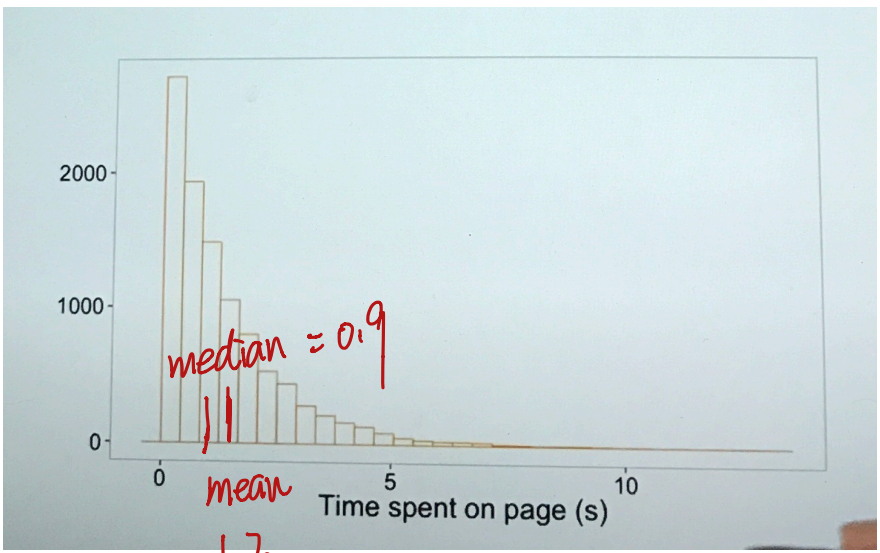
e.g. $\frac{P(\text{revenue-generating click})}{P(\text{any click})}$

① 为什么是指标簇

Means, median, and percentage (指标分布) ② 什么是频率

⇒ 用户在网页的某个页面花多长时间.



How to measure sensitivity and robustness?

1. running experiments or using experiment you already have

t. a versus a experiments to determine if they're too sensitive.

↳ It's an experiment where you don't change anything. you just compare people who saw the same thing to each other.

2. sort of a retrospective analysis of your logs: 对你记录的回顾性分析

if you don't have experiment data or you can't run new experiments, you can look back at the changes you know you made to your site, and see if the metrics you're interested in actually moved in conjunction with those changes.

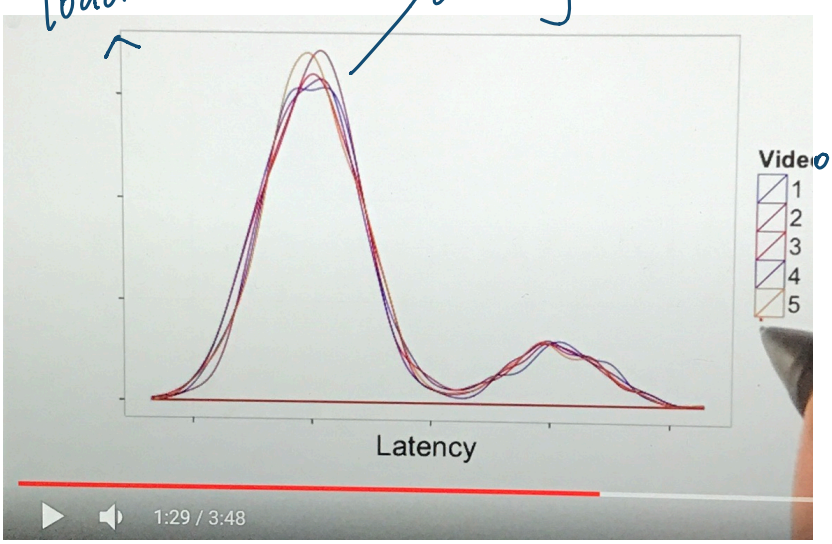
or look at the history of the metric and see if you can find a cause for any major changes that you see.

Measuring sensitivity and robustness

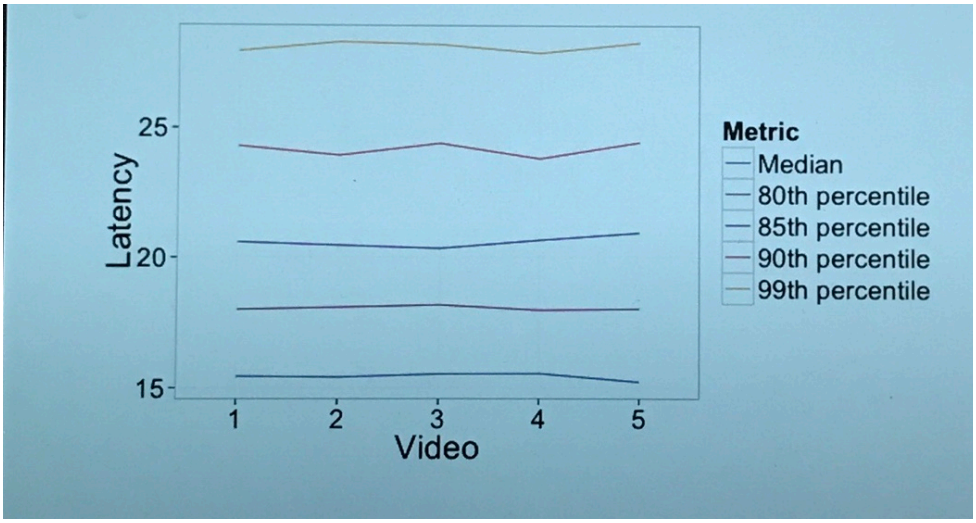
① - choose summary metric for latency of a video

load time

density line



有2个高峰，1个加载时间相对长



90^{th} 与 99^{th} 是 Z shape

↳ not robust enough

接近直线

Mid and 80^{th} is not sensitive enough.

② look at the actual experiment.

↳ distribution for experimental video

Absolute & relative difference

1. 确定如何计算这个 comparison:

现在获得的是实验组的数据 + control data

方法: 1. difference

又称 percent change

2. relative change 而不是 absolute change

优: choose one practical significance boundary to get stability over time.

applicable situation: seasonality, system is changing

缺点: variability.

Absolute vs. relative difference

Suppose you run an experiment where you measure the number of visits to your homepage, and you measure 5000 visits in the control and 7000 in the experiment. Then the absolute difference is the result of subtracting one from the other, that is, 2000. The relative difference is the absolute difference divided by the control metric, that is, 40%.

Relative differences in probabilities

For probability metrics, people often use percentage points to refer to absolute differences and percentages to refer to relative differences. For example, if your control click-through-probability were 5%, and your experiment click-through-probability were 7%, the absolute difference would be 2 percentage points, and the relative difference would be 40 percent. However, sometimes people will refer to the absolute difference as a 2 percent change, so if someone gives you a percentage, it's important to clarify whether they mean a relative or absolute difference!

Variability: { 确定实验规模
分析置信区间 和得出结论

使用 count 或 probability, only dealing with the variability of a single measurement, or a constrained one, in case of probability.

使用 ratios or percentiles, compute the variability empirically. 跟据经验

Calculating variability.

To calculate a confidence interval, you need:

- Variance (or standard deviation)
- Distribution

Binomial distribution: 在两方面利用了这是一个二项分布的事实:

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ ← 利用是二项分布的事实得出标准误差的公式

$m = z^* \cdot SE$ ← 我们假设这是一个正态分布.

△ N越大，二项分布越趋近于正态分布.

type of metric	distribution	estimated variance
probability	binomial (normal) 根据中心极限定理	$\frac{\hat{p}(1-\hat{p})}{N}$
mean	normal (if sample size is larger)	$\frac{\hat{\sigma}^2}{N}$

△注意样本方差与实际指标方差的区别:

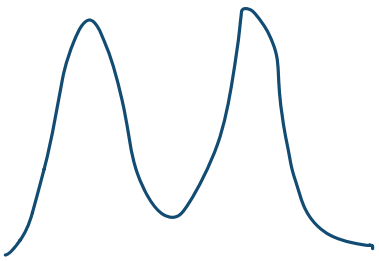
样本方差: 所有数据点的方差. take each data point and collect the variance of them.

实际指标方差: (variance of the actual metric): if you were to collect a new sample, how would you expect this metric to vary

median/percentile	depends on the distribution of the underlying data	depends on the distribution of the underlying data

other summary metrics may be harder to analyze:

e.g. median - could be non-normal if data is non-normal



e.g. lantancy is a bimodal distribution.

mean是正态分布，而mid不是

好估计出位数的方差，你需

要假设基础数据的分布情况

根据你的假设，你可以轻松

估计其分析方差.

	↳对于demographic data	各个变量的方差之和
count /difference	normal (maybe)	Var(X) + Var(Y)
rates	poisson	$\bar{X}$ 平均数
e.g. $\frac{\hat{p}_{exp}}{\hat{p}_{cont}}$ instead of $\hat{p}_{exp} - \hat{p}_{cont}$ (实验组 / 对照组)		
ratios	depends on 分母和分子的分布	depends on 分母/分子的分布.

Confidence interval for a mean

Measure: Mean number of homepage visits per week

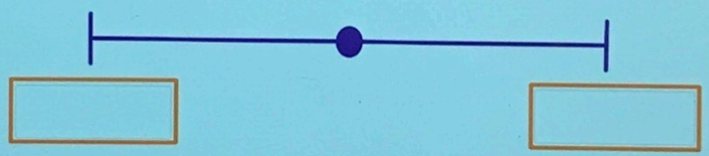
$N_1 = 87,029$
$N_2 = 113,407$
$N_3 = 84,843$
$N_4 = 104,994$
$N_5 = 99,327$
$N_6 = 92,052$
$N_7 = 60,684$

$\bar{N} = \frac{N_1 + \cdots + N_7}{7}$

$\sigma = SD(N_1, \ldots, N_7)$ → use spreadsheet or programming language

$SE = \frac{\sigma}{\sqrt{7}}$

95% confidence interval



$\bar{N} = 91762.29$ $\sqrt{7} = 2.646$

$\sigma = 17014.80$

$SE = 6430.99$

$$\left(\bar{X} - z \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Nonparametric answer

nonparametric method: 无需对分布情况作假设，也能分析数据

e.g. sign test 符号检验.

抽取20天的 AB test，有15天实验组测量数据高于对照组.

使用二项分布的方法计算结果的实际发生概率，并检查实验组和对照组之间是否有差异。

缺点：不能估计效应的大小

优点：easy to do.

当得出结论有差异：

Choose 1: summary statistic distribution is nice and normal

→ use a normal confidence interval with the variants you estimated empirically.

Choose 2: data is funnier / want to be really robust

→ compute a non-parametric confidence interval.

实证.

Empirical variability

针对复杂的指标，使用实证预测方差，而不是通过分析计算.

计算指标方差时，需要对数据潜在的分布有一个假设.

对于复杂指标，分布会很复杂，所以使用实证

eg. google 使用 A/A实验从而通过实证预估所有指标的差异性。

A/B：{ A → 对照
B → 实验

A/A { A → 对照
A → 另一个对照

A/A → 你测量的任何区别是因为潜在的差异性. e.g. system 或 user population, what users are doing.

边际效益递减

there's a diminishing return as you run more A/A test. The key rule is that standard deviation is going to be proportional to the square root of the number of sample. 标准偏差与样本量的平方根

我正确。

bootstrap = 自助法 — you take that big sample, and randomly divvy it up into a bunch of small samples and do the comparison within those random subsets

A/A . v.s. bootstrap

使用 A/A is because if your experiment system is itself complicated, it's actually a very good test of your system.

使用 bootstrap : push towards more complicated metrics or running more and more features

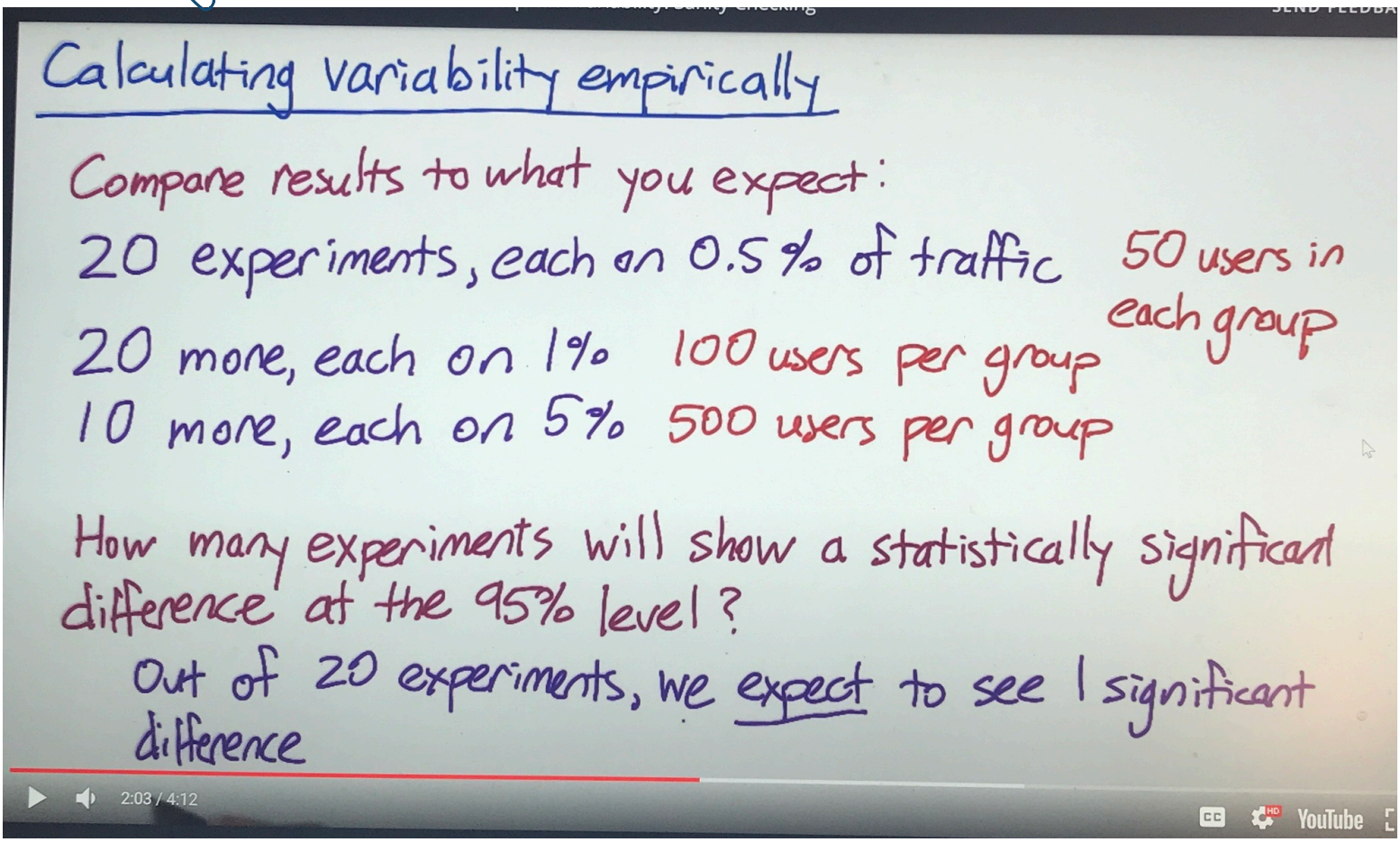
Calculating variability empirically

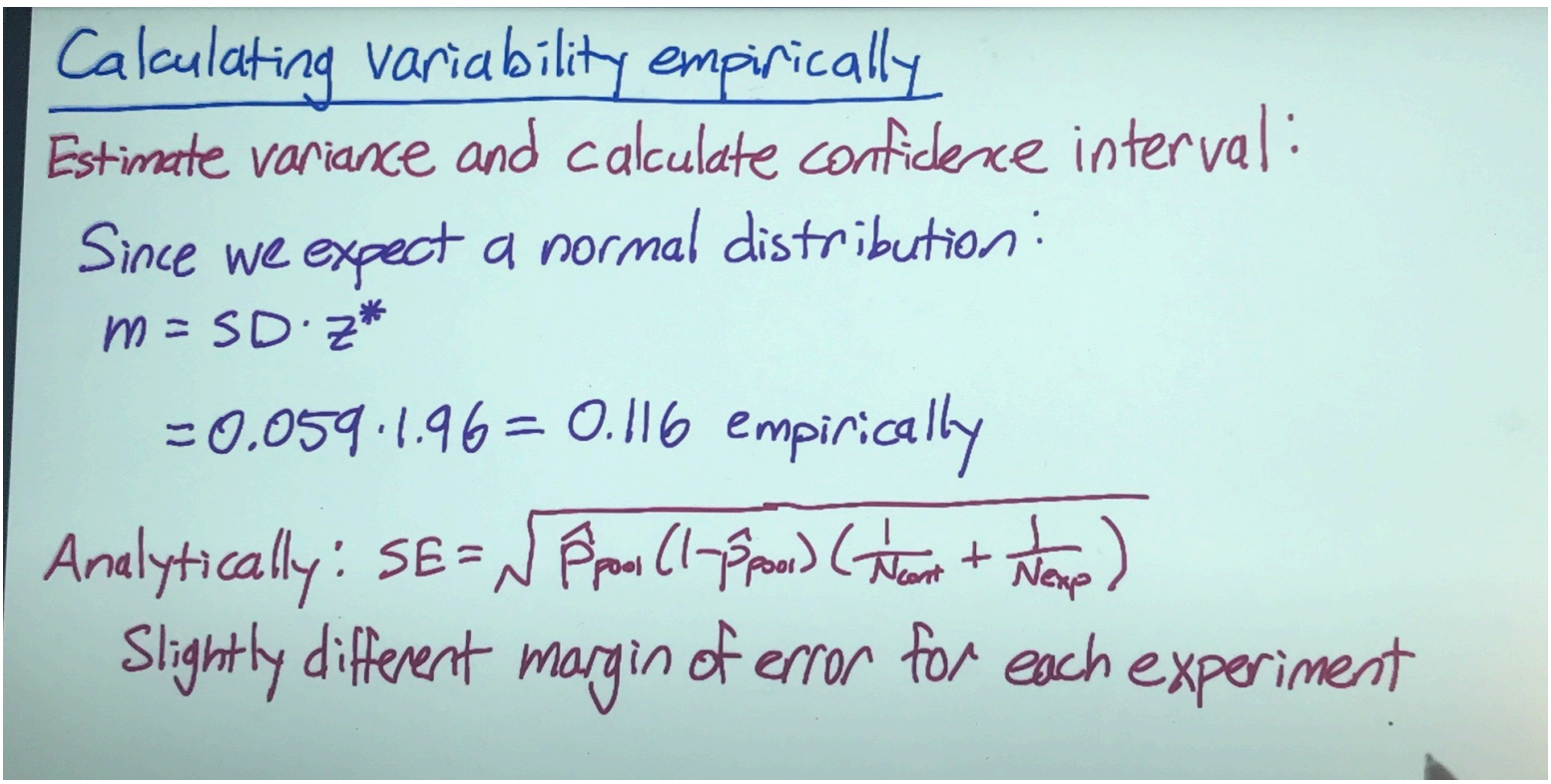
Look at A/A tests on click-through-probability

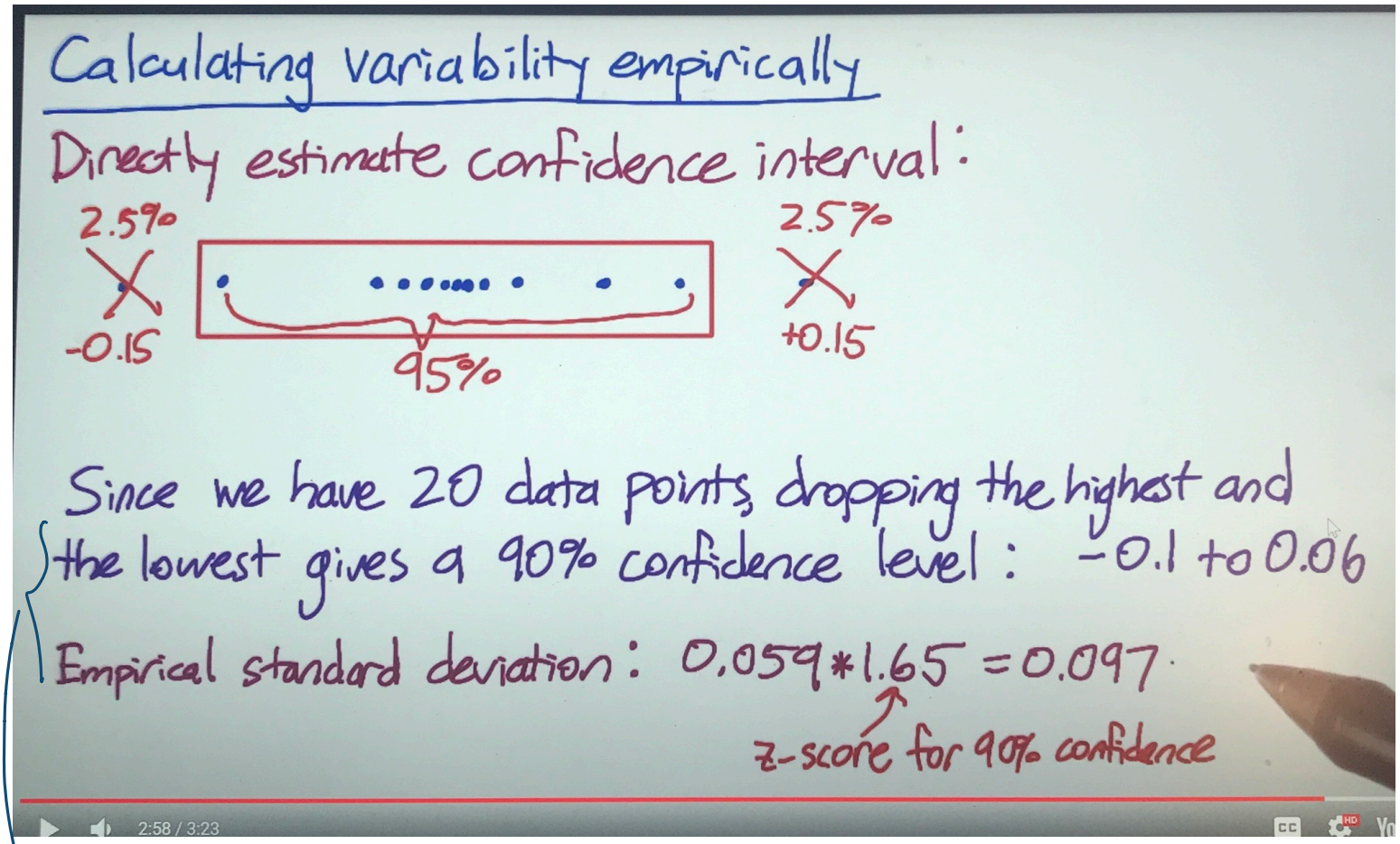
Uses of A/A tests:

1. compare results to what you expect (sanity check)
2. estimate variance and calculate confidence

3. directly estimate confidence interval







2个数据不太相似是因为我们只有20个 data point

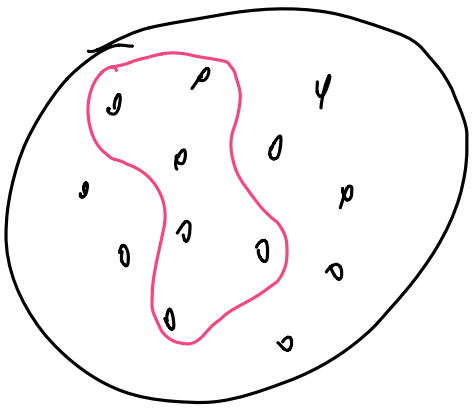
Bootstramping:

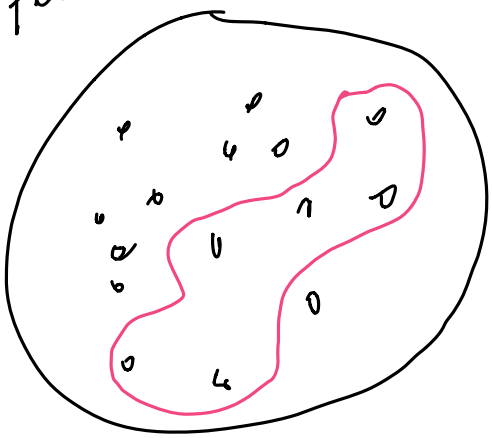
你运行一次实验，在excel sheet中，每个实验每一组我们只显示一个数字一定为CTR，实际上这些数字是由单独为很多 data point 计算得来为一组很多页面浏览量与点击数据。

↳ take random sample from each side 计算 CTR based on use as a "simulated experiment"

repeat to get many experiment

one experiment





Calculating a confidence interval empirically:

- For each experiment, calculate the difference in click-through-probability between the two groups

Calculate the standard deviation of the differences, and assume metric is normally distributed.

Calculate an empirical confidence interval, making no assumptions about the distribution

confidence level = 95%

Use average of diffs as point estimate

-0.06

0.08

View Intro

0.0043

VIEW ANSWER

SUBMIT ANSWER

A/A test data

$m = 0.0364 \times 1.96 =$

Sort the difference 从小到大

第二个小的值，第二个大的值

个

$95\% \times 40 = 38$ min 和 max 忽略

Variability summary

1. different metric have different variability

if variability is so high for a metric, not practical to use

2. To compute the variability, we need to understand the distribution of the underlying data. { analytical techniques
empirical techniques