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: H0

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

Hipótese Alternativa: H1

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 |

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: 9.06091 > 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 -*
* *amostra\_B\_click.csv* -

Hipóteses:

Hipótese Nula: A média de “*clicks”* em ambas as amostras são iguais.

H0: μ = μ0

Hipótese Alternativa: As médias são diferentes.

H0: μ ≠ μ0

Procedimento:

**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 sobre diversos filmes, incluindo “*imdb\_score*“ e “*facenumber\_in\_poster*”.