Date: Tuesday, February 5, 2019

Bayesian A/B Testing

* Classic A/B Testing (aka Bucket testing)

Setup:
- Two experimental treatments A & B
- Collection of randomly sampled individuals
- n_A are exposed to treatment A
- n_B are exposed to treatment B

Original example: Drug efficacy.
(Inspired by clinical trials)

Treatment A is say a blood pressure medicine and treatment B is a placebo.

QUESTION: Is the blood pressure medicine working?

↳ Measurement: Blood Pressure of every individual X_j on day 31.

- Let μ_A = True mean blood pressure of group A
  μ_B = " " B

Test the efficacy of the drug by investigating the hypothesis test:

$H_0: \mu_A = \mu_B$ (It didn't work)

↳ Recast just a bit: $\delta = \mu_B - \mu_A$

↳ If the drug is "useful", then we expect $\delta > 0$.

* Frequentist approach:

Take $\bar{X}_A = \frac{1}{n_A} \sum_{j \in A} X_j$

$\bar{X}_B = \frac{1}{n_B} \sum_{j \in B} X_j$

Run a t-test: $T = \dfrac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{1}{n_A}\sum_{j\in A}(X_j - \bar{X}_A)^2 + \frac{1}{n_B}\sum_{j\in B}(X_j - \bar{X}_B)^2}}$

↳ Pooled standard deviation of the group

In the frequentist perspective μ_A and μ_B are fixed unknown constants, so we estimate them and look at the sampling distributions of the estimates.

* Bayesian approach

μ_A and μ_B are events that have a corresponding probability distribution (i.e. μ_A and μ_B are random!)

Prior beliefs on μ_A and μ_B are specified by

$$\mathbb{P}(\mu_A);\ \mathbb{P}(\mu_B)$$

Data: $X_j,\ j \in A,B$ w/ distribution $f(X_j \mid \mu_A, \mu_B)$

Posterior: $\mathbb{P}(\mu_A \mid X_j, j\in A);\ \mathbb{P}(\mu_B \mid X_j, j\in B)$

To answer our question of efficacy, we need to investigate the posterior distribution of

$$\delta = \mu_B - \mu_A$$

↳ key idea: If the distribution of δ is stochastically greater than 0, drug A is useful.

↳ every possible value is always positive (best case scenario) distribution is shifted to the right of 0.

New York Times changes the title of their stories w/ A/B t. :o

* In the frequentist approach, we have a p-value to characterize significant difference.

* Bayesians though have probabilities on δ (much stronger)!

In our example, we'd like to know $\mathbb{P}(\delta > 0 \mid \underset{\sim}{X}, n_A, n_B)$

NOTE: δ is called the average treatment effect (ATE)

Example: User experience & click-thru rate → conversion rate (actually buy a product)

Version A vs Version B (think ~~Netflix~~ Amazon)

Version A: Shown to N_A people

n_A people purchase a product

$p_A = \frac{n_A}{N_A}$ = conversion rate

Version B: Shown to N_B people

n people purchase a product

$p_B = \frac{n_B}{N_B}$ = conversion rate

Experimental setup: A unique visitor comes to Amazon. With probability 1/2, this visitor is shown version A and w/ prob 1/2 shown version B. Amazon keeps running this procedure until a desired number of visits have occurred.

QUESTION: Which version of the website led to a higher conversion rate?

⇒ To answer this, we look at posterior distributions for p_A, p_B, and $\delta = p_A - p_B$

Data: We observe conversion rates for each of our versions. Call them $\hat{p}_A$ and $\hat{p}_B$.

We also know $N_A = 1500$, $N_B = 750$

$\hat{p}_A = 0.05$; $\hat{p}_B = 0.04$

Model: $n_A \sim Bin(N_A = 1500, p_A)$

$n_B \sim Bin(N_B = 750, p_B)$

p_A, p_B : "True" conversion rates.

Prior: $p_A \sim U(0,1)$

$p_B \sim U(0,1)$

No other info except we know they [lie ...] between 0, 1

Posteriors: $p_A \mid \hat{p}_A, N_A$

$p_B \mid \hat{p}_B, N_B$

} Get posterior of δ and answer question!

To answer our question, we calculate:

$$\mathbb{P}(\delta \geq 0 \mid \hat{p}_A, \hat{p}_B, p_A, p_B, N_A, N_B) = 0.983$$

↳ weight of the histogram to the right of 0.

$$\mathbb{P}(\delta < 0 \mid \dots) = 0.017$$

Good evidence that A is better for conversion than B.

A little more about posterior inference

There are many ways to use the posterior distribution once you have it:

1) Probabilities of the parameter of interest.

   Ex: For A/B test, $\mathbb{P}(\delta > 0 | X)$

2) Credible intervals: the Bayesian analog to confidence intervals.

   A $(1-\alpha)100\%$ credible interval $[a,b]$ s.t. $\mathbb{P}(\theta \in [a,b] | X) = 1-\alpha$

3) Maximum à posteriori (MAP) estimate:

   The value of θ that has the highest posterior probability (i.e. the mode of $\theta|X$).

Some comments about goodness of fit

* Recall that our overall aim is to develop a model for how the data we observe was generated!

* This is the same as any other statistical or ML model, so the same metrics can be used to evaluate our model/performance.

* As an example, we can split data into training vs test. Build Bayesian model on training, an simulate possible values for the test set. Finally, compare simulated values with observed values using your favorite metric.

* Similarly, to evaluate performance within the training sample, we can simulate in-training data and compare w/ what we observed using similarity metrics (ex: MSE, Kolmogorov, Smirnoff, etc.)

About the test:

Fill in the blank, multiple choice, Short answer

- Types of models to use — distribution)
  how to use & parameters.
- Basic probability questions (result of calculation)
- Bayesian modeling
- No coding!