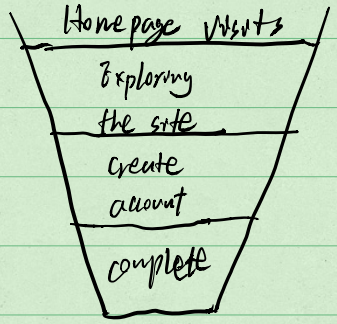
Case : Audacity that create online finance course

I. Case overview

Customer funnel



- Experiment:
  Change "start Now" from blue to pink

- Hypothesis: This change will increase how many students explore Audacity's course
  First step to second step.

II. Refining the hypothesis (choose a metric)

1. Which metric to use?

- x: Total number of courses completed (Take to long)
- x: Number of clicks (need to use percentage)

√: $\frac{\text{Number of Clicks}}{\text{Number of page views}}$ = Click through rate (CTR)

√: $\frac{\text{Unique visitors who click}}{\text{Unique visitors to page}}$ = Click through prob. (CTP) (Better than CTR)

2. Updated hypothesis:

Changing the "start Now" button from orange to pink will increase CTP of the button

3. Choose between CTR and CTP

- Rate: Measure usability of a site. User can click many button on a page, it show how often do they actually find that button
- Prob: Measure the total impact. If just want to know how often user go the second page, use prob. because you don't want to count if user double-click, reload. etc

4. Repeated measures for CTP

eg: visitors = 1000, unique clicks = 100 ⇒ CTP ≈ 10%

III. Review of Statistics

1. Binomial distribution

① Two types of outcomes

② Independent events

③ Identical distribution (P same for all)

For number: mean: np std: $\sqrt{np \cdot (1-p)}$

For prob: mean: p, std: $\sqrt{\frac{p \cdot (1-p)}{n}}$

- Example of not identical:
  - x: drawing 20 cards from shuffled desk: red and black ∵ limited number of cards

- check if binomial
  - V: roll a die 50 times: 6 or other
  - x: click on a search result page: click or not click
    - ∵ the next click will be influenced by previous click (find necessary information or not)
  - V: students complete of course after 2 months: complete or not
  - x: purchase of item with a week: purchased or not
    - ∵ related to number of items in previous purchase.

2. Confidence interval

① Benefit for knowing distribution: have formula for sample SE and can calculate

② meaning of 95% CI:

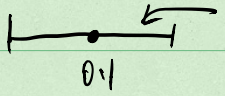
If we repeated sampling over and over again, we would expect the interval we construct around our sample mean to cover the true value in population 95% of the time.

③. calculating CI:

Center of interval: $\hat{p} = \frac{x}{N}$ (# users who click / # users)

$\hat{p} = \frac{100}{1000} = 0.1$

To use normal: check $N \cdot \hat{p} > 5$ and $N \cdot (1-\hat{p}) > 5$



(z distribution)

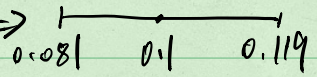
z: z-score for $N(0,1)$

$z_{0.05}$, two side: $-1.96, 1.96$

{ Normal approximation: $z * SE$

Binomial distribution: $z * \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$

$\Rightarrow m = 0.019 \Rightarrow$



3. Hypothesis Testing

need to calculate: P(results due to change)

H_0: $p_{cont} = p_{exp}$ or $p_{exp} - p_{cont} = 0$

H_A: $p_{exp} - p_{cont} \neq 0$

Measure $\hat{p}_{cont}$ and $\hat{p}_{exp}$. calculate $P(\hat{p}_{exp} - \hat{p}_{cont} \mid H_0)$

compute the probability that this difference would have arisen by chance, if H_0 were true

Want to Reject Null if the prob. is small enough (< 0.05)

4. How to compare two samples: pooled standard error

① Reason for using it:

∵ Have two samples with different number of observations

∴ Need to choose a standard error that gives us a good comparison of both samples ⇒ calculate pooled standard error

② Calculation:

X_{cont}, X_{exp}; N_{cont}, N_{exp}

$$\hat{p}_{pool} = \frac{X_{cont} + X_{exp}}{N_{cont} + N_{exp}}$$

$$SE_{pool} = \sqrt{\hat{p}_{pool} \cdot (1-\hat{p}_{pool}) \cdot \left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$$

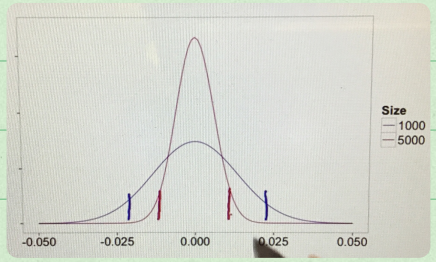
$\hat{d} = \hat{p}_{exp} - \hat{p}_{cont}$, $H_0: d = 0$, $d \sim N(0, SE_{pool})$

- If $\hat{d} > 1.96 \cdot SE_{pool}$ or $\hat{d} < -1.96 \cdot SE_{pool}$, we reject null

IV. Design the experiment

1. Size v.s. power trade-off

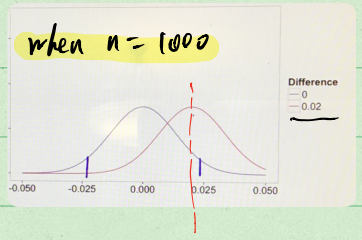
① How page views affect sensitivity



- α: P(reject null | null is true)
- β: P(fail to reject null | null false)

$\alpha = 0.05 \Rightarrow$ area outside boundary : 0.05

↑ That is, if new mean in the range, we fail to reject null



when size is small, it's already practically significant but not statistically significant.

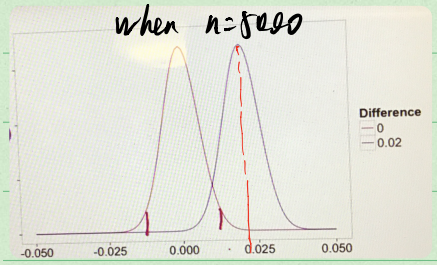
$\Rightarrow \beta$ is pretty high

- α: prob to launch a bad experiment
- β: prob to fail to launch a good experiment that actually did have a difference you care about.

- Small sample: α low, β high

(80%)

- $1-\beta$: sensitivity what a high level of sensitivity at the practical significant boundary



More data:

When there is a true difference, much less likely to fall within the range of

⇒ more likely to reject null when there is a difference

- Large sample: given α & same, β lower ⇒ power higher

2. practically / substantive significant

① Def: From business perspective, what size of change matters to us

② Reason: investment, want for next change that do better

In previous case, practically significant is 2%

③ Purpose: repeatability

- statistical significant is about repeatability: When set up experiment, want to get the guarantee that these results are repeatable, so it's statistically significant
- Also want to see a change in experiment that you are interested in from a business standpoint so it's practically significant, it's also statistically significant.

∴ Need to size experiment appropriately, such that the statistically significant bar is actually lower than the practically significant bar.

3. Calculating number of page views

Need to know: use online calculator

- Base line conversion rate (before experiment)
- Minimum detectable effect (practical significant level)
- absolute or relative change (will introduce later)
- statistical power (sensitivity): $1-\beta$ = P(reject null | null is false)
- significance level: α

4. How number of page views varies changes:

① Higher CTP in control group (but less than 0.5) : Increase Page Views

$SE = \sqrt{\frac{p(1-p)}{N}}$ $\sqrt{0.1*0.9} = 0.3$ ⇒ $\sqrt{0.5*0.5} = 0.5$

∴ need to keep SE same, P↑, N↑

② Increase practical significant level (d_{min}) : Reduce page views

Larger change is more easy to detect

③ Increase confidence level ($1-\alpha$) (α↓) : Increase page views

need to be more certain that a change has occurred before you reject the null ⇒ more conservative

α: P(reject null | null is true), β: P(fail to reject null | null is False)

∴ when α↓, β↑, then sensitivity $(1-\beta)$↓

③' ∴ n↑, sensitivity↑ (∵ more page views ⇒ narrow distribution)

④ Higher sensitivity : Increase page views

V : Analyze the results

1. Is the difference due to random variation:

need to calculate confidence interval

2. Calculate CI:

$N_{cont} = 10,072$ $N_{exp} = 9886$ $d_{min} = 0.02$

$X_{cont} = 974$ $X_{exp} = 1242$ Confidence level = 95%

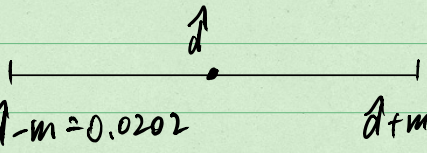
$$\Rightarrow \hat{p}_{pool} = \frac{974+1242}{10,072+9886} = 0.111$$

$$SE_{pool} = \sqrt{0.111 \cdot (1-0.111) \cdot \left(\frac{1}{10072} + \frac{1}{9886}\right)} = 0.00445$$

$$\hat{d} = \frac{X_{exp}}{N_{exp}} - \frac{X_{cont}}{N_{cont}} = 0.0289$$

$\hat{d}$: difference in two groups

$m = SE_{pool} * 1.96 = 0.0087$: margin of error

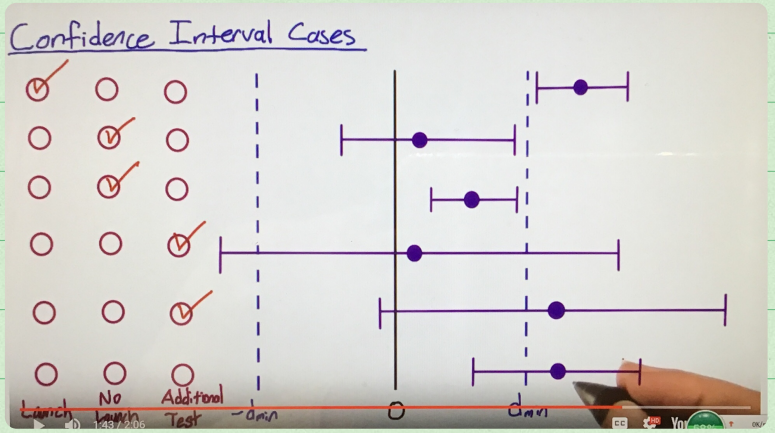


$\because \hat{d} - m > d_{min}$ $\therefore$ we should launch the change

3. Wether to launch an experiment

- CI all higher than d_{min} : launch
- CI all lower than d_{min} : no launch
- CI very large, part of it is larger than d_{min} : need additional test (with more power)

(Power = sensitivity $= 1 - \beta$)



4. What to do when last three cases happended

- Sometimes no time for a new A/B testing
- Need to communicate to decision-makers when they are going to have to make a judgement, and take a risk, because the data is uncertain. They're going to use other factors, like strategic business issues, or other factors besides the data