Y = variavel resposta, target, dependente, output  
X = variável preditora, independente, input

Nos métodos de classificação, que no *machine learning* fazem parte dos métodos de análise supervisionada, a variável resposta/*target/label*, ou por você conhecida sobre outra nomenclatura, em geral é uma variável categórica de **nível nominal**.

Citando alguns exemplos, como nos problemas clássicos, na base de dados **Íris** [link](https://www.kaggle.com/uciml/iris) (possui 3 categorias - Species: S*etosa, Versicolor, Virginica*), no **Titanic**[link](https://www.kaggle.com/c/titanic/data) (possui 2 categorias - Survived: *yes, no*), no **Sonar** - [link](https://www.kaggle.com/ypzhangsam/sonaralldata) ( possui 2 categorias - Class: *Rock, Mine*). É importante deixar claro que as categorias também podem ser de **nível ordinal e sem limites de quantidades.**

Antes de continuar, gostaria de relembrar os níveis de mensuração que existem sobre os dados:

1. Escala Nominal: é o nível mais simples e limitado de representação dos dados, onde as observações são divididas em categorias não hierarquizadas segundo um ou mais de seus atributos. Os números atribuídos às categorias servem apenas para sua identificação. Com escala nominal a única operação possível é a contagem. Exemplo: Sexo, tipo sanguíneo, cor.
2. Escala Ordinal: em geral, os valores de uma escala ordinal podem ser expressos por postos, que são números indicando uma ordenação segundo um processo de comparação em relação a determinada característica. Assim, por exemplo respostas a um tratamento médico (fraco, regular, bom, ótimo), ordenamento de Opiniões, Percepções, Preferências, Atitudes, Nível de instrução, Classe social, Estado patológico, podem ser representados através de uma escala ordinal.
3. Escala Intervalar: neste tipo de escala os dados são de forma quantitativa com a mesma unidade de mensuração, e permite o posicionamento de valores em relação a um ponto arbitrário. Nesse nível, é possível fazer comparações de diferenças entre quaisquer duas medições, mas não podemos concluir quanto à magnitude absoluta das medições. Por xemplo, as temperaturas medidas em graus Centígrados e Fahrenheit (o ponto zero é diferente nas duas escalas) são exemplos de escalas de intervalos.
4. Escala Razão: possui as mesmas propriedades da escala intervalar além de informar a magnitude absoluta das medições. Nessa escala, a quantificação é produzida a partir da identificação de um ponto zero fixo e absoluto como origem. Permite, também, a comparação de valores em termos absolutos. Exemplos: peso, altura, renda, volume de vendas.

Voltando ao papo anterior e supondo que estamos em um problema de classificação, uma das etapas comumente realizada é aplicação do cálculo de correlação entre as variáveis preditoras/independentes/*features*, entre elas mesmas e também com a variável resposta/*label*.

***Aqui é iniciado a provocação juntamente com a ideia de escrever esse texto.***

***"É correto utilizar os métodos tradicionais de correlação quando nossa variável resposta é de nível nominal?"***

Existem 3 (três) métodos de correlações mais conhecidos e frequentemente utilizados, que são a **correlação de Pearson**, **correlação de Spearman** e **correlação de Kendall**. Cada uma delas possuem suas particularidades e melhores casos de utlização. Mas também, elas possuem algo em comum: ***NÃO PODEM SER UTILIZADAS COM VARIÁVEIS CATEGÓRICAS A NÍVEL NOMINAL.***

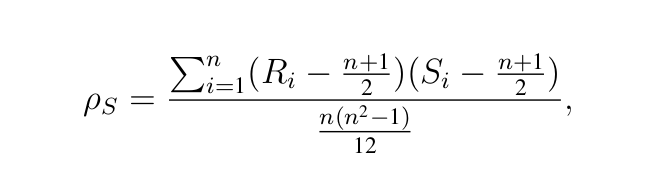
Descrevendo cada uma delas, temos que:

1. Correlação de Pearson: exige que as duas variáveis sejam **quantitativas** de nível intervalar ou razão, é obtido como resultado da seguinte expressão

Não foi fornecido texto alternativo para esta imagem

onde (x\_i , y\_i) são pares de n observações das variáveis x e y com médias x̄ e ȳ, respectivamente. O valor numérico de ρ varia de −1 a +1. Quando ρ estiver próximo de −1, indica uma correlação negativa forte e, próximo de +1, uma correlação positiva forte e um valor de ρ próximo a 0 indica que não há correlação ou pode implicar uma relação não linear entre as variáveis.

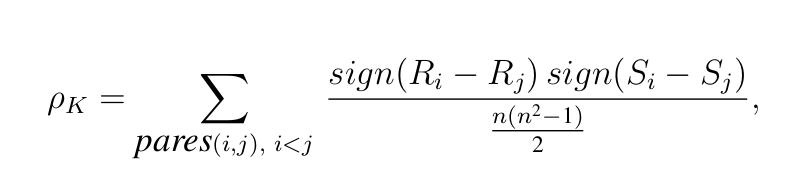
2. Correlação de Spearman (ρ de Spearman): é uma alternativa não-paramétrica para o coeficiente de correlação de Pearson quando este não puder ser usado, mede o grau de associação (ou o grau de relação monotônica) entre duas variáveis de no mínimo, em escala **ordinal**. A correlação é obtida através da equação a seguir:



onde R\_i e S\_i são os postos das variáveis e n é o número de observações. O coeficiente de Spearman também varia de −1 a +1. Uma das vantagens dessa correlação é que ela não é sensível a *outliers*.

OBS.: Um posto (rank) é um número atribuído a um item individual da amostra segundo sua ordem nos itens ordenados. A ordenação dos itens é feita segundo algum critério, como do menor para o maior ou do melhor para o pior.

3. Correlação de Kendall (τ de Kendall): exige que escala de mensuração seja no mínimo **ordinal** de modo que se possa atribuir postos às observações.



onde "sign( )"significa o sinal da diferença entre parênteses, isto é, sign(a − b) = +1 se a > b; ou −1 se a < b.

O τ de Kendall e o ρ de Spearman servem propósitos muito semelhantes. De maneira geral, o τ de Kendall tem níveis de significância mais confiáveis em amostras pequenas do que o ρ de Spearman, mas o valor absoluto do τ de Kendall tende a ser um pouco menor do que o do ρ de Spearman.

**Diante da descrição das correlações acima e do questionamento do texto, sabendo que a variável resposta/*label*é geralmente uma variável categória de nível nominal, temos que diante das métricas mensionadas anteriormente não se pode fazer um estudo de correlação.**

Pela própria definição, uma variável categórica nominal é caracterizada pelo nome da categoria à qual pertence, por exemplo, marca de cerveja preferida, status civil (solteiro, casado, divorciado etc), tem carro ou não, etc. Variáveis desse tipo não podem ser quantificadas por números.

Embora as variáveis categóricas não sejam quantificáveis numericamente, é muito comum representá-las por códigos numéricos, do tipo 0 = homem e 1 = mulher; ou 1 = solteira, sem filhos, 2 = casada, sem filhos, 3 = casada, com filhos, 4 = mãe solteira, 5 = divorciada, sem filhos, 6 = divorciada, com filhos, etc, mas os códigos numéricos são apenas uma maneira abreviada de se referir às categorias e não têm qualquer sentido considerá-los como valores em um estudo de correlação.

Qualquer tentativa de correlacionar valores numéricos verdadeiros, que representem quantidades medidas ou observadas sobre variáveis, com códigos numéricos, usados apenas para rotular variáveis categóricas, levará a análises de correlação completamente sem sentido.

Dessa forma serão indicados algumas soluções abaixo para uma obtenção de resultados corretos sobre seus dados.

**Solução 1 :**

**O que se costuma fazer quando uma das duas variáveis é categórica do tipo nominal, e a outra podendo ser de qualquer nível de mensuração, é aplicar um teste de associação chamado de teste do χ2 (Teste Qui-Quadrado).**

O teste Qui-Quadrado é um teste de hipóteses, no qual possui como hipótese nula uma não associação entre duas variáveis, isto é, que as variáreis são independentes entre si. Para isso é necessário que as categorias de cada variável sejam mutuamente excludentes, ou seja, que cada observação só pertença a uma categoria.

Mesmo assim, para aplicação desse teste, é necessário satifazer algumas pressuposições, como: (i) escala de mensuração é no mínimo nominal, (ii) as frequências esperadas na tabela de contingência não sejam inferiores a cinco.

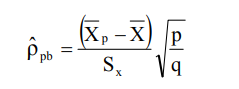
**Solução 2 (ATENÇÃO: apenas quando a variável resposta/*label* é dicotômica/binária):**

Neste caso especial, é possível calcular o coeficiente de correlação entre duas variáveis, sendo uma **numérica** e a outra categórica **nominal**, isso é possível quando a variável categórica tem apenas **dois** valores que cobrem todos os casos possíveis, por exemplo, *masculino/feminino,*Aprovado/Negado, *Sim/Não*. Uma variável desse tipo é chamada de dicotômica (por assumir apenas dois valores).

A correlação que pode ser utilizada neste caso é a

* Correlação Ponto-Bisserial: o coeficiente de correlação ponto bisserial é derivado do coeficiente de correlação linear de Pearson. Como já mencionado, esse método é indicado quando uma das variáveis é dicotômica e a outra contínua.  Ele é interpretado da mesma forma que ρ de Pearson. O estimador do coeficiente de correlação ponto bisserial foi obtido a partir do estimador do coeficiente de correlação linear de Pearson, e é obtido da seguinte forma:

Sejam X e Y duas variáveis, tal que, X é variável numérica e Y é a variável dicotômica, assim temos



em que x̄\_p é a média dos valores de X (quando Y=1), x̄ é a média total da amostra, S\_x é o desvio padrão total da amostra, p é a proporção de casos do grupo superior (quando Y=1), q é a proporção de casos do grupo inferior (quando Y=0).

Existem outros coeficientes de correlação, mas não serão abordados nesse texto, entre eles, temos:

1. Coeficiente de correlação bisserial: usado quando deseja-se calcular uma correlação entre uma variável contínua que obedece a uma distribuição normal e que é artificialmente **dicotomizada** e uma outra variável numérica.
2. Coeficiente de correlação *phi*: usado nas situações, em que as variáveis são medidas em nível **nominal**, sendo as duas variáveis **dicotômicas**. Nesses casos, não é possível a utilização de *nenhum dos métodos vistos anteriormente*. O estimador do coeficiente de correlação *phi* também foi obtido a partir do estimador do coeficiente linear de Pearson.
3. Coeficiente de correlação entre variáveis dicotômica e ordinal (rank): em seu livro "*Contributions to correlational analysis*", Wherry indica a utilização desse coeficiente quando uma das variáveis é dicotômica e a outra ordinal (rank). O seu estimador também foi obtido a partir do coeficiente de correlação linear de Pearson.

**Conclusão**

Nesse texto, foi feito um questionamento interessante sobre onde é possível utilizar correlações, principalmente no contexto de variável nominal, que geralmente são as existentes nas variáveis resposta dos classificadores e métodos supervisionados.

Dessa forma concluímos que é possível utilizar o coeficiente de correlação "adequado" o tomar os devidos cuidado no momento de calcular a correlação, e que existem propostas e adaptações para variados cenários.

Outro ponto importante também, é que dentre os fatores que afetam o coeficiente linear de Pearson, pode-se citar o tamanho da amostra, principalmente quando é pequeno. Dessa forma, além dos cuidados com os níveis das variáveis, há que se atentar para a questão do tamanho da amostra das variáveis envolvidas na análise.

<https://github.com/yukioandre/Python/blob/master/ANOVA.ipynb>

<https://www.kaggle.com/code/evitaginiyatullina/one-way-anova-comparison/notebook>

They were dataplor, foursquare, ifood, google, (will check). Hubster = consolidate data from multiple delivery services

0.3578283486354659

0.7720700884020208