Universidade Federal do Rio de Janeiro
Departamento de Ciência da Computação

Trabalhando com Grande Volume de Dados
MAB102

Aplicações de Estatística

Nome: Patrícia de Andrade Kovaleski
DRE: 113090316

Questão 1: Nova feature para um site

Descrição:

Um portal de notícias implementou um play automático de vídeos em suas páginas, visando manter seus usuários por mais tempo no site. Deseja-se descobrir se vale a pena manter esta feature pois seu custo por visualização é de 0.005 centavos, e cada minuto adicional de navegação gera em média 5 centavos de lucro.

Dados:

- *Populacao_tempo.csv* - dados contendo os tempos de acesso de um mês anterior a implementação da feature.
- *Amostra_tempo.csv* - dados contendo os tempos de acesso do último mês, após a implementação feature.

Hipóteses:

Hipótese Nula: H_0

Pertencem a mesma população, ou seja, a feature não causou mudanças significativas.

Hipótese Alternativa: H_1

Não pertencem a mesma população, ou seja, a feature mudou o estado da população.

Procedimento:

Para determinarmos se vale a pena ou não manter a nova feature no site precisamos primeiro verificar se houve alguma mudança significativa na duração média de cada acesso. Dessa forma, podemos calcular o impacto econômico dessa mudança baseado nos custos de sua manutenção e o lucro gerado pelos minutos a mais de visualização.

Após analisar todos os dados contidos nos respectivos *datasets*, as médias calculadas foram:

População	**Amostra**
1.99830	3.47206

Tabela 1: *Médias calculadas para os dados da população e da amostra*

Agora, usaremos o teste da hipótese para verificar se de fato estamos avaliando uma população diferente da original. Ou seja, é preciso avaliar se os dados coletados sobre os tempos de acesso após a implementação da feature são diferentes por pertencerem a uma nova realidade do problema ou se apenas representam uma das variações da situação original.

Utilizando *alpha* = 0.05, podemos rejeitar a hipótese nula, pois o valor obtido para *Z* é maior que o limite tabelado: 9.06 > 1.96

Sabendo que houve um impacto com a implementação da nova feature, precisamos agora avaliar econômicamente se ela deve ou não ser mantida.

Custo gerado a mais pela visualização da feature: 5.00000 centavos
Lucro gerado pelos minutos a mais de visualização: 7.36881 centavos

Resposta:

Vemos que a nova feature deve ser mantida pois gera 2.36881 centavos de lucro!

Questão 2: Clicks do site

Descrição:

Uma métrica comum em sites de e-commerce é o número de clicks que um usuário efetua durante a navegação. Um grupo de marketing quer fazer uma campanha de um novo produto, entretanto não sabe se deve o apresentar apenas na página ou como um pop-up.

Visando responder esse problema, dois grupos foram selecionados. Para o primeiro grupo, foi apresentado apenas a tela com o produto. Para o segundo, foi apresentado a tela com o *pop-up*. Assim, deseja-se descobrir se faz diferença utilizar o *pop-up* ou não.

Dados:

- *amostra_A_click.csvi* – Dados sobre o click de usuários sem o *pop-up*
- *amostra_B_click.csv* – Dados sobre o click de usuários com o *pop-up*

Hipóteses:

Hipótese Nula: H_0

Ambas as amostras vieram da mesma população, ou seja, o *pop-up* não causou impacto o suficiente nos dados.

Hipótese Alternativa: H_1

As amostras vieram de populações diferentes, ou seja, o *pop-up* influenciou o compartamento da população.

Procedimento:

A primeira coisa a ser analisada neste problema é se realmente faz diferença apresentarmos o marketing em forma de *pop-up* ou não. Para isto, usaremos o Teste do Chi-Quadrado para duas amostras.

Neste teste, construímos uma tabela com os valores observados e esperados para cada grupo e classe da população. Com estes valores, calculamos o valor de X_n^2 , que nos dará a informação necessária para aceitarmos ou rejeitarmos a hipótese nula.

Valores observados:

	Classe A	**Classe B**
Yes	301	387
No	682	621

Tabela 2: *Valores observados para as amostras A e B*

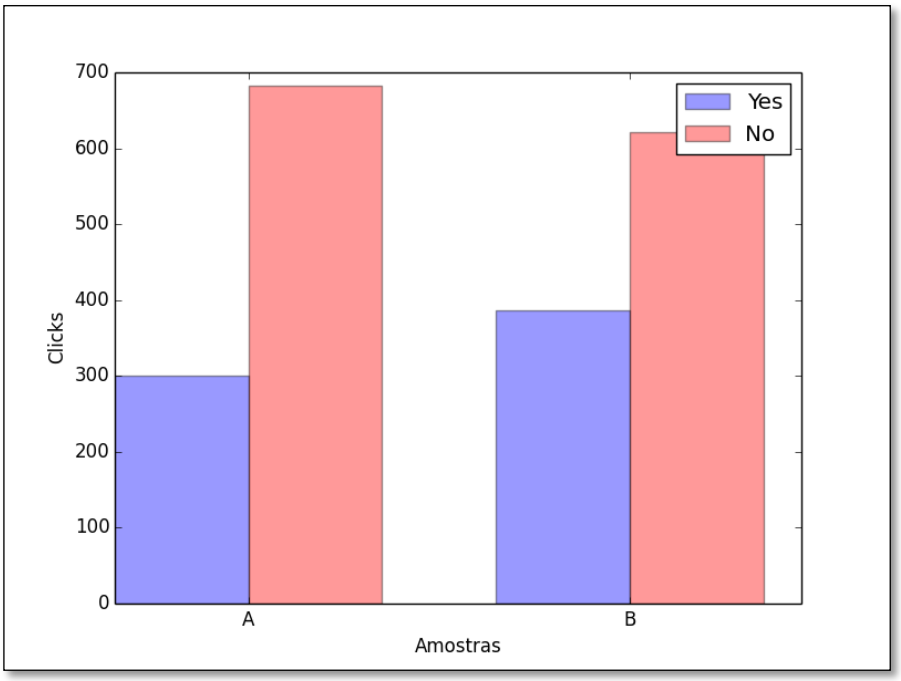

Imagem 1: *Comparação entre os valores observados para as amostras A e B*

Valores esperados:

	Classe A	**Classe B**
Yes	339.68056	348.31944
No	643.31944	659.68056

Tabela 3: *Valores esperados calculados para as amostras A e B*

Com os valores acima, e considerando um *alpha* = 0.05, temos que as amostras não pertencem a mesma população, pois o valor de Z é maior que o valor tabelado: 13.294 >= 3.841. Assim, nós rejeitamos a hipótese nula.

Resposta:

Ao rejeitarmos a hipótese nula nós afirmamos que a utilização do *pop-up* faz diferença no anúncio do marketing em um site. Analisando os valores observados para cada classe, descritos na Tabela 2 e na Imagem 1, vemos que o uso do *pop-up* resultou em um maior número de clicks na publicidade, logo, ele deve ser utilizado pela campanha de marketing.

Questão 3: Produtor de cinema

Descrição:

Um produtor acredita que deve-se construir um poster para um filme utilizando a maior quantidade de atores possíveis. O fato se deve a uma correlação existente entre a nota no IMDB e o número de faces existentes nos posters. Deseja-se verificar a existência dessa correlação, ajudando o produtor a conseguir a melhor nota no IMDB possível.

Dados:

- *movie_metadata.csv* – Dados diversos sobre filmes, incluindo "*imdb_score*" e "*facenumber_in_poster*".

Hipóteses:

Hipótese Nula: H_0

Existe uma correlação entre a nota no IMDB e o número de faces existentes nos posters.

Hipótese Alternativa: H_1

Não existe uma correlação entre a nota no IMDB e o número de faces existentes nos posters.

Procedimento:

Para verificarmos a existência ou não de uma correlação entre os atributos desejados precisamos ser capazes de interpretar os dados do arquivo corretamente. Para isso, devemos realizar um pequeno pré-processamento, tratando os dados de forma que eles se adequem à nossos objetivos.

Neste caso, analisamos apenas as linhas com o mesmo número de células e que não possuam valores nulos para os atributos avaliados: "*imdb_score*" e "*facenumber_in_poster*".

De um total de 5043 registros, foram considerados válidos apenas 4961.

Com os dados em mãos, podemos calcular o coeficiente de correlação de Pearson entre eles:

$$r = 0.06393661.$$

Este valor indica que não há correlação entre a nota no IMDB e o número de faces existentes nos posters. A Imagem 3 abaixo ajuda a visualizar a falta de correlação entre os atributos através da forma como os dados estão distribuídos no gráfico.

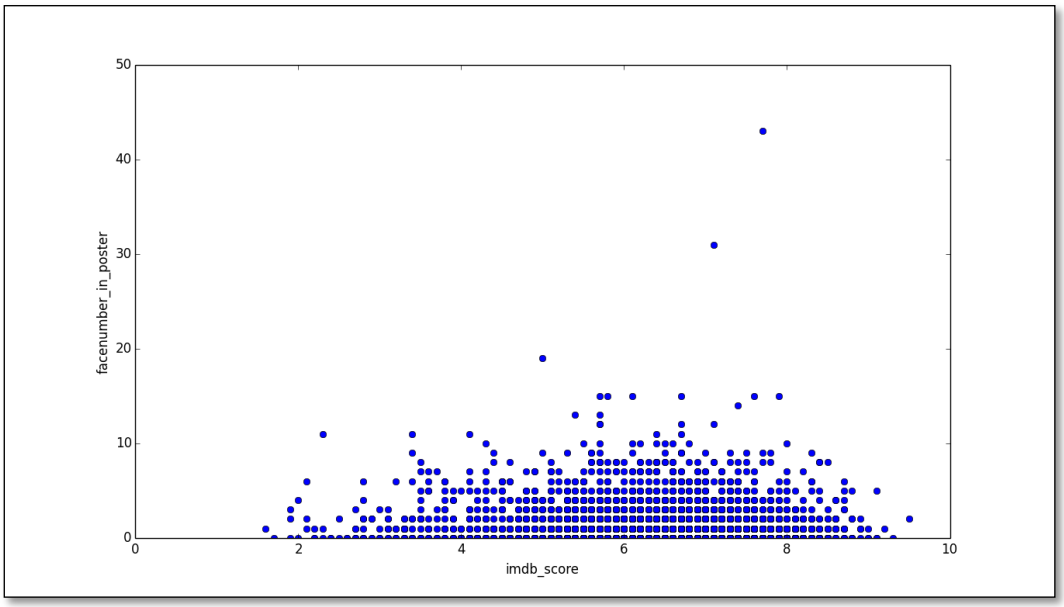

Imagem 2: *Comparação entre os valores de "imdb_score" e "facenumber_in_poster" para cada registro.*

Resposta:

Não há correlação entre a nota no IMDB e o número de faces existentes nos posters. Logo, o produtor precisará pensar em outras formas de obter a melhor nota possível no IMDB.