



Estatística para Experimentos Controlados na Web

Testes A/B e Multi Armed Bandits

Ricardo Giglio | giglio@chaordic.com.br

Henrique Grolli | henrique@chaordic.com.br

Apresentação

Oi! :]

O curso tem como objetivo apresentar métodos para a realização de **Experimentos Controlados na Web** e será dividido em quatro partes.



Programa



Aula 1A | 05.08.2014 | 19h: Fundamentos

Apresenta o **conceito** de um teste A/B, e discute algumas motivações e **boas práticas**, sendo principalmente fundamentada na nossa **experiência** com experimentos controlados realizados na Chaordic.

Aula 1B | 05.08.2014 | 21h: Métodos estatísticos I - Paramétricos

Apresenta os métodos estatísticos **tradicionalmente usados** em testes A/B, inclusive pelas **ferramentas padrão** existentes no mercado como a Optimizely e o Google Analytics.

Programa



Aula 2A | 07.08.2014 | 19h:

Métodos estatísticos II - Não Paramétricos:

Discute quando os métodos tradicionais são válidos, e apresenta **alternativas não paramétricas** para os casos onde os métodos tradicionais paramétricos são inadequados.

Aula 2B | 07.08.2014 | 21h:

Making out like a Bandit

Introduz o conceito de **Multi-Armed Bandits**, e o contrasta com os testes A/B. O objetivo é que o participante entenda o trade-off entre 'exploitation' e 'exploration' que define um teste do tipo Bandit.

Objetivos

Entender o que é, para que serve, quando e como usar um
experimento controlado na Web

Ter o ferramental e conhecimento necessário para
implementar um teste A/B na web com o pouco esforço
usando o que já está disponível no seu site

Passará sempre a duvidar da **HIPPO**
(Highest Paid Person Opinion)

Objetivos

Entender como **testes estatísticos tradicionais** são calculados

Aprender a checar se as **suposições** dos testes tradicionais valem para o seu caso específico

Aprender **métodos alternativos** quando os tradicionais forem inadequados

Ser capaz de **diferenciar os conceitos** de teste A/B e de Multi Armed Bandits

Objetivos

Saber derivar o Teorema de Bayes a partir de definições simples de probabilidade

Entender como é construído um Bandit por meio de uma aplicação utilizando a **distribuição Beta** e o procedimento de reamostragem **Thompson Sampling**

Levar para casa receitas **ready-to-plug-and-use** referentes aos pontos trabalhados no curso

Fundamentos

O que é um A/B test?

Boas práticas e a nossa experiência

O que é necessário para realizar um A/B?

Infra e implementação

Estatística

Métodos Paramétricos

Lei dos grandes números

Análise de médias - estatística 't'

Análise de taxas - estatística 'Z'

Intervalos de confiança

Testes de hipótese

Não Paramétricos

A pesada suposição de normalidade

Meus dados não são normais! E agora?

Intervalos de confiança: Bootstrapping

Diferença de médias: Rank-sum

Multi Armed Bandits

O que é e quando usar um Bandit?

O que muda na implementação de um Bandit?

Conceitos básicos de estatística Bayesiana

Distribuição Beta e Thompson Sampling

Fundamentos

Experimentos Controlados

(Charles) Darwin grew the plant for experiments, and he carefully cross-fertilized some flowers and self-fertilized others. When he grew the seeds, he found that the hybrids were bigger and stronger than the purebreds.

He had a clever, and at that time novel, idea. Since slight differences in soil or light or amount of water could affect the growth rates, **he planted the seeds in pairs** – one cross-pollinated seed and one self-pollinated seed in each pot.

Fundamentos

Experimentos Controlados

Existem ao menos
dois grupos
uma métrica de sucesso

Ex: Homeopatia realmente faz efeito?

Medicamento (tratamento) e placebo (controle)

Outros nomes: randomized experiments, split tests,
parallel flights, ...

Fundamentos

Experimentos
Controlados

A/B test
experimento controlado online

A/A test
experimento de controle que serve apenas para
validar a metodologia

Bandit
experimento controlado dinâmico

Fundamentos

Nossa Experiência

Recently, several studies have been pointing out some pitfalls with regard to the evaluation of recommender systems that are based purely on off-line prediction performance. ... it is also evident in literature that off-line evaluations themselves are of limited scope when one deals with real commercial applications, because there is no guarantee that prediction performance necessarily translates into business value.

Case study on the business value impact of personalized recommendations on a large online retailer

<http://dl.acm.org/citation.cfm?id=2366014>

Fundamentos

Boas Práticas

Teste padrão: A/B/C/D

A – alternativa 1 – 25%

B – alternativa 1 – 25%

C – alternativa 2 – 25%

D – alternativa 2 – 25%

Assim existem 4 testes A/B (AC, AD, BC, BD) e 2 testes A/A (AB, CD) simultâneos, possibilitando validar resultados e métodos, além de A+B vs C+D

Fundamentos

Boas Práticas

Ramp-up

Iniciar um experimento expondo apenas uma pequena parcela dos usuários ao tratamento, e então gradualmente aumentar essa porcentagem

Fundamentos

Boas Práticas

One accurate measurement is worth more than a thousand expert opinions - *Admiral Grace Hopper*

Enlightened trial and error outperforms the planning of flawless execution - *David Kelly*

Almost any question can be answered cheaply, quickly and finally, by a test campaign. And that's the way to answer them, not by arguments around a table.

Go to the court of last resort: the buyers of your products.

Claude Hopkins, Scientific Advertising, 1922

Fundamentos

The
HIPPO



Fundamentos

O que testar?

Tamanho, posição e cor de botões **call-to-action**

Quantidade de campos em um formulário

Imagens em páginas de descrição de produto

Ofertas promocionais

Fundamentos

Métricas de Sucesso

Testes funcionam melhor se bem definidos *a priori*

Lembre: Um experimento é composto por:
ao menos duas alternativas
uma métrica de sucesso

Click Through Rate (CTR)

Conversion

Revenue per visitor

Implement.

A implementação possui três pontos:

Divisão dos Grupos

Consistência (manutenção do estado)

Coleta

Implement.

Divisão dos Grupos

A divisão dos grupos deve ser feita antes de tudo.

Basicamente sortear para a qual grupo o usuário pertence.

Podemos fazer o controle da divisão do lado do servidor ou do lado do cliente.

Implement.

Divisão dos Grupos

Divisão do lado do Servidor

Primeiro request do usuário faz com que o servidor atribua um grupo a ele usando o **método de randomização adequado.**

php: openssl_random_pseudo_bytes

node: crypto.pseudoRandomBytes

Java: SecureRandom

Python: os.urandom

Any unix: /dev/urandom

Implement.

Divisão dos
Grupos

Controle do lado do Client

Depende da geração de números aleatórios confiáveis, o ideal é que um serviço no servidor exista para isso. O controle de quando isso vai ser pedido e como isso vai ser usado fica no cliente.

Implement.

Consistência

Após a atribuição de um grupo ao usuário é importante para a validade do teste que o usuário pertença "sempre" ao mesmo grupo.

Session cookie, week cookie ou forever cookie.

Durante uma mesma sessão é quase sempre o suficiente para a maioria dos testes. Isso vai depender do perfil dos usuários e duração do teste.

Implement.

Coleta

Para processarmos o resultado do teste precisamos coletar as informações que geram as métricas e marcá-las com os grupos.

É como se estivéssemos comparando dois sites diferentes.

Implement.

Coleta

O ideal é que a coleta e marcação dos dados seja feita como parte fundamental da coleta dessas informações.

Caso isso não aconteça hoje, há uma solução.

Implement.

SHOW ME THE CODE!!!

https://github.com/chaordic/academy-controlled_experiments/blob/master/example/index.html

Fundamentos

Estatística

Por que não podemos apenas comparar as médias?

Qual a altura média dos integrantes da Chaordic?

Não tem tempo de medir todo mundo? Faça uma amostra e calcule a média R: 160cm

E se, por acaso, sua amostra apenas tiver pessoas com menos de 170cm? Faça outra amostra e calcule a média R:180cm

As médias das duas amostras são diferentes... E agora?

O fator sorte deve ser dimensionado

Fundamentos

Estatística

Hipóteses nula (H_0) e alternativa (H_A)

A H_0 é de que não existe diferença entre as alternativas com relação à métrica de sucesso e que qualquer diferença observada durante o experimento é apenas flutuação aleatória (sorte)

Nível de significância (ex: 95%)

A probabilidade de falhar em rejeitar H_0 quando ela é verdadeira

Intervalo de confiança

Medida de confiança em uma estimativa. Ex: Se o experimento for repetido 100 vezes, em 95 casos a métrica estará dentro do intervalo de confiança

Métodos Paramétricos

Lei dos grandes números

Análise de médias - estatística 't'

Análise de taxas - estatística 'Z'

Intervalos de confiança

Testes de hipótese

Métodos Paramétricos

Da Binomial
à Normal

The Basics

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TheBasics.py

Flipping Coins

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/FlippingCoins.py



Hallgrímskirkja Reykjavík



Testes de hipótese

One-sample vs two-sample

mean of A is higher than 0.5 (one-sample)

mean of A is higher than mean of B (two-sample)

One-tailed vs two-tailed

mean of A is higher than 0.5 (one-tailed, one-sample)

mean of A is equal to 0.5 (two-tailed, one-sample)

mean of A is equal to mean of B (two-tailed, two-sample)

Análise de médias

Estatística 't' - hard mode

Teste 't' para a média de uma amostra

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesisMath.py#L15

Teste 't' para comparar as médias de duas amostras

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesisMath.py#L44

Análise de taxas

Estatística 'Z' - hard mode

Teste 'Z' para uma proporção

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesisMath.py#L79

Teste 'Z' para comparar duas proporções

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesisMath.py#L100

Análise de médias

Estatística 't' - python mode

Teste 't' para a média de uma amostra

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesis.py#L14

Teste 't' para comparar as médias de duas amostras

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesis.py#L52

Análise de taxas

Estatística 'Z' - python mode

Teste 'Z' para uma proporção

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesis.py#L68

Teste 'Z' para comparar duas proporções

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/TestingHypothesis.py#L88

Métodos Paramétricos

Intervalos de confiança

Three-sigma rule

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/ConfidenceIntervals.py#L15

Normal confidence intervals

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/ConfidenceIntervals.py#L32

Quantiles

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/ConfidenceIntervals.py#L43

obrigado!

CHAORDiC
you'll like 



Estatística para Experimentos Controlados na Web

Testes A/B e Multi Armed Bandits



Ricardo Giglio | giglio@chaordic.com.br
Henrique Grolli | henrique@chaordic.com.br

Não Paramétricos

A pesada suposição de normalidade

Meus dados não são normais! E agora?

Intervalos de confiança: Bootstrapping

Diferença de médias: Rank-sum

Não Paramétricos

Suposição de Normalidade

Dados aleatórios não normais

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/NonParametricMethods.py#L20

Teste para a suposição de normalidade

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/NonParametricMethods.py#L28

Não Paramétricos

Bootstrapping

Meus dados não são normais! E agora?

Sorteio com reposição

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/NonParametricMethods.py#L47

Bootstrapping - Intervalos de confiança não
paramétricos

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/NonParametricMethods.py#L44

Não Paramétricos

Rank-sum

Rank-sum (Mann-Whitney U test)

Conceito

A ordem, e não os valores, é que importa

Teste de hipótese não paramétrico para a diferença
de médias

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/NonParametricMethods.py#L72



Multi Armed Bandits

O que é e quando usar um Bandit?

O que muda na implementação de um Bandit?

Conceitos básicos de estatística Bayesiana

Distribuição Beta e Thompson Sampling

Multi Armed Bandits

O quê?

Em um A/B mostramos alternativas aleatoriamente seguindo uma proporção fixa durante todo o experimento

Em um Bandit alternativas que vêm performando melhor durante o experimento são escolhidas mais vezes

Multi Armed Bandits

O quê?

Em um A/B analisamos os resultados após o experimento terminar e decidimos pela alternativa vencedora

Em um Bandit a alternativa vencedora ganha gradativamente mais espaço durante o período de experimento

Multi Armed Bandits

Implement.

Vamos revisitar nosso exemplo de A/B

Multi Armed Bandits

Estatística
Bayesiana

Apêndice

Multi Armed Bandits

Distribuição Beta

Números aleatórios da Distribuição Beta

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/BayesianStatistics.py#L15

Comparação de duas amostras da Beta

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/BayesianStatistics.py#L24

Multi Armed Bandits

Thompson Sampling

Thompson Sampling

https://github.com/chaordic/academy-controlled_experiments/blob/master/stats/ThompsonSampling.py

Para cada alternativa, atribuir um **número aleatório** da distribuição Beta usando como parametros os números de **sucessos e fracassos observados** até o momento.

Mostrar a alternativa com maior número aleatório

obrigado!

CHAORDiC
you'll like 



Apêndice

Estatística Bayesiana

Introduction to probabilities

Bayes's theorem

The cookie problem

Diachronic interpretation

Probability mass functions

Prior distributions

Likelihood functions

Posterior distributions

Apêndice

Introd. to Probabilities

Probability is a number defined in $[0, 1]$ which represents the degree of belief in something

The probability of tails in a fair coin is 0.5

The probability of seeing a seven in a six-sided die is 0

The probability of the sun rising tomorrow is 1

The (regular) **probability** of something (A) is represented as:

$$p(A)$$

Apêndice

Introd. to Probabilities

Conditional probability is a number defined in $[0, 1]$ which represents the degree of belief in something given that other things happened

The **conditional probability** of A given B is represented as:

$$p(A | B)$$

Will I have a heart attack next year?

Every year, 200K brazilians have a heart attack

Population in Brazil is 200M

$$p(\text{heart attack}) = 0.001$$

However, I am not a regular brazilian

My style: I smoke, I drink heavily, and I love bacon

$$p(\text{heart attack} | \text{my style}) > p(\text{heart attack})$$

Apêndice

Introd. to Probabilities

The **conjoint probability** is a number defined in [0, 1], and is a fancy way to say two things are true

The **conjoint probability** of A and B happening is represented as:

$$p(A \text{ and } B)$$

I throw two dice, what is the conjoint probability that the output of both dice are 3?

$$p(A \text{ and } B) = p(A)p(B) = 1/6 * 1/6 = 1/36$$

This is **only true** because A and B are **independent**

What happens to the conjoint probability of A and B if they are not independent?

$$p(A \text{ and } B) = p(A)p(B | A) = p(B)p(A | B)$$

Suppose

$$p(\text{woman}) = 0.5, p(\text{long hair}) = 0.5$$

$$p(\text{long hair} | \text{woman}) = 3/4$$

Then

$$p(\text{long hair and woman}) > p(\text{long hair})p(\text{woman}) = 1/4$$

$$p(\text{long hair and woman}) = p(\text{woman})p(\text{long hair} | \text{woman})$$

$$p(\text{long hair and woman}) = 3/8$$

Apêndice

Introd. to Probabilities

We know that the conjoint probability of two dependent (A and B) events happening is:

$$p(A \text{ and } B) = p(A)p(B | A)$$

We also know that the conjoint probability is commutative

$$p(A \text{ and } B) = p(B \text{ and } A)$$

$$p(A)p(B | A) = p(B)p(A | B)$$

Dividing both sides by, say, $p(A)$ yields

$$p(B | A) = p(B)p(A | B)/p(A)$$

And that's Bayes's theorem =D

Apêndice

The cookie problem

There are 2 bowls of cookies (B1 and B2)

B1 has 30 vanilla (V) and 10 chocolate cookies (C)

B2 has 20 vanilla (V) and 20 chocolate cookies (C)

I choose a bowl at random, and then a cookie at random from it:

It is a vanilla cookie!

Problem:

What is the probability that it came from bowl 1?

We know that:

$$p(V | B1) = 3/4$$

But we don't know

$$p(B1 | V)$$

Apêndice

The cookie problem

Bayes's theorem:

$$p(B1 | V) = p(B1)p(V | B1)/p(V)$$

By parts:

chose one of the two bowls at random

$$p(B1) = 1/2$$

probability of a V given bowl 1

$$p(V | B1) = 30/40 = \frac{3}{4}$$

probability of a V regardless of the bowl

$$p(V) = (20 + 30)/(40 + 40) = 5/8$$

$$p(B1 | V) = p(B1)p(V | B1)/p(V)$$

$$p(B1 | V) = (1/2 * 3/4) / (5/8) = 3/5$$

Given that $p(B1 | V) > p(B2 | V)$, there is more likelihood that the cookie came from bowl 1

Apêndice

Diachronic Interpretation

Diachronic means something happening over time

This interpretation allows one to update his/hers beliefs of an hypothesis (H) given new data (D) observed

The cookie problem revisited

H = cookie came from bowl 1

D = it is a vanilla cookie

$$p(H) = 1/2$$

$$p(H | D) = 3/5$$

Apêndice

Diachronic Interpretation

$$p(H | D) = p(H)p(D | H) / p(D)$$

$p(H | D)$

is the **posterior** information, that is, information we have after we observe new data

$p(H)$

is the **prior** information, that is, information we have before we observe new data

$p(D | H)$

is the **likelihood** of observed data given that hypothesis is true

$p(D)$

is the probability of data under any hypothesis, it acts as a **normalizing constant**

Apêndice

Diachronic Interpretation

Simplifying assumptions:

All the hypotheses $H[i]$, $i = 1, \dots, N$ are

mutually exclusive: only one is true

collective exhaustive: at least one is true

If the set of N hypotheses in a problem obey both rules above, it is called a **suite**

If the set of hypotheses is a suite:

we can use the law of total probability to compute the normalizing constant

$$p(D) = \sum \{n=1 \text{ to } N\} p(H[n])p(D | H[n])$$

In the cookie problem

$$p(D) = p(H[1])*p(D | H[1]) + p(H[2])*p(D | H[2])$$

$$p(D) = 1/2 * 3/4 + 1/2 * 1/2 = 5/8$$

Apêndice

Prob. mass functions

Bayes's theorem revisited

Consider a suite of N hypotheses

$$p(H[i] | D) = p(H[i])p(D | H[i]) / \sum \{n=1 \text{ to } N\} p(H[n])p(D | H[n])$$

Then, the set of $p(H[i] | D)$, $i=1,\dots,N$ gives us the probability mass function of the suite

You can think of a PMF as a Python dict mapping hypotheses (keys) to their corresponding probabilities (values)

Ex: six-sided fair die PMF

$$\text{pmf}[1] = 1/6, \text{pmf}[2] = 1/6, \dots, \text{pmf}[6] = 1/6$$

Apêndice

Prior Distribution

Consider an alternative (ex: red button) which you know nothing a priori with regard to its CTR

Then, any CTR from 0 to 1 is equally likely

$$\text{pmf}[0] = \text{pmf}[0.01] = \dots = \text{pmf}[0.99] = \text{pmf}[1] = 1/101$$

Approximation from a continuous to a discrete process

This is the **prior distribution** represented as a PMF before we observe any data

Apêndice

Likelihood Functions

Here we are only interested in CTR-like
(0 or 1) measures

A Bernoulli trial is one where
a success 's' (1) occurs with probability 'p', and
a failure 'f' (0) occurs with probability '1-p'
no other thing can happen (that is, ' p ' + ' $1 - p$ ' = 1)

If a random variable is drawn ' n ' times from a Bernoulli process, the resulting distribution is Binomial

$$p(s | p = x) = \text{bin_coef}(n, s) * x^s * (1-x)^{n-s}$$

As it can be seen when $n = 1$

$$\text{Bernoulli}(p) = \text{Binomial}(1, p)$$

Apêndice

Likelihood Functions

Short digression - Factorial

ex: $5! = 5 * 4 * 3 * 2 * 1 = 120$

By convention, $0! = 1$

Short digression - Binomial coefficient

$\text{bin_coef}(n, k) = n!/[k!(n-k)!]$

also known as 'nchoosek' (from n choose k)

Binomial-Bernoulli relation revisited

$p(s | p = x) = \text{bin_coef}(n, s) * x^s * (1-x)^{n-s}$

if $n = 1$ and $p = 0.5$

$p(s | p = 0.5) = 1 * 0.5^s * (0.5)^{1-s}$

Think about it as a single throw of a coin

$p(1 | p = 0.5) = 1 * 0.5^1 * (0.5)^{1-1} = 0.5$

Apêndice

Posterior Distributions

To choose a **posterior distribution**, it is convenient to use one that has the chosen likelihood function as its conjugate prior

our likelihood function is binomial, which is a conjugate prior of the **Beta distribution**

Short digression - Beta distribution

two parameters: α and β

$$p(p = x) = [x^{\alpha-1} * (1-x)^{\beta-1}] / B(\alpha, \beta)$$

Where $B(\alpha, \beta)$ is the Beta function given by

$$B(\alpha, \beta) = (\alpha-1)!(\beta-1)! / (\alpha+\beta-1)!$$

Apêndice

Posterior Distributions

To choose a **posterior distribution**, it is convenient to use one that has the chosen likelihood function as its conjugate prior

our likelihood function is binomial, which is a conjugate prior of the **Beta distribution**

Short digression - Beta distribution

two parameters: α and β

$$p(p = x) = [x^{\alpha-1} * (1-x)^{\beta-1}] / B(\alpha, \beta)$$

Where $B(\alpha, \beta)$ is the Beta function given by

$$B(\alpha, \beta) = (\alpha-1)!(\beta-1)! / (\alpha+\beta-1)!$$

Apêndice

Posterior Distributions

$$p(p = x) = [x^{\alpha-1} * (1-x)^{\beta-1}] / B(\alpha, \beta)$$

Illustrations

$$\alpha=1, \beta=1$$

$$p(p = x) = [x^0 * (1-x)^0] / B(1,1)$$

$$p(p = x) = 1 / 1 = 1$$

that is, an uniform distribution

$$\alpha=2, \beta=2$$

$$p(p = x) = [x^1 * (1-x)^1] / B(2,2)$$

$$p(p = x) = x * (1-x)$$

$$\alpha=2, \beta=5$$

$$p(p = x) = [x^1 * (1-x)^4] / B(2,5)$$

$$p(p = x) = x * (1-x)^4$$