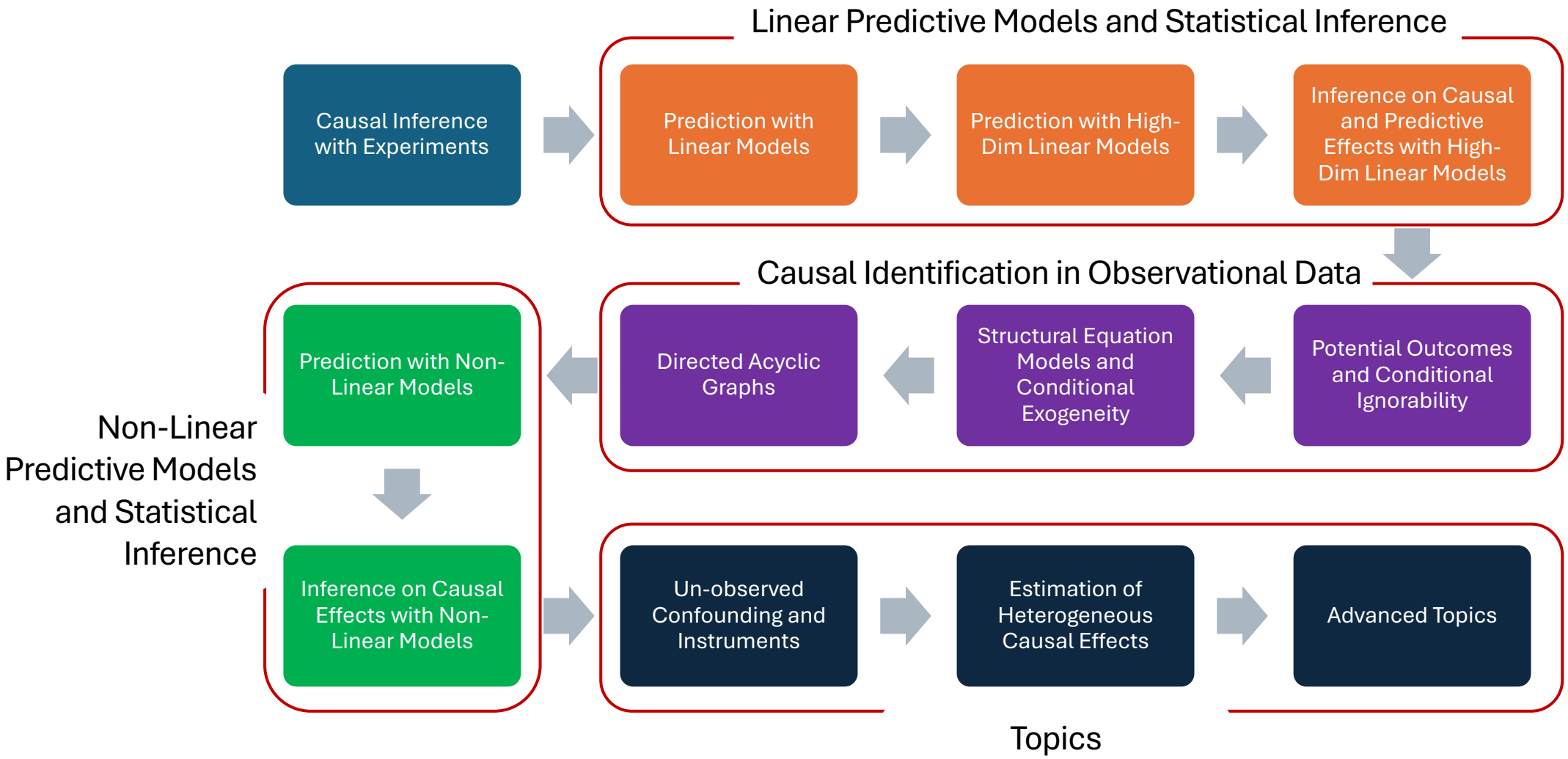
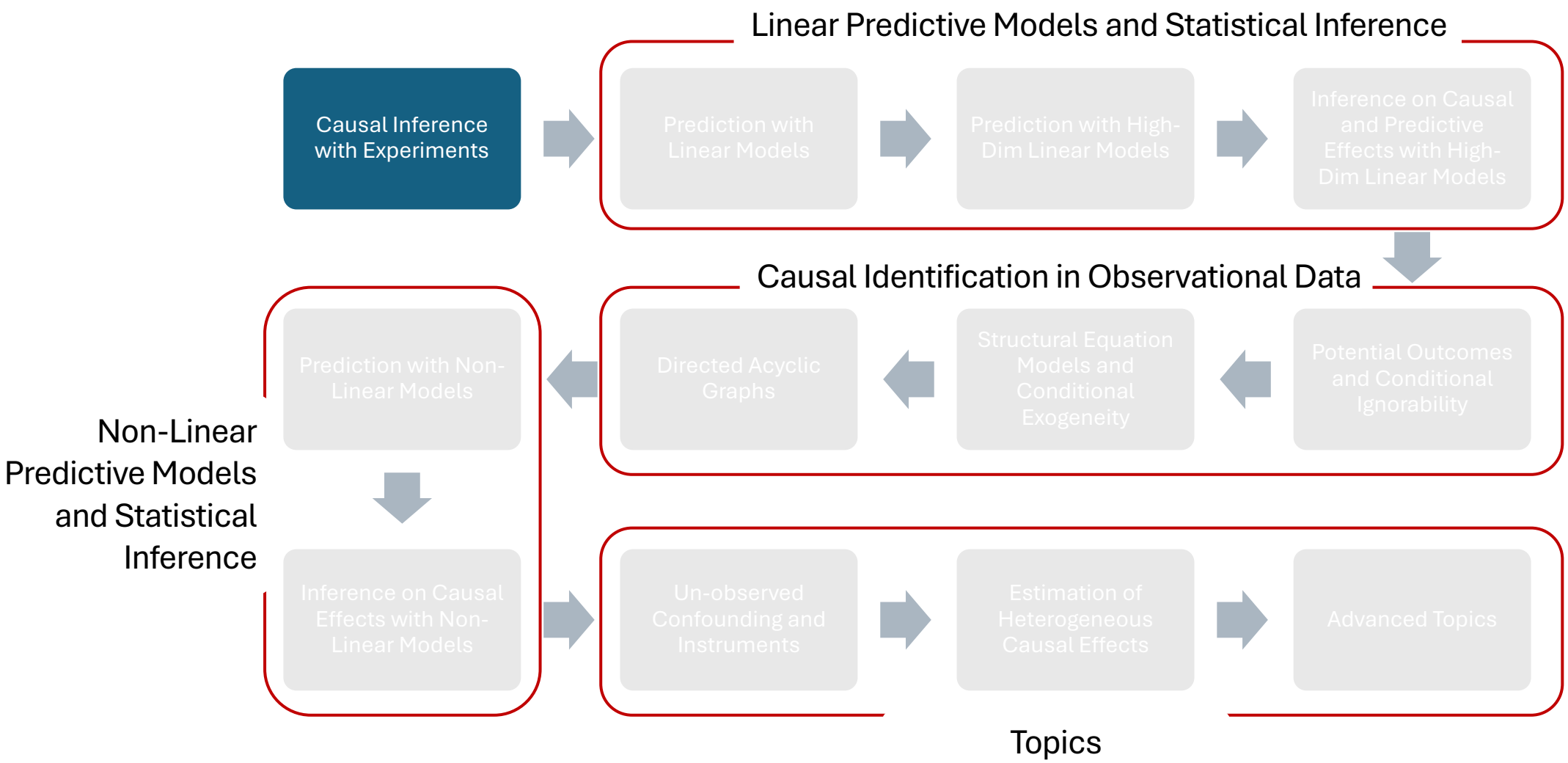
MS&E228: Lecture 2
Causality via Experiments

Vasilis Syrgkanis
Assistant Professor
Management Science and Engineering
(by courtesy) Computer Science and Electrical Engineering
Institute for Computational and Mathematical Engineering


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Non-Linear Predictive Models and Statistical Inference
Prediction with Non-Linear Models
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Estimation of Heterogeneous Causal Effects
Advanced Topics
Topics


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Prediction with Non-Linear Models
Non-Linear Predictive Models and Statistical Inference
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Estimation of Heterogeneous Causal Effects
Advanced Topics
Topics

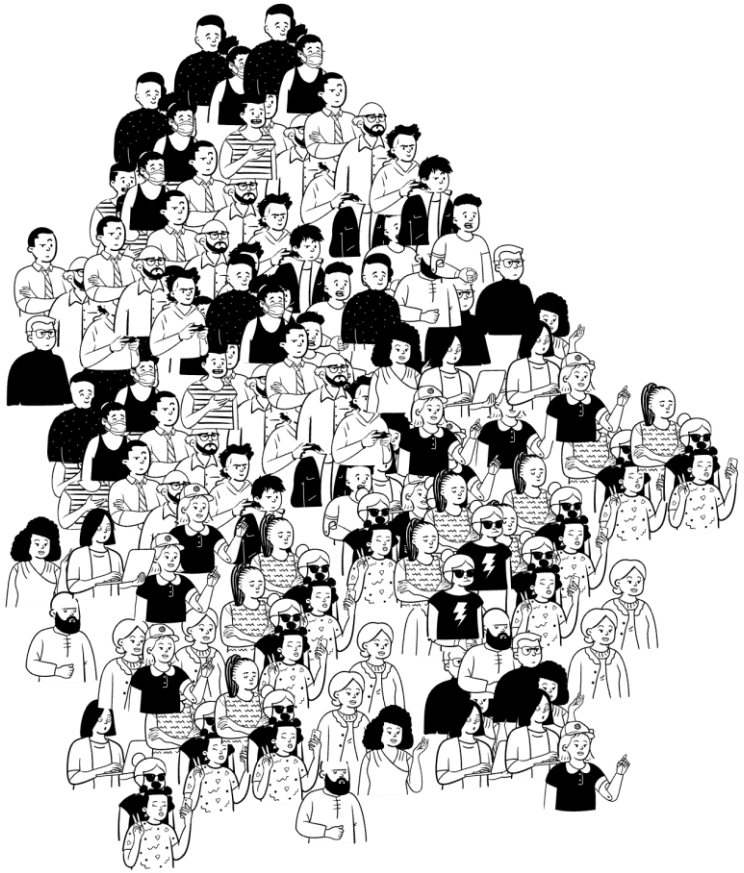
The Basics of A/B Testing

Randomization, Causality, Statistical Inference

The Mechanics

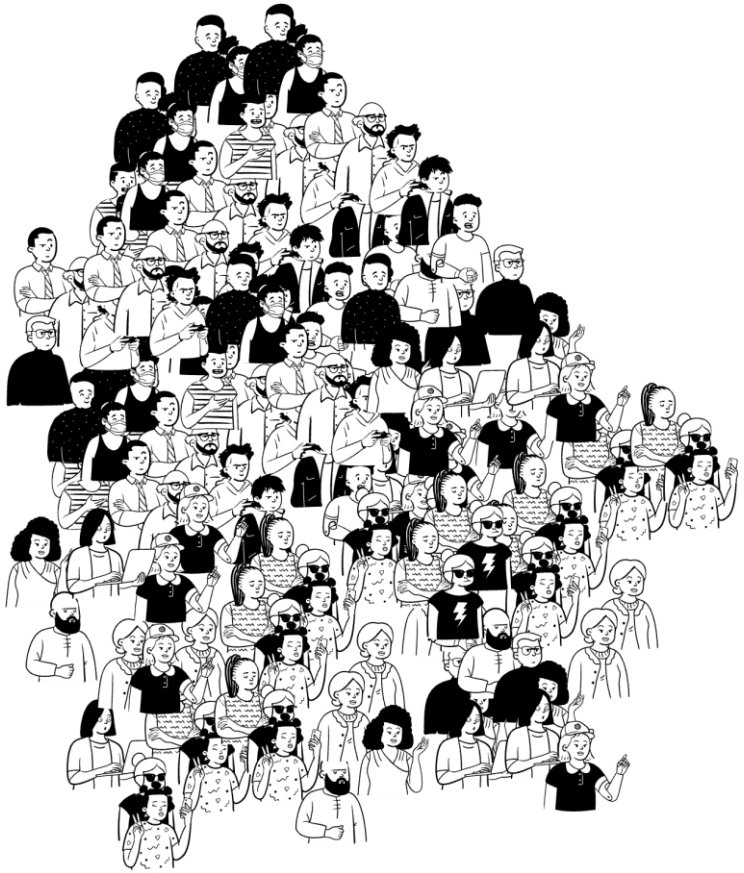
A/B Testing

user base



A/B Testing

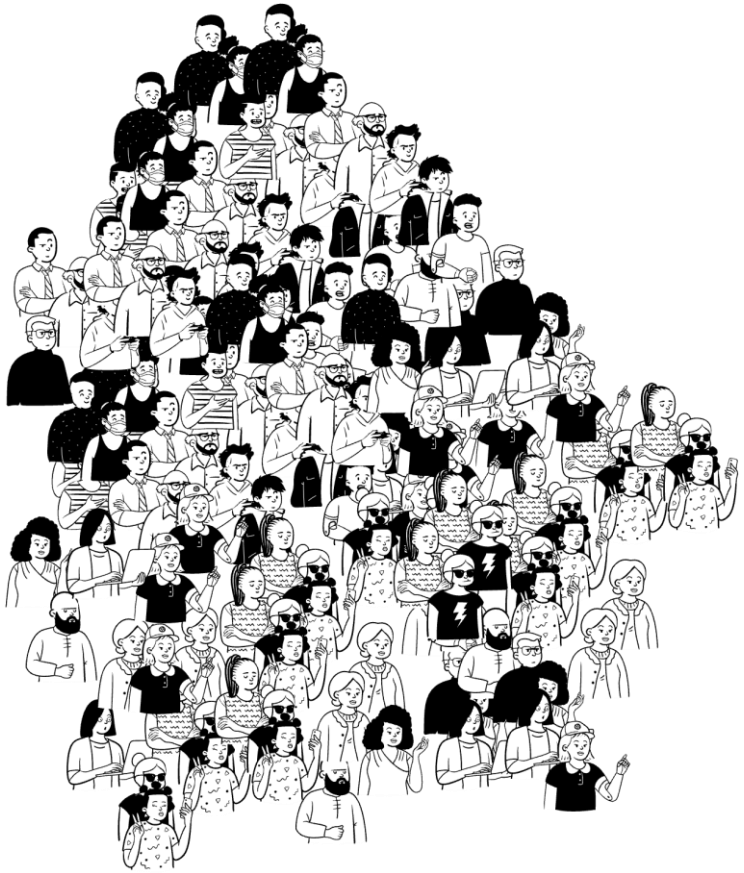
user base



sample

A/B Testing

user base

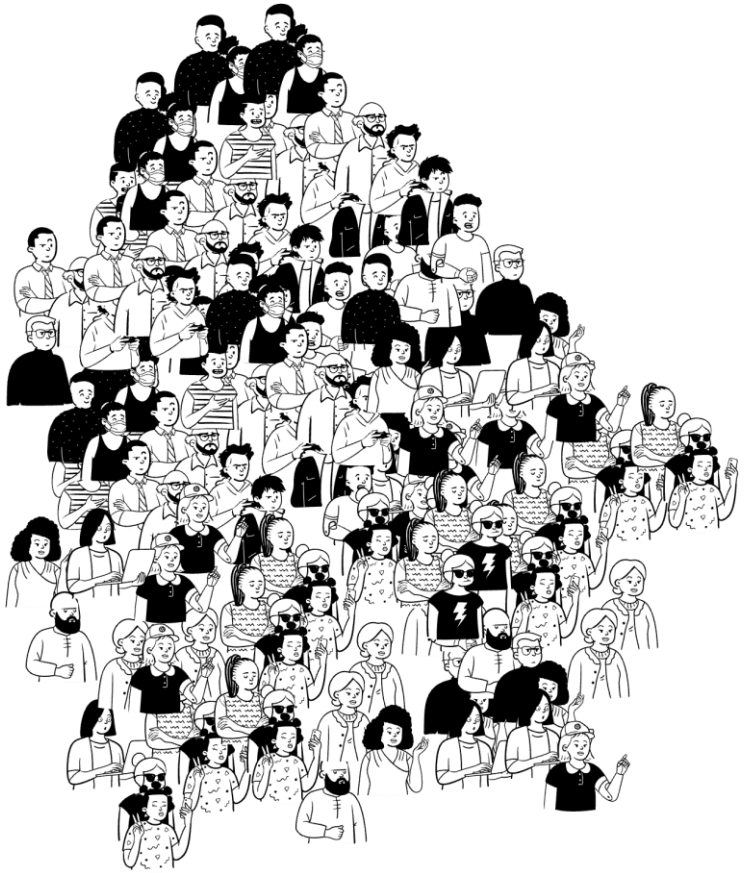


sample

flip a coin for each user

A/B Testing

user base

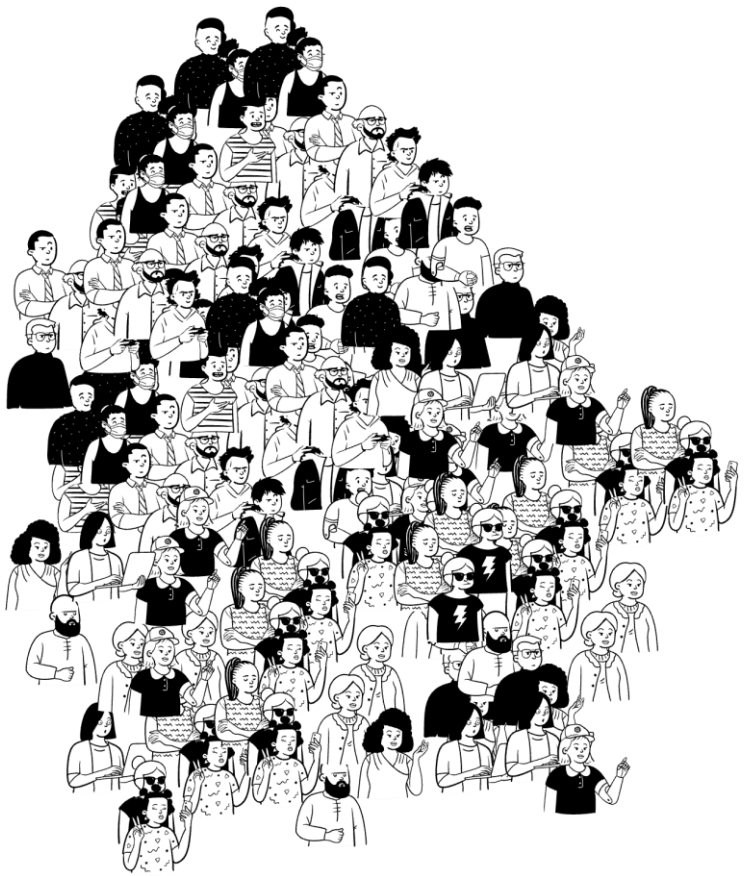


sample

split into groups based on coin

A/B Testing

user base



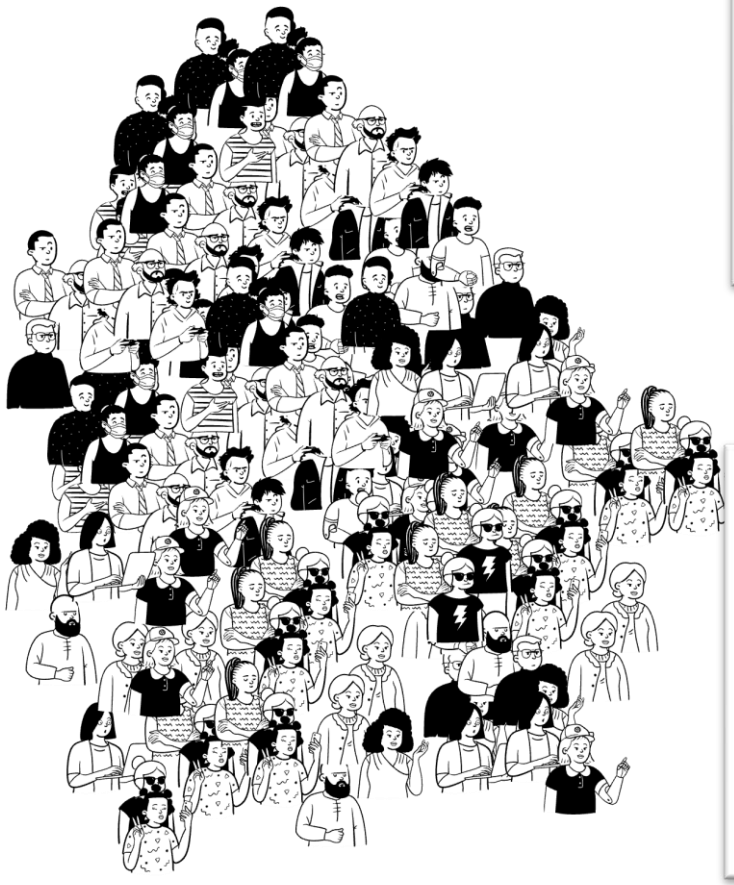
Group A

Group B

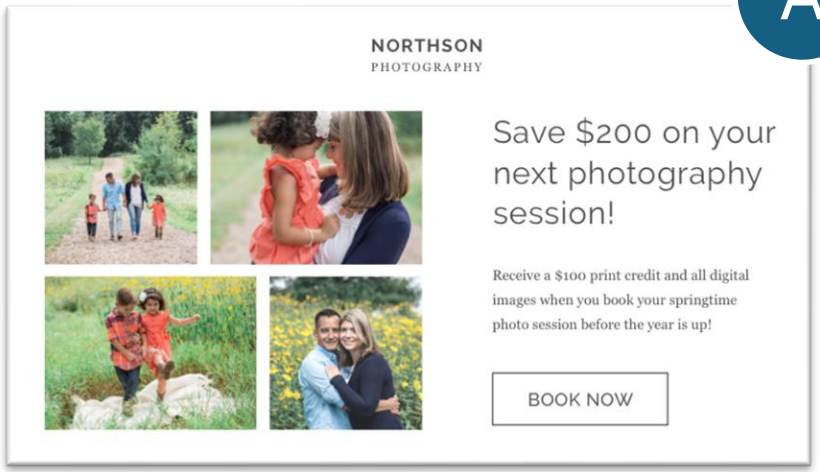
Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

user base

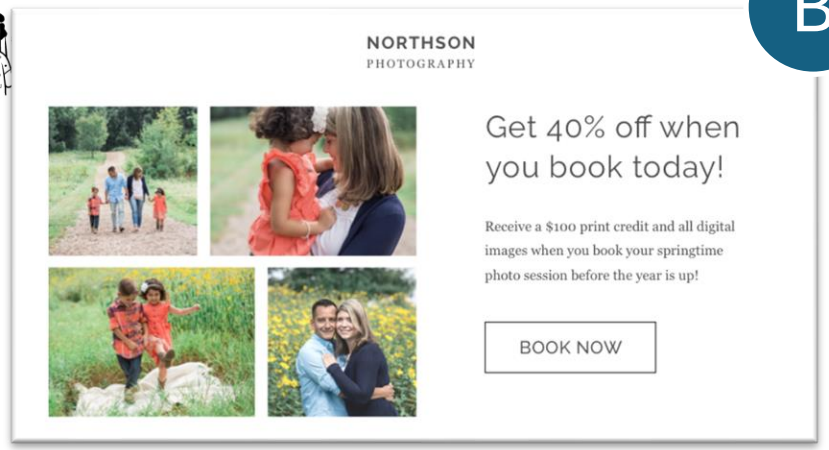


A



Group A

B

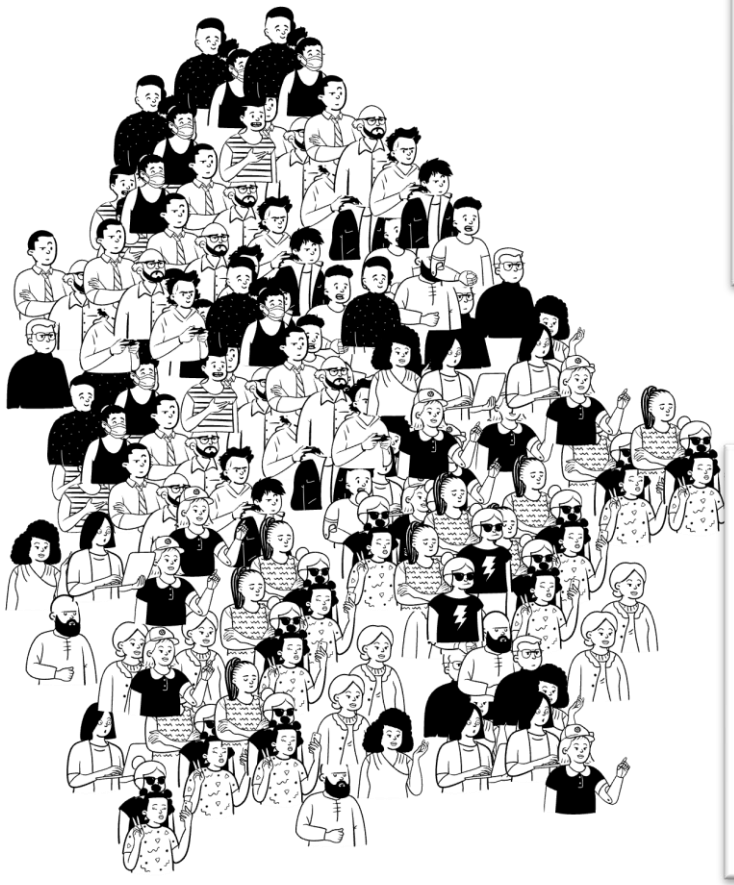


Group B

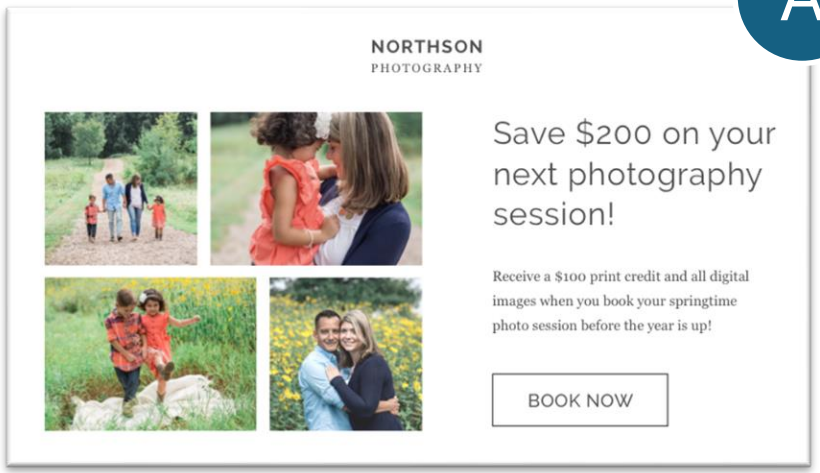
Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

user base

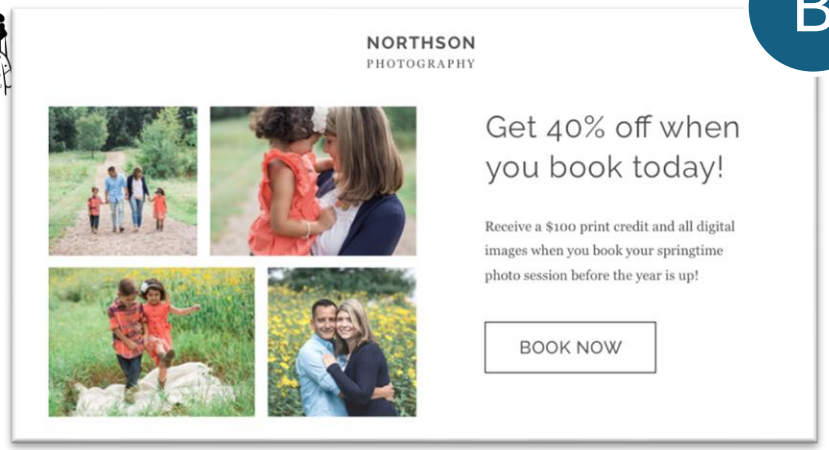


A



Group A

B

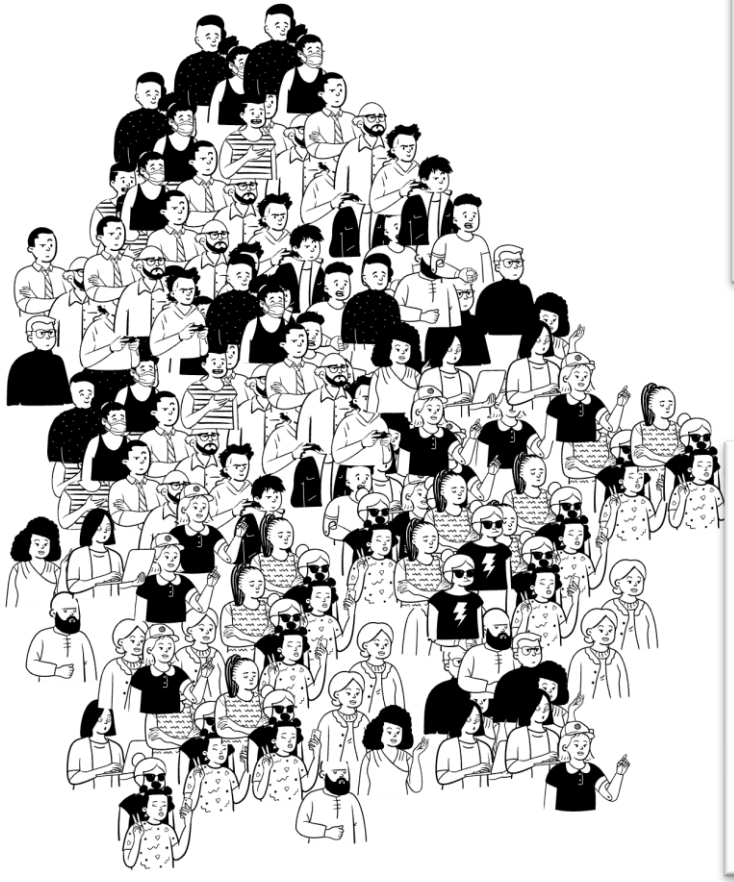


Group B

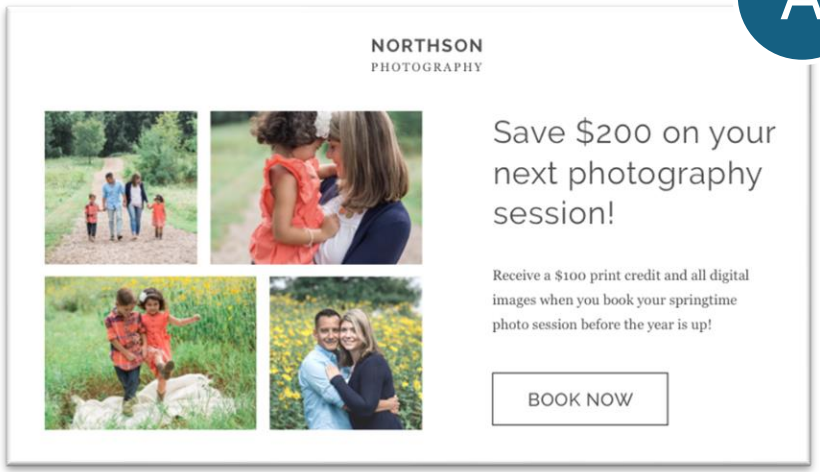
Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

user base

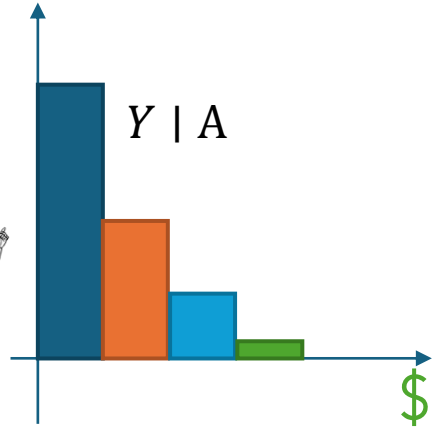


A



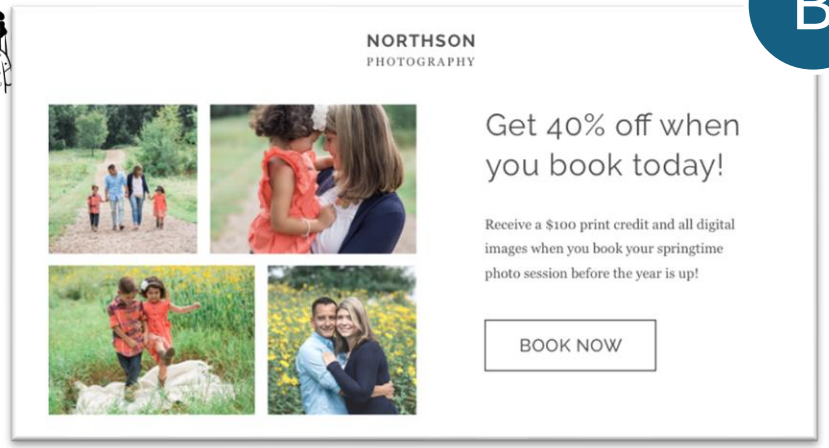
Group A

% of people



$\mu_A = 10\$$ (average spend)

B

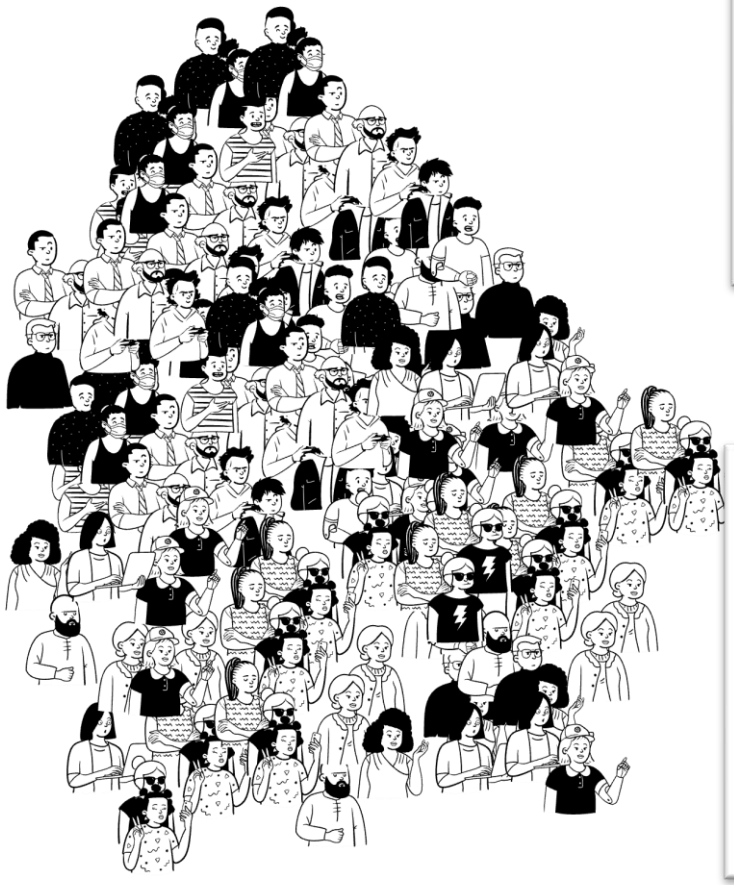


Group B

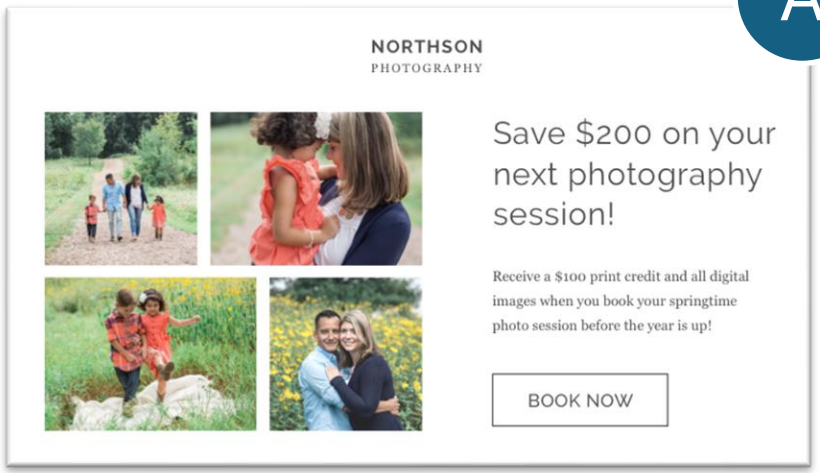
Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

user base

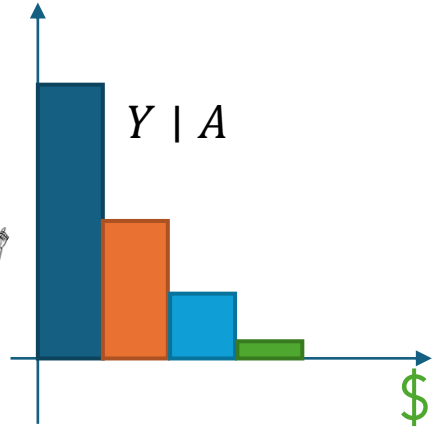


A



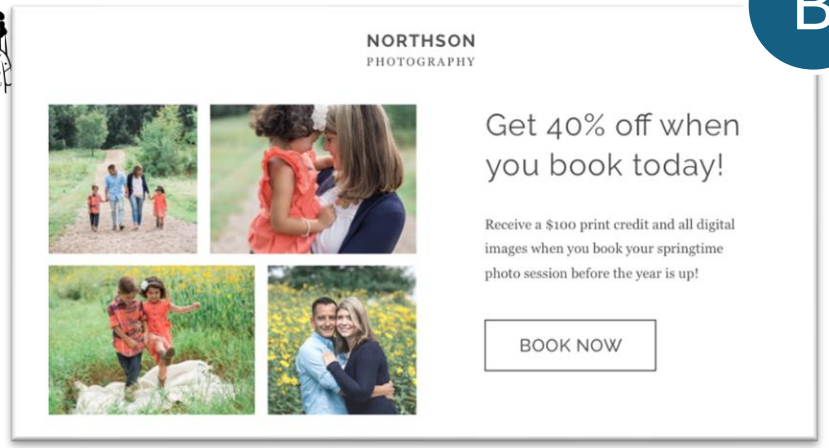
Group A

% of people



$\mu_A = 10\$$ (average spend)

B

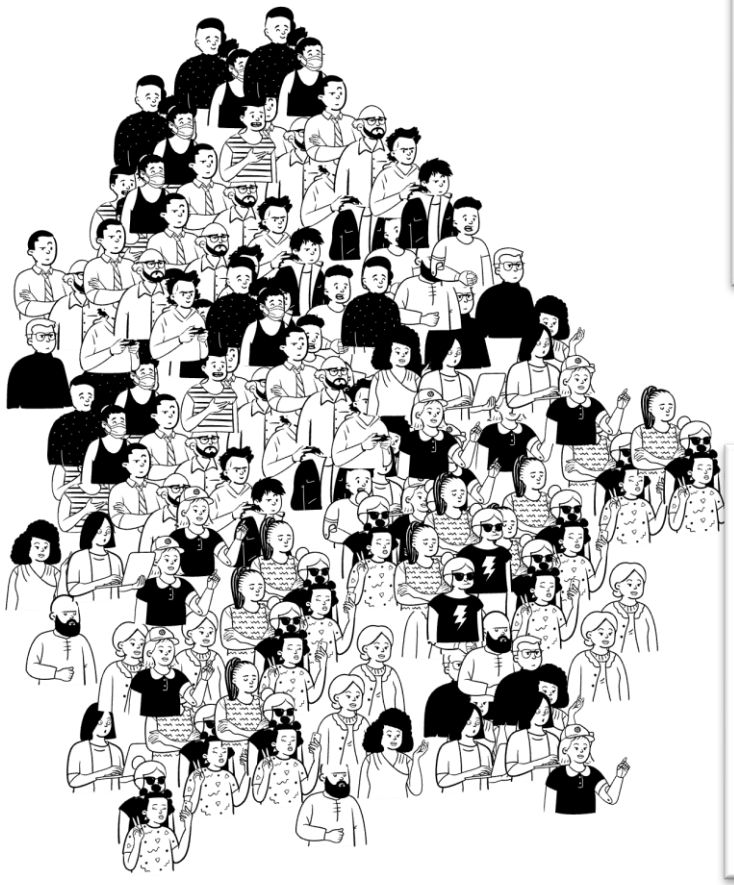


Group B

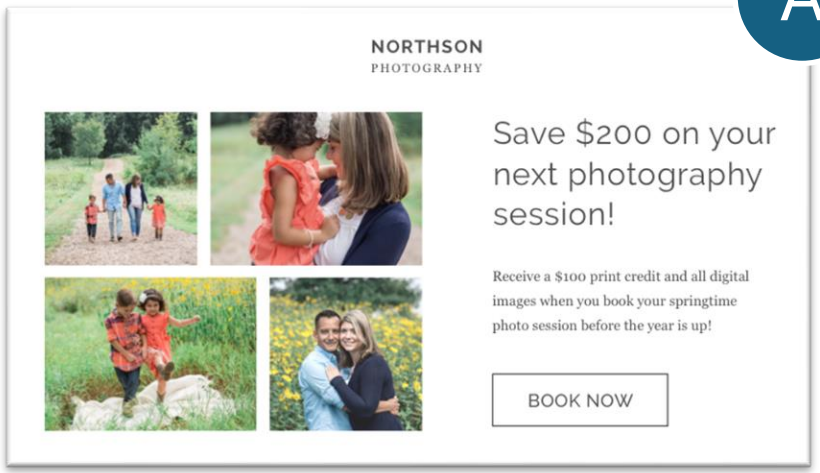
Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

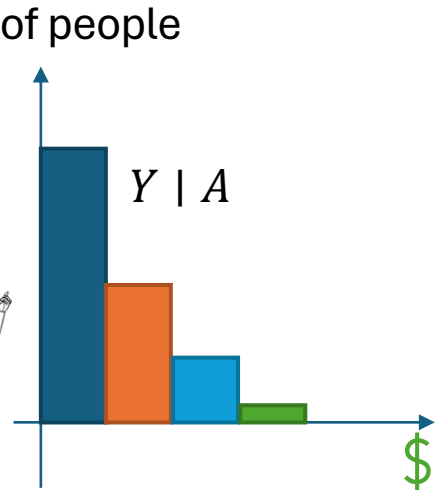
user base



A

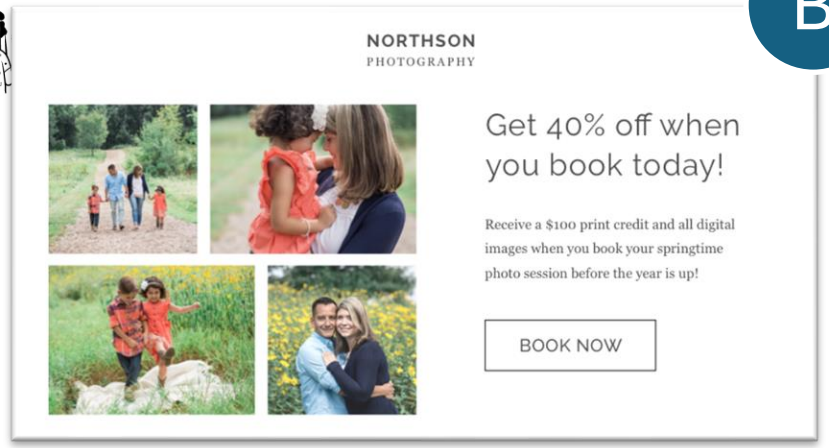


Group A

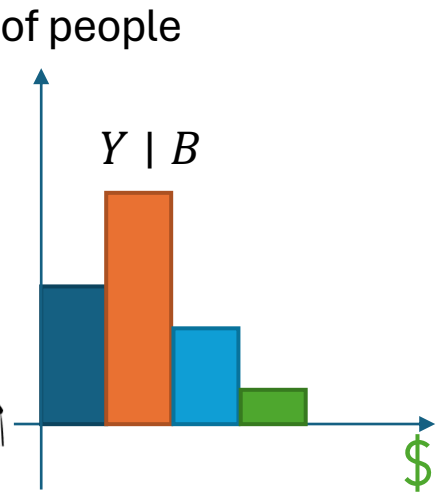


$\mu_A = 10\$$ (average spend)

B



Group B

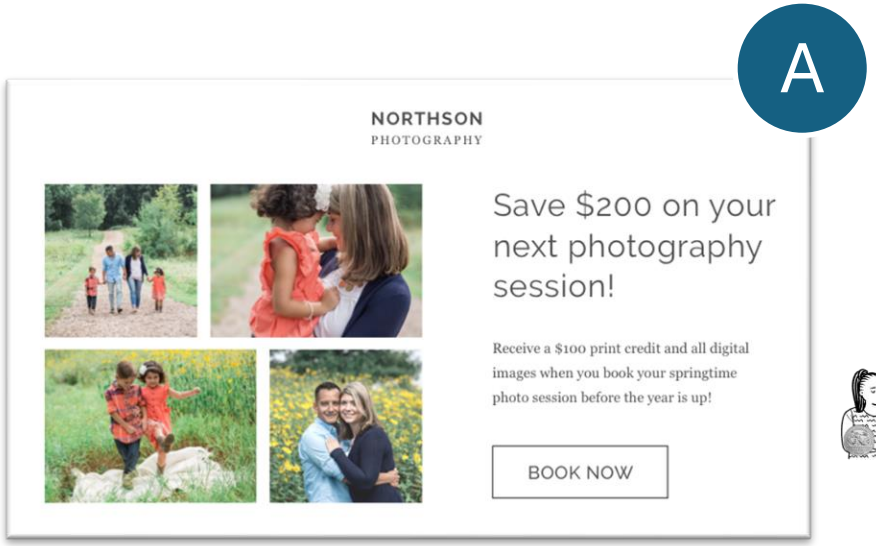


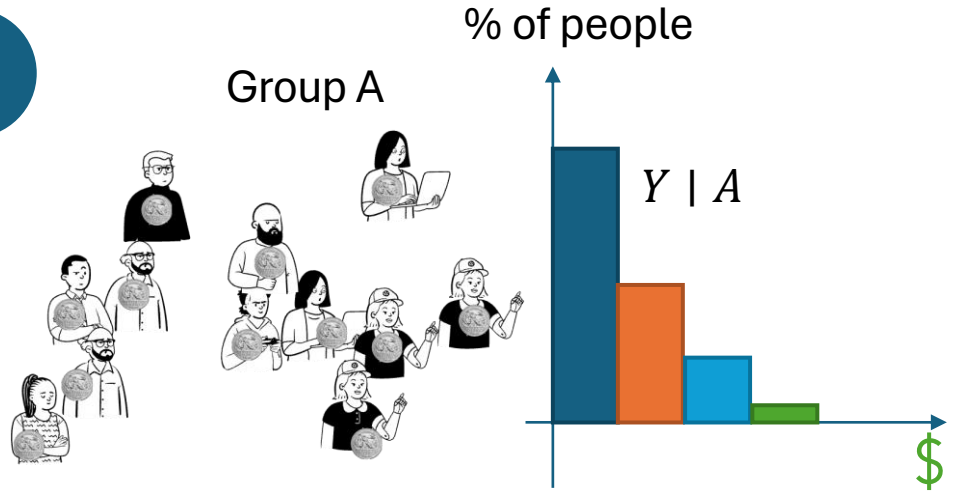
$\mu_B = 20\$$ (average spend)

Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

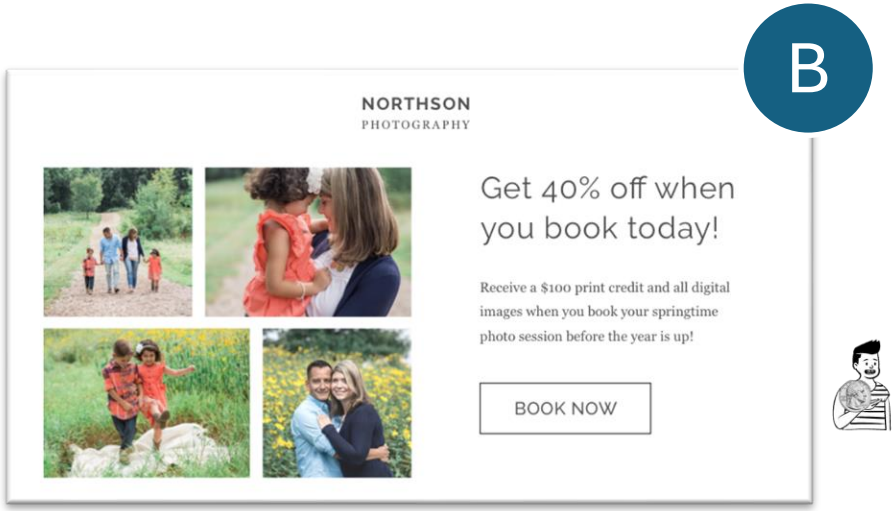
Control
Baseline
Status quo

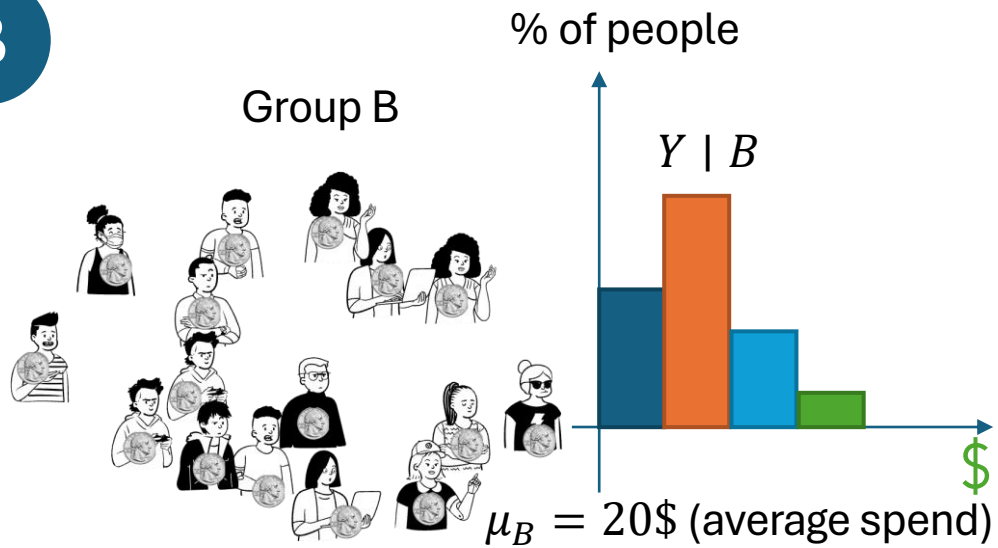




$\mu_A = 10\$$ (average spend)

Treatment
Innovation





$\mu_B = 20\$$ (average spend)

Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A/B Testing

Control

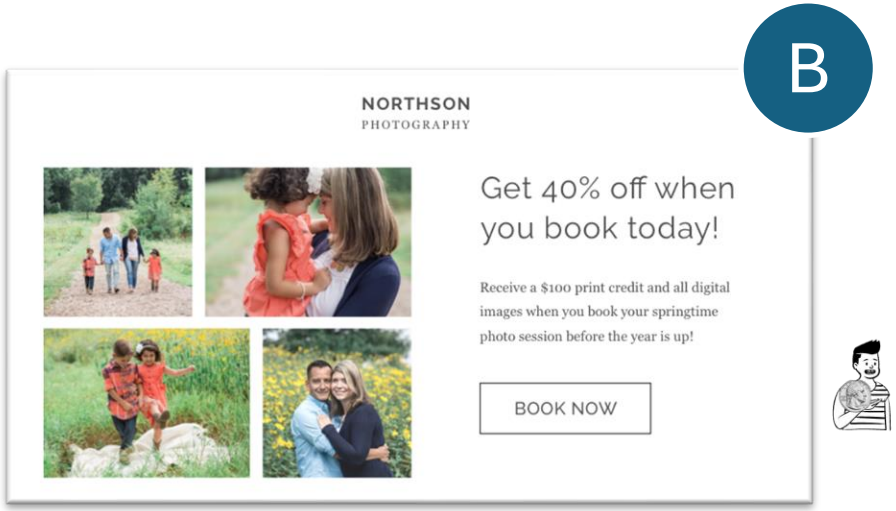
Baseline

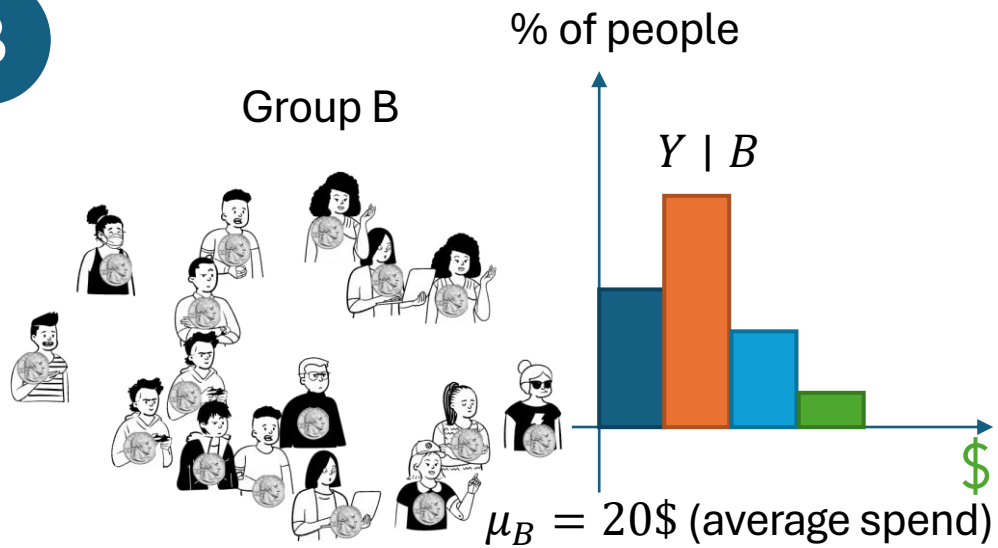
Status quo

$\mu_A = 10\$$ (average spend)

Treatment

Innovation



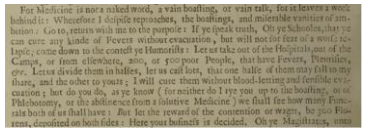


$\mu_B = 20\$$ (average spend)

Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

A Brief History of Experimentation

1662 — John Baptista Van Helmont proposes a randomized experiment for the effect of bloodletting


For Medicine is not a naked word, a vain boasting, or vain talk, for it leaves a work behind it: Wherefore I despise reproaches, the boastings, and miserable vanities of ambition: Come, return with me to the purpose: If ye speak truth, Oh ye Schools, that ye can cure any kinde of Fevers without evacuation, but will not for fear of a worse relapse; come down to the contest ye Humorists: Let us take out of the Hospitals, out of the Camps, or from elsewhere, 200, or 300 poor People, that have Fevers, Pleurisies, &c. Let us divide them in halves, let us cast lots, that one halfe of them may fall to my share, and the other to yours; I will cure them without blood-letting and sensible evacuation; but do you do, as ye know (for neither do I rye you up to the boasting, or of Phlebotomy, or the abstinence from a solative Medicine) we shall see how many Funerals both of us shall have: But let the reward of the contention or wager, be 300 Florens, deposited on both sides: Here your business is decided. Oh ye Magistrates, unto

1828 — Louis conducts a non-randomized controlled experiment showing strong evidence against bloodletting

1923 — Neyman publishes "On the Application of Probability Theory to Agricultural Experiments"

1925 — Fisher publishes Statistical Methods for Research Workers

1935 — Michigan Supreme Court authorizes RCTs as part of medical practice

Fisher publishes The Design of Experiments

1947 — First published medical RCT in Great Britain

1962 — FDA to prohibit the marketing of any new drug in the absence of "substantial evidence."

1969 — FDA: "…partially controlled studies are not acceptable evidence to support claims of effectiveness"

2000- — Abhijit Banerjee, Esther Duflo, and Michael Kremer, popularize the use of RCTs in economics (esp. for policy problems related to poverty).

10s of thousands of RCTs run annually by companies like Airbnb, Amazon, Booking.com eBay, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yahoo! and Yandex

Credit: Ron Kohavi, History of Controlled Experiments

RCTs are the gold standard for measuring the “causal effect” of a “treatment” on an “outcome”

Causality

Goal #1: Mathematical Definition of Causality

- We want to formally (mathematically) define the causal effect of a binary treatment on a scalar outcome of interest

Examples

- Effect of seed A vs B on crop yield
- Effect of completion of a job training program on observed wage
- Effect of drug vs placebo on overall survival
- Effect of an ad impression on conversion (purchase)
- Effect of eating avocados on longevity

Causal Effect

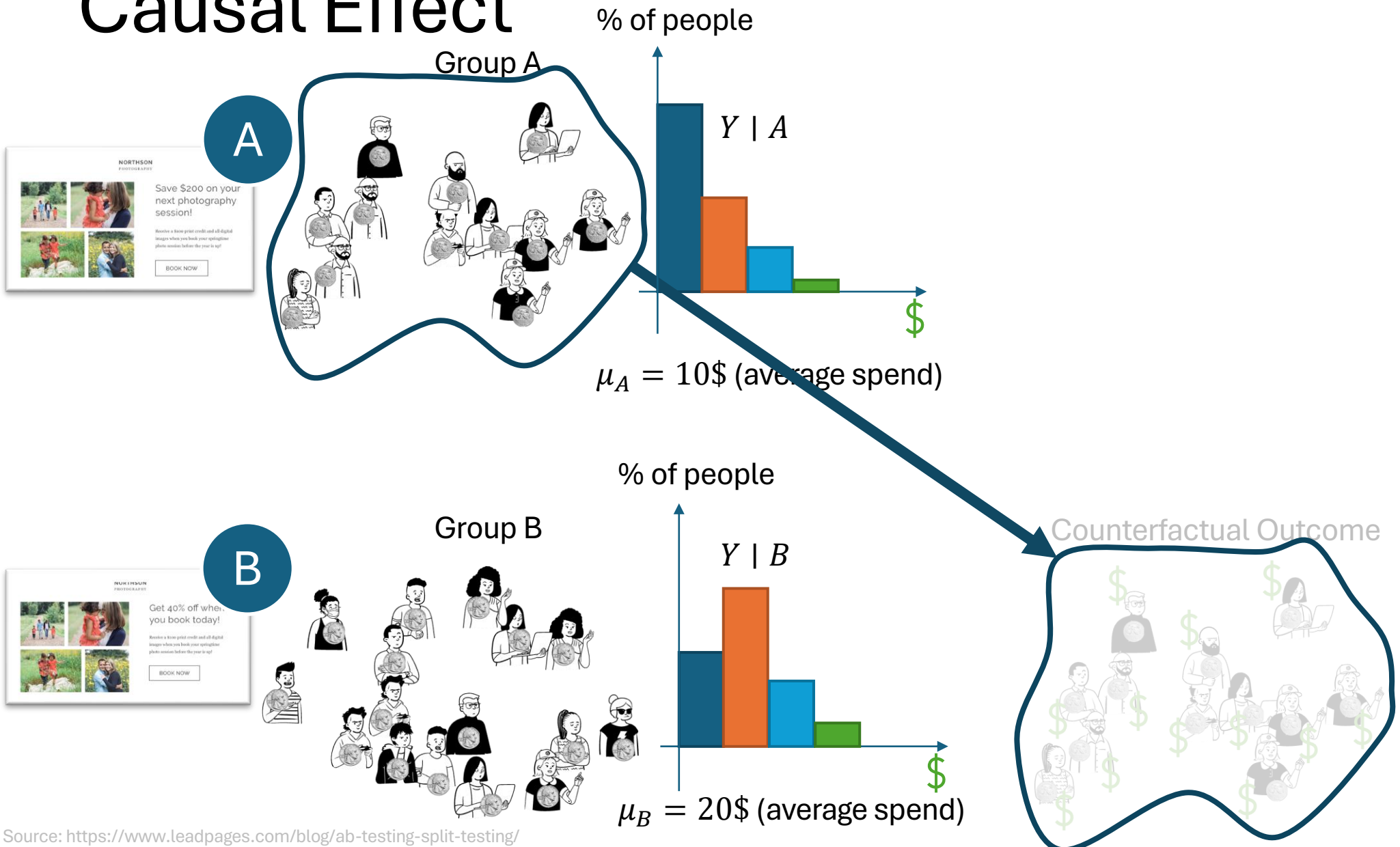


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Causal Effect

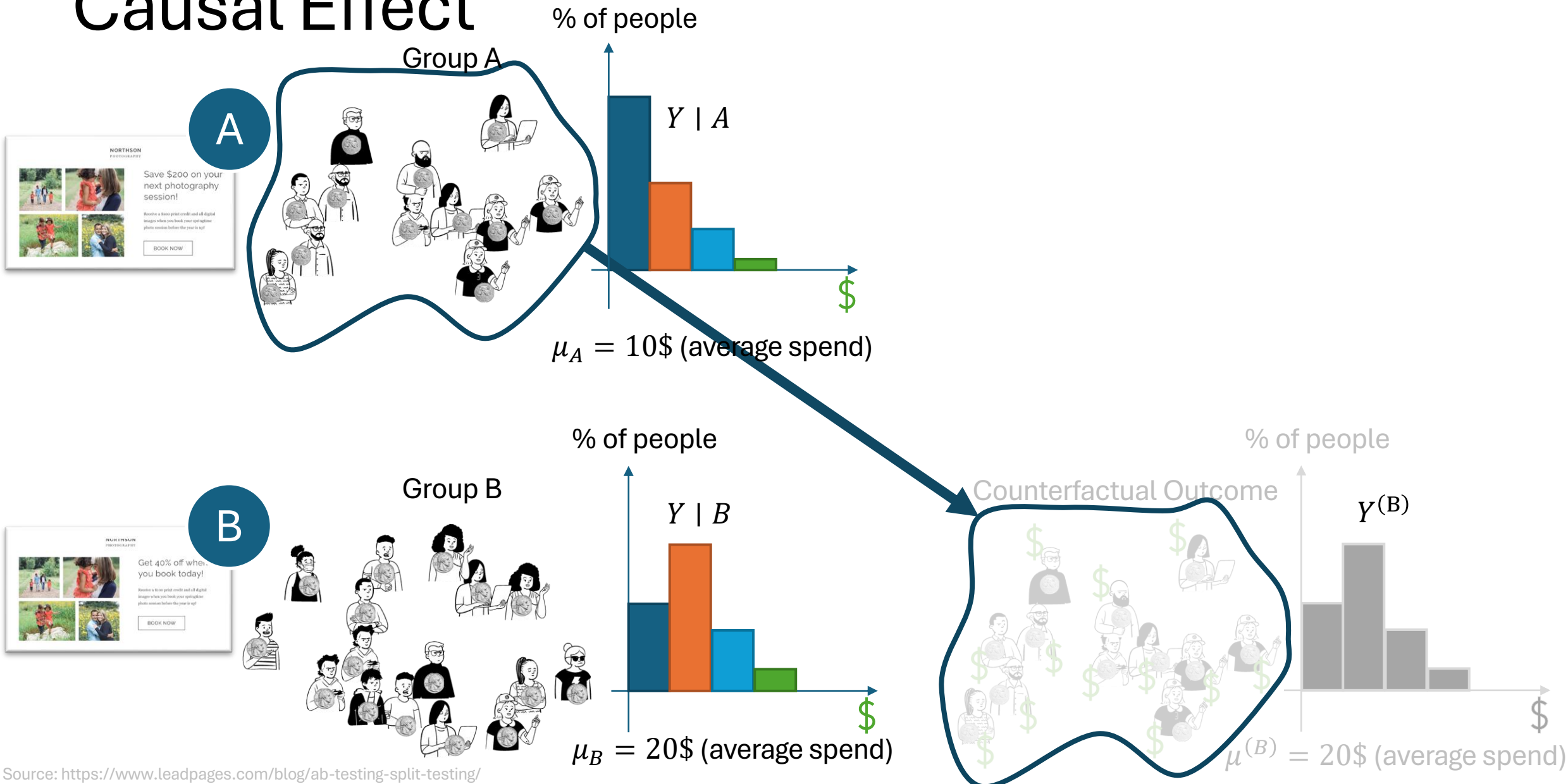


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Causal Effect

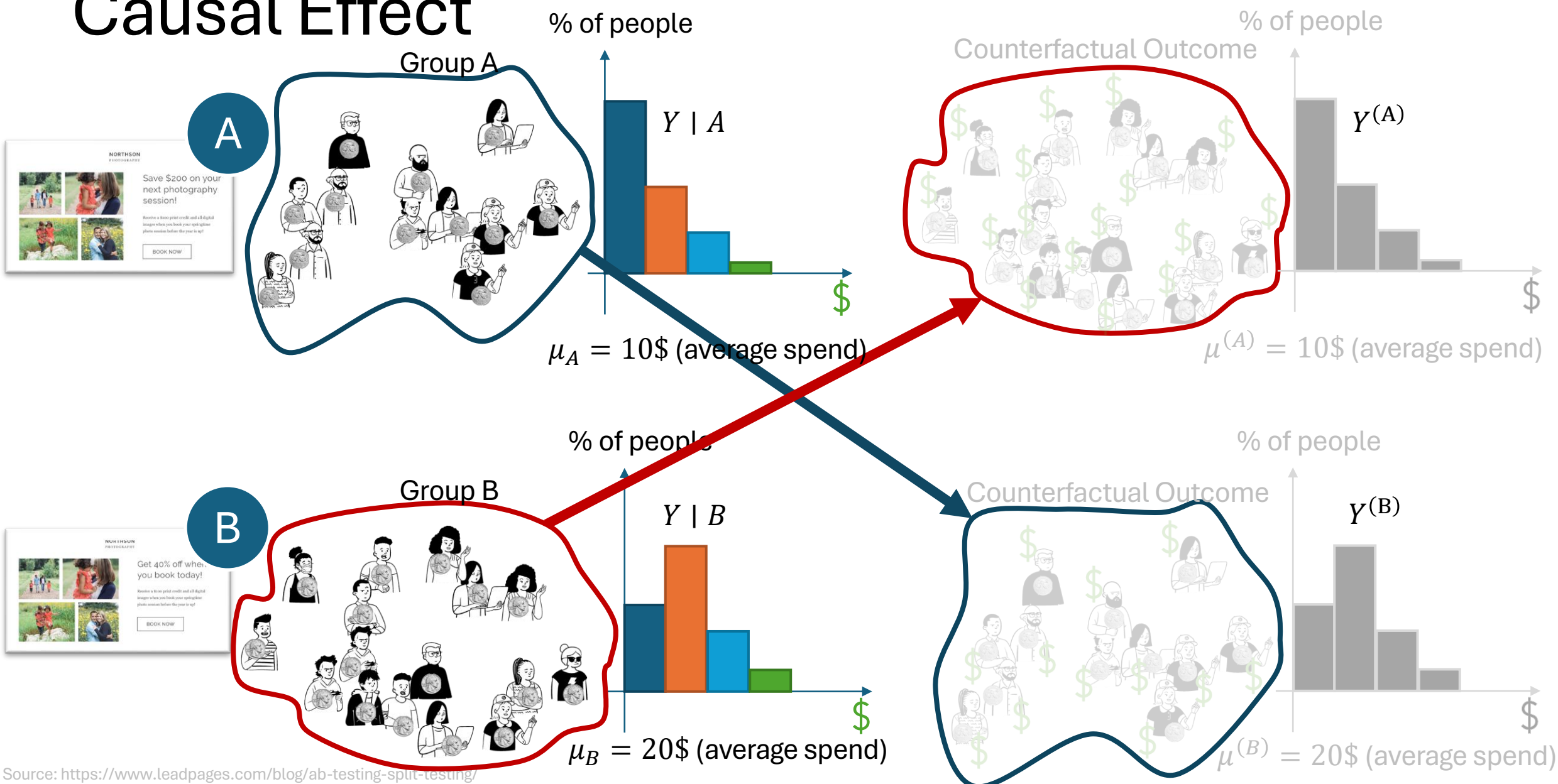


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Causality via Potential Outcomes

- Nature generates two latent (unobserved) random outcomes $Y^{(0)}, Y^{(1)}$
- $Y^{(d)}$: potential outcome that would have been observed if unit received treatment $d \in \{0,1\}$

Example

- $Y^{(0)}$: wage if you don't participate in a training program
- $Y^{(1)}$: wage if you participate in a training program

Causality via Potential Outcomes

- $Y^{(0)}, Y^{(1)}$ are called "counterfactuals" as they can never be simultaneously observed
- Fundamental problem of causal inference

Example

- We don't have two replicas of each unit
- We cannot observe your wage with the training program and without

Average causal effect (ATE)

$$\delta := E\left[Y^{(1)} - Y^{(0)}\right]$$

Treatment Assignment and Observed Data

- Each unit receives treatment $D \in \{0,1\}$, we observe
$$Y \equiv Y^{(D)}$$

- Given data (D, Y) what quantities can we "identify" $\equiv$ "measure if we had access to infinite data"

- Can we measure the ATE?

Randomization implies

$$Y|D = 0 \sim Y^{(0)}$$
$$Y|D = 1 \sim Y^{(1)}$$

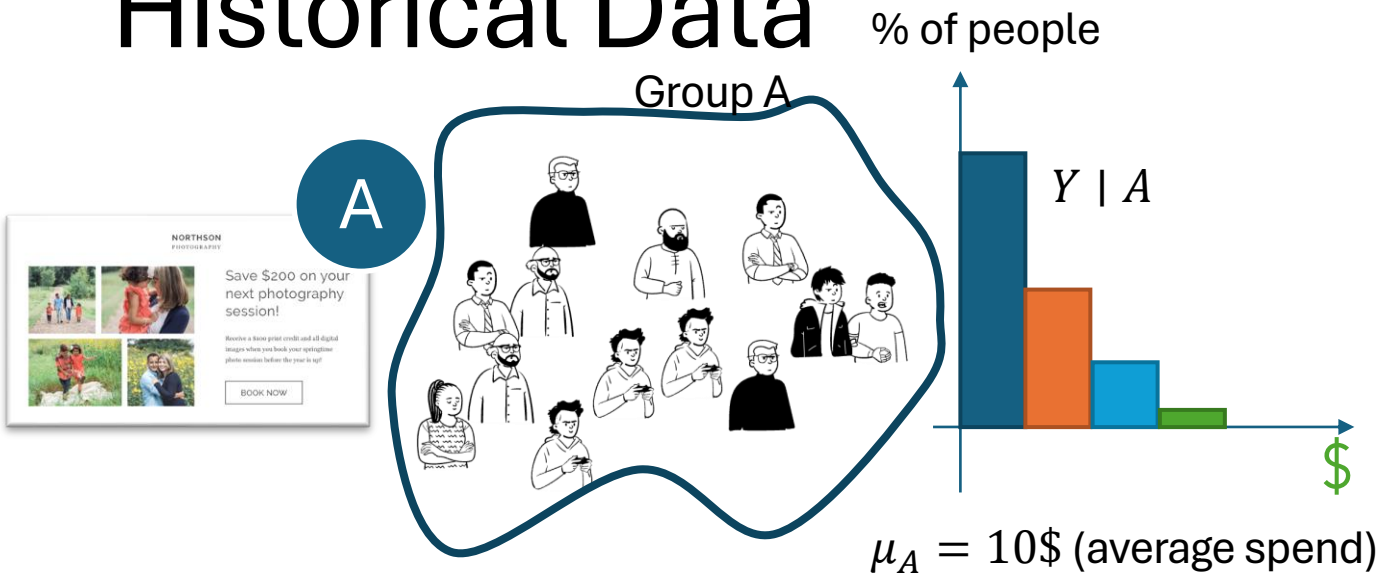
Aggregate differences between groups

$$\mathrm{E}[Y|D = 1] - \mathrm{E}[\mathrm{Y}|\mathrm{D} = 0]$$

Equal aggregate causal effect

$$E\left[Y^{(1)} - Y^{(0)}\right]$$

Historical Data



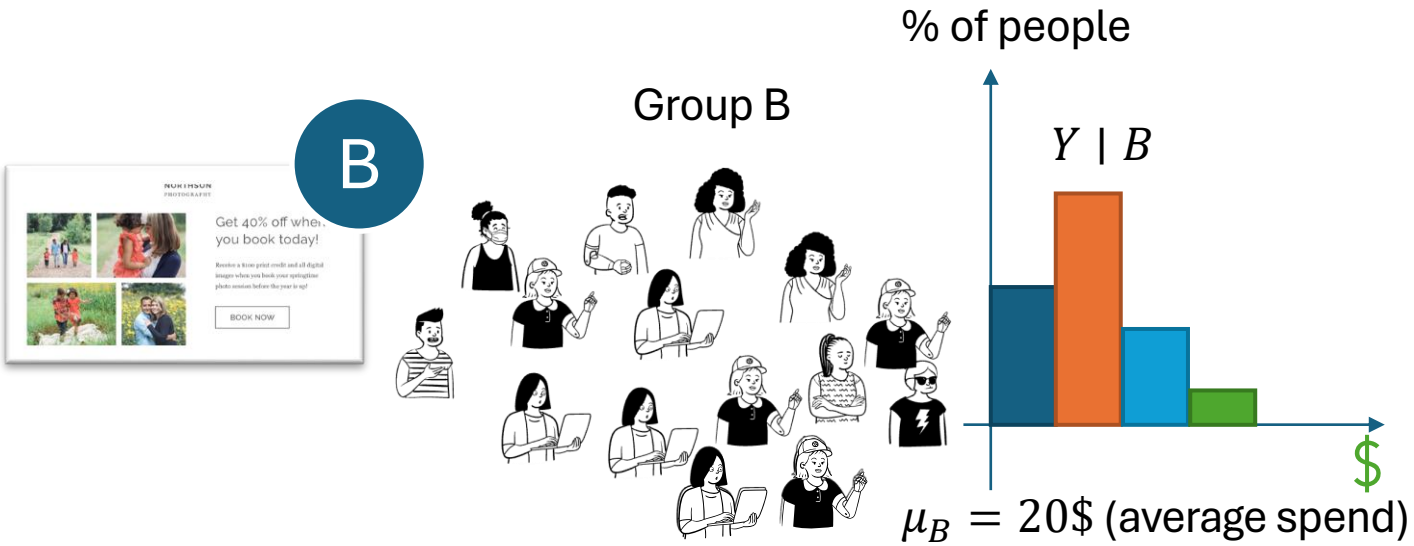


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Historical Data

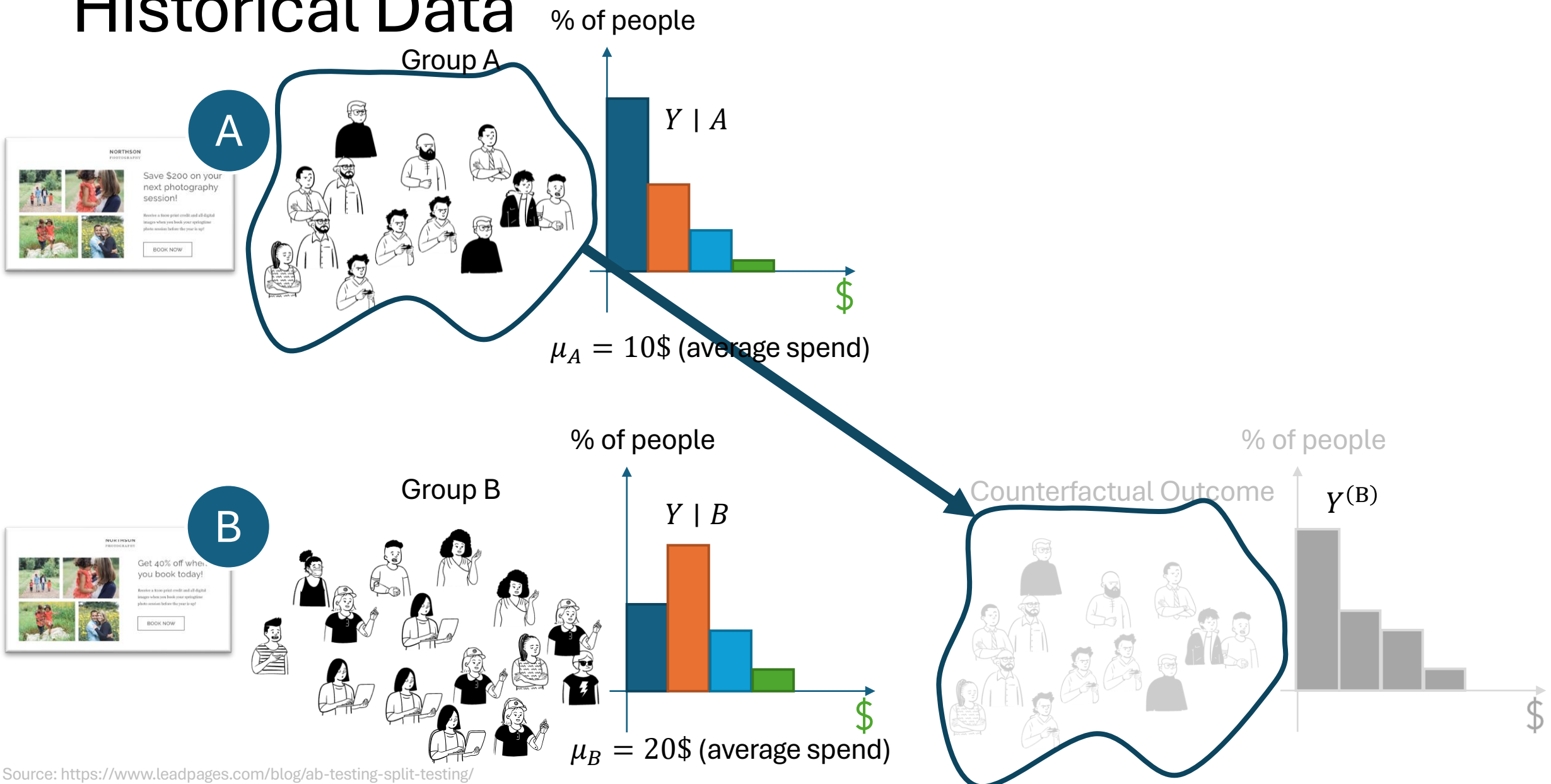


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Image source: https://www.linkedin.com/pulse/why-90-fortune-500-companies-now-using-microsoft-cloud-natasha-spurr/

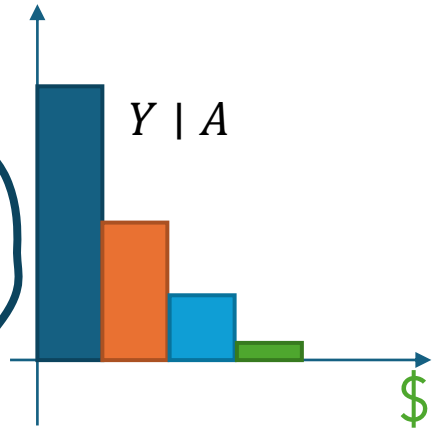
Historical Data

No cloud specialist

A

Group A

% of companies



$\mu_A = 10\$$ (average spend)

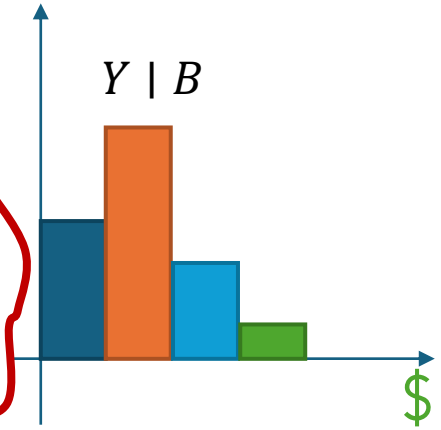
Assign Cloud Specialist

B

Group B

Fortune500

% of companies



$\mu_B = 20\$$ (average spend)

Image source: https://www.linkedin.com/pulse/why-90-fortune-500-companies-now-using-microsoft-cloud-natasha-spurr/

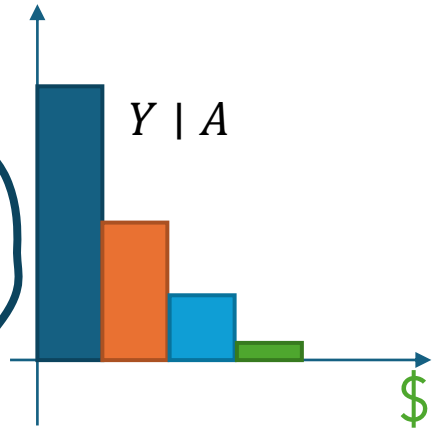
Historical Data

No cloud specialist

A

Group A

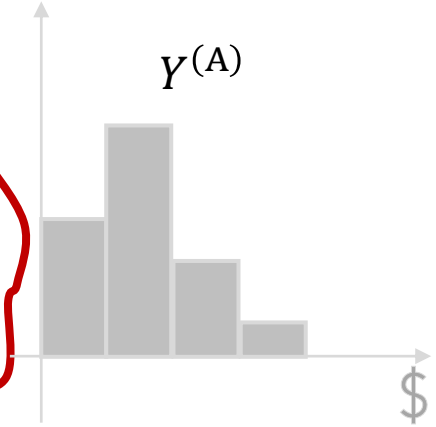
% of companies



$\mu_A = 10\$$ (average spend)

Counterfactual outcome

% of companies



$\mu^{(A)} = 20\$$ (average spend)

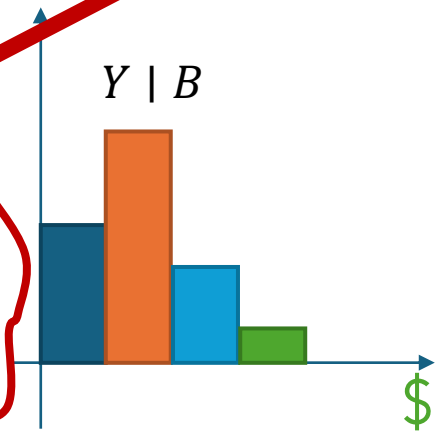
Assign Cloud Specialist

B

Group B

Fortune500

% of companies



$\mu_B = 20\$$ (average spend)

Image source: https://www.linkedin.com/pulse/why-90-fortune-500-companies-now-using-microsoft-cloud-natasha-spurr/

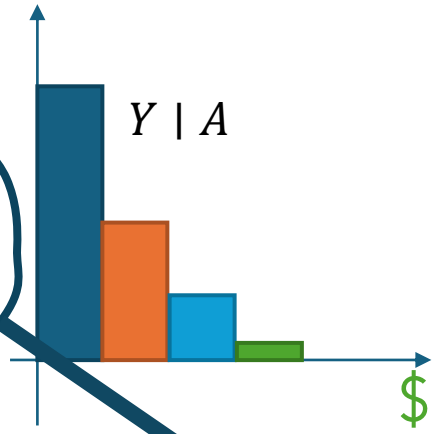
Historical Data

No cloud specialist

A

Group A

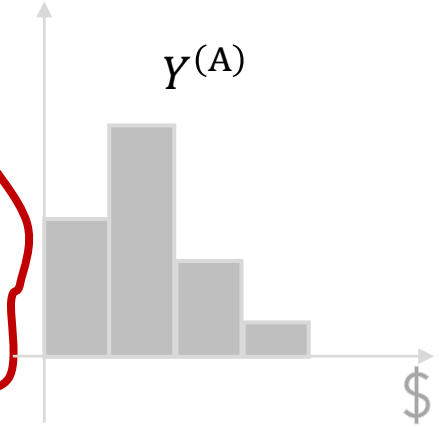
% of companies



$\mu_A = 10\$$ (average spend)

Counterfactual outcome

% of companies



$\mu^{(A)} = 20\$$ (average spend)

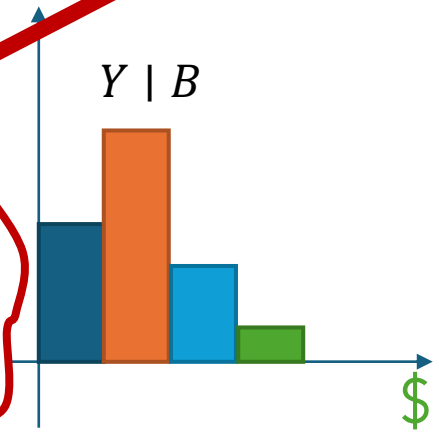
Assign Cloud Specialist

B

Group B

Fortune500

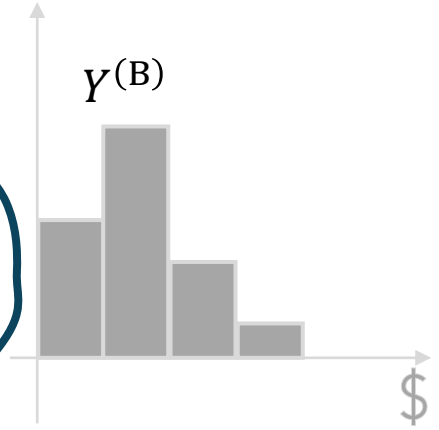
% of companies



$\mu_B = 20\$$ (average spend)

Counterfactual Outcome

% of companies



$\mu^{(B)} = 20\$$ (average spend)

Identification of ATE in RCTs

Suppose that treatment is randomly assigned (i.e. RCT) with $\Pr(D = 1) \in (0, 1)$

$$Y^{(d)} \perp\!\!\!\perp D$$

Average **observed** outcome in treatment group $d \in \{0,1\}$ recovers average **potential** outcome for treatment d

$$E[Y|D = d] = E\left[Y^{(D)}\middle|D = d\right] = E\left[Y^{(d)}\middle|D = d\right] = E\left[Y^{(d)}\right]$$

Average **predictive** effect recovers the average **treatment** effect

$$\begin{aligned}\pi &:= E[Y|D = 1] - E[Y|D = 0] \\ &= E\left[Y^{(1)}\right] - E\left[Y^{(0)}\right] =: \delta\end{aligned}$$

Limitations of RCTs

Externalities, Stability and Equilibrium Effects

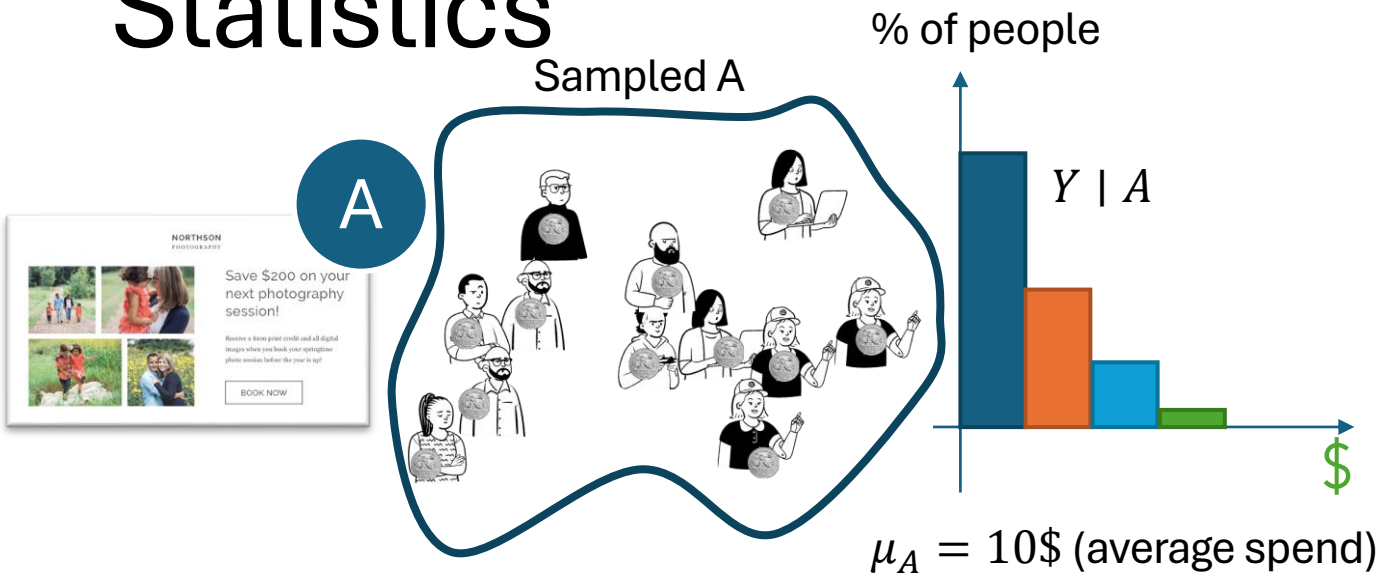
- Stable Unit Treatment Value Assumption (SUTVA)
- Implicit in our notation the potential outcome of a unit depends only on its own treatment
- This can be violated due to what is known as spillover effects, externalities or general equilibrium effects
- In vaccine trials: if large fraction of population is treated, then the effect of vaccinating an additional unity changes, due to herd immunity
- In labor: if we make an intervention that incentivizes a large fraction of population to attend college, the college-wage premium will likely decrease

Ethical, Practical and Generalizability Concerns

- **Ethical Concerns:** Many potential trials would correctly be judged unethical by a human subject trial review board; Key principles [78 Belmont report]: (i) respect for persons, (ii) beneficence, (iii) justice
- **Practical Concerns:** Practically infeasible due to high cost (e.g. expensive treatment, expensive data collection, low signal regime, long-term outcomes) or inability to directly randomize the treatment
- **Generalizability:** Local population used for RCT might not generalize to broader population


Statistics
289.33

Statistics



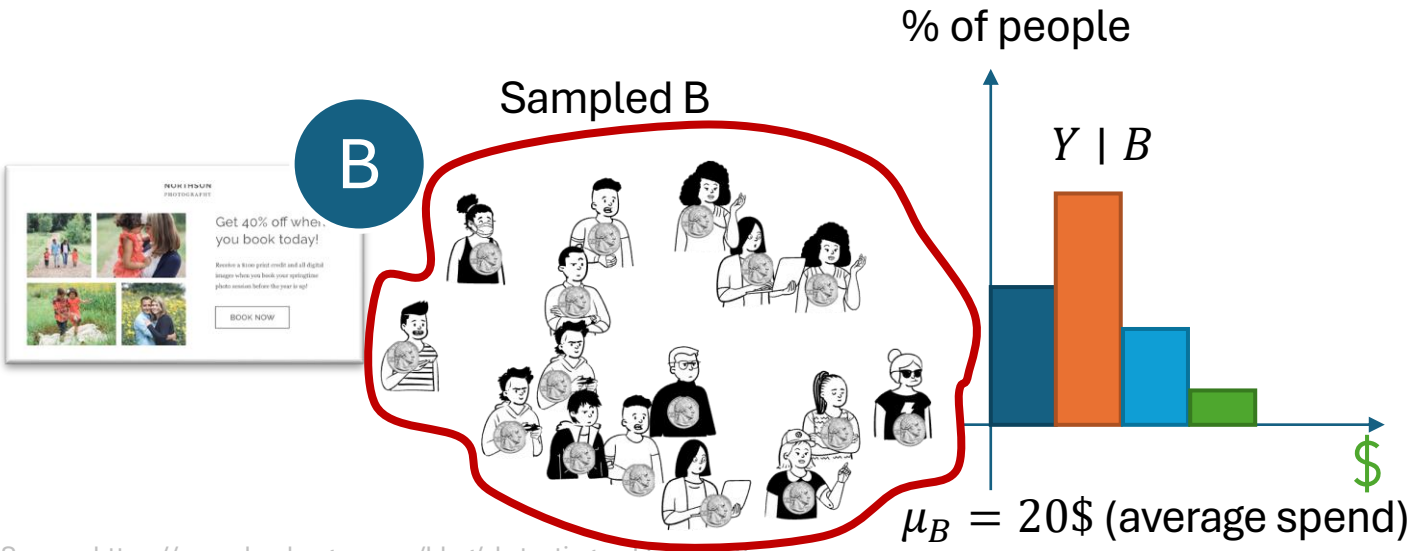
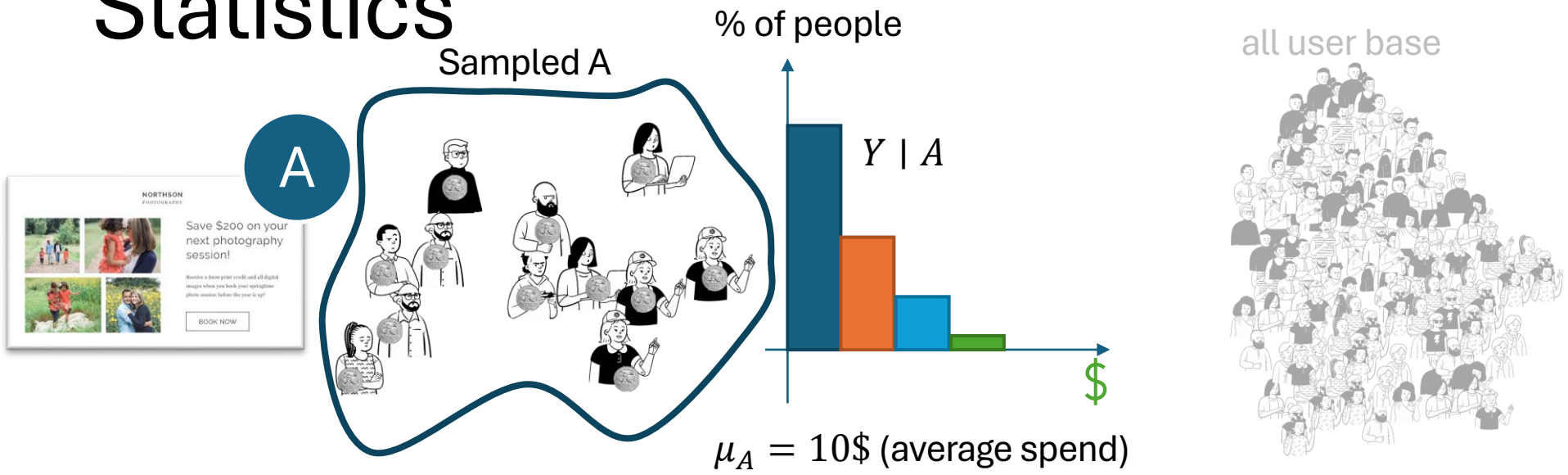


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Statistics



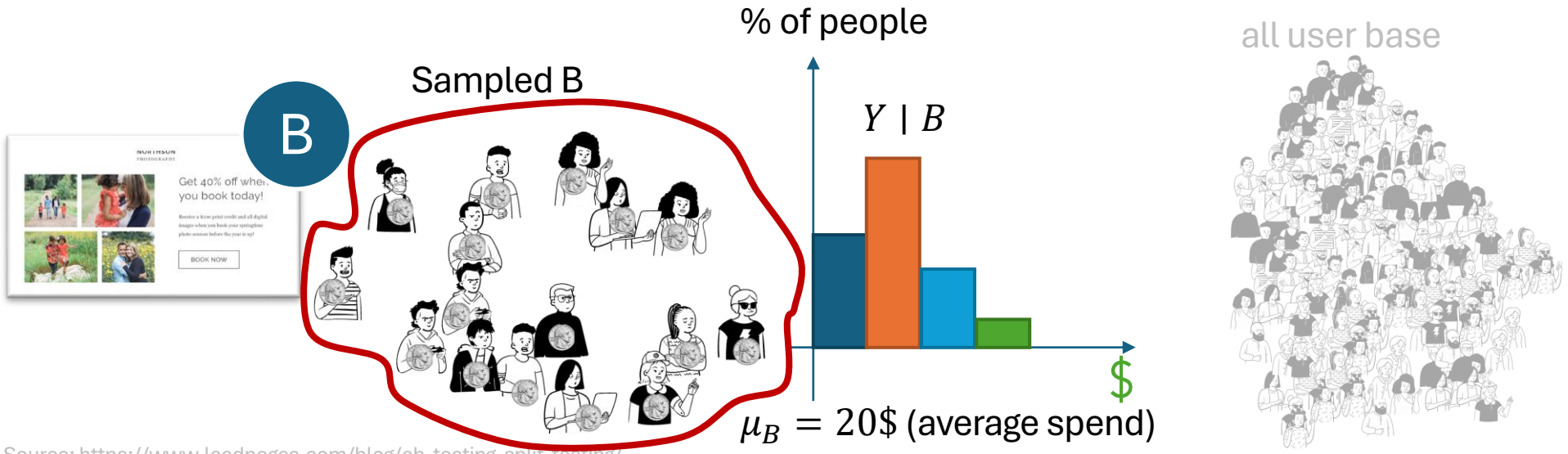


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Statistics

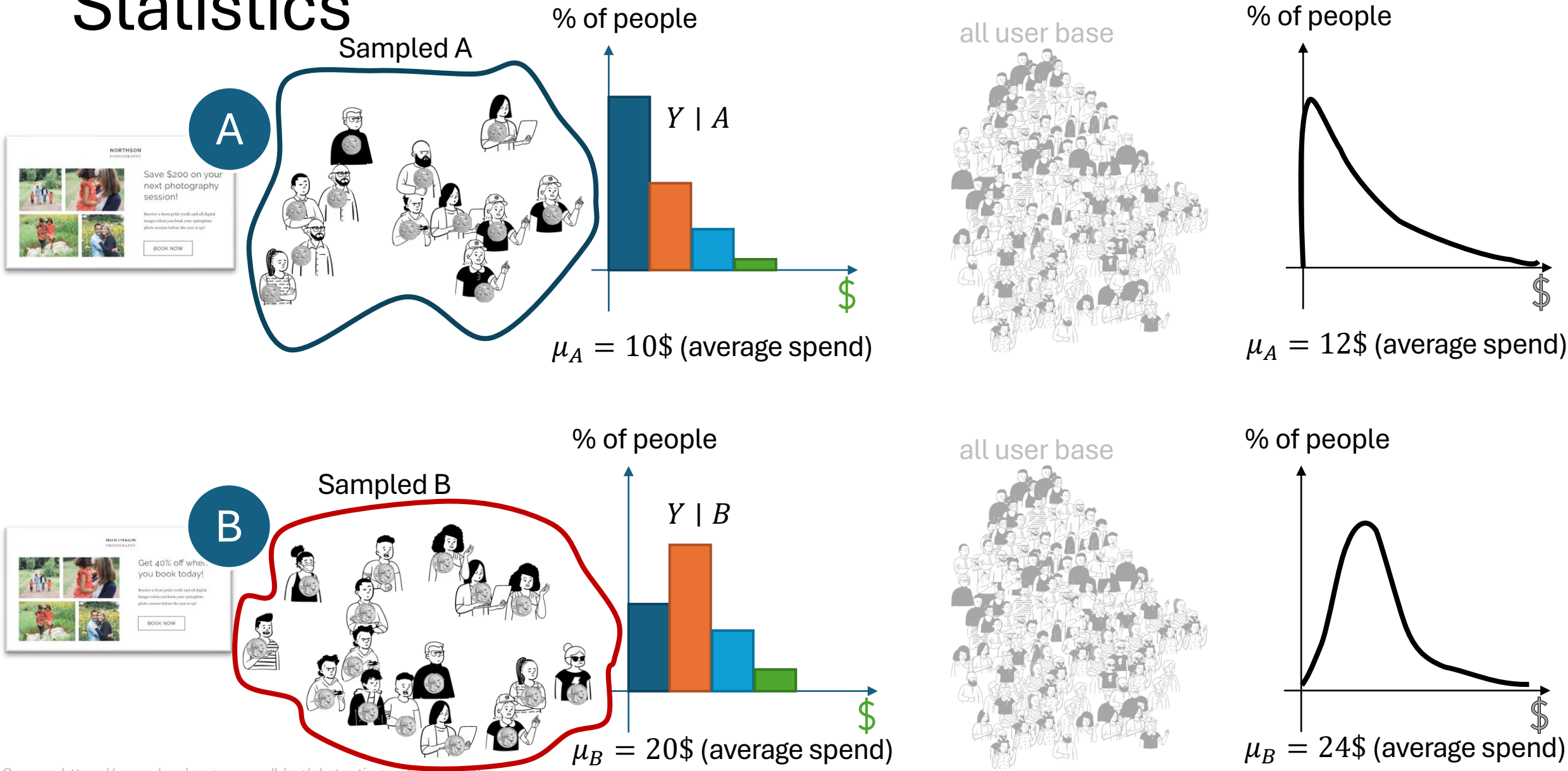
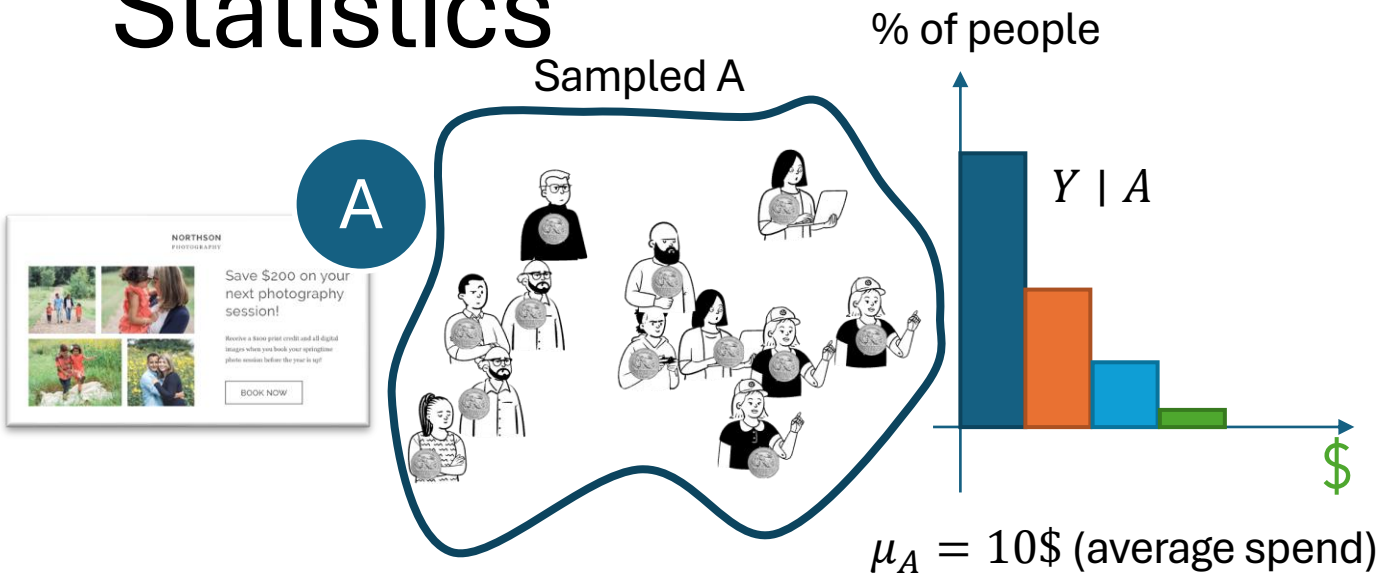


Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Statistics

with probability 95%

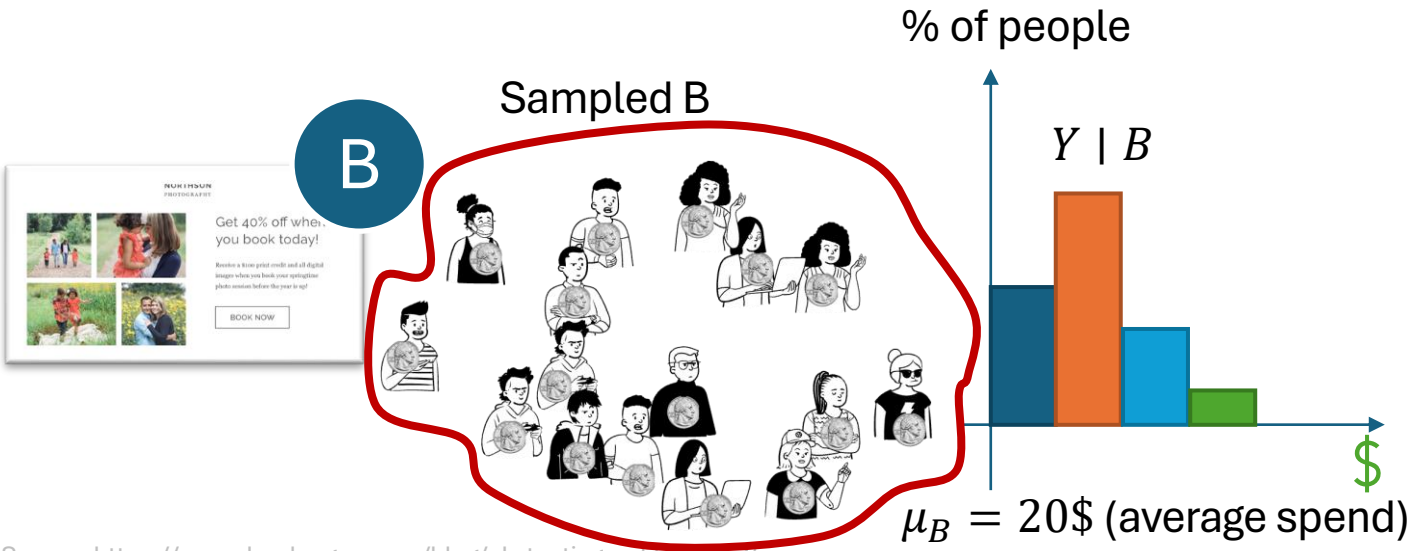


population mean

$$\mu_A \approx \hat{\mu}_A \pm 2\sqrt{\frac{\mathrm{Var}(Y|A)}{N_A}}$$

sample mean

sampling error



population mean

$$\mu_B \approx \hat{\mu}_B \pm 2\sqrt{\frac{\mathrm{Var}(Y|B)}{N_B}}$$

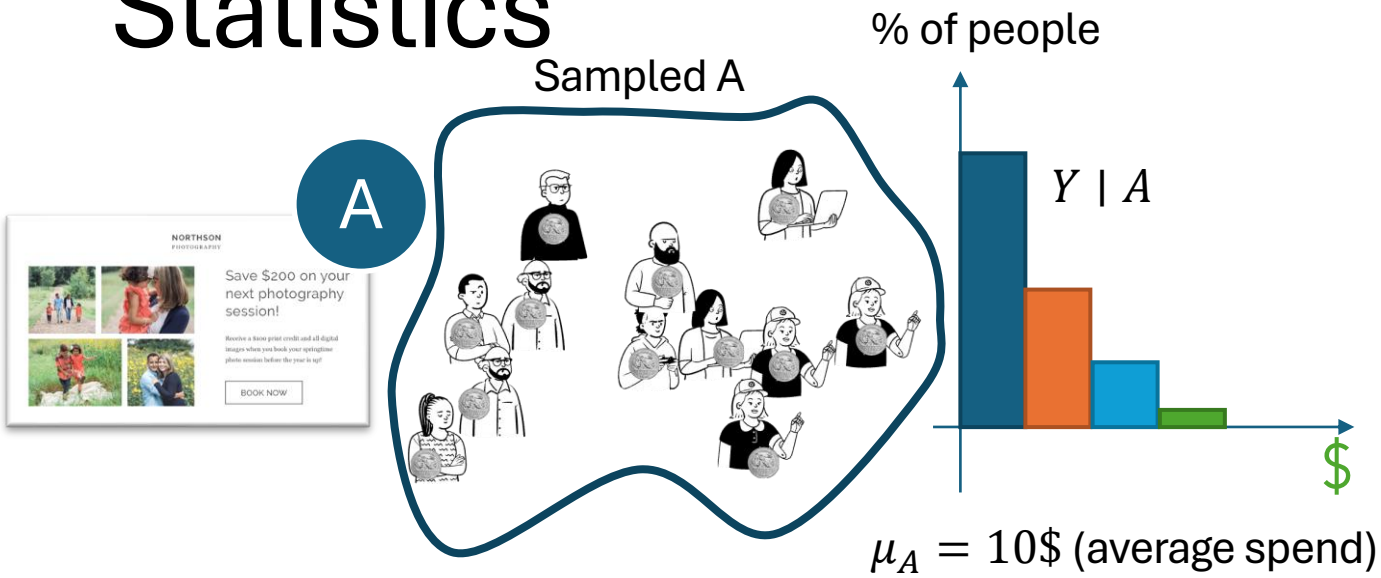
sample mean

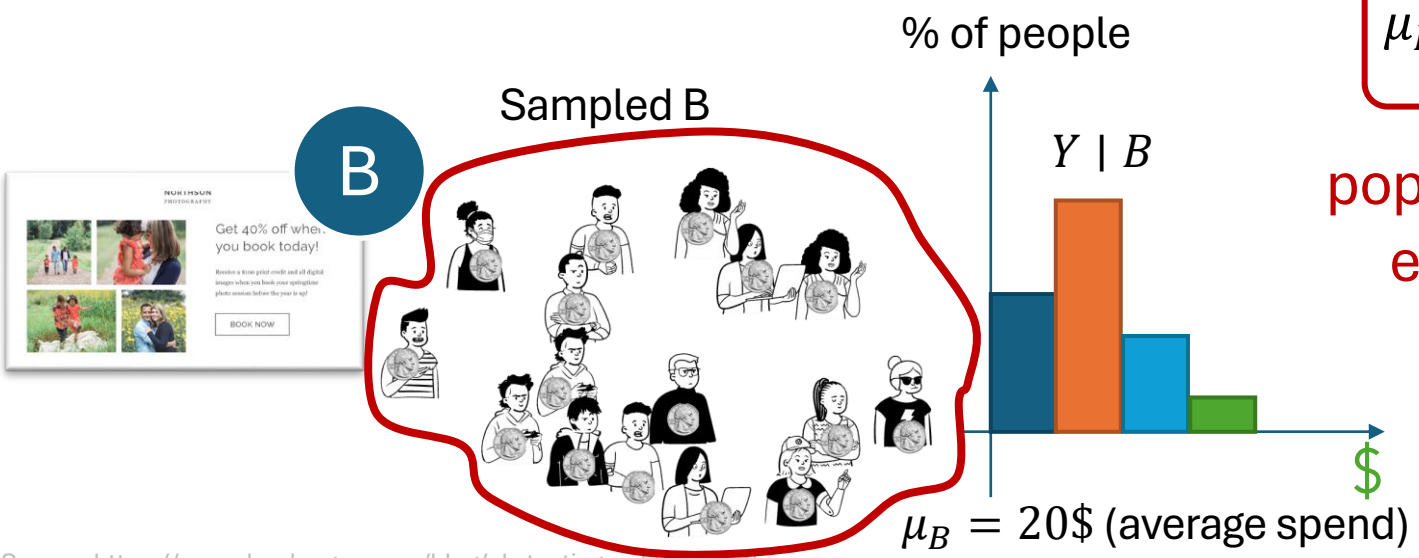
sampling error

Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

Statistics

with probability 95%





$$\boxed{\mu_B - \mu_A} \approx \boxed{\hat{\mu}_B - \hat{\mu}_A} \pm \boxed{2\sqrt{\frac{\mathrm{Var}(Y|A)}{N_\mathrm{A}} + \frac{\mathrm{Var}(Y|B)}{N_\mathrm{B}}}}$$

population effect | sample effect | sampling error

Image Source: https://www.leadpages.com/blog/ab-testing-split-testing/

More formally

$X \overset{a}{\sim} Y$ means that as $n \to \infty$:

$$\sup_{R \in \mathcal{R}} |P(X \in R) - P(Y \in R)| \approx 0$$

where $\mathcal{R}$ set of all hyper-rectangles

Under mild regularity conditions

$$\sqrt{n}\{\hat{\mu}_d - \mu_d\}_{d \in \{0,1\}} \overset{a}{\sim} N(0, V)$$

where

$$V = \begin{pmatrix} \dfrac{\mathrm{Var}(Y|D=0)}{P(D=0)} & 0 \\ 0 & \dfrac{\mathrm{Var}(Y|D=1)}{P(D=1)} \end{pmatrix}$$

Hence

$$\sqrt{n}(\hat{\delta} - \delta) \overset{a}{\sim} N(0, V_{11} + V_{22})$$

Proof Sketch

Trivially we can write $\mu_d = E\left[Y^{(d)}\right]\frac{E_n[1(D=d)]}{E_n[1(D=d)]}$

$$\hat{\mu}_d - \mu_d = \frac{E_n\left[Y^{(d)}\,1(D=d)\right]}{E_n[1(D=d)]} - \mu_d = \frac{E_n\left[\left(Y^{(d)} - E\left[Y^{(d)}\right]\right)1(D=d)\right]}{E_n[1(D=d)]}$$

By Law of Large Numbers (LLN)

$$\hat{\mu}_d - \mu_d \approx \frac{E_n\left[\left(Y^{(d)} - E\left[Y^{(d)}\right]\right)1(D=d)\right]}{P(D=d)}$$

Difference is average of the i.i.d. mean zero r.v.s: $\frac{\left(Y_i^{(d)} - E\left[Y^{(d)}\right]\right)1(D_i=d)}{P(D=d)}$

With zero covariance and variance: $\frac{E\left[\left(Y_i^{(d)} - E\left[Y^{(d)}\right]\right)^2 1(D_i=d)\right]}{P(D=d)^2} = \frac{Var(Y|D=d)}{P(D=d)}$

Statement follows by Central Limit Theorem (CLT)

Variance estimate

- Same statement also holds with consistent estimate of variance

$$\hat{V} = \begin{pmatrix} \dfrac{\mathrm{Var_n}(Y|D=0)}{E_n[1-D]} & 0 \\ 0 & \dfrac{\mathrm{Var_n}(Y|D=1)}{E_n[D]} \end{pmatrix}$$

Confidence Interval

- $X \overset{a}{\sim} Y \equiv \sup_{[\ell,u]} |P(X \in [\ell,u]) - P(Y \in [\ell,u])| \approx 0$

- If we consider $[\ell,u]$ the $\left(\frac{\alpha}{2}, 1-\frac{\alpha}{2}\right)$ quantile of $N\left(0,\hat{V}\right)$ then

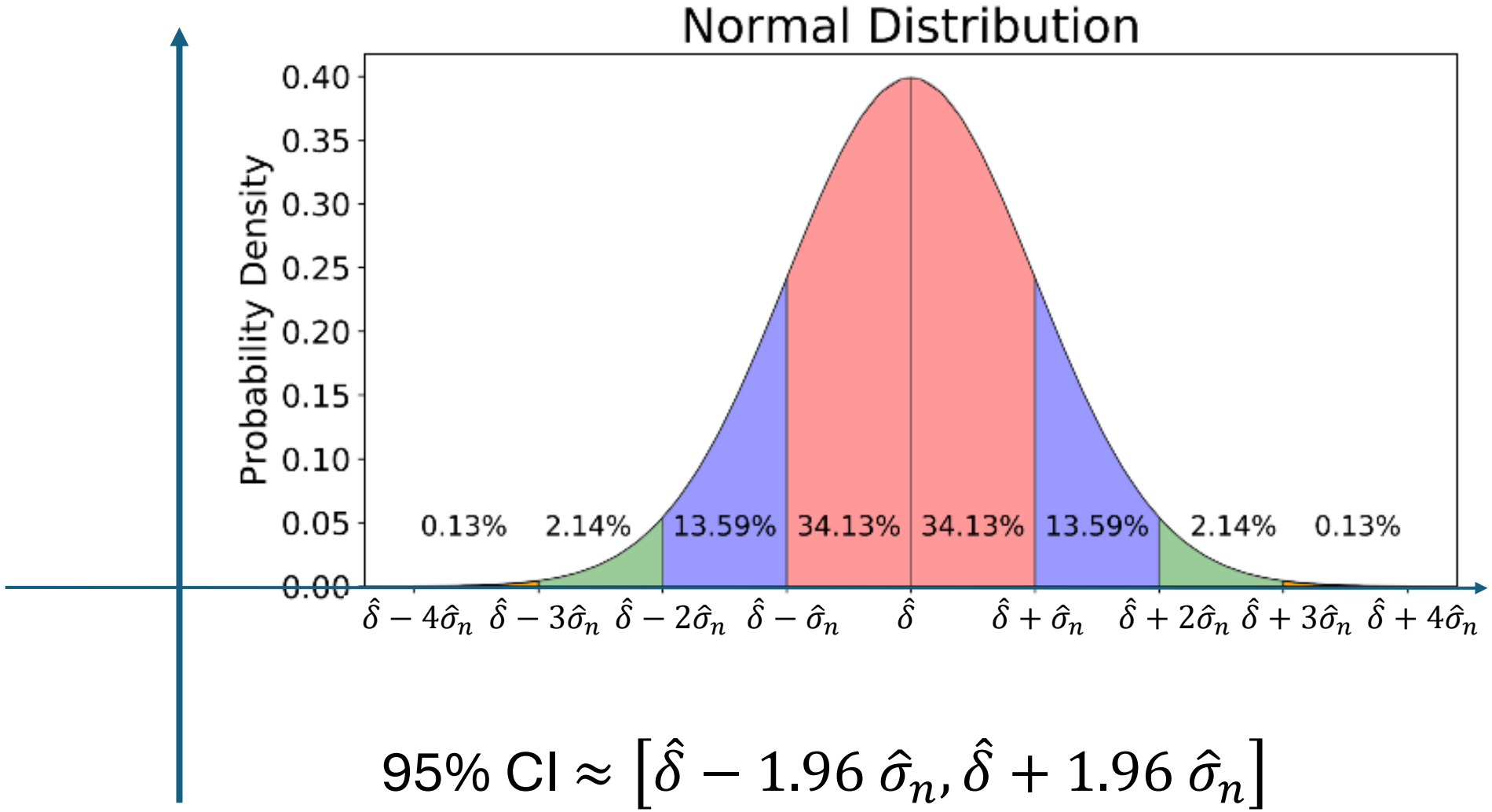
$$P\left(\sqrt{n}\left(\hat{\delta}-\delta\right) \in [\ell,u]\right) \approx 1-\alpha$$

- Equivalently, let z_α the α quantile of $N(0,1)$ and $\boxed{\hat{\sigma}_n = \sqrt{\hat{V}/n}}$ then:

Standard error

$$P\left(\delta \in \left[\hat{\delta} - z_{1-\frac{\alpha}{2}}\hat{\sigma}_n, \hat{\delta} + z_{1-\frac{\alpha}{2}}\hat{\sigma}_n\right]\right) \approx 1-\alpha$$

Confidence Interval



$$95\% \text{ CI} \approx \left[\hat{\delta} - 1.96\,\hat{\sigma}_n, \hat{\delta} + 1.96\,\hat{\sigma}_n\right]$$

Let's try it out!

Relative effect

- Many times (e.g. vaccine trials) we are interested in relative effect

$$RE = \frac{E\left[Y^{(1)} - Y^{(0)}\right]}{E\left[Y^{(0)}\right]} = \frac{\mu_1 - \mu_0}{\mu_0}$$

- We can construct a plug-in estimate

$$\widehat{RE} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\mu}_0} = \frac{\hat{\mu}_1}{\hat{\mu}_0} - 1$$

Delta method: for any function f, with $G = \nabla f(\theta)$

$$\sqrt{n}\left(f(\hat{\theta}) - f(\theta)\right) \approx G\sqrt{n}(\hat{\theta} - \theta) \overset{a}{\sim} N(0, G'VG)$$

- By delta method with $G = \left(-\mu_1/\mu_0^2, 1/\mu_0\right)$

$$\sqrt{n}(\widehat{RE} - RE) \overset{a}{\sim} N\left(0, \frac{\mu_1^2 V_{11}}{\mu_0^4} + \frac{V_{22}}{\mu_0^2}\right)$$

Example: Pfizer Vaccine

Efficacy Endpoint Subgroup	BNT162b2 N[a]=19965 Cases n1[b] Surveillance Time[c] (n2[d])	Placebo N[a]=20172 Cases n1[b] Surveillance Time[c] (n2[d])	Vaccine Efficacy % (95% CI)[e]
Overall	9 2.332 (18559)	169 2.345 (18708)	94.6 (89.6, 97.6)
Age group (years)			
16 to 17	0 0.003 (58)	1 0.003 (61)	100.0 (-3969.9, 100.0)
18 to 64	8 1.799 (14443)	149 1.811 (14566)	94.6 (89.1, 97.7)
65 to 74	1 0.424 (3239)	14 0.423 (3255)	92.9 (53.2, 99.8)
≥75	0 0.106 (805)	5 0.109 (812)	100.0 (-12.1, 100.0)

Approximate Confidence Interval

- Outcomes are binary $y \in \{0,1\}$
- Distribution of outcome for each d is Bernoulli with success p_d
- For each subpopulation, mean outcome $\hat{p}_d = \frac{\text{Cases}_d}{N_d}$ is estimate of p_d

$$\text{Vaccine Efficacy } (VE) = -\widehat{RE} = \frac{\hat{p}_0 - \hat{p}_1}{\hat{p}_0}$$

- An estimate of the variance of y is $\hat{p}_d(1 - \hat{p}_d)$
- 95% confidence interval can be derived by delta method

Let's try it out!

Pre-Treatment Covariates

Heterogeneity, Checks, and Precision

Pre-Treatment Covariates

- Assume we have covariates W that correspond to variables determined prior to treatment assignment (e.g. age, income)
- How can we use them?

- Heterogeneity: how does the effect vary with these covariates

- Formalized by the Conditional Average Treatment Effect (CATE)

$$\delta(W) := E\left[Y^{(1)} - Y^{(0)} \middle| W\right]$$

Identification of CATE under Random Assignment

Suppose that treatment is randomly assigned (i.e. RCT) with $\Pr(D = 1) \in (0, 1)$

$$\left(Y^{(d)}, W\right) \perp\!\!\!\perp D$$

Then conditional **observed** outcome in treatment group $d \in \{0,1\}$ recovers conditional **potential** outcome for treatment d

$$E[Y|D = d, W] = E\left[Y^{(d)}\middle|D = d, W\right] = E\left[Y^{(d)}|W\right]$$

Hence, conditional **predictive** effect recovers the CATE

$$\begin{aligned}\pi(W) &\coloneqq E[Y|D = 1, W] - E[Y|D = 0, W] \\ &= E\left[Y^{(1)}|W\right] - E\left[Y^{(0)}|W\right] =: \delta(W)\end{aligned}$$

If we only care about ATE are co-variates useful?

Co-variates for Sanity Check

- Since treatment is supposed to be independent of co-variates W

$$W|D = 1 \sim W|D = 0$$

- For instance, $E[W|D = 1] = E[W|D = 0] = E[W]$
- D does not predict any covariate
- Equivalently, D is not predictable by any covariate

$$D|W \sim D$$

- Can test conditions on samples to find violations of random assignment
- These are typically referred to as co-variate balance tests

Co-variates for Precision

- Suppose variance of y is large but can be explained largely by W

- Then we can use W to remove all the explained variation from y
- Then perform our ATE analysis on the remnant variation

- This is oftentimes performed in practice via ordinary linear regression of y on the vector $(1, D, W)$ (after centering W, i.e. $E[W] = 0$)

Is this consistent?

- Suppose that the conditional expectation function (CEF) of the outcome is indeed linear, with $(D, 1, W)$

$$E[\,Y \mid D, W\,] = D\alpha + \alpha_0 + W'\beta$$

- Then note that

$$E[Y(0)] = E\big[E[Y|D = 0, W]\big] = \alpha_0$$
$$E[Y(1)] = E\big[E[Y|D = 1, W]\big] = \alpha + \alpha_0$$

- Baseline outcome is coefficient associated with the intercept 1
- Average effect is coefficient associated with treatment D
- Next lecture: this does not require the linear CEF assumption

Appendix

Analysis of Variance (ANOVA)

Sneak peek into some material from next lecture

Variance of Estimate

- The OLS theory that we will cover in the next section yields

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{a}{\sim} N(0, V_\alpha) \qquad V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- ϵ is residual outcome:

$$\epsilon = y - D\alpha - \alpha_0 - W'\beta \qquad E[\epsilon | D, W] = 0 \text{ (by Linear CEF)}$$

- $\widetilde{D}$ is residual treatment (removing whatever is linearly predictable from $(1, W)$)

$$\widetilde{D} = D - E[D]$$

Variance without adjustment

- OLS theory that we will cover in the next section yields

$$V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- ϵ is residual outcome: $\epsilon = y - D\alpha - \alpha_0 - W'\beta$ with $E[\epsilon | D, W] = 0$
- Two means estimate is equivalent to OLS without W. OLS theory gives variance $\bar{V}_\alpha$ of same form but with residual:

$$\bar{\epsilon} = y - D\alpha - \alpha_0 = W'\beta + \epsilon$$

$$\begin{aligned} E\left[\bar{\epsilon}^2 \widetilde{D}^2\right] &= E\left[(W'\beta + \epsilon)^2 \widetilde{D}^2\right] \\ &= E\left[(W'\beta)^2 \widetilde{D}^2\right] + E\left[\epsilon^2 \widetilde{D}^2\right] + 2E\left[\beta' W \epsilon \widetilde{D}^2\right] \\ &= E\left[(W'\beta)^2 \widetilde{D}^2\right] + E\left[\epsilon^2 \widetilde{D}^2\right] + 2E\left[\beta' W E[\,\epsilon \mid D, X\,] \widetilde{D}^2\right] \\ &= E\left[(W'\beta)^2 \widetilde{D}^2\right] + E\left[\epsilon^2 \widetilde{D}^2\right] \end{aligned}$$

$$\bar{V}_\alpha \geq V_\alpha$$

Variance of OLS estimate with extra co-variates (adjusted) is weakly smaller than two-means estimate (un-adjusted)

Heteroskedasticity Robust Variance

- Variance formula

$$V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- is valid even when the linear CEF assumption is violated
- Inference is asymptotically valid!
- Important to note that this formula is known as the "heteroskedasticity robust variance formula" (HC0)
- Many software packages make the simplification that the residual ϵ is independent of D, W, leading to $V_\alpha = E\left[\epsilon^2\right]/E\left[\widetilde{D}^2\right]$. This is incorrect in most cases!

Precision Beyond Linear CEF

- The precision statement invoked the property that the residual of the OLS regression of y on D, X is mean zero conditional on D, X
- If linear CEF is violated, then all we know is the orthogonality property

$$E\left[\epsilon \begin{pmatrix} D \\ 1 \\ W \end{pmatrix}\right] = 0 \left[\text{FOC of: } \min_{\alpha,\alpha_0,\beta} E\left[(y - D\alpha - \alpha_0 - W'\beta)^2\right]\right]$$

- This is not sufficient to argue that the cross-term vanishes

$$E\left[\beta' W \epsilon \widetilde{D}^2\right] = E\left[\beta' E[\, W\epsilon \mid D\,]\widetilde{D}^2\right]$$

- Note that we only need that:

$$E[\, W\epsilon \mid D\,] = 0$$

Let's try it out!

OLS with Interactive Terms

- It is advisable that instead of running OLS of y on $D, 1, W$ we also include interaction terms, i.e. y on $D, 1, W, DW$ [Lin'13]
- In the absence of any model assumptions, the coefficient of D and of the intercept, recover the ATE and the mean baseline outcome
- These interactive terms enforce the residual ϵ of OLS to satisfy the stronger orthogonality property with $X = (1, W)$

$$E\left[\epsilon \begin{pmatrix} X \\ DX \end{pmatrix}\right] = 0$$

- $E[\epsilon DX] = 0 \Rightarrow E[\,\epsilon X \mid D = 1\,] = 0 \Rightarrow E[\,\epsilon X \mid D = 0\,] = 0$
- Interaction term in $\bar{V}_\alpha$ is zero and we get that $\bar{V}_\alpha \geq V_\alpha$ without assumptions
- OLS with interactive terms always has weakly smaller variance than two means estimate!

Let's try it out!

Even if you only care about ATE, if you have p covariates and $p \ll n$ run OLS with interactive terms!

Guaranteed improved precision, plus can uncover potential dimensions of heterogeneity

Example: Heterogeneous Effects

- Suppose that:

$$E[\,y \mid D, X\,] = D\underbrace{\boxed{\alpha}}_{\text{ATE}} + \alpha_0 + D\underbrace{\boxed{W'\gamma}}_{\text{effect modifier}} + W'\beta$$

- What OLS is estimating is the solution to:

$$E\left[\left(y - D\tilde{\alpha} - \tilde{\alpha}_0 - W'\tilde{\beta}\right)\begin{pmatrix} D \\ X \end{pmatrix}\right] = 0$$

$$E\left[\left(D\alpha + \alpha_0 + DW'\gamma + W'\beta - D\tilde{\alpha} - \tilde{\alpha}_0 - W'\tilde{\beta}\right)\begin{pmatrix} D \\ X \end{pmatrix}\right] = 0$$

- Since $E[W] = 0$ and $W \perp D$: $\alpha = \tilde{\alpha}, \alpha_0 = \tilde{\alpha}_0$
- But $\tilde{\beta} = \beta + E[D]\gamma$
- Residual of OLS is $\epsilon = \widetilde{D}W'\gamma + \nu$ with $E[\nu|D, W] = 0$
- Then interaction term is:

$$\mathrm{E}\left[\beta' W \epsilon \widetilde{D}^2\right] = E\left[\widetilde{D}^3\right] \beta' E[WW']\gamma \neq 0$$