Além do Teste A|B convencional

Outras maneiras de experimentação

Weslley S. Patrocinio
@wespatrocinio
(LinkedIn | Twitter | GMail | Telegram | Skype)

Quem sou eu?

- BSc. em Física e MSc. em Física Aplicada (USP);
- 2 anos de experiência com Pesquisa em Nanotecnologia (semicondutores);
- 4 anos de experiência em Desenvolvimento de Software;
- Desde 2014 trabalhando com *Machine Learning*, *Data Science* e correlatos;

Atualmente, atuo como **Gerente de Engenharia (Software & *Machine Learning*) na OLX Brasil**, como foco em desafios de Qualidade & Segurança da plataforma.

PONTO IMPORTANTE!

As opiniões compartilhadas nesta apresentação são minhas e **não** representam as opiniões da empresa onde trabalho ;)

Um breve contexto de Desenho de Experimentos

Histórico - Um novo velho mundo

- Assim como outros temas ressuscitados por Big Data, Design de Experimentos não é algo novo:
  - *The Design of Experiments*: Ronald Fisher (1935);
  - *Solutions Manual Design and Analysis of Experiments (1st Ed.)*: Douglas C. Montgomery (1976);
- Em algumas perspectivas, *Design* de Experimentos se mistura com Inferência Estatística. E aí partimos para publicações do século XIX;
- Áreas que, por muito tempo, foram distantes do mundo digital foram precursoras em vários avanços na área: agricultura, psicologia, entre outras;
- Mais recentemente, com a alta volumetria e diversidade de dados disponíveis nas instituições (empresas, universidades, entre outros), experimentação & inferência estatística tornaram-se ferramentas populares e acessíveis;

Alguns conceitos importantes antes de continuar a falar sobre experimentos

EXPERIMENTO

Um experimento ou ensaio é um procedimento planejado para obter novos fatos, negar ou confirmar hipóteses ou resultados obtidos anteriormente.

TRATAMENTO

(ou VARIÁVEL INDEPENDENTE ou FATOR)

Um tratamento é uma condição imposta ou objeto que se deseja medir ou avaliar em um experimento.

UNIDADE EXPERIMENTAL

(ou PARCELA)

Onde é feita a aplicação do tratamento. É a unidade experimental que fornece os dados para serem avaliados.

ERRO EXPERIMENTAL

Em todo experimento, ocorre sempre uma variação ao acaso entre observações de um mesmo tratamento.

REPETIÇÃO

O número de vezes que um tratamento aparece no experimento.

VARIÁVEL RESPOSTA

(ou VARIÁVEL DEPENDENTE)

Uma variável/característica que espera-se apresentar variação por conta do tratamento aplicado.

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Passos para planejar um experimento

1. Conheça o problema;
2. Escolha de fatores e níveis;
3. Escolha da variável resposta;
4. Escolha do *design* do experimento (repetições, variações de tratamento, etc.);
5. Execução e condução do experimento;
6. Análise de dados;
7. Conclusões e recomendações;

Normalmente é um grande erro planejar um único, grande e abrangente experimento para iniciar um estudo

Montgomery, 1991

Por que precisamos ir além do Teste A|B?

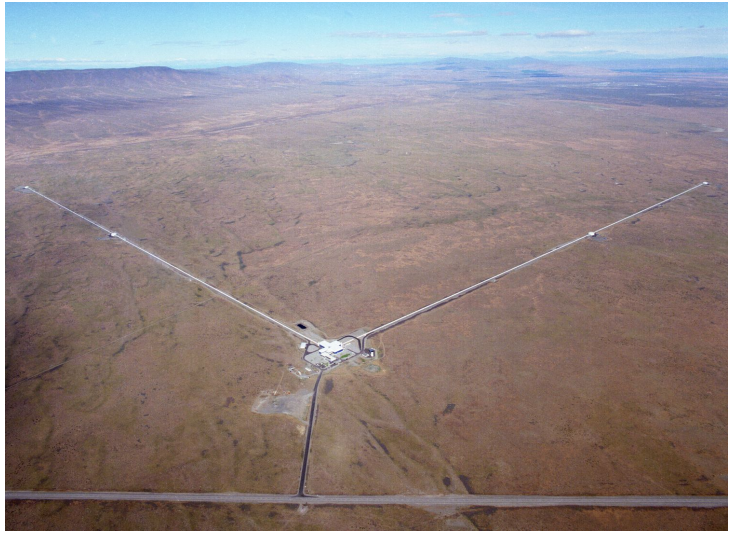
O que estou chamando de Teste A|B convencional?

- Amostragem aleatória de usuários;
- Separação de grupo controle e variante com o novo tratamento;
- Variável resposta binomial;
- Avaliação do experimento por p-valor e intervalo de confiança;

O **experimento** deve ser ajustado ao **problema**, e não o contrário.

LIGO - *Laser Interferometer Gravitational-Wave Observatory*

- Projeto fundado em **1992**;
- Operou **entre 2002 e 2010 sem sucesso** na detecção de ondas gravitacionais;
- Nascimento do *Advanced* LIGO em 2008, que começou a operar em 2015;
- A **detecção** de ondas gravitacionais foi **anunciada em 2016**, resultando no prêmio Nobel de Física de 2017 para os pesquisadores que lideraram a iniciativa;
- **Foram 25 anos, USD 620 milhões, +900 cientistas e +40 instituições envolvidas na iniciativa com um todo**;



Mas quanto eu preciso investir nos meus experimentos?

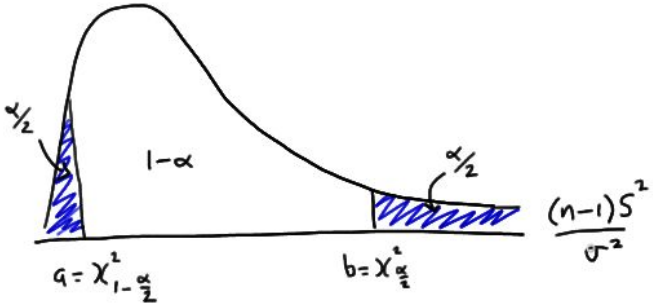
Depende de quão **importante** ele é para você!

Outras abordagens de experimentação

que podem se ajustar melhor aos seus problemas

Testes A|B para métricas não-binomiais

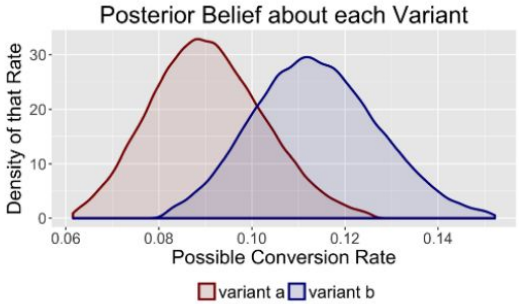
- Alguns tratamentos podem afetar não apenas a ocorrência de um evento, mas também algum atributo dele;
- É possível fazer um experimento onde a significância é calculada a partir da variância da sua métrica contínua de interesse;
    - Exemplo: *conversion rate* VS *ARPU;*
- Com isso, você minimiza riscos causados por *outliers* e afins, que poderiam ser mais impactantes uma análise pós-experimento;



Confidence Intervals for Variances

Testes A|B Bayesianos

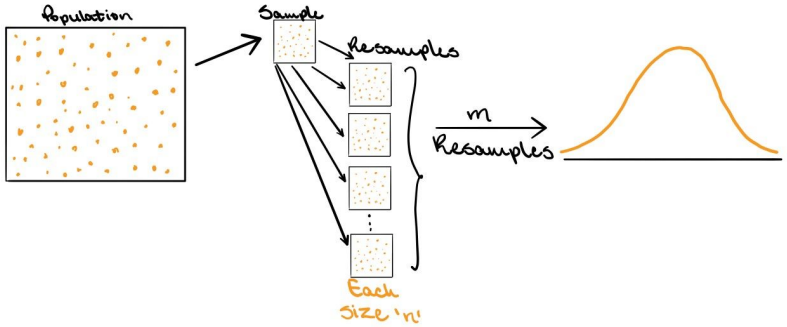
- O método frequentista é "cruel" com testes cuja melhoria na métrica de interesse é pequena e o p-valor/intervalo de confiança é limítrofe, pois favorece a hipótese nula;
    - Quem trabalha com experimentos de conversão que envolve receita diretamente já deve ter sentido essa dificuldade;
- Ao explorar a distribuição de probabilidade de uma variante ser vencedora, podemos medir os eventuais riscos/custos de assumir uma variante como vencedora;
- Neste tipo de teste, usa-se a diferença do *likelihood* das duas distribuições, e as respectivas magnitudes para definir o critério de parada;



The Power of Bayesian A/B Testing

Bootstrapping

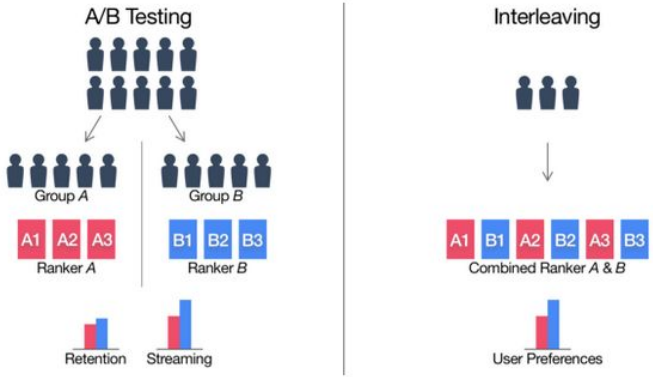
- Em alguns casos, o tamanho de amostra disponível não é suficiente para obter resultados estatisticamente significativos, ou então há suspeita de algum tipo de viés na amostragem;
- Não é necessário, como primeira abordagem, re-executar o experimento. Você pode simular N experimentos a partir do original usando a reamostragem;
- Usando a lei dos grandes números, você pode obter estatísticas de sua população a partir das estatísticas de cada *resampling* realizado;



Bootstrapping Statistics. What it is and why it's used.

Interleaving

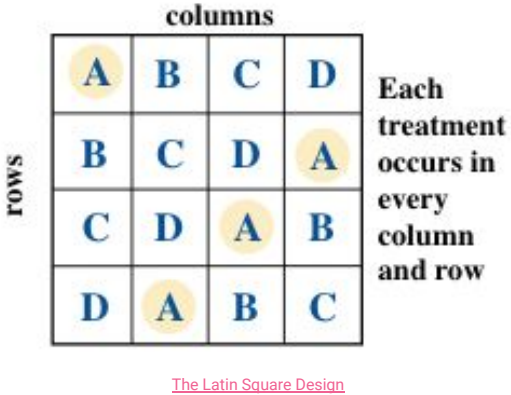
- Em iniciativas de personalização de conteúdo, Testes A|B para medir um impacto de uma nova ordenação tendem as ser custos em tempo e tamanho de população;
  - Principalmente se você deseja testar várias ordenações possíveis;
- Uma maneira alternativa é fazer um *interleaving*, misturando as duas ordenações para uma mesma população - convergindo mais rápido a preferência dos usuários por uma ordenação;
- Após saber a melhor ordenação, você executa um Teste A|B para medir apenas o efeito em assinaturas, vendas e afins;



Innovating Faster on Personalization Algorithms at Netflix Using Interleaving

Quadrados latinos

- Existem situações onde você suspeita de possíveis variações externas às suas amostras, e que isso pode estar afetando a medida do efeito do tratamento;
- Para não ter que executar um mesmo experimento N vezes e levar um tempo longo, você pode usar mão de Quadrados Latinos!
- Resumidamente, você divide seu domínio em diversos blocos, e aloca seus tratamentos aleatoriamente sobre os blocos (garantindo não-repetições em uma mesma linha e coluna);
    - Não se preocupe pois existe estatística específica para isso, já incorporada em diversos *softwares* por aí;



The Latin Square Design

O **experimento** deve ser ajustado ao **problema**, e não o contrário.

Muito obrigado!

Dúvidas?

Weslley S. Patrocinio
@wespatrocinio
(LinkedIn | Twitter | GMail | Telegram | Skype)