APPLIED ROOTS

Applied Statistics: Statistical Testing & Experiments

APPLIED ROOTS

Welcome & Thank you

APPLIED ROOTS

Four Sessions

1) 9:00 AM – 11:00 AM

2) 11:30 AM – 1:00 PM

3) 2:30 PM – 4:30 PM

4) 5:00 PM – 6:30 PM

APPLIED ROOTS

(Q) will these videos be available later?

- Yes, as short-er videos over next few days

- Youtube; AppliedAI; Diploma

APPLIED ROOTS

Pre-reqs:

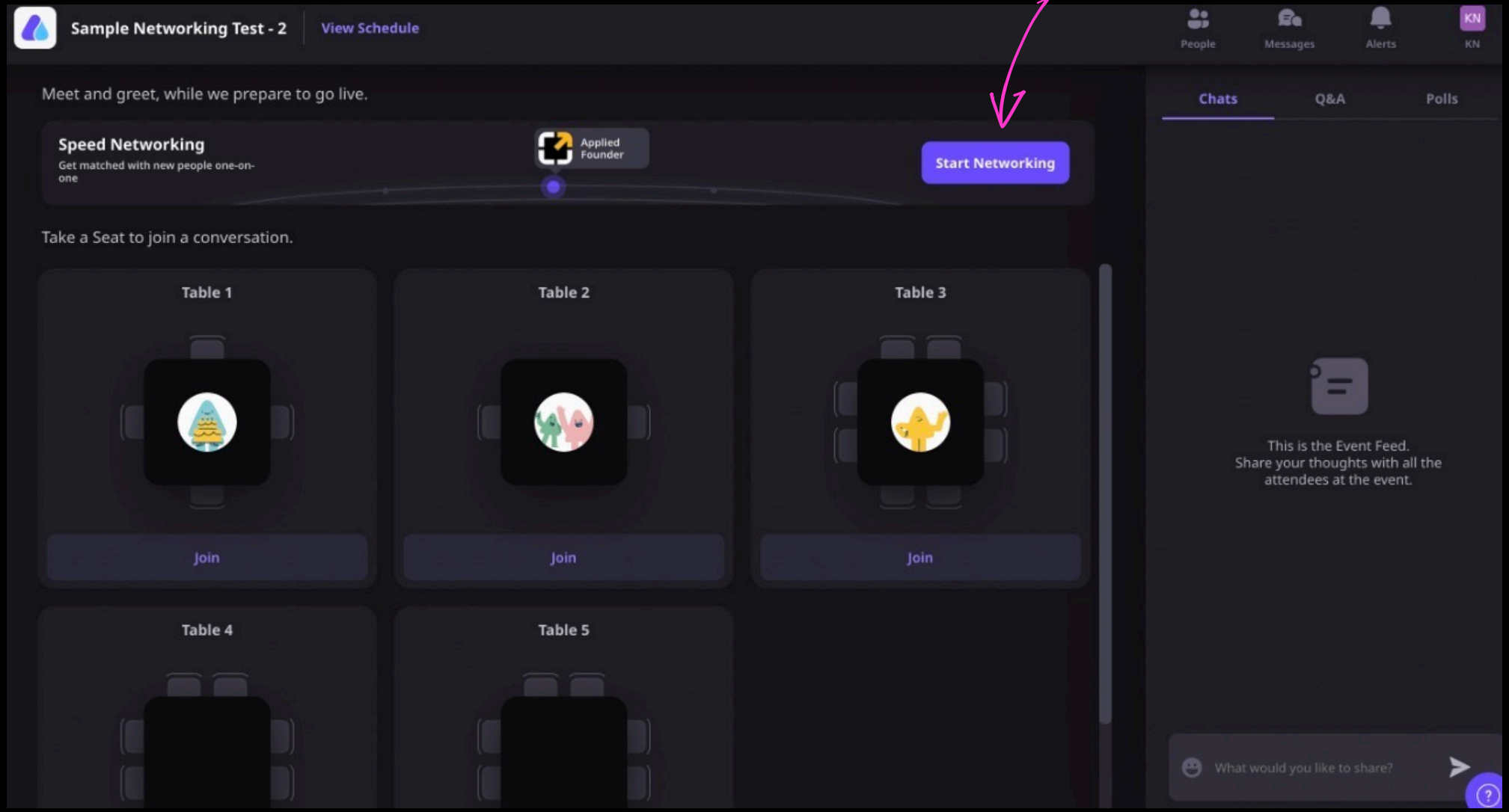
1. Probability (11 & 12th class)
2. Python programming basics

APPLIED ROOTS

Interactive - Learning :

- Quiz Question (timed: 2min / 5min)
  - ↳ scenario-based; Math ; code ; MSQ

- Think - Pair - share
  - ↳ (Speed Networking)


Sample Networking Test - 2
View Schedule
People
Messages
Alerts
KN
Meet and greet, while we prepare to go live.
Speed Networking
Get matched with new people one-on-one
Applied Founder
Start Networking
Take a Seat to join a conversation.
Table 1
Table 2
Table 3
Join
Join
Join
Table 4
Table 5
Chats
Q&A
Polls
This is the Event Feed.
Share your thoughts with all the attendees at the event.
What would you like to share?

APPLIED ROOTS

Instructor – Participant interaction:

- "Q&A"; voting (end of each session)
- "Raise hand" + Mic

Any technical glitches?

↳ chat window

→ we will reconnect shortly

Participants from various backgrounds

↳ we will try & accomadate

APPLIED ROOTS

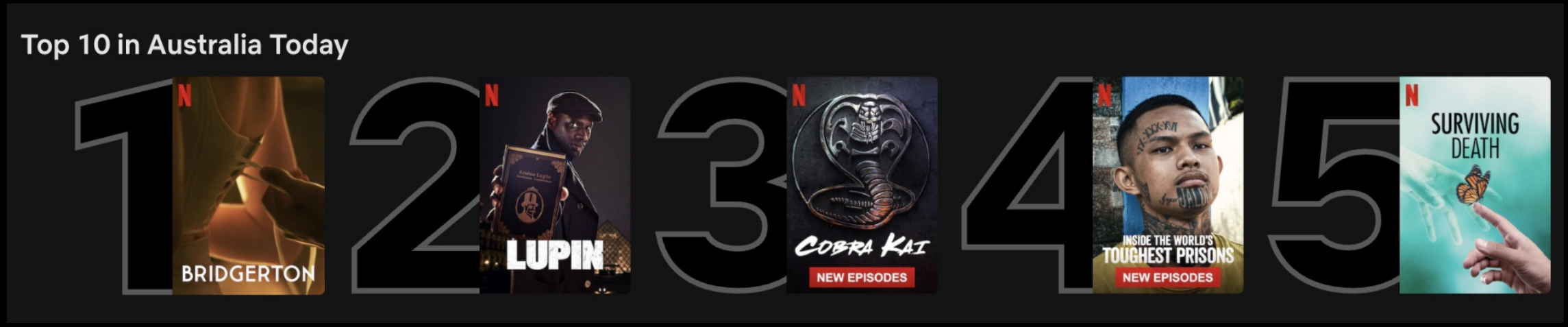
Questions in Q&A

↳ Organization

APPLIED ROOTS

Statistical Testing & Experiments

Real-world Problem @ Netflix



https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962

Does showing Top-10 (in a country) impact time spent on Netflix

Netflix

↳ more time spent ⇒ more renewals

⇓

more revenue

A/B testing

↳ Randomized Controlled Trails (RCT)

→ bucket testing

→ split-run test

Netflix Members

RANDOM

Version ‘A’ (Control)

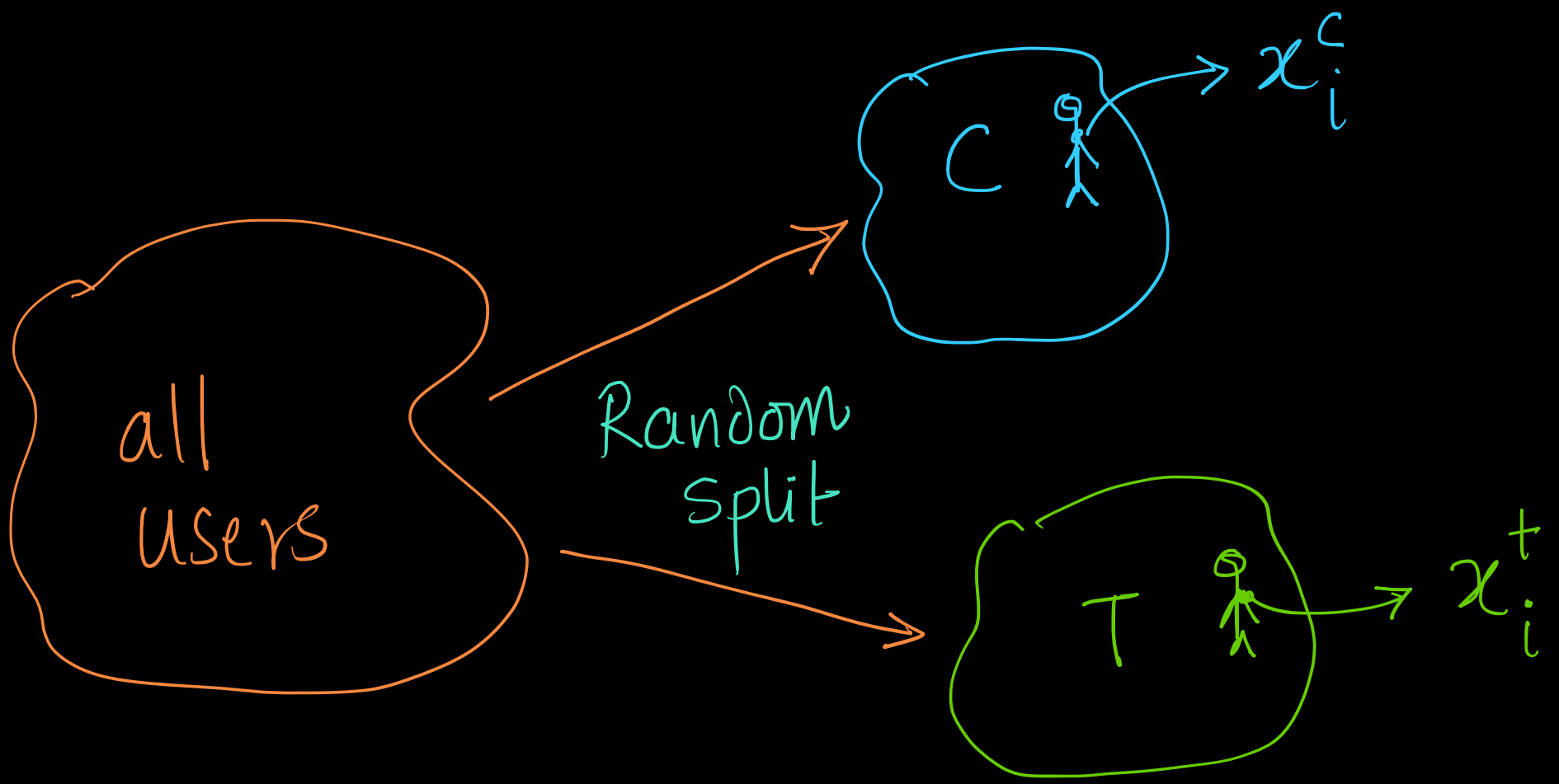
No Top-10

Version ‘B’ (Test)

Compare member behavior

Top-10

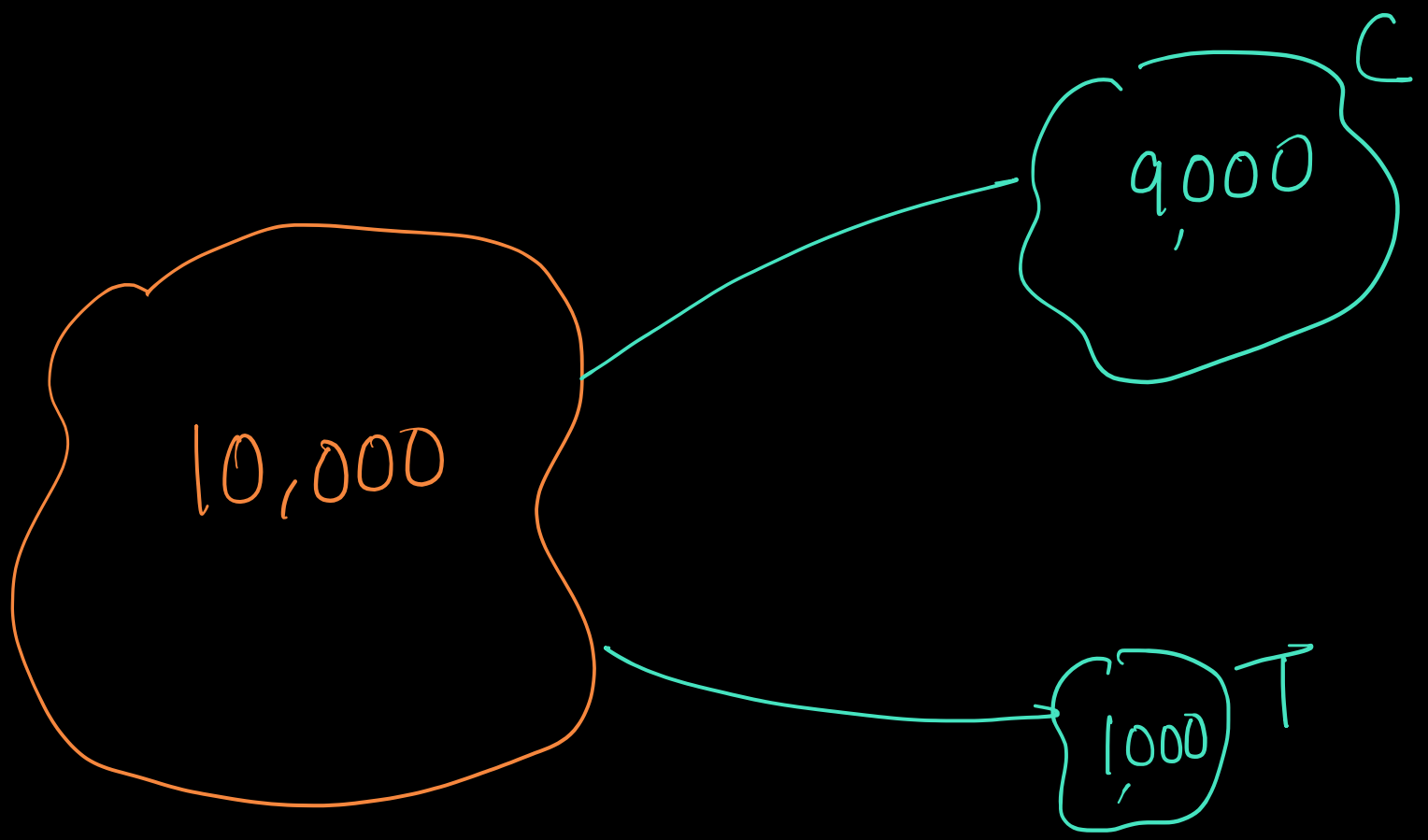
https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962



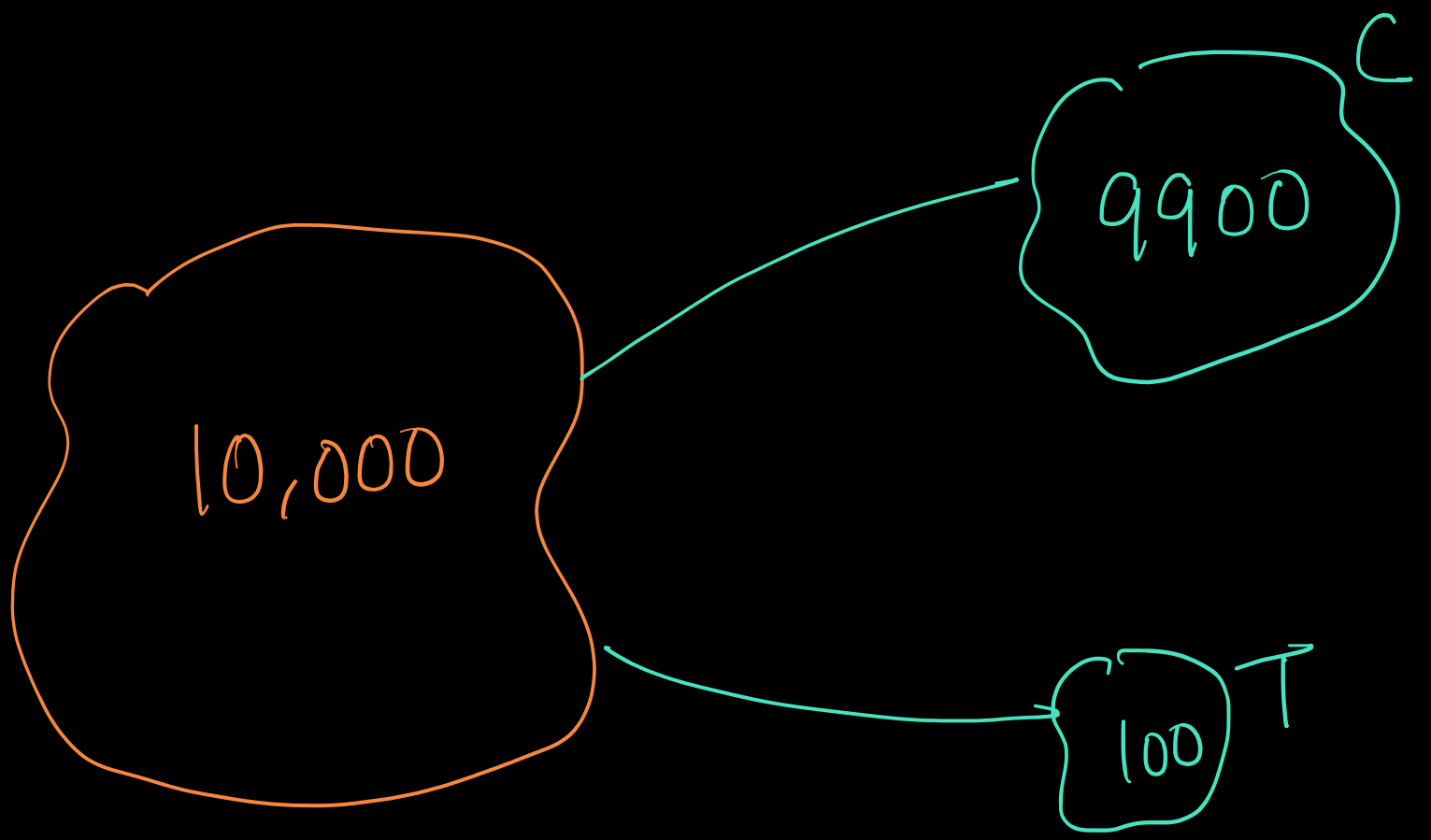
$x_i^{c/t}$: time spent on Netflix

Cost of experimentation

Treatment (B) could be terrible



Small impact



minimal impact

APPLIED ROOTS

(Q) How do we randomly split users into 90:10 ratio using code

↳ write the pseudo-code & paste it in chat

[3 min]

(sol)

For each user u:

$r = \text{random}()$ # uniform r.v

if $r \leq 0.9$:

assign u to Control

else:

assign u to treatment

APPLIED ROOTS

Collect data:

$X^c : x_1^c, x_2^c, \ldots x_n^c$

$X^t : x_1^t, x_2^t, \ldots x_m^t$

n & m can be different

Is X_t "greater" than X_c ?

↓

what does this mean?

APPLIED ROOTS

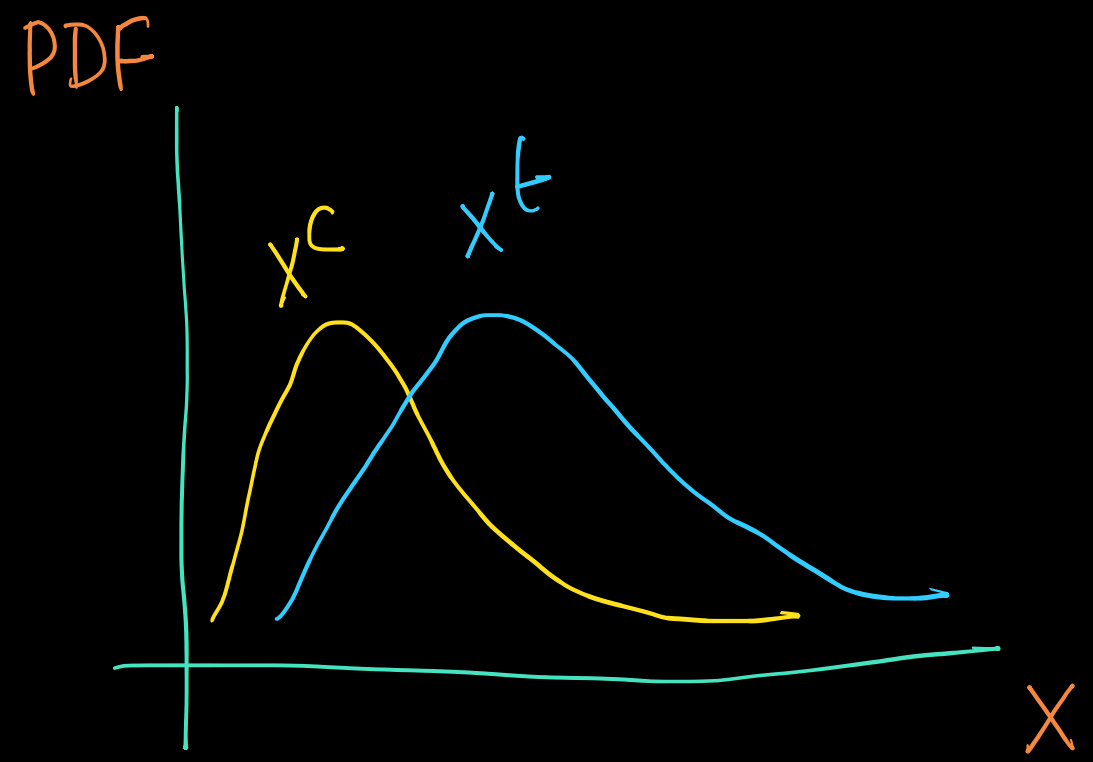
1. $\text{mean}(x^t) > \text{mean}(x^c)$

2. $\text{median}(x^t) > \text{median}(x^c)$

3. $tp99(x^t) > tp99(x^c)$

APPLIED ROOTS

4



many possible & valid methods

Simplest:

median (x^t) vs median (x^c)

(+) robust to outliers

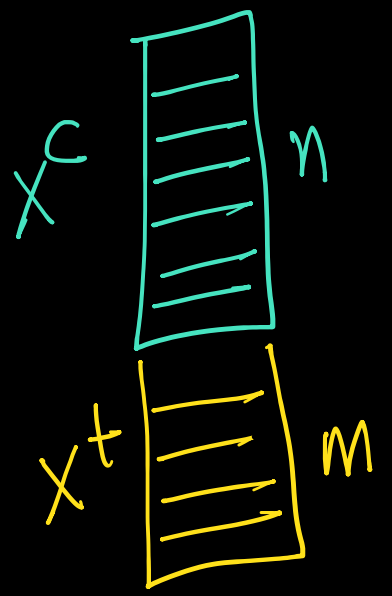
empirical observation:

median (xt) − median (xc)

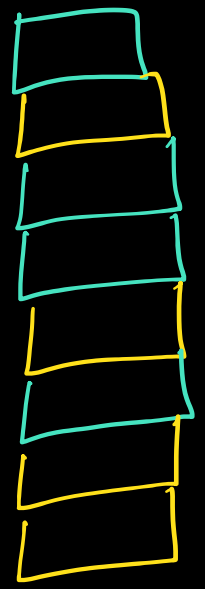
= 20 minutes (let)

↳ median-diff-observed

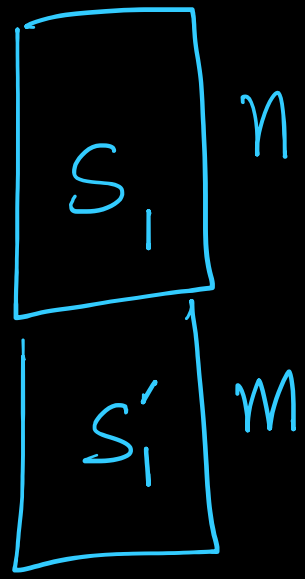
Permutation Testing:



Combine



shuffle /Permute



Split 1

median (s_1)

- median(s_1)

median -diff-1

Why combine, shuffle & split

(***) ↳ simulating if control & treatment data come from the same distribution

median-diff-1

↳ one possible median-diff if C & T are the same with no difference

Repeat k-times

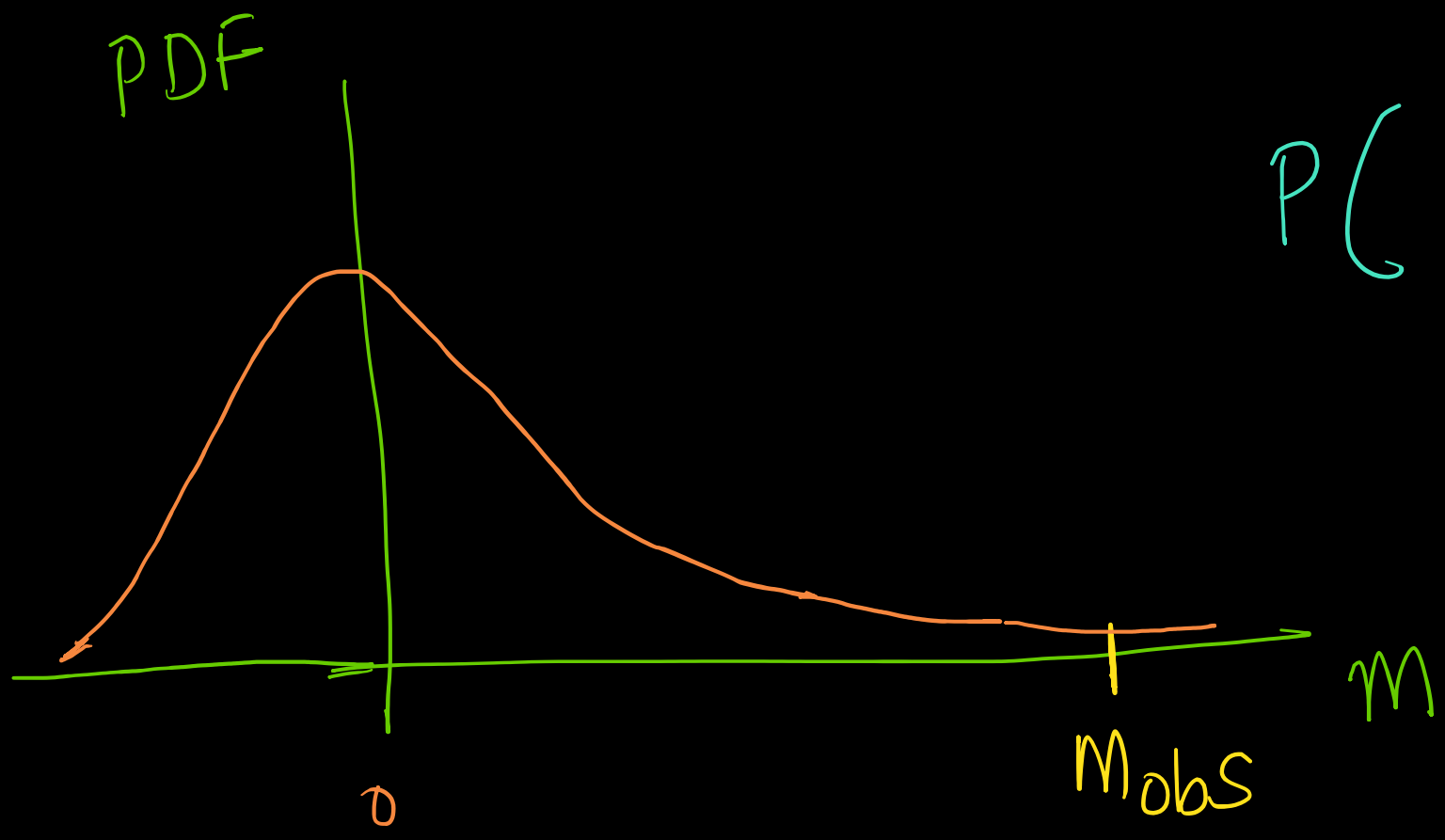
median-diff-1 (M_1)

median-diff-2 (M_2)

⋮

median-diff-k (M_k)

median-diff-observed (M_{obs})

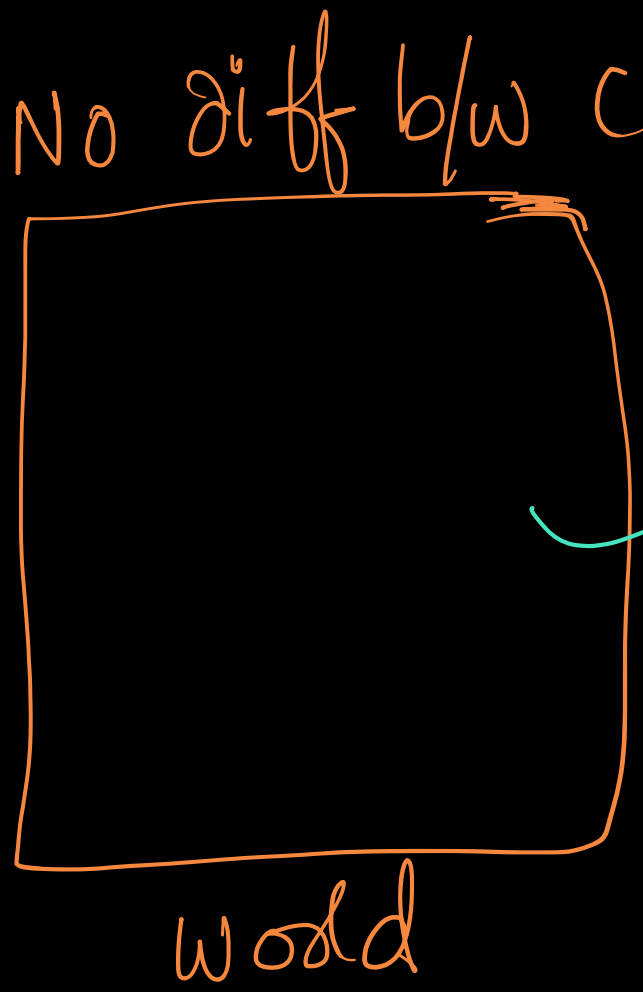


$$P(m \geq m_{obs}) = 0.02$$

Conditional-prob:

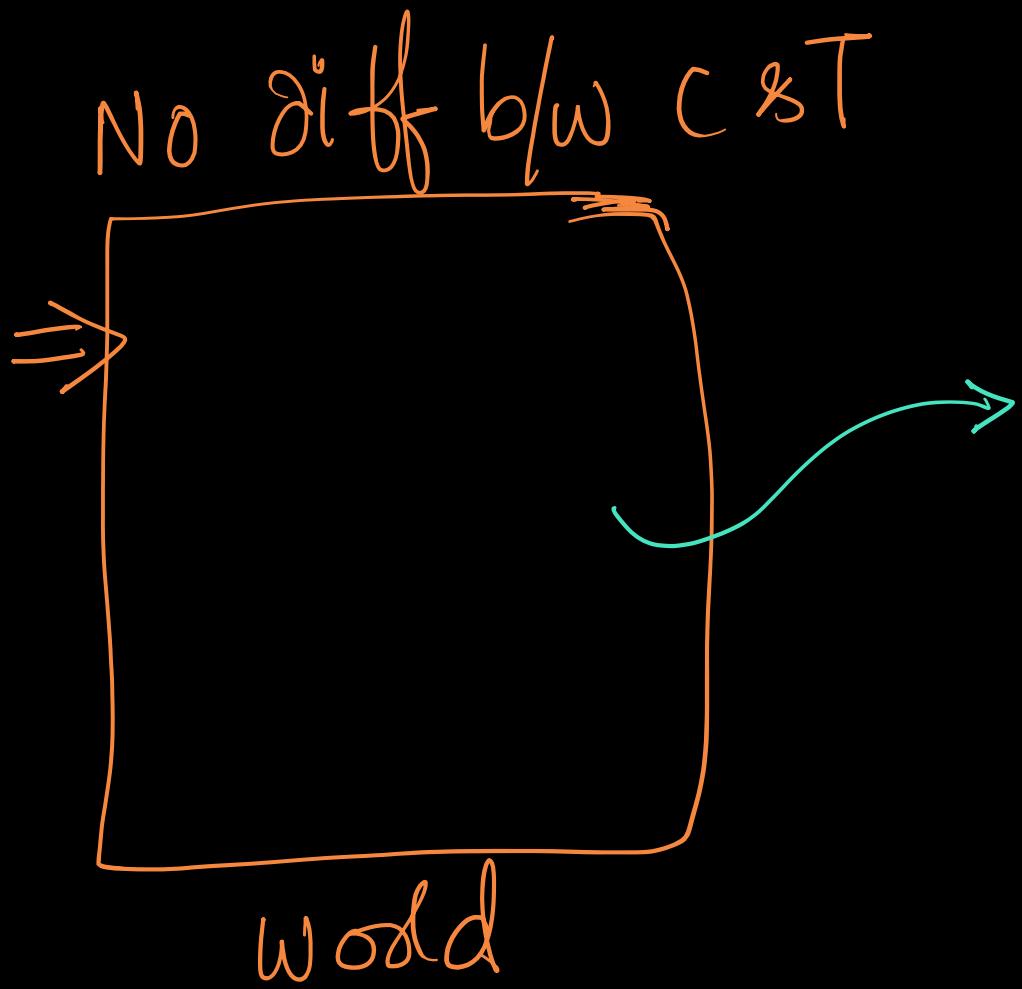
$$P(\text{observing a median-diff} \geq m_{obs} \mid \text{no difference in C \& T})$$

$$= 0.02 = 2\%$$



$p(m \geq 20)$ is very small (2%)

But we observed this in our A/B test



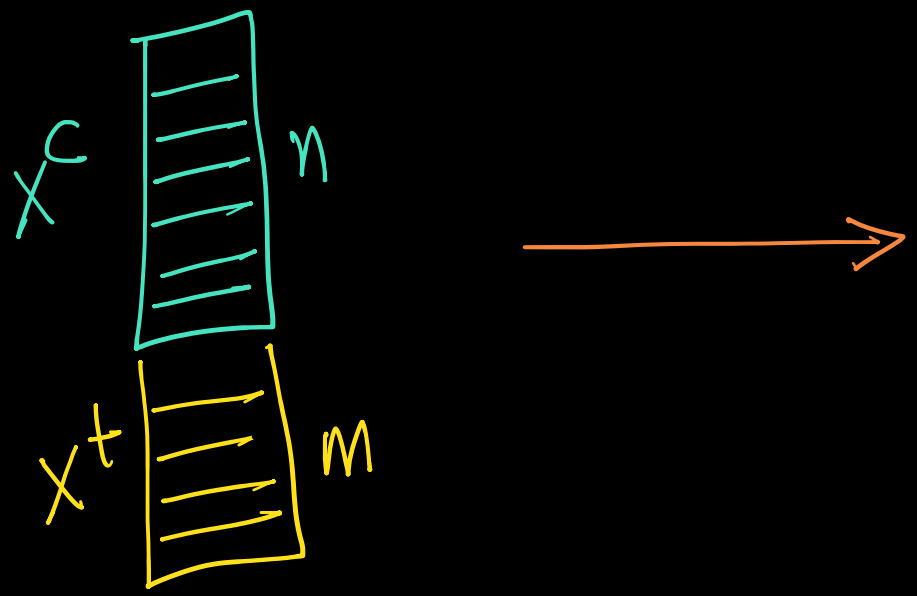
this world is not likely

$\Rightarrow$ C & T 's medians differ

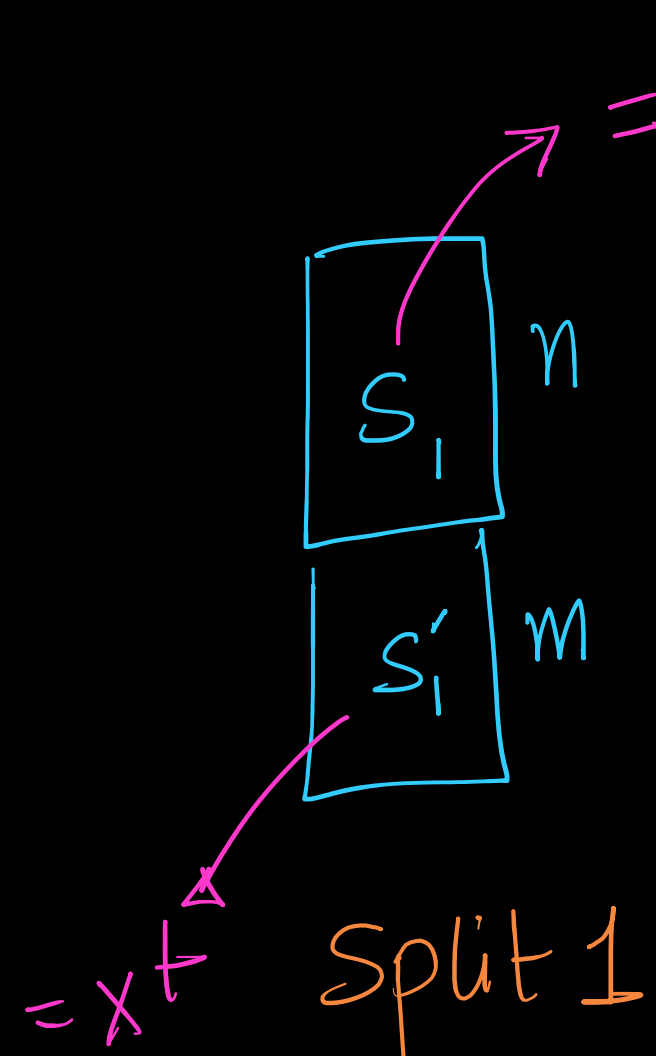
(Q) what would happen if we don't shuffle/permute

↳ provide an answer with justification

No-shuffle



Combine



Split 1

median (s_1)

$-$ median(s_1)

median $-$ diff -1

$\parallel$

Mobs

(Q) How would you shuffle in code?

↳ provide the idea/pseudo-code

Do NOT give the python function.

1, 2, 3, ------ n+m

↓

Permutation (using recursion)

↓

3, 2, 6, 4, -----

Hypothesis Testing (Terminology)

H_0: Null Hypothesis

↳ X^C & X^T are the same

APPLIED ROOTS

Ha: Alternative hypothesis $X^T > X^C$

Test statistic: median (x^t) - median (x^c)

APPLIED ROOTS

p-value:

$$p\left(\text{observing a value as extreme as the empirically observed value} \mid H_0\right)$$

↳ $$p\left(m \geq m_{obs} \mid H_0\right) = 2\%$$

Significance-level: (α)

α : 5% (typically)

if p-value $< \alpha$:

reject H_0

else

Accept H_0

(e.g)

$$P(m \geq m_{obs}) = 0.02 < \alpha = 5\%$$

$$\Downarrow$$

reject H_0 @ 5% significance level

* Accept H_0 @ 1% significance level

Errors:

	H_0 is True	H_a is True
accept H_0	✓	Type II (FN) (β)
accept H_a	Type I (FP)	✓ ($1-\beta$) = Power

Power

↳ how often do we accept H_a when H_a is True

(Q) why did we choose the
Ho as $X^c \approx X^t$?

↳ Provide short explanation in the chat.

if we choose

$H_0: X^c \not\approx X^t$

↳ how do we simulate this H_0?

$H_0 : X^c \approx X^t$

↓

can be simulated using Permutation + Resamplings

APPLIED ROOTS

Code for Permutation Testing

https://colab.research.google.com/drive/1Lp5LW2TiJOtSFBZd37uqMmAghpMLvQRp?usp=sharing

(Q) what happens if $m = n$ = small (100)

& everything else stays the same

$\alpha = 5\%$

$K = 100$

(Q) what happens when k - increases to 1,000 ?

$$m = n = 100$$

$$\alpha = 5\%$$

(Q) what if

$$M = 10000$$

$$n = 1000$$

$$\alpha = 5\%$$

$$k = 100$$

(Q) Medical domain

$m \approx n \approx 30$

(Q) if $n = 30$; $m = 20$

what is the maximum value of k ?

APPLIED ROOTS

$n + m = 50$

_ _ _ _ - - - - _

1 2 3 4 - - - - - 50

$50!$

No - test is perfect

APPLIED ROOTS

Permutation test cannot compare

Assumption:

↳ observations are exchangable

↓

True for Randomized trials

- Super easy to experiment with
- No distributions or complex assumptions
- Test-statistic flexibility

APPLIED ROOTS

Variations:

default

new

which art-work generates more clicks?

https://netflixtechblog.com/selecting-the-best-artwork-for-videos-through-a-b-testing-f6155c4595f6

- Binary variable 1: click 0: noclick

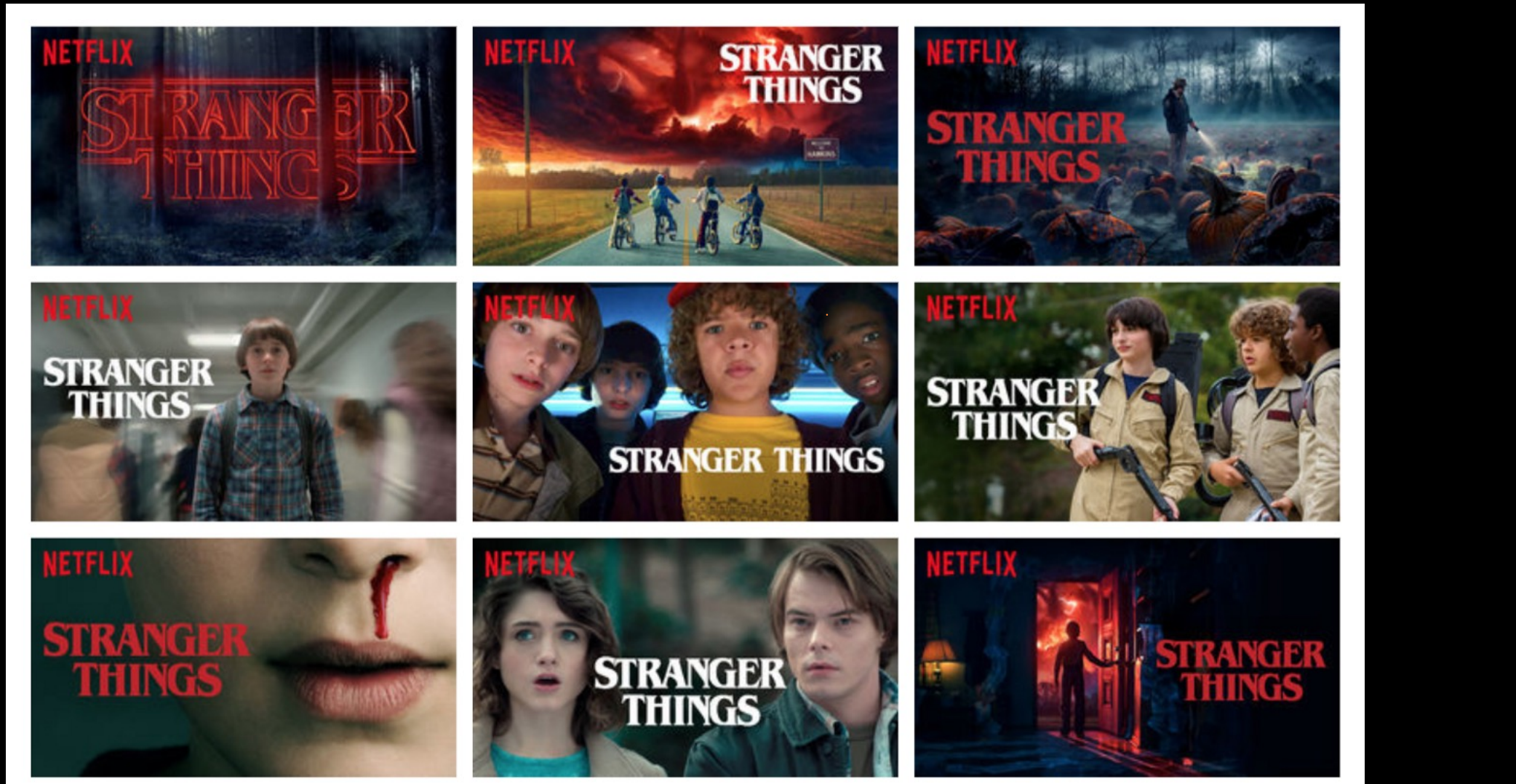
Test - statistic

↳ Click Through rate

Permute - resample - compare

what if we have 9 alternatives ?

APPLIED ROOTS



- A/B Tests can be wasteful

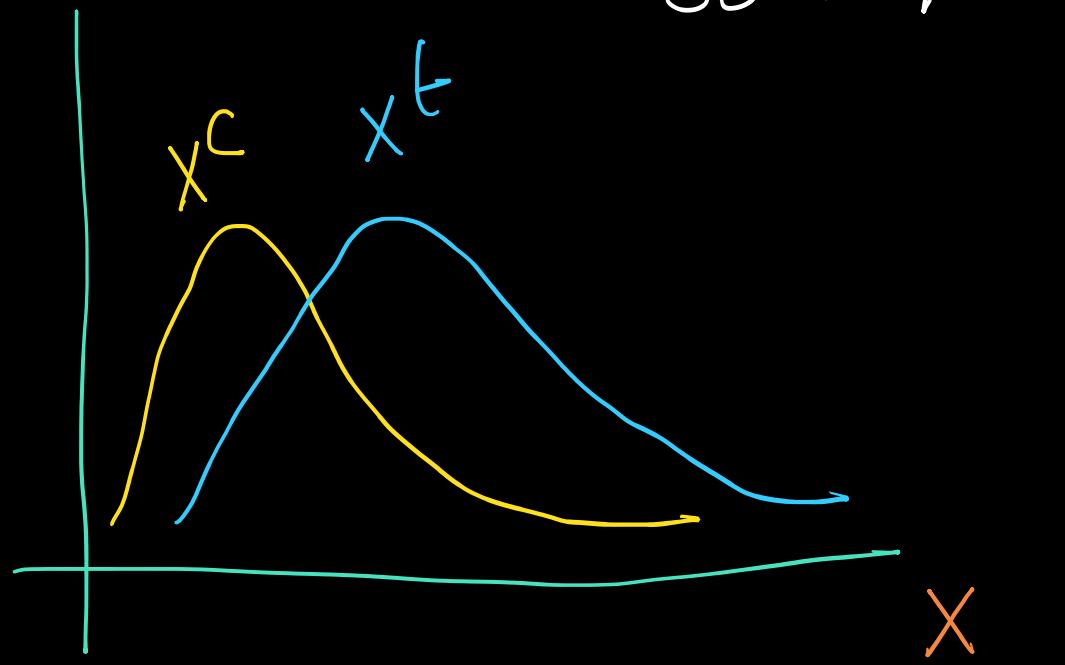
9 x 5% ⇒ 45%

[Multi-arm bandits (later)]

Mann - whitney - U test

- Mann - whitney - wilcoxon (mww)
- rank-sum test
- very - popular

H_0 $x^t \neq x^c$



APPLIED ROOTS

Assumptions

1. x^c & x^t's observations are independent of each other

2. x^c_i & x^t_i values are ordinal $<, =, >$

APPLIED ROOTS

3 $H_0: X^c \approx X^t$

$H_a: X^c \not\approx X^t$

test-statistic (U)

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} s(x_i^c, x_j^t)$$

$$s(x_i^c, x_j^t) = \begin{cases} 1 & \text{if } x_i^c > x_j^t \\ 1/2 & \text{if } x_i^c = x_j^t \\ 0 & \text{if } x_i^c < x_j^t \end{cases}$$

↳ s_{ij}

APPLIED ROOTS

- compare every pair x_i^c & x_j^t

x^c	x^t
1	1
6	6
9	9

$$(1/2 + 0 + 0)$$
$$+ (1 + 1/2 + 0)$$
$$+ (1 + 1 + 1/2)$$

$$= 4.5$$

$$= \frac{m \times n}{2}$$

Compare every pair x_i^c & x_j^t

x^c	x^t
1	2
6	12
9	18

$$(0+0+0) + (1+0+0) + (1+0+0) = 2$$

Compare every pair x_i^c & x_j^t

x^c	x^t
1	0
6	1
9	2

$$(1 + 1/2 + 0) + (1 + 1 + 1) + (1 + 1 + 1) = 7.5$$

if n & $m \geq 20$; under H_0

$$U \sim \text{Normal}(\mu_U, \sigma_U)$$

$$\mu_U = \frac{n \times m}{2}$$

$$\sigma_U = \sqrt{\frac{nm(n+m+1)}{12}}$$

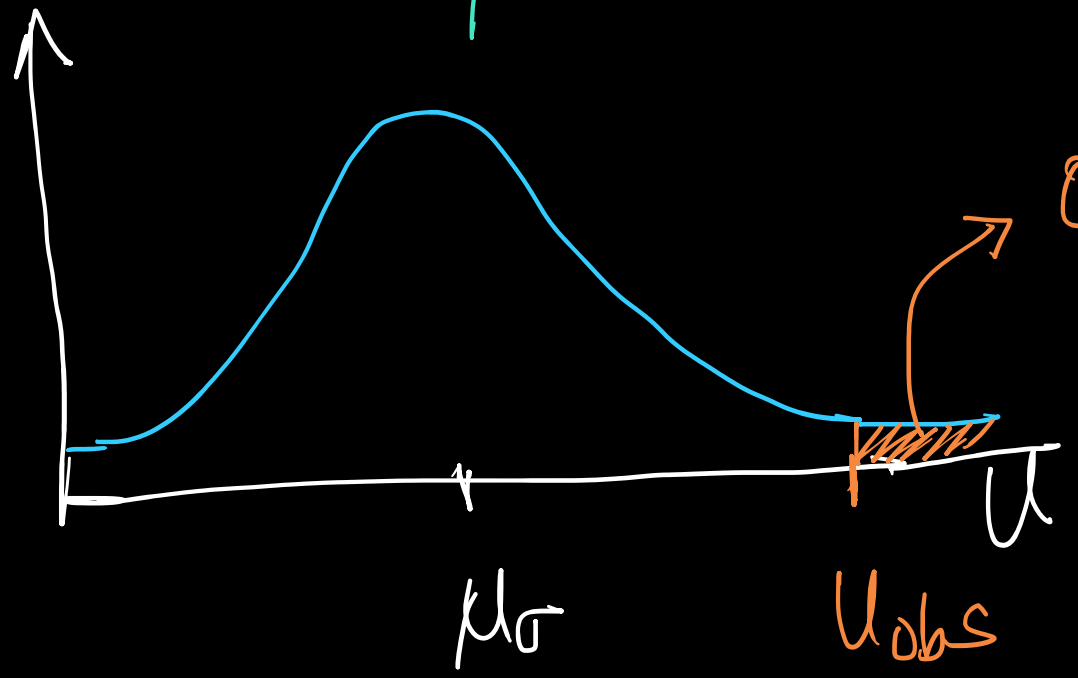
Hypothesis testing:

① $H_0: X^c \approx X^t$ $\quad H_a: X^c \not\approx X^t$

② Compute U_obs

APPLIED ROOTS

(3) $U \sim Norm(\mu_U, \sigma_U)$ under H_0



APPLIED ROOTS

④ if $\alpha = 5\%$

reject H_0

APPLIED ROOTS

Note:

↳ We know the distribution of test-statistic

$P(\text{test-statistic} > \text{obs} \mid H_0)$ is easy to compute

APPLIED ROOTS

Permutation Testing

↳ P(test statistic > obs | H_0)

Simulation (resampling)

APPLIED ROOTS

MWW

- widely used
- non-parametric
- every pair is compared
- cannot say $x^c > x^t$ or $x^t > x^c$

(Q) what is the max value of test-statistic U & when does that occur?

x^c	x^t
10	1
20	2
30	3

$$(1+1+1)$$
$$+(1+1+1)$$
$$+(1+1+1)$$
$$=9$$

$m \times n$

APPLIED ROOTS

Python - code

(Q) Small sample sizes

Permutation vs MWW

(median)

Hypothesis Testing (recap)

- Ho & Ha
- assumptions
- test-statistic & its distribution
- accept or reject

Multiple - Testing

$$x^1 \sim x^2$$
$$x^1 \sim x^3$$
$$x^1 \sim x^4$$
$$\vdots$$
$$x^1 \sim x^n$$

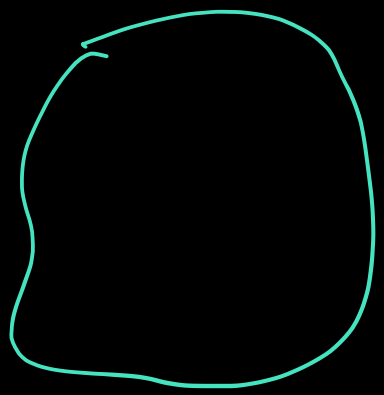
$\alpha = 5\%$

n is large

"ERRORS"

HOLM - BONFERRI Correction

out-of-sample randomized tests



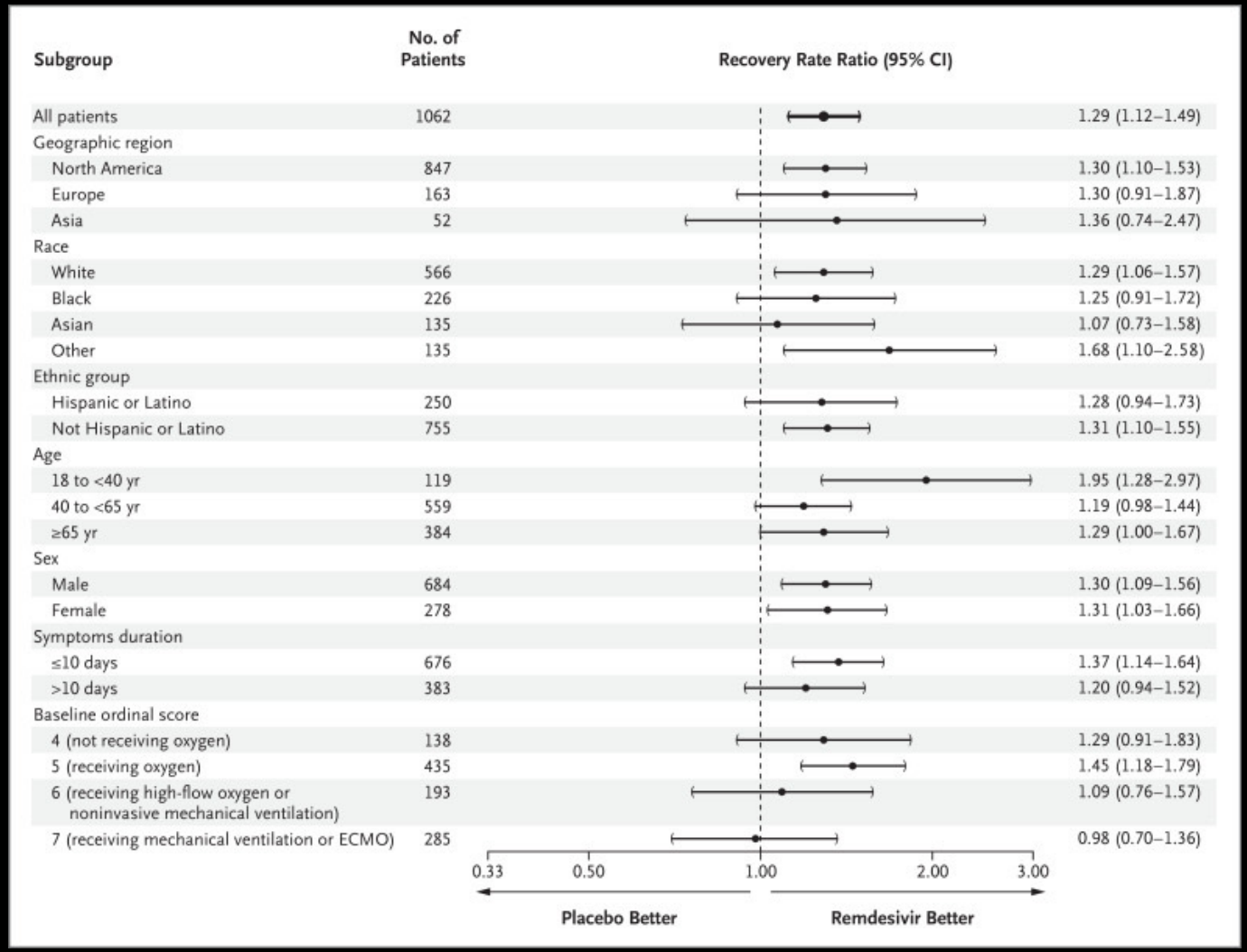
A (build hypotheses using this)

B (test on this)

Boot-strapping & Confidence Intervals

APPLIED ROOTS

https://pubmed.ncbi.nlm.nih.gov/32445440/



Subgroup	No. of Patients	Recovery Rate Ratio (95% CI)
All patients	1062	1.29 (1.12–1.49)
Geographic region		
North America	847	1.30 (1.10–1.53)
Europe	163	1.30 (0.91–1.87)
Asia	52	1.36 (0.74–2.47)
Race		
White	566	1.29 (1.06–1.57)
Black	226	1.25 (0.91–1.72)
Asian	135	1.07 (0.73–1.58)
Other	135	1.68 (1.10–2.58)
Ethnic group		
Hispanic or Latino	250	1.28 (0.94–1.73)
Not Hispanic or Latino	755	1.31 (1.10–1.55)
Age		
18 to <40 yr	119	1.95 (1.28–2.97)
40 to <65 yr	559	1.19 (0.98–1.44)
≥65 yr	384	1.29 (1.00–1.67)
Sex		
Male	684	1.30 (1.09–1.56)
Female	278	1.31 (1.03–1.66)
Symptoms duration		
≤10 days	676	1.37 (1.14–1.64)
>10 days	383	1.20 (0.94–1.52)
Baseline ordinal score		
4 (not receiving oxygen)	138	1.29 (0.91–1.83)
5 (receiving oxygen)	435	1.45 (1.18–1.79)
6 (receiving high-flow oxygen or noninvasive mechanical ventilation)	193	1.09 (0.76–1.57)
7 (receiving mechanical ventilation or ECMO)	285	0.98 (0.70–1.36)

Why has healthcare moved to C.I ?

↳ Intervals over point estimates

↓

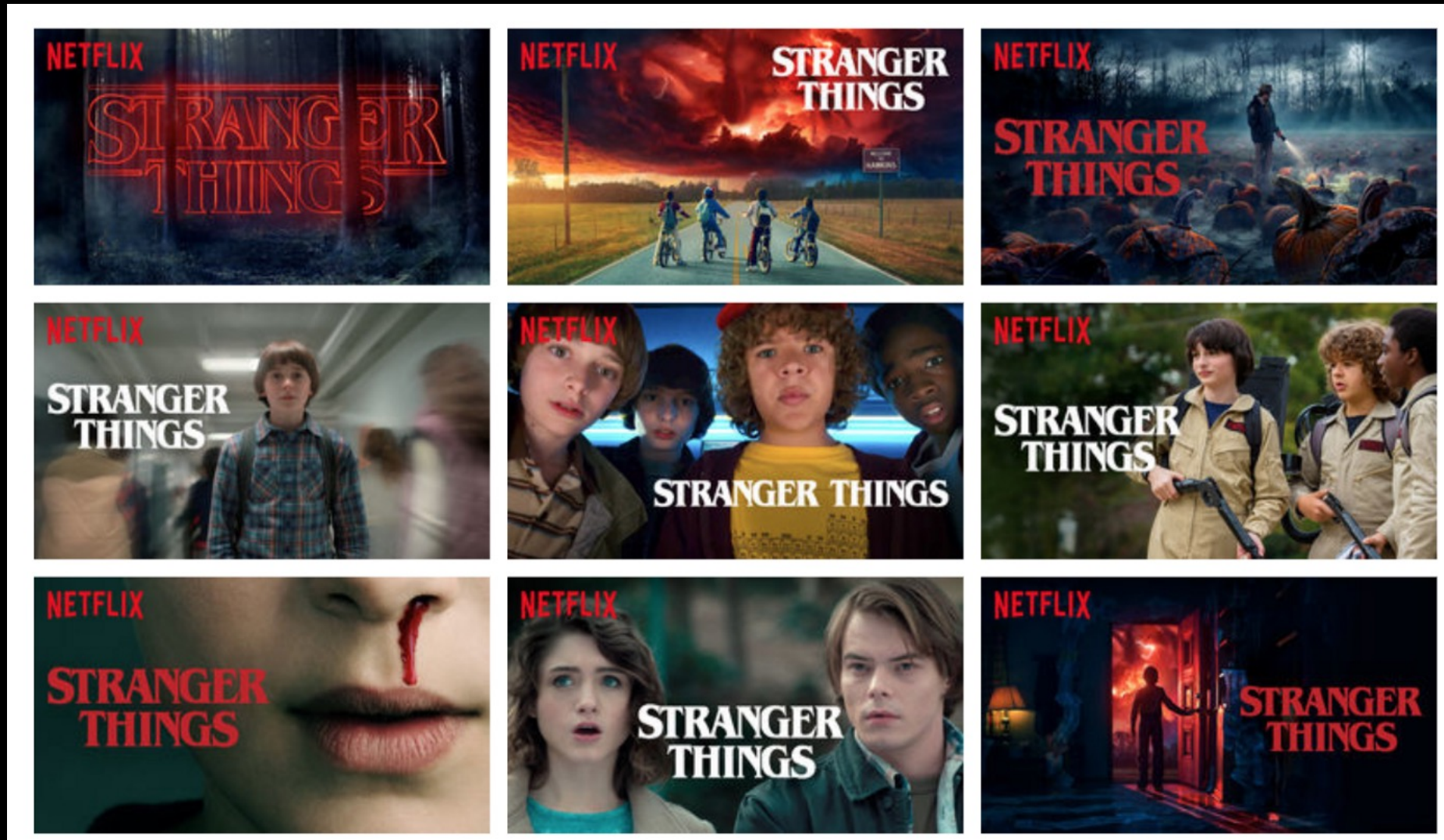
(distribution)

1000's of simultaneous experiments

APPLIED ROOTS

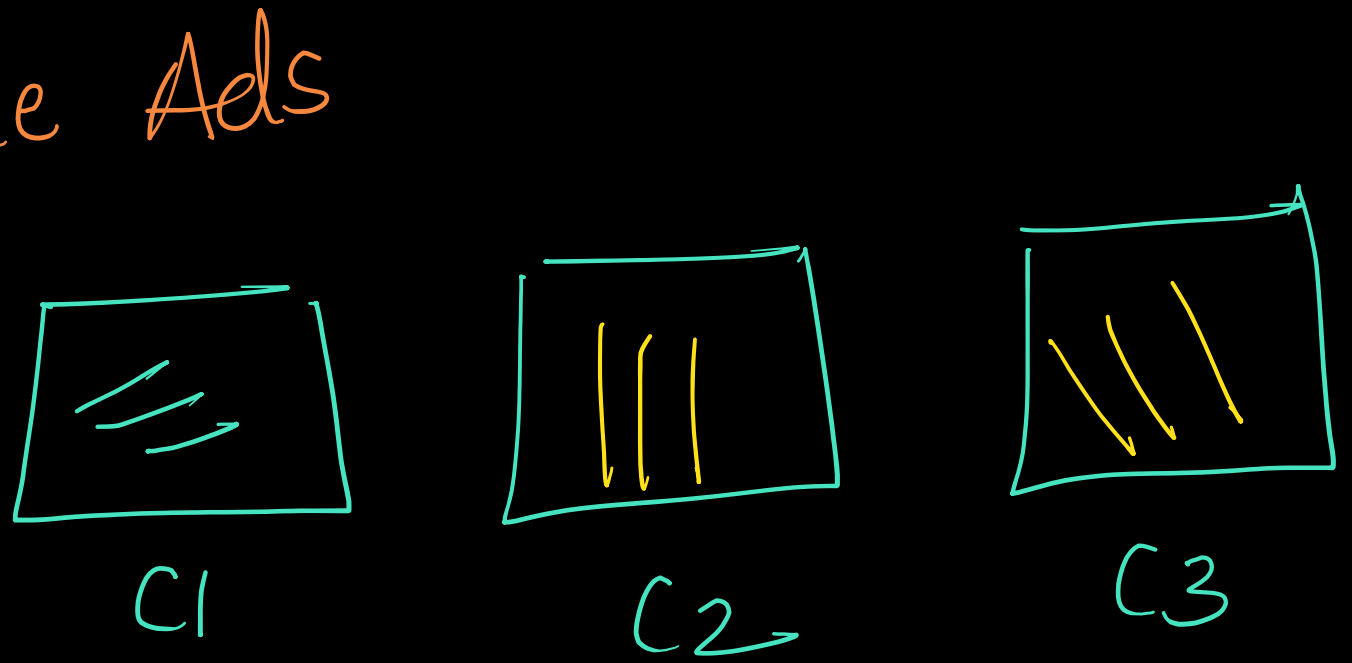
Multi-arm-Bandits

APPLIED ROOTS



APPLIED ROOTS

Google Ads



APPLIED ROOTS

https://towardsdatascience.com/beyond-a-b-testing-multi-armed-bandit-experiments-1493f709f804

APPLIED ROOTS

MAB

- ϵ-greedy
- Bayesian Thompson Sampling
- Contextual Bandits

APPLIED ROOTS

Chi-square Test

t-test

Dive-deep into
Multi-arm-bandits

} Future Sessions

APPLIED ROOTS

Additional References for t-test & Chi-square test

https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample#error-probabilities-and-power

https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/chi-square-distribution-introduction

https://youtu.be/WXPBoFDqNVk

https://machinelearningmastery.com/how-to-code-the-students-t-test-from-scratch-in-python/