655683834849
p_value =  0.08234531803549583

Normal
D =  0.044531849851028094
p_value =  0.018572422090608276

Lognormal
D =  0.04056142112151151
p_value =  0.041130681450034956
```

Para o VO2 máximo as distribuições aceitas são a Gamma e a Weibull. Conforme foi observado por todos os métodos até então, a Gamma é que mais se adéqua a empírica do VO2 e mais uma vez isso é observado pelo fato dela possuir um p-valor na ordem de dez vezes maior que a da Weibull.

A exponencial apresentou para todas as distribuições p-valor igual a zero, dada a sua enorme discrepância em relação a empírica.

Análise de dependência entre as variáveis, modelo de regressão

Nesse item buscamos verificar se havia dependência entre VO2 máx e alguma outra variável (peso, idade e carga final). Para tal efetuou-se o scatter plot V02 x Idade, V02 x Peso e VO2 x Carga Final, utilizando a função:
data.plot(kind='scatter',x='Nome_da_Variavel',y='Outra_Variavel'), do pandas.
Além disso, foram calculados os parâmetros do modelo de regressão por meio da função np.polyfit() - que internamente usa o método de mínimos quadrados - do numpy e o coeficiente de correlação de Pearson, esse último dado pela fórmula:

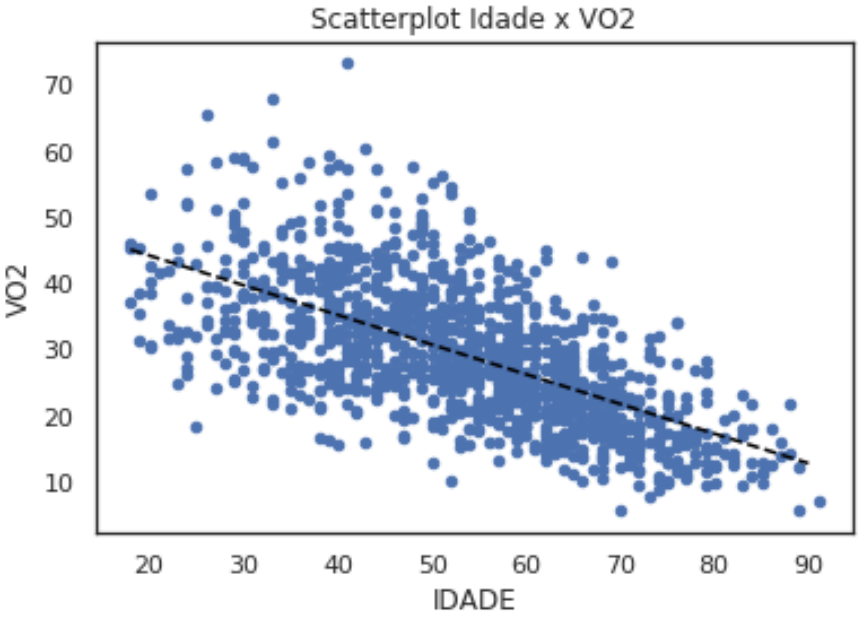
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}},$$

sendo “x” e “y” as variáveis em questão, “r” o coeficiente propriamente dito e “n” o número de amostras.

1-Idade x VO2

Conforme pode ser observado pelo gráfico em seguida, o padrão da plotagem formado pela idade e VO2 lembram uma reta decrescente. Isso se reflete no valor obtido para o coeficiente de correlação, que foi de -0.6300720192503417, onde o sinal negativo reflete o fato dela ser decrescente e cujo módulo nos informa de se tratar de uma correlação expressiva. De fato, à medida em que se envelhece o metabolismo tende a cair. Como as reações químicas envolvidas no metabolismo tem como principal reagente o oxigênio, é de esperar que a taxa máxima do mesmo decaia com o passar

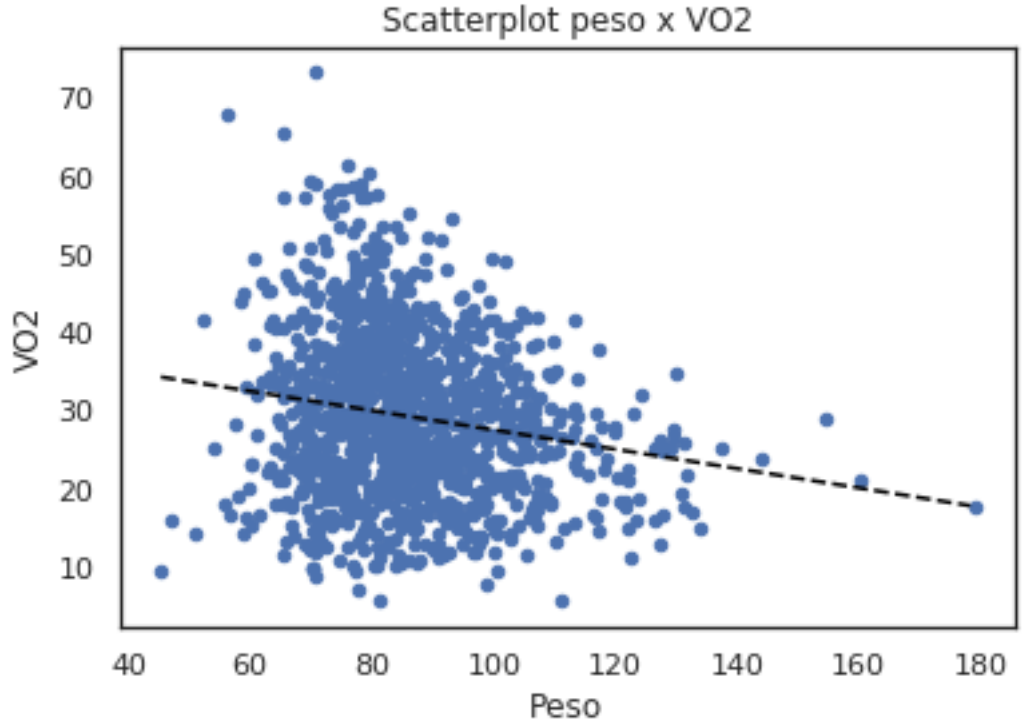
do tempo. Sendo assim, é possível usar o modelo de regressão linear nesse caso, cujos parâmetros são -0.44852097 para o coeficiente angular e 53.2968391 para o linear.



Observa-se ainda que acima da reta há alguns pontos bem afastados, por conta dos outliers previstos pela boxplot (taxas acima de 56.93 ml/kg.min), pessoas que possuem uma taxa aeróbica e preparo físico maior.

2-Peso x VO2

O padrão formado pela scatter plot se assemelha a um círculo nesse caso. Isso nos leva a crer que o coeficiente de correlação seja muito pequeno, e de fato é: `-0.17440061829630796.` Conclui-se portanto que o peso e a taxa de VO2 não estão correlacionadas: por exemplo, um atleta e uma pessoa sedentária podem possuir pesos parecidos, embora taxas de V02 completamente distintas.

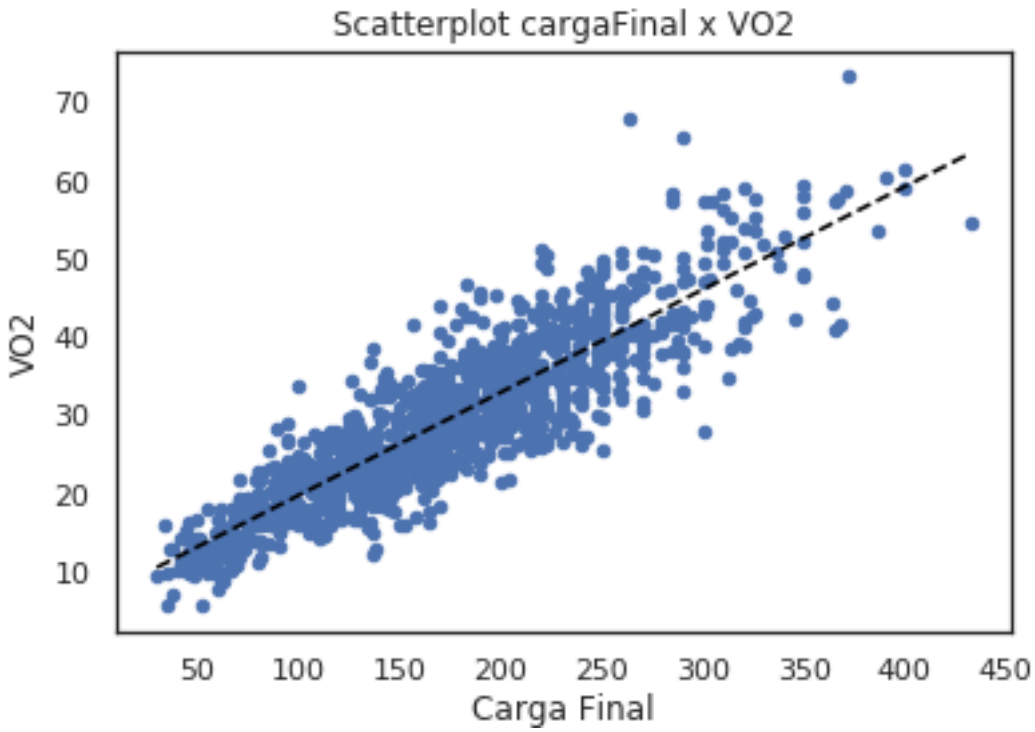


Nesse caso, conforme se observa da figura ao lado, aplicar o modelo de regressão não nos fornece nenhuma informação útil

3-Carga Final x VO2

Nesse último caso, o padrão do scatter plot se assemelha bastante ao de uma reta, ainda mais do que aquele observado para a idade, já que os pontos estão mais próximos da reta obtida pela regressão linear usando mínimos quadrados. Dessa forma, conforme era de se esperar, o coeficiente de correlação é bem perto de um: 0.8783256094059618. Sendo assim, das 3 variáveis comparadas, a carga final é aquela que nos permite inferir mais precisamente a respeito do VO2 máximo de um paciente.

Quanto aos parâmetros obtidos pela regressão, obtivemos os seguintes valores: 0.13153934 para o coeficiente angular e 6.73424783 para o linear.



Inferência Baysiana

Como a variável com maior coeficiente de correlação é a carga final, será ela que usaremos nesse item para efetuar a inferência Baysiana.

Primeiramente vamos dividir a carga final em 20 intervalos iguais. Escolheu-se esse valor pois no histograma obtido no 1 item tínhamos 16 bins e queremos um múltiplo de 4 mais próximo de forma a obter intervalos bem definidos ao invés de dízimas periódicas. Aqui a probabilidade_do_intervalo será a probabilidade de um certo individuo possuir carga final dentro de um intervalo. Como, não sabemos o valor exato de antemão, contaremos a frequência dos valores que se situam nesse intervalo para obter a pmf da probabilidade_do_intervalo de forma empírica, dividindo depois a frequência pela quantidade total de amostras.

Para obter a condicional, isto é, a probabilidade da VO2 ser menor do que 35 dado que a carga final está naquele intervalo contaremos quantos indivíduos se enquadram, ao mesmo tempo, nesses dois critérios. Então dividiremos pelo espaço amostral, isto é, a frequência de indivíduos que possui a carga final nesse intervalo. Em outras palavras:

Seja A o conjunto de indivíduos que possuem VO2 < 35

Seja B o conjunto de indivíduos que possuem carga final nesse intervalo

Então $P(A \mid B) = P(A,B) / P(B)$

Mas P(A,B)/P(B) = Numero total de indivíduos que estão na interseção de A e B / Numero de indivíduos do conjunto B.

Além disso, o numerador será P(A|B)*P(B). Para achar o P(A), isto é, P(VO2<35), aplicaremos o Teorema da Probabilidade Total, onde P(A)= Somatório de P(A|B)*P(B), para todo B. Portanto, para calcular a posterior, basta dividirmos o numerador de Bayes por P(A)

Podemos aplicar o mesmo método analogamente para VO2 >= 35.

```
hipotese VO2 max < 35
```

	Bayes Num	Posterior	hipotese	likelihood(<35)	prior
0	0.020478	0.028369	(30.0, 50.1)	1.000000	0.020478
1	0.051195	0.070922	(50.1, 70.2)	1.000000	0.051195
2	0.057167	0.079196	(70.2, 90.3)	1.000000	0.057167
3	0.075085	0.104019	(90.3, 110.4)	1.000000	0.075085
4	0.104949	0.145390	(110.4, 130.5)	1.000000	0.104949
5	0.100683	0.139480	(130.5, 150.6)	0.975207	0.103242
6	0.092150	0.127660	(150.6, 170.7)	0.907563	0.101536
7	0.085324	0.118203	(170.7, 190.8)	0.793651	0.107509
8	0.064846	0.089835	(190.8, 210.9)	0.666667	0.097270
9	0.042662	0.059102	(210.9, 231.0)	0.500000	0.085324
10	0.017918	0.024823	(231.0, 251.1)	0.225806	0.079352
11	0.005973	0.008274	(251.1, 271.2)	0.148936	0.040102
12	0.001706	0.002364	(271.2, 291.3)	0.068966	0.024744
13	0.000853	0.001182	(291.3, 311.4)	0.050000	0.017065
14	0.000853	0.001182	(311.4, 331.5)	0.052632	0.016212
15	0.000000	0.000000	(331.5, 351.6)	0.000000	0.008532
16	0.000000	0.000000	(351.6, 371.7)	0.000000	0.005119
17	0.000000	0.000000	(371.7, 391.8)	0.000000	0.002560
18	0.000000	0.000000	(391.8, 411.9)	0.000000	0.001706
19	0.000000	0.000000	(411.9, 432.0)	0.000000	0.000853

```
0.7218430034129693
```

hipotese VO2 max >= 35

	Bayes Num	Posterior	hipotese	likelihood(>=35)	prior
0	0.000000	0.000000	(30.0, 50.1)	0.000000	0.020478
1	0.000000	0.000000	(50.1, 70.2)	0.000000	0.051195
2	0.000000	0.000000	(70.2, 90.3)	0.000000	0.057167
3	0.000000	0.000000	(90.3, 110.4)	0.000000	0.075085
4	0.000000	0.000000	(110.4, 130.5)	0.000000	0.104949
5	0.002560	0.009202	(130.5, 150.6)	0.024793	0.103242
6	0.009386	0.033742	(150.6, 170.7)	0.092437	0.101536
7	0.022184	0.079755	(170.7, 190.8)	0.206349	0.107509
8	0.032423	0.116564	(190.8, 210.9)	0.333333	0.097270
9	0.042662	0.153374	(210.9, 231.0)	0.500000	0.085324
10	0.061433	0.220859	(231.0, 251.1)	0.774194	0.079352
11	0.034130	0.122699	(251.1, 271.2)	0.851064	0.040102
12	0.023038	0.082822	(271.2, 291.3)	0.931034	0.024744
13	0.016212	0.058282	(291.3, 311.4)	0.950000	0.017065
14	0.015358	0.055215	(311.4, 331.5)	0.947368	0.016212
15	0.008532	0.030675	(331.5, 351.6)	1.000000	0.008532
16	0.005119	0.018405	(351.6, 371.7)	1.000000	0.005119
17	0.002560	0.009202	(371.7, 391.8)	1.000000	0.002560
18	0.001706	0.006135	(391.8, 411.9)	1.000000	0.001706
19	0.000853	0.003067	(411.9, 432.0)	1.000000	0.000853

0.27815699658703075

	Bayes Num1	Posterior 1	hipotese	lk(<35)	lk(>=35)	predict	prior
0	0.000000	0.028369	(30.0, 50.1)	1.000000	0.000000	0.000000	0.020478
1	0.000000	0.070922	(50.1, 70.2)	1.000000	0.000000	0.000000	0.051195
2	0.000000	0.079196	(70.2, 90.3)	1.000000	0.000000	0.000000	0.057167
3	0.000000	0.104019	(90.3, 110.4)	1.000000	0.000000	0.000000	0.075085
4	0.000000	0.145390	(110.4, 130.5)	1.000000	0.000000	0.000000	0.104949
5	0.002560	0.139480	(130.5, 150.6)	0.975207	0.024793	0.003458	0.103242
6	0.009386	0.127660	(150.6, 170.7)	0.907563	0.092437	0.011800	0.101536
7	0.022184	0.118203	(170.7, 190.8)	0.793651	0.206349	0.024391	0.107509
8	0.032423	0.089835	(190.8, 210.9)	0.666667	0.333333	0.029945	0.097270
9	0.042662	0.059102	(210.9, 231.0)	0.500000	0.500000	0.029551	0.085324
10	0.061433	0.024823	(231.0, 251.1)	0.225806	0.774194	0.019218	0.079352
11	0.034130	0.008274	(251.1, 271.2)	0.148936	0.851064	0.007042	0.040102
12	0.023038	0.002364	(271.2, 291.3)	0.068966	0.931034	0.002201	0.024744
13	0.016212	0.001182	(291.3, 311.4)	0.050000	0.950000	0.001123	0.017065
14	0.015358	0.001182	(311.4, 331.5)	0.052632	0.947368	0.001120	0.016212
15	0.008532	0.000000	(331.5, 351.6)	0.000000	1.000000	0.000000	0.008532
16	0.005119	0.000000	(351.6, 371.7)	0.000000	1.000000	0.000000	0.005119
17	0.002560	0.000000	(371.7, 391.8)	0.000000	1.000000	0.000000	0.002560
18	0.001706	0.000000	(391.8, 411.9)	0.000000	1.000000	0.000000	0.001706
19	0.000853	0.000000	(411.9, 432.0)	0.000000	1.000000	0.000000	0.000853

Bayes numerator total: 0.7218430034129693
Total predict:0.1298487184296515

Já para a previsão fizemos:
Pegamos a tabela obtida para VO2 < 35. Então multiplicou-se a posterior para VO2<35 pela likelihood da VO2 > =35. Por fim, somamos o resultado dessa multiplicação para cada um dos intervalos de carga final e assim obtivemos a previsão.

https://github.com/FelipeSchreiber/Estat-stica-TrabalhoFinal/tree/master/TrabalhoFinal

[Fontes]
https://medium.com/@rrfd/what-is-maximum-likelihood-estimation-examples-in-python-791153818030
https://stat-d.si/mz/mz11.1/Nwobi2014.pdf
https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2927&context=etd
https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm
http://www.iesc.ufrj.br/cursos/regressao/aula_24/3%20Diagnostico%20Regressao%202009_alunos.pdf
http://www.portalaction.com.br/inferencia/62-teste-de-kolmogorov-smirnov
http://www.scielo.br/pdf/jbpneu/v41n5/pt_1806-3713-jbpneu-41-05-00485.pdf