A/B test.

- Can only test A or B, can't tell you what is missing.
- Need clear control, and clear metrics.

1. what A/B testing can do and can't do.

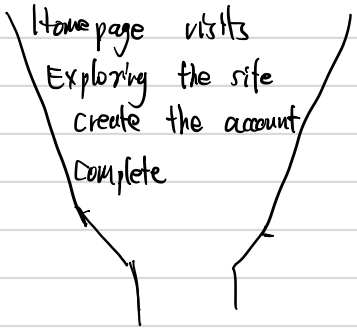
can't

- if too long or don't have data (buy a car, or refer)
- Suprisingly emotional. (New logo, needs long time to watch.)

can do

- test layout : clear control, clear metrics

- Customer funnel



- End to End.

1. Choose a metric

① Initial the hypothesis : 'changing the 'start now' button from orange to pink will increase the number of students exploring the site

② Refine the hypothesis :

- Which metric to use?
  - X. Total numbe of courses completed : will take too much time to get a result
  - X. Number of clicks : total number or fraction?
  - number of clicks / number of page views. : click-through rate / CTR.
  - ✓ unique visitors who click / unique visitors to page : click-through probability

③ updated hypothesis: Change the 'start now' button from orange to pink will increase the click-through-probability of the button

Click-through-rate vs click-through-probability
- if want to measure the usability of the button, use ··rate
  because users have a variety of different places on the page that they can choose to click on and rate will show how often they actually find the button
- If you want to know how they would get into the next level, use ·· probability.
  because you need to exclude the probability of double click, reload and so on

2. Review statistics

① which surprise you most? 150 900.? why?

a. Binomial distribution.:

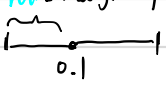
- with mean $= p$, standard deviation $= \sqrt{\frac{p(1-p)}{N}}$

When can use binomial
- 2 types of outcomes
- Independent.
- Identical distribution
  - p for all.

b. confidence interval

$\hat{p} = \frac{\#\text{ of users who clicked}}{\#\text{ of users.}} = \frac{100}{1000} = 0.1$ (N)

m = margin of error



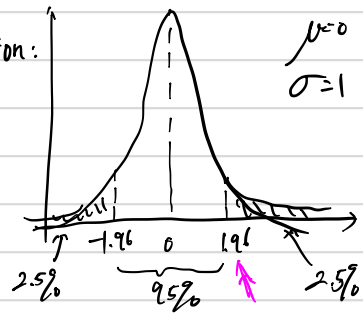
To use normal: check $N\cdot\hat{p} > 5$ and $N(1-\hat{p}) > 5$

$m = z^* SE$

$= z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$

z: confidence level, SE: standard error

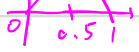
z distribution:



→ 1.96 for 97.5%

$m = 0.019$

$\hat{p}(1-\hat{p})$



- p away from 0.5 $p(1-p)$ ↓ m ↓
- N ↑ m ↓

two-tailed test ⇒ 99% → 1% → 0.5% ⇒ 99.5% ⇒ z = 2.58

exe: N=2000, X=300. confidence level = 99%. ⇒

$\hat{p} = \frac{300}{2000} = \frac{3}{20} = 0.15$

$m = 2.58 \times \sqrt{\frac{0.15 \times 0.85}{2000}} \approx 0.017$

④ Establish statistical Significance

- Hypothesis testing: How likely it is that your results occured by chance.
  a. null hypothesis / baseline : no difference of probability between our experiment and control group

  b. Alternative hypothesis :

⇓.

P_{con} P_{exp}.

a. Null hypothesis : H_0 $P_{con} = P_{exp} \Rightarrow P_{con} - P_{exp} = 0$

b. Alternative hypothesis H_A : $P_{con} \neq P_{exp} \Rightarrow P_{con} - P_{exp} \neq 0$

c. measure P_{con}, and P_{exp}
calculate $P(\hat{P}_{exp} - \hat{P}_{con} \mid H_0) = \alpha$

d. reject null if α is smaall enough
α is same type of significance threshold as a confidence interval
reject null if $\alpha < 0.05$

⑤. Compare two samples
X_{cont} X_{exp}. N_{cont}, N_{exp}

Pool probability $\hat{p} = \dfrac{X_{cont} + X_{exp}}{N_{cont} + N_{exp}}$

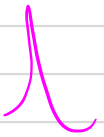
$$SE_{pool} = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$$

$\hat{d} = \hat{p}_{exp} - \hat{p}_{cont}$

$H_0 : \hat{d} = 0$ $\hat{d} \sim N(0, SE_{pool})$

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

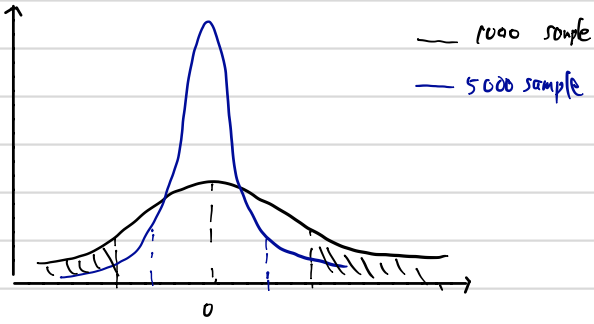
N 越大.
SE 越小
越陡.
当 $\hat{d} > 1.96 \times SE$
或 $\hat{d} < -1.96 \times SE$ 时.
reject null



③ design : How many pages we need to get a statistically significant result.

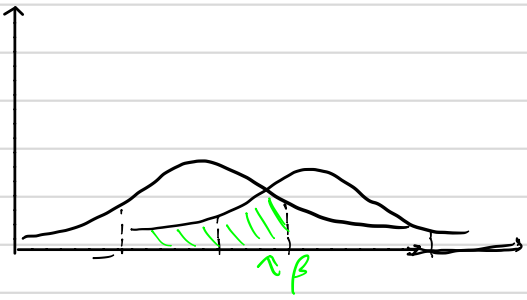
α = P(reject null | null true)

β = P(faill to reject null | null false)



small sample : α low : unlikely to launch a bad exp

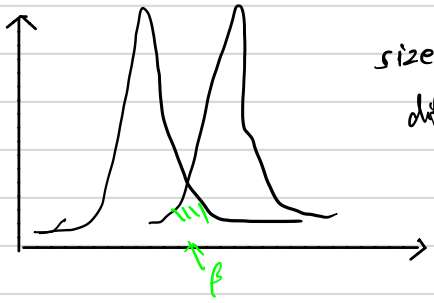
β high : likely to fail to launch an experiment that actually did have a difference



there is a difference

$1-\beta$ = sensitivity

often 80%



size = 5000

difference = 0.02

β is much smaller

Longer sample : α same

β lower

Lesson 2. Policy and Ethics for experiments.

1. Risk , → minimal risk.

2. Benefit . → improve product, understand human better.
   in education, medicine risks are higher but benefits higher

3. Alternative: What other choice do participants have

4. Data sensitivity / Privacy : What data is being collected, what is the expectation of privacy and confidentiality.

Lesson 3. Choosing and characterizing metrics

- Define
- Build Intuition
- Characterize

1. Define.
   a. Think How to use the metric. { · Invariant checking : unchanged
   · Evaluation { business : revenue, market share, # of users
   detailed : How long user stayed

2. Choose metric.

- difficulty: · Don't have access to data.
  · takes too long (whether get a job finally, which is a long process)

eg: Rate of returning to 2nd course → takes too long.
Average hapiness of shoppers → No access to the data.
Probability of finding infor via search → No access to the data

- method: . find metrics from the process of user ⇒ funnel.
  . combine several metrics together

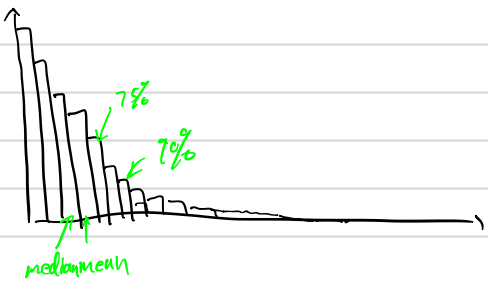
- filter spam data segament data in different dimensions. eg. in weekly, monthly, yearly, in country

Question: suspect there is an issue on tracking double click on mobile platform.
· Need the click-through rate and click-through-probability of these two platform

- Summary metrics: { Poisson distribution.
  { Pareto distribution: 80% of society wealth is hold by 20% of people
  { heavy tailed: the mean probably doesn't work well, even may be infinite

- Category of summary metrics:
  - Sums and counts : eg: # of users who visited page
  - Mean, median, and percentile: mean age of users who complete a course
    median latency of a page load
  - Probability and rate: Probability of outcome 0 or 1.
  - Ratios : $\frac{P(\text{revenue-generating click})}{P(\text{any click})}$

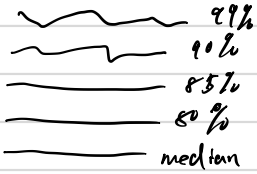
- Means, medians, and percentile



75th percentile
90 th percentile → if the user find useful info

- Sensitivity and robustness of the metric
  - not robust: mean may be influenced badly by extremely outliers. eg: video load latency is much longer for users with bad connection
    ⇓
    median is more robust, but if you only affect a fraction of your users, there may not be change on median. this is <u>not sensitivity</u>
    ⇓
    use 90% or 99%

- How to measure robust and sensitivity:
  latency of a video.
  ⇓.
  ① retrospective. & comparable riders, • different metrics: median, 80%, 85%, 90%, 99%.



90%, 99%. not robust.

② result of experiments:

80%, median not sensitive.

- Absolute or Relative difference
  5% vs 7%: absolute: 2 percentage points.
  relative: 40%

3. Variability

① To calculate a confidence interval, you need:
- variance (or standard deviation)
- Distribution

Binomial distribution

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$

$m = z \times SE$

standard deviation : $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

variance : $Var = \sigma^2$

standard error : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

type of metric	distribution	estimated variance
probability	binomial (normal)	$\frac{\hat{p}(1-\hat{p})}{N}$
mean	normal	$\frac{\hat{\theta}^2}{N}$
median / percentile	depends	depends
count / difference	normal (maybe)	$Var(X) + Var(Y)$
rates	Poisson	$\bar{x}$
ratios	depends	depends.

② Non-parametric methods vs

③ A versus A experiment : control A and another control A, no chages
test for the sensitivity and robustness of metrics
↳ if your experiment system is itself complicated, it's a very good test for your system : is your randomization truely random? do you have any other issues like biase

↓ if can't do many A versus A.
Use bootstrap : divide big sample into small ones, and compare ⓑ between the subsets

④ Calculate variability empirically.

- Look at A/A test on click-through-probability

Use of A/A test:

- Compare results to what you expect (sanity check)
- Estimate variance and calculate confidence
- Directly estimate confidence interval

→ Since we expect a normal distribution: $m = SD * Z$

$= 0.0069 \times 1.96$

Analytically: $SE = \sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})(\frac{1}{N_{cont}} + \frac{1}{N_{exp}})}$

if doesn't look like normal distribution

↓

Directly estimate confidence iterval