P7: Projetar um Teste A/B

Vagner Sanches Vasconcelos
vsvasconcelos@gmail.com
Setembro-2017

RESUMO

Este trabalho apresenta os resultados obtidos da análise do Teste A/B do qual os dados do experimento encontram-se aqui.

1 DESIGN DO EXPERIMENTO

A definição das variáveis pode ser acessada aqui.

1.1 Análise das Variáveis

A variável *Number of cookies* é uma boa métrica Invariante uma vez que os *cookies* são criados antes do aluno clicar no botão *start free trial*, desta forma ela será a mesma para ambos os grupos (Controle e Experiência);

Devido suas características, a variável *Number of userids* não é uma boa métrica Invariante e nem de Avaliação, uma vez que o número de usuários que se inscrevem no teste gratuito depende dos resultados dele e pode ser diferente em grupos de Controle e de Experiência;

Number of clicks é uma boa métrica Invariante porque acontece antes da janela de *Pop-up*, portanto, não vai mudar entre grupos de Controle e Experiência; já para Avaliação não é uma boa escolha devido sua condição imutável;

Assim como *Number of clicks*, a variável *Clickthroughprobability* é uma boa métrica Invariante, uma vez que isso acontece antes do teste, sendo assim independente deste e não vai mudar entre os grupos, característica essa que faz dela não ser uma boa escolha como métrica de Avaliação;

Gross conversion é uma boa métrica de Avaliação porque depende dos resultados do teste e mostra seus efeitos; não é uma boa métrica Invariante devido à sua estrutura dependente.

Retention é uma boa métrica de Avaliação porque é dependente do teste, já para métrica Invariante não é uma boa opção visto que o número de alunos inscritos e pagos são afetados pelos resultados do teste.

Net conversion também é uma boa métrica de Avaliação porque ela muda de acordo com os resultados dos testes e explica se perguntar aos alunos o tempo disponível reduz o número de alunos frustrados que deixaram o teste gratuito ou não; desta forma não é uma boa métrica Invariante.

1.2 Métricas Escolhidas

1.2.1 Métricas Invariantes.

a) *Number of cookie*s (NCo),

b) *Number of clicks (NC)*, e

c) Clickthroughprobability (CTP).

1.2.2 Métricas de Avaliação

a) *Gross conversion (GC)*, e

b) *Net conversion (NtC)*.

1.2.3 Resultados Esperados (hipóteses)

Espera-se que neste experimento o número de alunos inscritos no *Start free trial* diminua, isto é, que a métrica *Gross conversion* caia com significância estatística para o grupo de Experiência, diminuindo custos para a empresa, sem no entanto diminuir o faturamento da companhia, isto é, a métrica *Net conversion* não diminuirá estatisticamente significativa para esse mesmo grupo.

1.3 Calculando o Desvio Padrão Estimado (SE)

Os cálculos foram realizados conforme a Tabela de Valores de Referência - *baseline*, considerando o tamanho da amostra (*sample size*) de 5000 visualizações de páginas (*pageviews*).

Tabela 1: Tabela de valores de referência

Variável	Valor
Unique cookies to view page per day	40000
Unique cookies to click "Start free trial" per day	3200
Enrollments per day	660
Click-through-probability on "Start free trial"	0.08
Probability of enrolling, given click	0.20625
Probability of payment, given enroll	0.53
Probability of payment, given click	0.1093125

1.4 Métricas de Avaliação

Número de *Unique cookies* que clicam no botão *Start free:* (N)

$$N = \frac{5000 * 3200}{40000} = 400$$

Assumindo que ambas as métricas de Avaliação possuem distribuição binomial, o desvio padrão estimado(SE) pode ser calculado pela fórmula:

$SE = \sqrt{\frac{p*(1-p)}{N}}$,na qual p é a probabilidade de ocorrência do evento.

Desta forma, tem-se os valores de SE para ambas as métricas de Avaliação.

Tabela 2: SE das Métricas de Avaliação

	Gross conversion	Net conversion
N	400	400
p	0.20625	0.1093125
SE	0.0202	0.0156

Como ambas as métricas de Avaliação (GC e NC) possuem o mesmo denominador (Nco), a variância pode ser estimada analiticamente.

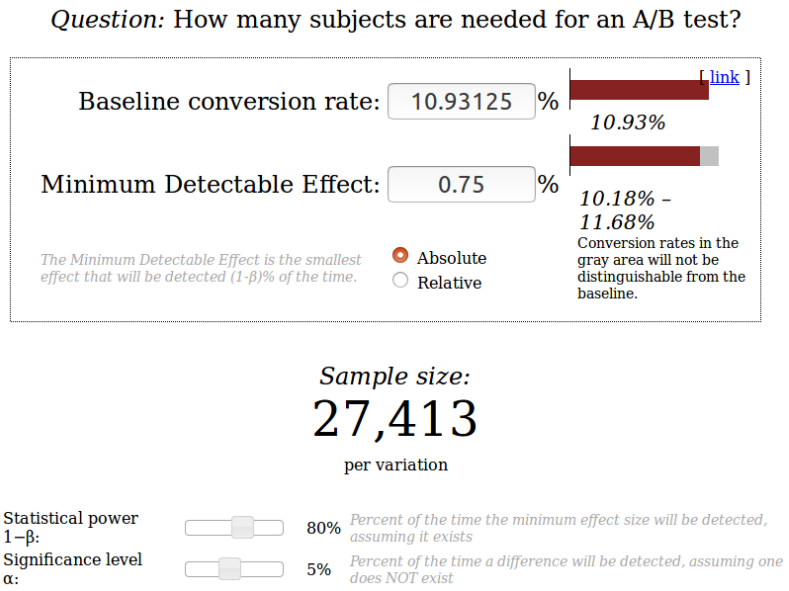
1.5 Amostragem

1.5.1 Número de amostras vs. Power

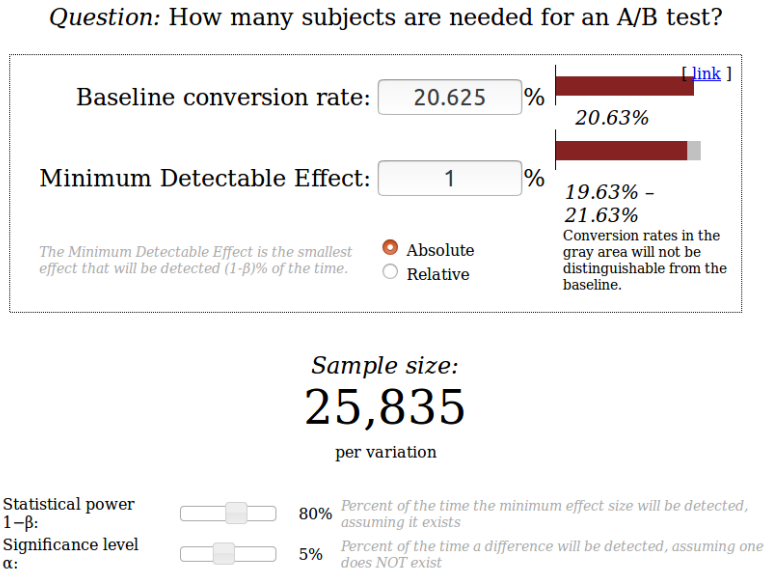
A correlação de *Bonferroni* não foi utilizada nesta análise.

Utilizando o portal Evan's Awesome A/B Tools, mais precisamente a funcionalidade cálculo da amostragem, determinou-se o tamanho da amostra, conforme segue:

- Para *Net Conversation*: 27.413 *pageviews*



- Para *Gross Conversion*: 25.835 *pageviews*



Considerando Click-*through*-probability on Start free trial = 8% (Tab. 1), calculou-se a quantidade de visualizações para a métrica de Avalização que necessita de maior quantidade (*Net Conversation*) e como há dois grupos o valor foi multiplicado por 2, desta forma:

$$Quantidade(pageviwes) = \frac{27413 * 2}{8} = 685325$$

1.5.2 Duração x Exposição

Considerando inicialmente a utilização de todo o tráfego do site (40.000 *cookies*/dia), o experimento duraria:

$$Duração_{experimento} = \frac{685.325}{40.000} \sim 18\, dias$$

Adotando um tempo considerado bom para a realização do experimento em 20 dias, tem-se que visualizar as 685.325 *pageviwers* neste período, assim:

$$pageviwers \;=\; \frac{685.325}{20} \;\sim\; 34.267$$

O que representa:

$$Tráfego \;=\; \frac{34267}{40000} \sim 85.7\,\%$$

Este não é um experimento de alto risco para a empresa, uma vez que: o valor adotado para a fração de 85.7% representa um *trade-off* entre um tempo considerado aceitável para a realização do experimento (20 dias) e uma folga na utilização do tráfego do site; além disto, a mudança proposta durante o teste não é no conteúdo do site e sim no processo de inscrição o que não impactará a experiência do usuário, uma vez que se trata de apenas uma autorreflexão sobre sua disponibilidade de dedicação de tempo ao curso.

2 ANÁLISE DO EXPERIMENTO

2.1 *Sanity Checks*

Os dados para os cálculos estão nesta planilha.

2.1.1 Cálculo para as variáveis *Number of cookie*s (Nco) e *Number of clicks (NC)*.

Com base nos dados da planilha, foram totalizados (somados) os dados dia a dia, para cada variável, obtendo:

Tabela 3: Síntese da planilha de dados

		Controle	Experimento
Pageviews	Total	345543	344660
	Com clicks	17293	17260
Clicks		28378	28325
Enrollments		3785	3423
Payments		2033	1945

Assim, tem-se:

$$SD_{NCO} = \sqrt{\left(\frac{0.5 * 0.5}{344660} + 345543\right)} = 0.0006$$, desvio padrão.

$\hat{p}_{NCO} = \frac{345543}{345543+344660} = 0.5006$, valor observado.

$Lower\,bound_{NCO} = 0.5 - 1{,}96 * 0.0006 = 0.4988$, limite de confiança inferior.

$Uper\,bound_{NCO} = 0.5 + 1{,}96 * 0.0006 = 0.5012$, limite de confiança superior.

Como o valor observado encontra-se dentro do intervalo de confiança, o *Sanity Test* está aprovado para Nco.

Realizando a mesma lógica para *Number of clicks* (NC), tem-se que:

$SD = 0.0021$

$\hat{p} = 0.5005$

$Lower\,bound = 0.4959$

$Uper\,bound = 0.5041$

Como o valor observado encontra-se dentro do intervalo de confiança, o *Sanity Test* está aprovado para NC.

2.1.2 Cálculo para a variável *Clickthroughprobability* (CTP)

Da Tab. 3, a média de *Clicks/Pageviews* para os grupos de Controle e Experiencia são respectivamente:

$p_{Controle} = \frac{28378}{345543} = 0{,}0821$

$p_{Experiência} = \frac{28325}{344660} = 0.0822$

A probabilidade do *Pooled* é:

$p_{Pooled} = \frac{28325+28378}{344660+345543} = 0{,}082154091$

A estimativa do desvio padrão do *Pooled* é:

$SE_{Pooled} = \sqrt{0{,}082154091 * (1 - 0{,}082154091) * (\frac{1}{344660} + \frac{1}{345543})} = 0{,}0006610608$

E os intervalos de confiança (95%) serão:

$Lower\,bound = -1{,}96 * 0{,}0006610608 = -0{,}0013$

$Uper\,bound = 1{,}96 * 0{,}0006610608 = 0{,}0013$

O valor observado será:

$\hat{d} = 0.0821824407 - 0.0821258136 = 0.0001$

Como o valor observado está dentro do intervalo de conferência, a métrica em questão foi aprovada no *Sanity Test.*

3 ANÁLISE DOS RESULTADOS

3.1 Efeitos do *Size Tests*

3.1.1 *Gross conversion* (GC)

$$p_{Controle} = \frac{3785}{17293} = 0.2189$$

$$p_{Experiência} = \frac{3423}{17260} = 0.1983$$

$$\hat{d} = 0.1983198146 - 0.2188746892 = -0.0206$$

$$p_{Pooled} = \frac{3423 + 3785}{17260 + 17293} = 0.2086$$

$$SE_{Pooled} = \sqrt{0.2086070674 * (1 - 0.2086070674) * (\frac{1}{17260} + \frac{1}{17293})} = 0.0044$$

E os intervalos de confiança (95%) serão:

$$Lower\,bound = -0.020554874 - 1.96 * 0.004371675 = -0.0291$$

$$Uper\,bound = -0.020554874 + 1.96 * 0.004371675 = -0.0120$$

O intervalo de confiança não contém o zero, portanto, pode-se confiar que não houve uma mudança, logo GC é estatisticamente significativa. Considerando o limite de significância prático de 0.01 (dmin= 0.01), a variável em questão também passa no teste de significância prática uma vez que ela está fora do intervalo de confiança calculado.

3.1.2 *Net conversio*n (NtC)

$$p_{Controle} = \frac{2033}{17293} = 0.1176$$

$$p_{Experiência} = \frac{1945}{17260} = 0.1127$$

$$\hat{d} = 0.1126882966 - 0.1175620193 = -0.0049$$

$$p_{Pooled} = \frac{1945 + 2033}{17260 + 17293} = 0.1151$$

$$SE_{Pooled} = \sqrt{0.1151274853 * (1 - 0.1151274853) * (\frac{1}{17260} + \frac{1}{17293})} = 0.0034$$

E os intervalos de confiança (95%) serão:

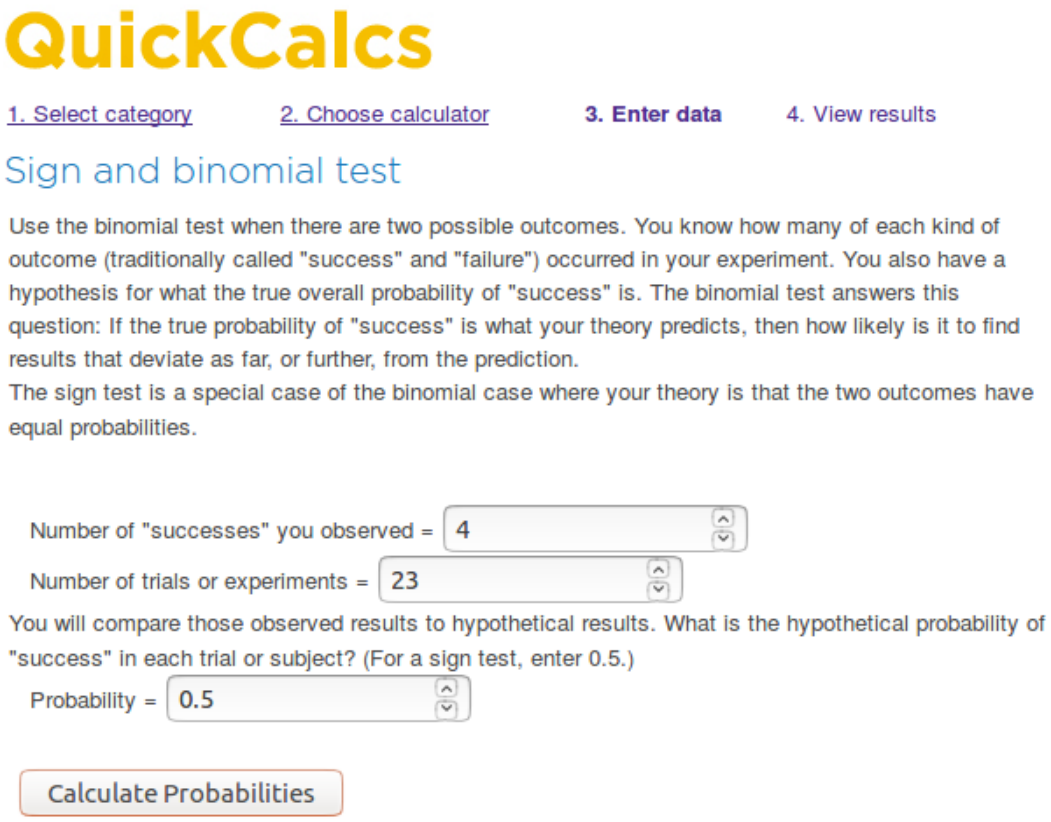
$$Lower\,bound = -0{,}0048737227 - 1.96 * 0.0034341335 = -0.0116$$

$$Uper\,bound = -0{,}0048737227 + 1.96 * 0.0034341335 = 0.0019$$

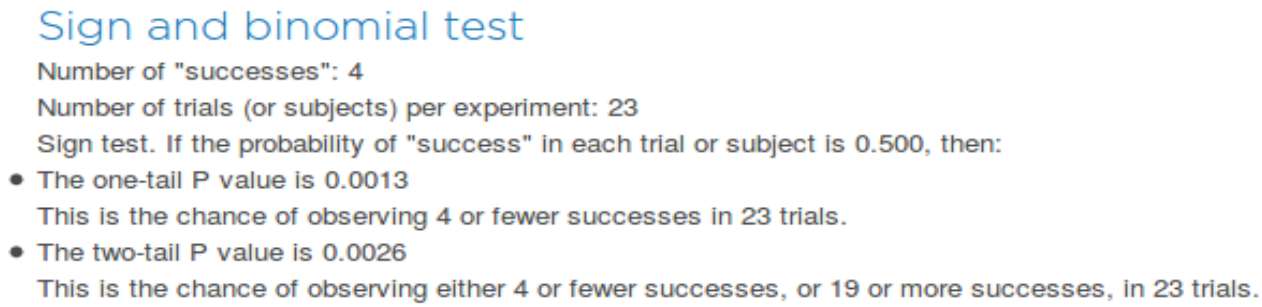
Como o intervalo de confiança contém o zero, NC não é estatisticamente significativa. Considerando o limite de significância prático de 0.0075 (dmin= 0.0075), a variável em questão também não passa no teste prático uma vez que o limite prático encontra-se dentro do intervalo de confiança.

3.2 *Sign Tests*

A realização destes testes foram realizados com auxilio da ferramenta on-line disponível no portal graphpad.com.



3.2.1 Gross conversion



Como o resultado (*p-valor* = 0.0026) é menor que o nível de significância (0.05) conclui-se pela significância estatística.

3.2.2 *Net* Conversion

Sign and binomial test

Number of "successes": 10

Number of trials (or subjects) per experiment: 23

Sign test. If the probability of "success" in each trial or subject is 0.500, then:

- The one-tail P value is 0.3388

  This is the chance of observing 10 or fewer successes in 23 trials.

- The two-tail P value is 0.6776

  This is the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials.

Como o *p-valor* de 0.6776 é maior que o nível de significância (0.05) conclui-se que o resultado é não significativo.

4 SUMÁRIO

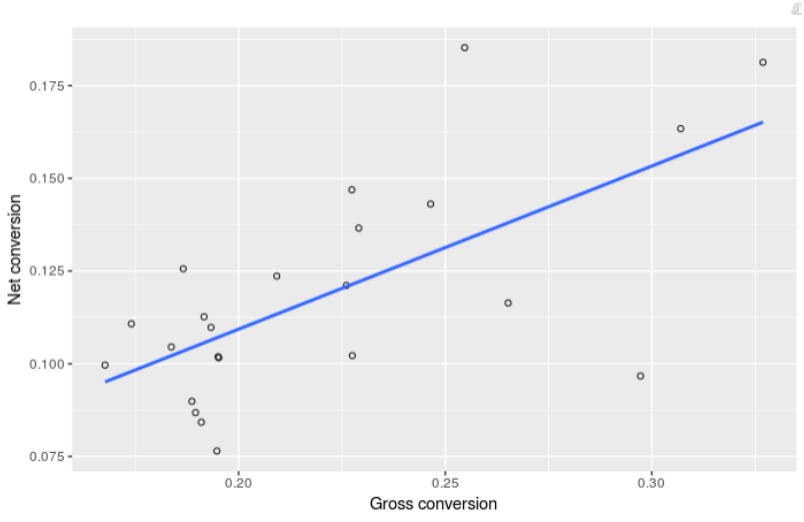
O quadro abaixo apresenta um resumo de todos os testes realizados para cada uma das métricas e os seus resultados.

	Number of cookies	*Number of clicks*	*Clickthrough-probability*	*Gross conversion*	*Net conversion*
Sanity Test	Não Rejeitado	Não Rejeitado	Não Rejeitado	----	----
Significância Estatística	----	----	----	Não Rejeitado	Rejeitado
Significância Prática	----	----	----	Não Rejeitado	Rejeitado
Sign Test	----	----	----	Não Rejeitado	Rejeitado

Quadro Resumo

Não existiu nenhuma discrepância entre os resultados dos testes do Efeito de Tamanho e dos testes de Sinal, ambos apresentaram que *Gross conversion* é estatisticamente significativa, e *Net conversion* não é.

A justificativa pela não utilização da correção de *Bonferroni* é que as métricas de avaliação adotadas são dependentes – positivamente correlacionadas -, conforme observa-se na figura abaixo.



Além disto, a Correção de *Bonferroni*, segundo Field (2009), […] é uma correção simples, porém efetiva, mas tende a ser muito rígida quando muitos testes são executados. Armstrong (2014), completa: "[…] *Whether or not to use the Bonferroni correction depends on the circumstances of the study. It should not be used routinely and should be considered if: (1) a single test of the 'universal null hypothesis' (Ho) that all tests are not significant is required, (2) it is imperative to avoid a type I error, and (3) a large number of tests are carried out without preplanned hypotheses"*.

A tabela abaixo ilustra o efeito do número de hipóteses/variações no Erro Tipo I (Falso Positivo), no caso específico deste projeto temos duas hipóteses, logo, o erro seria de 9,8%, mas com a correção de *Bonferroni* cairia para 2,5%; contudo, como já observado, no nosso caso como as variáveis são dependentes o teste não é necessário.

Tabela 4: Efeito do nº de hipóteses no Erro Tipo I

*N*º hipóteses ou Variações	Probabilidade Erro Tipo I (Falso Positivo)	Correção de *Bonferroni* (α corrigido)	Intervalo de Confiança com Correção de *Bonferroni*
1	5,00%	0,050000	95,00%
2	9,80%	0,025000	97,50%
5	22,60%	0,010000	99,00%
8	33,70%	0,006250	99,40%
10	40,10%	0,005000	99,50%
41	87,80%	0,001220	99,90%

Um efeito colateral da diminuição do Erro Tipo I com a aplicação da Correção de Bonferroni é o aumento do Erro Tipo II (Falso Negativo).

5 RECOMENDAÇÕES

A recomendação é que a alteração não entre em produção.

Conforme os resultados obtidos pelos testes, a primeira hipótese foi comprovada, isto é, com a utilização do *Pop-u*p de alerta o número de estudantes inscritos diminui (*Gross conversion* diminuiu com significância estatística para o grupo de experiencia), o que pode reduzir custos para a empresa; contudo a segunda hipótese não pode ser comprovada estatisticamente, uma vez que a métrica *Net conversion* foi reprovada no teste nos testes, e ainda ela foi um pouco menor para o grupo de experiencia (valor observado = - 0.0049), isto é, se houve significância estatística a receita financeira com o alerta seria menor, o que não é desejado.

6 *FOLLOW-UP* DE UM EXPERIMENTO

O incentivo oferecido pela Udacity nos *NanoDegree* em devolver 50% do valor pago pelo aluno caso este conclua o curso em 1 ano é muito motivador; contudo, se o incentivo fosse de mais curto prazo poderia estimular ainda mais os alunos; por exemplo, ao término de cada parte do curso concluído no prazo programado o aluno ganharia o desconto de 50% na próxima parte do curso; com isto, o tempo médio para conclusão do curso diminuiria.

Desta forma, a hipótese nula do teste é: o incentivo de 50% ao término de cada parte do curso concluído no prazo estipulado não diminuiria o tempo médio de conclusão da respectiva parte do curso.

O experimento consiste em dividir alunos entrantes em dois grupos, de controle e de avaliação, sendo que o grupo de controle só teria direito ao desconto no final do curso, e o grupo de avaliação ao final de cada parte do curso.

Um risco neste experimento é o contato durante o curso – por meio de fóruns, slack, e-mail, etc – de alunos nos dois formatos: desconto por módulo; e desconto só no final do curso; isto, poderia ser encarado como algum tipo de privilégio e causar descontentamento em quem se achar prejudicado. Considerando esse risco, uma boa prática seria limitar o tráfego neste experimento, bem como não realizar outros testes simultaneamente.

O período de avaliação será o dobro do tempo da primeira parte (módulo) do curso. Exemplo: considerando que a primeira parte (módulo) exija 50h (aulas + projeto aprovado) e o curso exija 10h/semana de dedicação – teríamos 5 semanas, logo o experimento aconteceria em 10 semanas, sendo a amostra coletadas nas 5 primeiras semanas.

A nova estratégia poderá entrar em produção se após o término do experimento a métrica Tempo Médio de Conclusão do Módulo (parte do curso) diminuir estatisticamente e praticamente significantes, além da métrica Retention aumentar também estatisticamente e praticamente significantes.

A *unit of diversion* será a identificação do usuário entrante no sistema (*user-id*) – desta forma, pode-se rastrear toda a dinâmica do aluno no curso.

As métricas adotadas são: i) Invariante - número de usuários entrantes (*number of user-ids*); e ii) Avaliação - tempo médio de término da parte (módulo) do curso.

REFERÊNCIAS

[1] A/B Test Udacity

[2] A/B Testing – Niranjan Shetty

[3] When to use the Bonferroni correction

[4] The top 3 mistakes that make your A/B test results invalid

[5] A/B Testing - Sheena Yu