**基本概念**

treatment factor – 要检验的，比如奖励的门槛

treatment level – 门槛的不同等级

response variables – 去检验的结果，比如顾客的反应

background factor – 周围的因素，比如天气

biased factor – 对结果造成影响的background factor

confounded factor - 对结果造成影响的treatment factor

**哪些因素会对实验的有效性产生影响**

成熟度 – 实验对象越来越熟悉实验

测试 – 反复测试

自选择 – 如顾客主动加入会员

实验对象流失

相关性误认为是因果

jone herry效应 – 因为知道自己在实验组所以更加努力

丧气效应 -因为知道自己在实验组所以更加不努力

**what to test in ads industry?**

* targeting – customer segment
* frequency - impression#
* flighting – when to scheduling the ads
* channel
* content

**design an experiment**

1. have clear objective
2. define the metric (invariant checking vs evaluation)
   1. High level concept of metrics (e.g active users, 结合公司目标)
   2. Details (e.g. how do you define user activity-DAU…)
   3. Measurement
      1. Sums and counts (e.g. # of users who visited a page)
      2. Distributional metrics - means, medians, percentiles (do a retrospective analysis, if normal distribution, then mean or median, if skewed distribution, percentiles)
      3. Probabilities and rates
      4. Ratios (e.g. revenue/click)

2.4 one metric or a suite of metrics (The more things you test, the more likely you are to see significant difference just by chance. There is another technique called multiple comparison that adjusts your significance levels that accounts for how many metrics or tests you are doing)

a.       Long-term use short-term proxy

b.       Invariant is for sanity check

c.       Think about how spam/bots would influence your metric

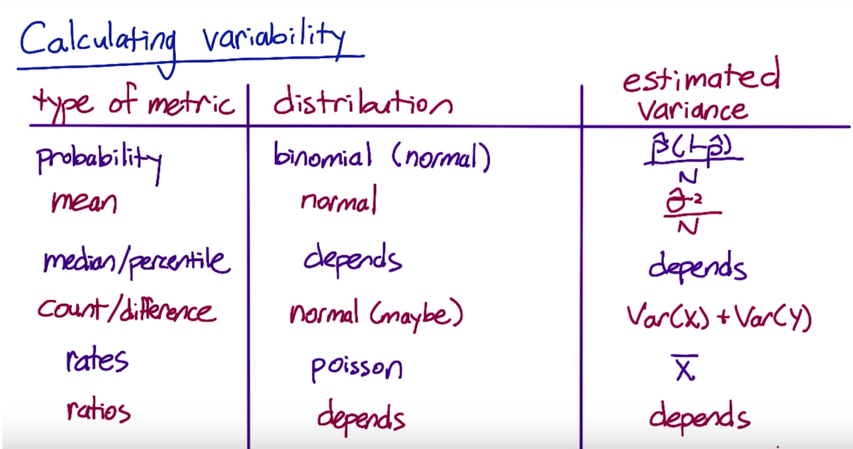
d.       Sensitivity and Robustness- when choosing metrics, make sure the directional change of the metric is in line with your expectation and the change is unlikely due to bot behavior and it would not take too long to evaluate.

e.g. mean is sensitive to outliers, median is robust but not sensitive to changes to small group of users. can use a/a test to measure the robustness

Calculating variability – 统计学or estimate the variance empirically with a/a test

如果样本量够大：n\*p>=5 and n\*(1-p)>=5

近似normal distribution



1. Choose population

Inter-user (两组用户)

Intra-user (一组用户，前后实验)

Control variables (demographic, behavior in the past…)

US only?

1. check unit of diversion – consistent experience, user visibility

User id

Anonymous id(cookie)

Event

Device id

IP address

1)User consistence –users visibility or Measure learning effect (if load is slower, will users use less, need to see across time)

2)Ethical consideration

3)variability – empirical variability may be much higher than analytical variability.

Because when unit of analysis (the denominator of metrics) is different from unit of diversion. E.g. unit of analysis is pageview and unit of diversion is cookie.

When you’re doing event-based diversion, every single event is different random draw. When you’re using cookie- based diversion, they are correlated,which increase the variability.

1. Control variables

randomized experiment,

when volume is high enough volume to avoid bias

in web traffic experiment. because you cannot control what traffic will come in,

可以后期过滤 如过滤新用户

Matched pair experiment

List potential variables

Is the data available?

Logical connection between control and target variables

Test correlation between control and target variables

Test correlation between control variables and delete repeated ones

How to select pairs---distance score

1. sample size

第一类错误α（false positive）不超过5%。也就是说，**Significance Level** = 5%。

第二类错误β（false negative）不超过20%。也就是说，**Statistical Power** = 1 -β = 80%。

**Baseline Rate**-做实验之前的指标数值。当baseline很大（接近1）或者很小（接近0）的时候，实验更容易检测出差别（power变大），如果保持power不变，那么所需要的样本数量变小。

**Minimum Detectable Effect**-参数越大（比如10%），说明我们期望实验能够检测出10%的差别即可。检测这么大的差别当然比较容易（power变大），所以保持power不变的情况下，所需要的样本量会变小。考虑business perspective，online有2% change就很高

根据软件算出的是每组样本量

1. duration

总样本量/每天量

考虑weekday/weekend不同

1. analyze the result

Sanity check (make sure test and control are comparable, e.g. test of mean for invariant variables with t test)

1. Make suggestion

a.       Do I have statistical/practical significance?

b.       Do I understand the change? Who is going to be impacted?

c.       Is it worth it? Cost vs benefit?

广告业

measurement metrics

* lift = (actual conversion-baseline)/baseline
* ROAS (return on ad spend) = revenue / cost

Example: post an ad on facebook

Randomized control trial (RCT)

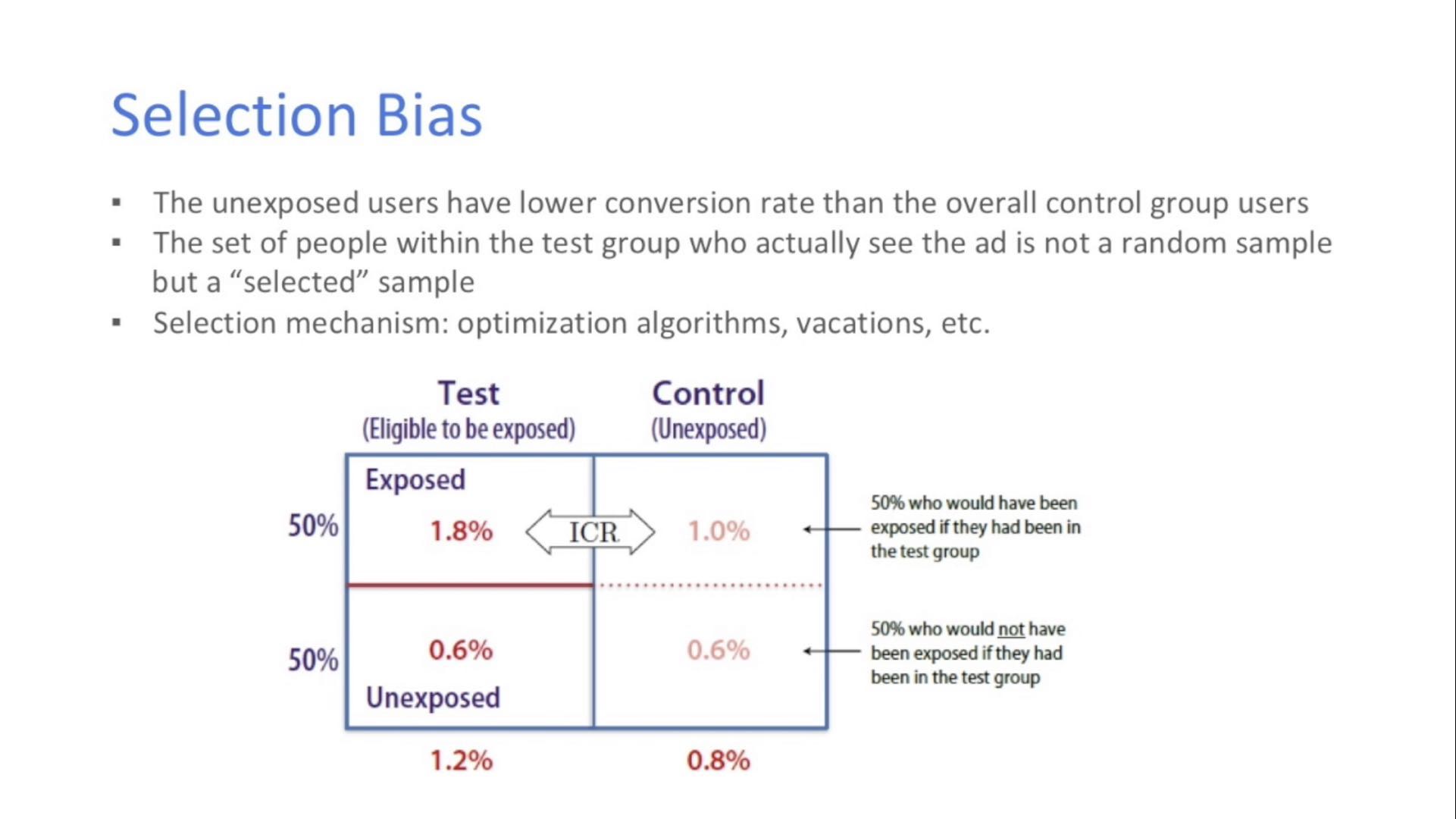
Step 1 Randomization check: control group and test group

User age, gender, married, length of time using FB, # of friends, # of FB uses in last 7 days/28 days

Step 2 serve ghost ads

if control group should get the ads, show them the second ads. If they shouldn’t get the ads, show them the original ads

step 3 calculation



step 4 减少selection bias的方法

individual – based comparison

* exact matching
* propensity score matching
* regression adjustment

market – based comparison

选取两个相似的市场

With A/A test, we can

1. Test sensitivity and robustness: if there is lots of variability, to sensitive
2. Compare result to what you expect (sanity check)
3. Estimate variance empirically and use your assumption about the distribution to calculate confidence
4. Directly estimate confidence interval without making any assumption of the data
5. a/b test duration

**pitfalls**

1. interaction: 两组互相影响，比如uber 司机乘客，facebook friends

解决方法：选取完全隔绝的两组(network, geo) then match comparable market in pairs

1. novelty effect

解决方法：only select new users

1. 需要较长时间才知道结果，如subscription retention

解决方法：use model to find short metrics can move the long term one

1. selection bias

解决方法：exact matching

1. test win but not launch it –labor cost, risk and bug
2. 多个metrics一起实验，容易出现significant
3. learning effect
4. 看到p value significant 就停止实验
5. Seasonal effect in experiments – airbnb traveling season
6. Cannibalization – because of the potential interaction of people within each experiment group.