Análise Inferencial de Dados apresentados pela Picmoney

Com foco na construção e interpretação de intervalos de confiança por meio de μ ou p

Após a análise descritiva de todas as planilhas apresentadas pela PicMoney, realizei o cálculo dos intervalos de confiança para a média populacional (μ) e para a proporção populacional (p), com o objetivo de obter estimativas mais precisas e estatisticamente confiáveis. Essa análise visa fornecer valores de intervalo de confiança com potencial valor estratégico para a empresa, permitindo uma interpretação mais sólida dos dados coletados.

Além disso, foi feita uma comparação entre diferentes níveis de confiança, de forma a observar a variação dos intervalos conforme o grau de incerteza adotado. Dessa forma, o estudo busca oferecer uma visão quantitativa e embasada que auxilie na tomada de decisões estratégicas.

Sendo assim, este documento será estruturado da seguinte maneira:

1. Reutilização dos dados já padronizados e remapeados na primeira entrega do projeto, garantindo a consistência e a continuidade das análises, com uma base de dados sólida, padronizada e coerente para a execução desta etapa.
2. Demonstração visual das ferramentas e padrões de código utilizados durante o processo de análise, assegurando transparência metodológica e uniformidade técnica.
3. Cálculo dos intervalos de confiança com os dados previamente processados, a fim de obter estimativas mais precisas e estatisticamente confiáveis, que possam representar valores estratégicos relevantes para a empresa.

A padronização descrita acima será realizada de maneira sequencial para cada uma das planilhas disponibilizadas pela PicMoney. Para a execução de todos os aspectos mencionados, foi utilizada a linguagem Python, com as seguintes bibliotecas: pandas, para leitura e manipulação dos dados; numpy, para gestão eficiente de arrays e listas; statsmodels.stats.proportion, para cálculo de intervalos de confiança de proporção populacional; e scipy.stats, para cálculo de intervalos de confiança da média populacional, como demonstrado no excerto a seguir:

```
#Importação da planilha que será utilizada e das bibliotecas
import pandas as pd
import numpy as np
from scipy import stats
from statsmodels.stats.proportion import proportion_confint
```

Além dos aspectos listados acima, também destaco que as planilhas foram renomeadas em prol de uma maior legibilidade e compreensão prática. PicMoney-Massa_de_Teste_com_Lojas_e_Valores-10000 linhas(1) foi

renomeado como compras.csv. PicMoney-Base_Simulada_-_Pedestres_Av__Paulista-100000 linhas foi renomeado como pavenue.csv. PicMoney-Base_de_Transa__es_-_Cupons_Capturados-100000 linhas foi renomeado como cupons.csv. PicMoney-Base_Cadastral_de_Players-10_000 linhas foi renomeado como dados_cadastrais.csv.

Por que a preferência por Python em relação ao R?

A escolha da linguagem Python como ferramenta principal para esta análise fundamenta-se em sua capacidade de gerenciar o ciclo de vida completo de um projeto de análise de dados, desde a padronização até a modelagem e visualização. Diferentemente do R, que é puramente estatístico, o Python é uma linguagem mais generalizada que, através de suas bibliotecas, oferece uma solução integrada, eficiente e mais versátil. Além disso, pelo fato de Python estar sendo usado em outras partes da aplicação, acredito que a opção por esta linguagem acaba oferecendo uma proposta mais eficaz, gerando prompts que podem ser utilizados em outras partes do código.

Análise dos dados presentes em cupons.csv:

A planilha cupons.csv apresenta duas variáveis numéricas, são elas valor do cupom, que descreve o valor do cupom utilizado por um determinado usuário do aplicativo, e repasse feito à PicMoney após a utilização desse cupom. Essas duas variáveis serão utilizadas para determinar Intervalos de Confiança para a média populacional (μ). Além das variáveis quantitativas, a planilha apresenta variáveis categóricas, são elas: data, hora, nome do estabelecimento, categoria do estabelecimento e tipo do cupom. Esses dados poderiam ser utilizados para determinar intervalos de confiança, principalmente de proporção, no entanto não serão tópicos abordados, pois não possuem um valor estratégico relevante para a estratégia da empresa. Para todas as análises, utilizaremos um padrão de valor de confiança de 95% ($Z = 1{,}96$).

Em Python, o cálculo dos intervalos de confiança e da média (μ) será feito da seguinte maneira:

```
confidence = 0.95 # Nível de confiança
data_cupon_value = df['valor_cupom'].dropna()
n_value = len(data_cupon_value) # Número total do conjunto
mean_value = data_cupon_value.mean() # Média (x̄) do conjunto
error_value = stats.sem(data_cupon_value) # Definir erro padrão do conjunto
gl_value = n_value - 1 # Grau de Liberdade do conjunto

#Cálculo do Intervallo de confiança
ic_valor = stats.t.interval(confidence,
                            gl_value,
                            loc=mean_value,
                            scale=error_value)
```

Esse padrão para determinação dos intervalos de confiança para a média será utilizado em todos os cálculos subsequentes.

Sendo assim, nossa análise busca responder duas questões fundamentais:

- Qual é a média do valor de cupom utilizado pelos usuários dentro da amostra, considerando um intervalo de confiança de 95%?
- Qual é a média do valor repassado à PicMoney dentro da amostra, considerando um intervalo de confiança de 95%?

Análise do Intervalo de confiança para valor do cupom:

$\boldsymbol{\bar{x} = 550{,}46}$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que o valor do cupom segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [548{,}88; 552{,}10]$$

Análise do Intervalo de confiança para repasse à PicMoney:

$\overline{\boldsymbol{x}} = 70{,}47$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que o valor repassado à PicMoney segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [69{,}91\,;\,71{,}04]$$

Conclusão sobre os intervalos e índices obtidos:

Com 95% de confiança, o valor médio real de todos os cupons usados pelos usuários está entre R$ 548,88 e R$ 552,10. Da mesma forma, o valor médio real repassado à PicMoney por cupom está entre R$ 69,91 e R$ 71,04.

Uma análise mais profunda sugere que o valor de repasse para a PicMoney representa, de forma consistente, aproximadamente 12,8% do valor total do cupom utilizado.

Análise dos dados presentes em compras.csv

A planilha compras.csv apresenta duas variáveis numéricas, são elas: valor_compra, que descreve o valor total da compra efetuada pelo usuário, e valor_cupom, que descreve o valor do cupom associado a essa compra. Essas duas variáveis serão utilizadas para determinar Intervalos de Confiança para a média populacional (μ). Além das variáveis quantitativas, a planilha apresenta variáveis categóricas, são elas: tipo_cupom, tipo_loja, local_captura e nome_loja. Esses dados poderiam ser utilizados para determinar intervalos de confiança, principalmente de proporção; no entanto, não serão tópicos abordados, pois não possuem um valor estratégico relevante para a estratégia da empresa. Para todas as análises, utilizaremos um padrão de valor de confiança de 95% ($Z = 1{,}96$).

Nossa análise busca responder a duas questões fundamentais:

- Qual é a média do valor de compra realizado pelos usuários dentro da amostra, considerando um intervalo de confiança de 95%?
- Qual é a média do valor de cupom recebido pelos usuários dentro da amostra, considerando um intervalo de confiança de 95%?

Análise de Intervalo de confiança para valor da compra:

$\overline{\boldsymbol{x}} = \mathbf{5}49{,}68$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que o valor da compra segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [544{,}57\ ;\ 554{,}80]$$

Análise de Intervalo de confiança para valor do cupom:

$\overline{\boldsymbol{x}} = 70{,}35$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que o valor dos cupons segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [68{,}57\ ;\ 72{,}14]$$

Conclusão sobre os intervalos e índices obtidos:

Com 95% de confiança, o valor médio real de cada compra (o "ticket médio") de todos os usuários está entre R$ 544,57 e R$ 554,80.

Dentro dessas compras, o valor médio real do cupom de desconto utilizado é de algo entre R$ 68,57 e R$ 72,14.

A conclusão estratégica é que os cupons de desconto utilizados representam, em média, cerca de 12,8% do valor total da compra.

Análise dos dados presentes em dados_cadastrais.csv:

A planilha dados_cadastrais.csv apresenta uma variável numérica: é ela idade, que descreve a idade dos usuários cadastrados na plataforma. Essa variável será utilizada para determinar o Intervalo de Confiança para a média populacional (μ). Além da variável quantitativa, a planilha apresenta variáveis categóricas, são elas: sexo, cidade_residencial, bairro_residencial e categoria_frequentada. Esses dados poderiam ser utilizados para determinar intervalos de confiança, principalmente de proporção; no entanto, não serão tópicos abordados, pois não possuem um valor estratégico relevante para a estratégia da empresa. Para todas as análises, utilizaremos um padrão de valor de confiança de 95% ($Z = 1{,}96$).

Nossa análise busca responder a uma questão fundamental:

- Qual é a média de idade dos clientes cadastrados dentro da amostra, considerando um intervalo de confiança de 95%?

Análise de Intervalos de confiança para idade dos usuários:

$\overline{\boldsymbol{x}} = \mathbf{5}2{,}79$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que a faixa etária segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [52{,}37\ ;\ 53{,}22]$$

Conclusão sobre os intervalos e índices obtidos:

Com 95% de confiança, a idade média real de *todos* os clientes cadastrados na plataforma (a população total) está entre 52,37 e 53,22 anos.

A conclusão principal é que a empresa tem uma estimativa extremamente precisa da idade do seu público-alvo. A média da amostra (52,79 anos) é um reflexo muito fiel da realidade, pois a margem de erro é de menos de meio ano (aproximadamente 0,42 anos).

Análise dos dados presentes em pavenue_pedestres.csv:

A planilha pavenue_pedestres.csv apresenta duas variáveis numéricas, são elas: idade, que descreve a idade do pedestre que circula pela Av. Paulista, e ultimo_valor_capturado, que descreve o valor do último cupom utilizado por esse pedestre (caso ele seja um usuário do aplicativo). Essas duas variáveis serão utilizadas para determinar Intervalos de Confiança para a média populacional (μ).

Além das variáveis quantitativas, a planilha apresenta variáveis categóricas de alto valor estratégico, como tipo_celular, sexo e, principalmente, possui_app_picmoney. Diferentemente das outras análises, aqui, será calculado um Intervalo de Confiança para a proporção populacional (p) da variável possui_app_picmoney, visando entender a penetração do aplicativo nesse local. Para todas as análises, utilizaremos um padrão de valor de confiança de 95% ($Z = 1{,}96$).

Nossa análise busca responder a três questões fundamentais:

- Qual é a média de idade dos pedestres dentro da amostra, considerando um intervalo de confiança de 95%?
- Qual é a média do último valor de cupom capturado (entre os que já utilizaram), considerando um intervalo de confiança de 95%?
- Qual é a proporção de pedestres que possuem o app PicMoney instalado dentro da amostra, considerando um intervalo de confiança de 95%?

Análise de Intervalos de confiança para idade dos pedestres:

$\overline{\boldsymbol{x}} = 42{,}97$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que a média etária dos pedestres analisados segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [42{,}87\,;\ 43{,}07]$$

Análise de Intervalos de confiança para último valor de cupom capturado:

$\overline{\boldsymbol{x}} = 254{,}79$

Após aplicação da fórmula $\mathrm{IC}(\mu) = \bar{x} \pm t_{\alpha/2,gl} \cdot \left(\frac{s}{\sqrt{n}}\right)$ em Python, os resultados demonstravam que o valor da compra segue o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [253{,}66\,;\ 255{,}92]$$

Análise de Intervalos de confiança para pedestres que utilizam o app:

Proporção da Amostra (p̂): 0.5996 = 59,96%

Após aplicação da fórmula $\mathrm{IC}(p) = \hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ em Python, os resultados demonstravam que a proporção de pedestres que utilizavam o aplicativo PicMoney segye o seguinte intervalo de confiança seguindo um nível de confiança de 95%:

$$IC = [59{,}65\%\,;\ 60{,}26\%]$$

Conclusão sobre os intervalos e índices obtidos:

Esta análise amostral de pedestres na Avenida Paulista revela três pontos-chave com 95% de confiança:

1. Penetração de Mercado: A penetração do aplicativo PicMoney entre os pedestres é extremamente alta e precisamente medida, situando-se entre 59,65% e 60,26%. Essencialmente, 6 em cada 10 pedestres no local são usuários.
2. Perfil Demográfico (Idade): O público pedestre no local é muito bem definido, com uma idade média real entre 42,87 e 43,07 anos.
3. Valor de Uso: Os usuários existentes que circulam pela avenida são de alto valor, com o valor médio real do último cupom utilizado por eles estando entre R$ 253,66 e R$ 255,92.

A principal conclusão estratégica é que a PicMoney tem uma penetração de mercado considerável (cerca de 60%) em um local de alto tráfego, composta por um público maduro (cerca de 43 anos) que utiliza ativamente cupons de valor significativo (cerca de R$ 254,79)