Atividade – Teste de hipótese

Seja o banco de dados, aluguel_bicicleta.csv, contendo informações sobre vários aluguéis de bicicleta em um determinado período.

O seu trabalho é verificar o seguinte: **o aluguel de bicicletas é dependente dos dias úteis ou não?** Em outras palavras, queremos saber se o número de bicicleta alugadas em dias de trabalho é diferente do número de bicicletas alugadas aos finais de semana.

Você pode realizar um teste de hipótese do estilo "**two-sample t-test**", como vimos na última segunda-feira. Claramente, neste caso os dois grupos de amostra, neste caso, seriam os aluguéis feitos aos finais de semana e os feitos nos dias da semana (observe a coluna '*workingday*').

Como todo t-test, precisamos calcular a média de algum lugar, correto? Observe na tabela, coluna 'count'. Ela indica a quantidade de bicicletas alugadas, a cada hora, em um determinado dia.

Tarefas a serem realizadas:

1) Defina a hipótese nula (H0)
2) Defina a hipótese alternativa (H1)
3) Qual o nível de significância adotado?
4) Calcule as estatísticas para o seu teste
5) Indique a decisão de se aceitar ou rejeitar a hipótese.

Tendo extraídos as duas duas amostras, algumas coisas importantes devem ser pensadas e executadas para se realizar o T-test corretamente.

1) **Será que ambas amostras têm o mesmo número de registros?** Se não tiver, você deve iguar o nro de registros em ambas amostras, eliminando registros da maior amostra.

2) **A variância das duas amostras é igual?** Nó exemplo visto em sala de aula, assumimos que elas eram iguais. Aqui você pode fazer o teste de Levene para verificar isso ! Este teste é facilmente realizável em R ou em Python.

3) **A distribuição do residual entre as duas amostras segue uma distribuição normal?** Para fazer isso basta subtrair uma amostra pela outra (lembre-se que a coluna de interesse e a coluna 'count'), normalizar os dados (calculando-se a diferença entra cada valor e a média e dividindo-se pelo desvio padrão) e verificar o formato se parece com uma distribuição normal. Ha outras formas de se verificar a 'normalidade' dos residuais: visualmente por meio do Q-Q teste, ou quantitativamente por meio do teste de Shapiro-Wilk.

Se estiver animado (o que está abaixo é opcional) você pode realizar uma filtragem ou pré-processamento dos dados para simplificá-lo. Aqui vão algumas dicas de pré-processamento ou preparação dos dados.

a) As colunas 'temp' e 'atemp' são correlacionadas? Se são, você pode eliminar a coluna 'atemp', por exemplo.

b) Existe alguma amostra que possui valores nulos? Se houve, você pode eliminá-las.

Dicas para realização: todas as operações que descrevi aqui podem ser feitas, cada qual, com uma linha de código em R ou mesmo em Python (usando Pandas, Numpy, from scipy import stats). Em Python, se quiser plotar as distribuições, pode usar matplotlib.

Qualquer dúvida, me escreva.

Obrigado

João.