L5.notes

2020年4月29日 星期三　　下午4:26

Two different types of learning effects

1. Change aversion
2. Knowledge effect

When users first encounter a change they will react, but will eventually adapt to a change.

4 Analyzing Results

4.1 Sanity Tests

One of the first things to do once you finish collecting experimental data is to analyze the invariants. This is done by calculating the values for one or more invariants on the test and control group, and check if the difference is statistically significant. For e.g. if the values for an invariant (say total # of cookies) are x and y, then calculate the se as $\sqrt{\frac{0.5*0.5}{x+y}}$, since one would expect the same number of cookies in both groups. Then calculate the margin as $1.96 * se$. If the margin is greater than $x/(x+y) - y/(x+y)$, then the difference of the invariant is insignificant. However if the difference is greater than the margin, then the difference is insignifiant and needs to be investigated further

An example is provided below:

```
control_event_ct = c(2451,2475,2394,2482,2374,1704,1468)
test_event_ct = c(2404,2507,2376,2444,2504,1612,1465)
control_total = sum(control_event_ct)
test_total= sum(test_event_ct)
p_cont = control_total/(control_total+test_total)
p_test = test_total/(control_total + test_total)
p_cont
```

```
## [1] 0.5005871
```

```
p_test
```

```
## [1] 0.4994129
```

```
se = sqrt(0.5*0.5/(control_total+test_total))
margin = 1.96*se
p_cf_min = 0.5 - margin
p_cf_max = 0.5 + margin
p_cf_min
```



```
## [1] 0.4944032
```

```
p_cf_max
```

```
## [1] 0.5055968
```

The most common reasons for sanity checks failing is data capture. Other reasons could be experimental set-up, for e.g., where there is a filter on the test but not on the control

4.2 Analysis with a Single Metric

```
# Data provided from test
Xs_cont = c(196, 200, 200, 216, 212, 185, 225, 187, 205, 211, 192, 196, 223, 19
2)
Ns_cont = c(2029, 1991, 1951, 1985, 1973, 2021, 2041, 1980, 1951, 1988, 1977, 2
019, 2035, 2007)
Xs_exp = c(179, 208, 205, 175, 191, 291, 278, 216, 225, 207, 205, 200, 297, 299
)
Ns_exp = c(1971, 2009, 2049, 2015, 2027, 1979, 1959, 2020, 2049, 2012, 2023, 19
81, 1965, 1993)

Xs_cont_sum = sum(Xs_cont)
Ns_cont_sum = sum(Ns_cont)
Xs_exp_sum = sum(Xs_exp)
Ns_exp_sum = sum(Ns_exp)

p_cont = Xs_cont_sum/Ns_cont_sum
p_exp = Xs_exp_sum/Ns_exp_sum

# Empirical standard error and count provided
empirical_se = 0.0062
empirical_ct = 5000
se = (sqrt(1/Ns_cont_sum + 1/Ns_exp_sum))*empirical_se/sqrt(1/empirical_ct + 1/
empirical_ct)

# Calculating the cf for the difference
d = p_exp-p_cont
margin = se*1.96
d_c95min = d - margin
d_c95max = d + margin

# Sign test
```



```
diff_sign = Xs_exp/Ns_exp - Xs_cont/Ns_cont
pos_diff = sum()
```

One thing to be wary of is Simpson's paradox, where the effect in aggregate may indicate one trend, and at a granular level may show an opposite trend.

4.3 Multiple checks

The more things you test, the more likely you are to see significant difference just by chance. This is a problem, but since it is not repeatable for the same metric across multiple attempts, there is a way out. One can do multiple runs of the experiment, or alternately bootstrap. There is another technique called multiple comparison that adjusts your significance levels that accounts for how many metrics or tests you are doing.

For e.g. if you had 10 metrics where you used a 95% confidence interval for each metric, what is the probability that one of the metrics will show up as a false positive?

```
p1 = 0.99
p_nofp = p1^10
p_fp = 1-p_nofp
p_fp
```

```
## [1] 0.09561792
```

As you increase the number of metrics, you can use a higher confidence level to overcome false positives.

A different method used in practice is Bonferroni correction. It has the advantages of being simple, makes no assumptions, and guaranteed to give $\alpha_{overall}$ as low as you have specified.

To use it, calculate

$$\alpha_{individual} = \frac{\alpha_{overall}}{n}$$

For e.g. if you want $\alpha_{overall}$ to be 0.05 and there are 5 metrics then $\alpha_{individual}$ will be $0.05/3 = 0.01666$

Bonferroni methods may be very conservative. Alternatives include closed testing procedure, Boole-Bonferroni bound and Holm-Bonferroni method. The $\alpha_{overall}$ above is often referred to as the familywise error rate (FWER). Another measure is the contol false discovery rate (FDR) defined as the (# false positives)/(#rejections). CDR makes sense if you have a large number (200) metrics.

An alternative to using multiple metrics is to use an 'Overall Evaluation Criterion' (OEC).

4.4 Gotchas

Effect may ramp out as you implement the change. There could be seasonal effects. For e.g. students on summer break have very different behavior than when they come back. Similarly during black friday and other holidays. One of the ways is to leave a small sample out as a hold-out to track them over time.

5 Summary

- Decision to launch change should be guided by business reasons and not just
- Opportunity cost of launching the change (engineering costs, user experience), based on not launching it

6 References

1. Additional Techniques for Brainstorming and Validating Metrics
2. Comparing 2 populations: Binomial and Poisson
3. Tests for Two Poisson Means
4. Final Project Template
5. Large Scale Validation and Analysis of Interleaved Search Evaluation
6. Overlapping Experimental Infrastructure: More, Better, Faster Experimentation

7 Glossary

change aversion: users averse to change reduce usage

interleaved experiment: an experiment in which you expose the same user to both A and B at the same time

interuser experiment: an experiment in which you expose users to either A or B.

intra-user experiment: is where you expose the same user to an experiment being on and of at different times.

novelty effect: users see something new and test out everything

retrospective analysis: look at historical data, observe changes, and conduct an evaluation

userflow: Shows the flow of users through the site, often referred to as a customer funnel.