

# PMDA

## Lecture 3: Designing A/B tests





# Slides on Canvas: module 3



# Housekeeping

- Professor office hours by appointment
- Problem Set # 1: Jan 27, 11.59 pm
- Problem Set # 2: Feb 11, 11.59 pm

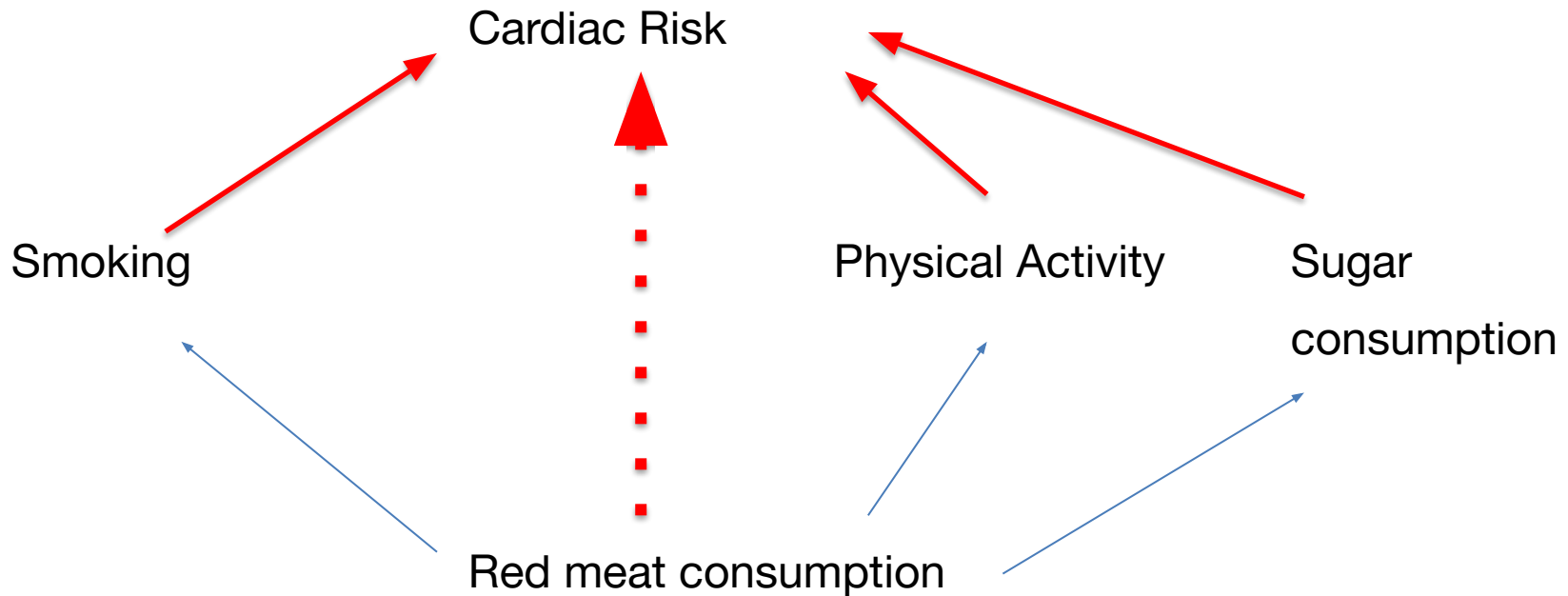
# Road map

- Re-cap: precision & causality
- A/B test design
  - Determinants of precision
  - Using control variables to improve precision

# Re-cap: two key concepts

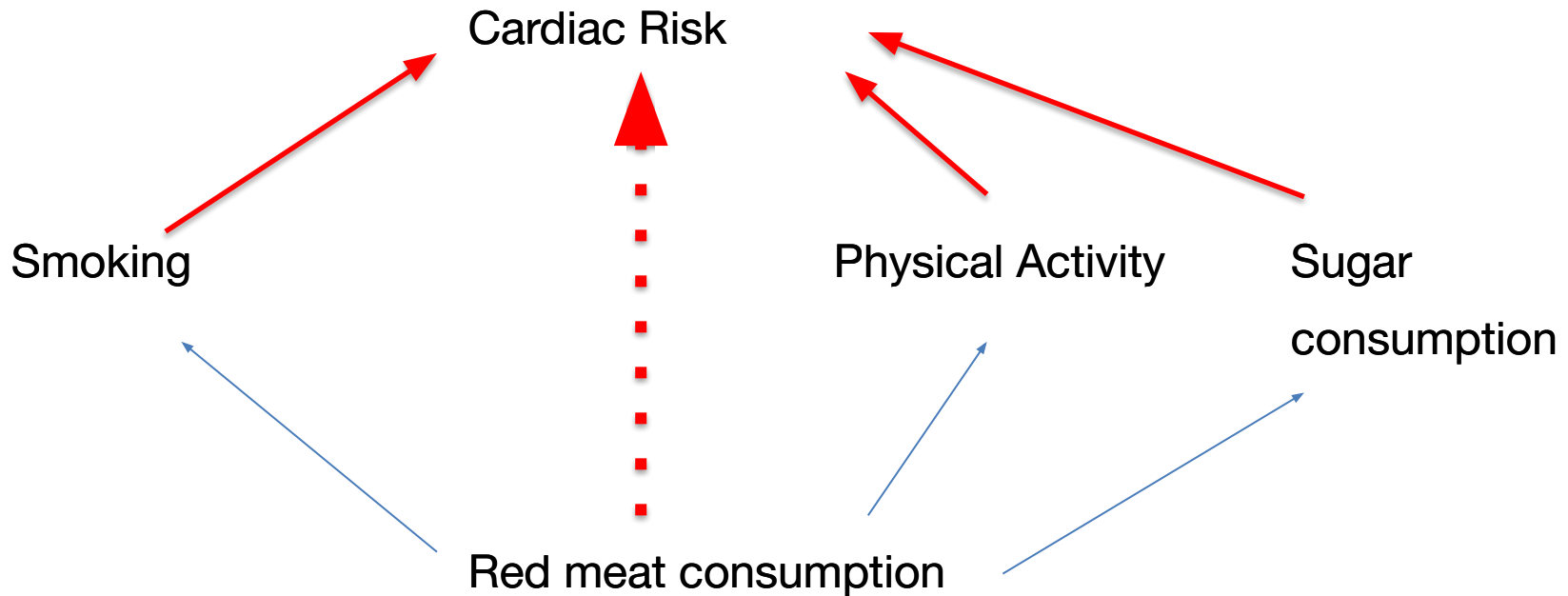
- Precision (week 1):
  - How variable is my estimate? i.e., how much should I trust it?
  - Can use information from one sample to assess how estimator behaves across samples
  - If it is highly variable across samples, we should trust it less
  - Quantify precision via Standard Error, T-stats, P-values
- Causality (week 2):
  - Does the coefficient on variable X capture the impact of manipulating “X and X alone” on the outcome Y?
- Relationship between the two:
  - Statistical significance and causality are separate concepts
  - Causality not a statistical concept, i.e., can't easily test for it
  - Causality established “by design” or “by assumption”

# Red meat causal graph in observational data



- Red arrows: Factors that directly impact cardiac risk
- Blue arrows: red meat correlates with factors that influence cardiac risk
- Object of interest: dashed red arrow (causal effect of red meat on cardiac risk)

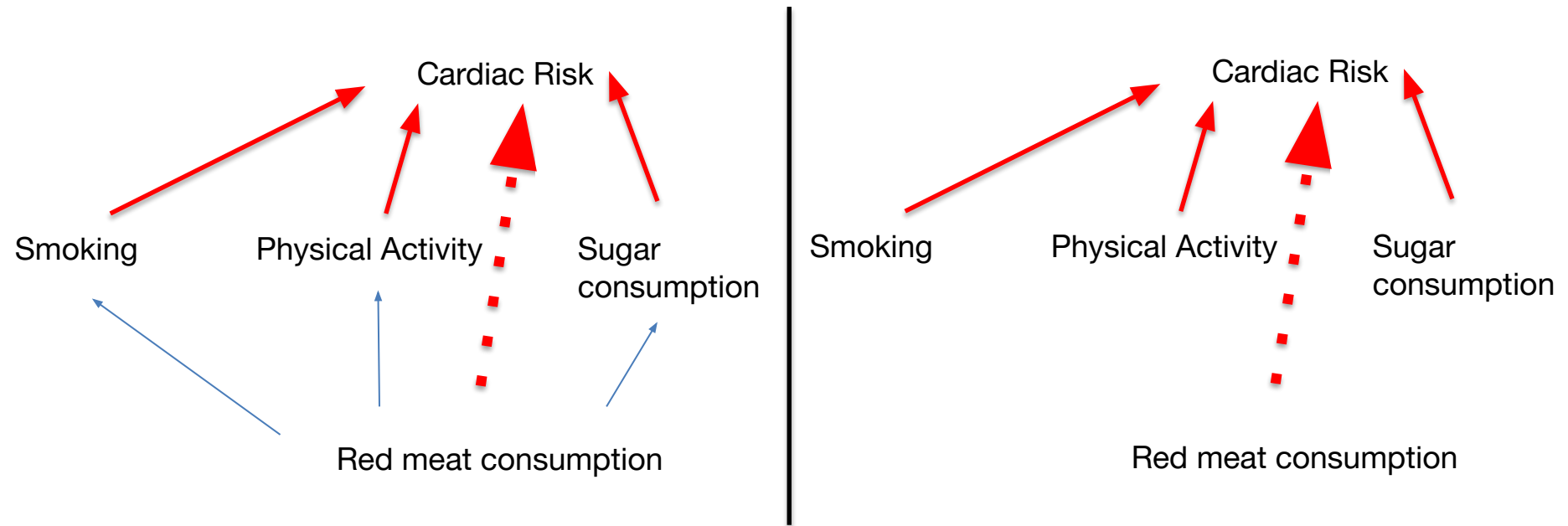
# Red meat causal graph in experimental data



**Q:** Assume we are able to assign red meat consumption randomly. How does the graph above change?

**A:** Blue arrows disappear! We can now measure the impact of varying red meat and red meat alone

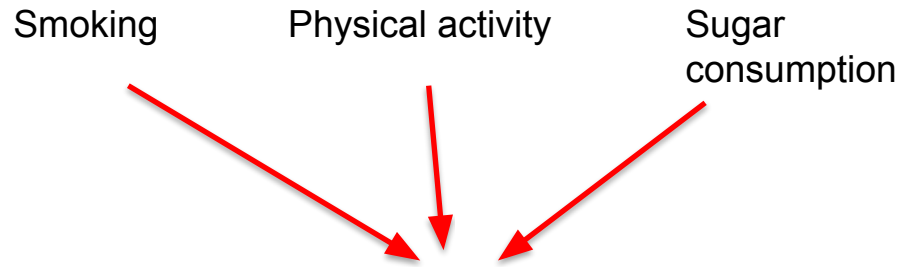
# “Testing” for no causality



- **Q:** How can you assess whether your data was generated from an experiment?
- **A:** Check whether red meat correlates with smoking, sugar, etc. i.e., test whether blue arrows are present or not
- **Note:** this is imperfect: causality requires uncorrelatedness with ALL variables, but we can only test subset ( → we can only falsify causality, never prove it)
- In practice, first thing you do with experimental data is to verify random assignment by testing if the treatment variable is uncorrelated with observables



# Causality and regression



- Hypothetical experimental data

$$CardiacRisk = \beta_0 + \beta_1 redmeat + e$$

- We would obtain an unbiased estimator of the causal effect of red meat

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{Cov(redmeat, CardiacRisk)}{Var(redmeat)} = \frac{Cov(redmeat, \beta_0 + \beta_1 redmeat + e)}{Var(redmeat)} \\ &= \beta_1 + \frac{Cov(redmeat, e)}{Var(redmeat)} = \beta_1 \end{aligned}$$

because redmeat is randomly assigned  $\rightarrow Cov(redmeat, e) = 0$  (this is the “no blue arrows” condition in the causal graph)

# Precision

- In week 1, we discussed the variance and standard error of the uni-variate regression slope coefficient

$$SE = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\frac{1}{N} \cdot \frac{s^2}{Var(X)}}$$

- From this expression we can derive t-stats and p-values in the usual way:
  - A standard test is for “statistical significance”. That is, a test of the null hypothesis that  $\beta_1 = 0$ . Under this hypothesis we have

The diagram illustrates the derivation of the test statistic. It starts with the normal distribution of the regression coefficient:  $\hat{\beta}_1 \sim N\left(0, \frac{1}{N} \frac{s^2}{Var(X)}\right)$ . A red arrow points from the '0' in the mean to the expression  $(\hat{\beta}_1 - 0)/SE$ , which is enclosed in a red box. Another red arrow points from the text 'Test statistic' below to this boxed expression. A red arrow also points from the text 'Under this hypothesis we have' in the list above to the '0' in the mean.

$$\hat{\beta}_1 \sim N\left(0, \frac{1}{N} \frac{s^2}{Var(X)}\right) \longrightarrow (\hat{\beta}_1 - 0)/SE \sim N(0, 1)$$

Test statistic

# Significance in A/B tests: some examples

1. We run an A/B test for a new ad (versus not showing the ad), the true effect of the ad is zero
  - Causal estimate b/c of random assignment
  - Coefficient will be close to zero and insignificant
2. A/B test for a new ad that has a positive effect, but small sample size
  - Causal estimate b/c of random assignment
  - Coefficient could be insignificant b/c of small sample
3. Observational data on ad exposure from large data set
  - No causal interpretation
  - Coefficient could be significant

Rest of today: we can get causal estimates from A/B test, but lack of significance is a potential problem: can we influence it?

# Yahoo A/B testing study

- With cooperation from Yahoo! and a major retailer, Lewis and Reiley (2014) implemented a large scale A/B test
- Randomize which users are exposed to ads (treated) and which users are not exposed (controls) and match them with the retailer's offline sales

Figure 2– Yahoo! Front Page with Large Rectangular Advertisement



# Statistical vs. economic significance

- A/B test shows that revenue per ad increases by \$0.05 but is not statistically significant
- Say ad costs \$0.04 to place (cost is often quite low)

→ Estimated effect leads to positive ROI, but result is not statistically significant

**Q:** What to do? Run the ad or not?

**A:** Since this is an experiment, you could try to manipulate the design to see if you can find economic significance (in this case, a positive ROI)



# Statistical vs. economic significance


- **Q:** In the ad example, what are the null and alternative hypotheses of interest if you want a positive ROI?
- **A:** A positive ROI means that the effect of the ad must be greater than the cost, so you put zero ROI under the null and positive ROI under the alternative

$$H_0 : \beta_1 = 0.04$$

$$H_1 : \beta_1 > 0.04$$

- **Q:** What can you do to affect the probability of rejection?
- **A:** The test statistic is  $t\text{-stat} = (\hat{\beta}_1 - 0.04) / SE$ .  $\hat{\beta}_1$  is an unbiased estimator because we have experimental data. The only thing we can manipulate is to **increase precision** of estimator so we are more likely to find significance

$$t - stat = \frac{\hat{\beta}_1 - 0.04}{SE} \longrightarrow \text{more likely to reject}$$

  
goes up                      goes down

# Power analysis

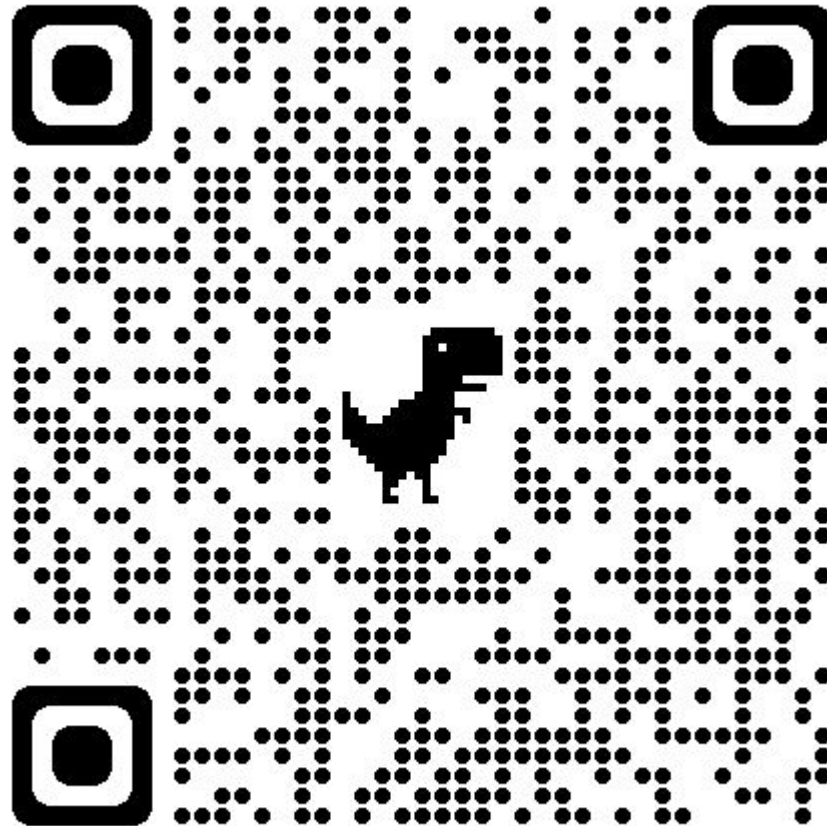
- What we discuss next is related to what is often called a **power analysis** for experiment design
- This refers to the power of a test

$$Power = Pr(\text{Reject } H_0 | H_1 \text{ is true})$$

- To be precise, you can only compute this probability if both  $H_0$  and  $H_1$  are equalities, which is not common in economics for  $H_1$
- Even for our  $H_1$ , we now see how to design an experiment that ensures that the test has power to reject the null of zero ROI if the true ROI is positive

# Attendance!

**Link will close soon!**



# Road map

- Re-cap: precision & causality
- A/B test design
  - **Determinants of precision**
  - Using control variables to improve precision

# How to increase precision in A/B tests

- To put yourself in the best position to find economic significance of your treatment (e.g., seeing an ad), you want to manipulate the precision of the estimator at the experiment design stage
- Remember the standard error of the univariate regression coefficient

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{N} \frac{s^2}{Var(X)}}$$

- We can thus increase precision (= lower SE) by choosing:
  1. Larger variance of the treatment variable  $X$
  2. Lower variance of regression residual  $s^2$
  3. Larger sample size  $N$



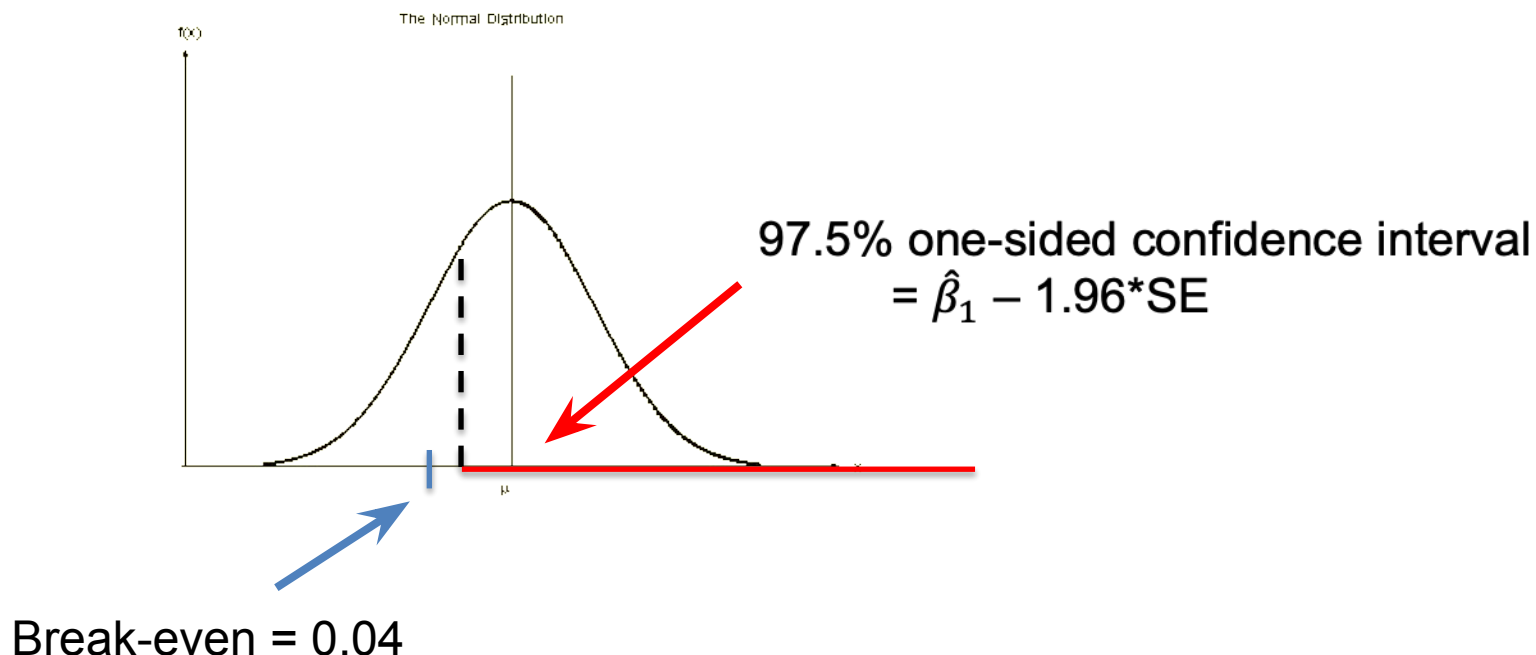
# How to increase precision in A/B tests in practice

1. Determine the SE you need in order to reject the null hypothesis. For a positive ROI in the example where an ad costs \$0.04, the hypothesis is
$$H_0 : \beta_1 = 0.04$$
$$H_1 : \beta_1 > 0.04$$
2. Pick a confidence level, e.g. 97.5% (which gives you the easy-to-use critical value 1.96 for a one-sided test)
3. You reject if  $(\hat{\beta}_1 - 0.04)/SE > 1.96$  which means  $SE < (\hat{\beta}_1 - 0.04)/1.96$ . So, given a preliminary estimate of  $\hat{\beta}_1$  choose a SE that satisfies this condition
4. Choose  $p = \text{prob of treatment}$  to maximize the variance of the treatment variable  $X$ . Since for a binary variable  $\text{Var}(X) = p^*(1-p) \rightarrow$  largest if we choose  $p = 0.5$
5. Get a preliminary estimate of  $s^2$  by running a multivariate regression with as many controls as you can (we will see later why this makes sense)
6. Given the chosen SE,  $\text{Var}(X)$  and  $s^2$ , choose the sample size as

$$N = \frac{s^2}{\text{Var}(X)} \frac{1}{SE^2}$$

# Graphical intuition for SE choice

- We have a preliminary estimate of ad effect  $\hat{\beta}_1$  and we profit if effect  $> \$0.04$



- We choose SE so that 0.04 is outside the 97.5% one-sided confidence interval for the true ad effect

# Example: Ad effectiveness and precision

- Pick SE for a significance level 97.5%
  - $SE < (0.05-0.04)/1.96 = 0.01/1.96 = 0.0051$
  - So choose  $SE = 0.005$  (get this quickly by dividing by 2:  $SE = 0.01/2$ )
- Pick  $p =$  probability of treatment  $= 0.5$  and a preliminary value for  $s^2$
- Pick the sample size:

$$N = \frac{s^2}{Var(X)} \cdot \frac{1}{SE^2} = 4 \cdot \frac{s^2}{0.005^2} = 160,000 \cdot s^2$$

# Road map

- Re-cap: precision & causality
- A/B test design
  - Determinants of precision
  - Using control variables to improve precision

# Using control variables in A/B tests

- So far we used a univariate regression:

$$Purchase = \beta_0 + \beta_1 \cdot treatment + e$$

- We can also use a multivariate regression

$$Purchase = \gamma_0 + \gamma_1 \cdot treatment + \gamma_2 X_2 + \gamma_3 X_3 + \cdots + u$$

- Where  $X_2$ ,  $X_3$ , etc. are control variables (such as consumer demographics)

**Q:** When including controls, does the treatment coefficient change?

**A:** No, since treatment is uncorrelated with controls b/c of randomization

Unlike in normal regressions, in A/B tests the reason for using controls IS NOT to remove omitted variable bias!



# Control variables and precision improvements

- So why are controls helpful in A/B tests? **To improve precision**
- Next week we will see a theorem (Frisch-Waugh theorem) that expresses the multivariate estimator in terms of sequential estimation. The theorem implies

$$\text{Var}(\hat{\gamma}_{1, \text{multi-var}}) = \frac{1}{N} \cdot \frac{\text{Var}(u)}{\widetilde{\text{Var}(\text{treatment})}}$$

- $\widetilde{\text{Var}(\text{treatment})}$  is the variance of the residual from a regression of treatment on controls:

$$\text{treatment} = a_0 + a_1 X_2 + a_2 X_3 + \cdots + \widetilde{\text{treatment}}$$

- $\text{Var}(u)$  = variance of multivariate regression residuals from the previous slide
- Let's compare the precision of the multivariate to that of the univariate estimator

$$\text{Var}(\hat{\beta}_{1, \text{uni-var}}) = \frac{1}{N} \cdot \frac{\text{Var}(e)}{\widetilde{\text{Var}(\text{treatment})}}$$

# Control variables and precision improvements

$$Var(\hat{\beta}_{1,uni-var}) = \frac{1}{N} \cdot \frac{Var(e)}{Var(treatment)}$$

$$Var(\hat{\gamma}_{1,multi-var}) = \frac{1}{N} \cdot \frac{Var(u)}{\widetilde{Var(treatment)}}$$

**Q:** How do  $Var(treatment)$  and  $\widetilde{Var(treatment)}$  relate to each other?

slido



# How do $\text{Var}(\text{treatment})$ and $\text{Var}(\sim\text{treatment})$ compare?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

# Control variables and precision improvements

$$\text{Var}(\hat{\beta}_{1,uni-var}) = \frac{1}{N} \cdot \frac{\text{Var}(e)}{\text{Var}(treatment)}$$

$$\text{Var}(\hat{\gamma}_{1,multi-var}) = \frac{1}{N} \cdot \frac{\text{Var}(u)}{\widetilde{\text{Var}(treatment)}}$$

**Q:** How do  $\text{Var}(treatment)$  and  $\widetilde{\text{Var}(treatment)}$  relate to each other?

**A:**

$$treatment = a_0 + a_1X_2 + a_2X_3 + \cdots + \widetilde{treatment}$$

They are (almost) the same because if treatment is randomly assigned the X-variables are uncorrelated with treatment so true coefficients  $a$ 's are zero

# Control variables and precision improvements

$$Var(\hat{\beta}_{1,uni-var}) = \frac{1}{N} \cdot \frac{Var(e)}{Var(treatment)}$$

$$Var(\hat{\gamma}_{1,multi-var}) = \frac{1}{N} \cdot \frac{Var(u)}{Var(\widetilde{treatment})}$$

**Q:** How do  $Var(e)$  and  $Var(u)$  relate to each other?



slido



How do  $\text{Var}(e)$  and  $\text{Var}(u)$  relate to each other?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

# Control variables and precision improvements

$$\text{Var}(\hat{\beta}_{1,uni-var}) = \frac{1}{N} \cdot \frac{\text{Var}(e)}{\text{Var}(treatment)}$$

$$\text{Var}(\hat{\gamma}_{1,multi-var}) = \frac{1}{N} \cdot \frac{\text{Var}(u)}{\text{Var}(\widetilde{treatment})}$$

**Q:** How do  $\text{Var}(e)$  and  $\text{Var}(u)$  relate to each other?

**A:**  $\text{Var}(u) < \text{Var}(e)$  because the multivariate regression has more explanatory power

# Justification for preliminary estimate of $s^2$

- We saw that when you design the experiment you need to pick a preliminary value for  $s^2$ , the variance of the residual in the regression you eventually want to run to establish the treatment effect
- We just saw that the regression you eventually want to run is a regression with controls, because this will improve the precision of the estimate of  $\gamma_1$

$$Purchase = \gamma_0 + \gamma_1 \cdot treatment + \gamma_2 X_2 + \gamma_3 X_3 + \cdots + u$$

- Before you run the experiment, you can get a preliminary estimate of  $s^2 = \text{Var}(u)$  by running the regression of Purchase on  $X_2, X_3$  etc. on observational data, if you have access to it
- If you don't have any data besides the pilot you run to get a preliminary estimate of  $\gamma_1$ , you can use the  $s^2$  from the multivariate regression in the pilot

# Summary: A/B testing

- Understand the power of randomization
  - Simple method to make causal statements → reliable, scalable, transparent, easy to communicate
  - Only important issue: **randomly assign treatment** variable
- Learn how to run more precise A/B tests
- Three ways to affect precision
  - Maximize  $\text{Var}(X)$  by splitting sample equally into treatment/control
  - Reduce variance of regression residual through control variables
  - Choose sample size (after optimizing along other two margins)

# Break & workshop setup

- We will take a break here !
- But first ... some logistics for the workshop:
  - You can work on the exercise as a team
  - Workshop is not graded, just for your own understanding
  - You should fill out the Jupyter notebook with code & answers
  - We will reconvene for last 15-20 minutes or so to discuss answers
  - Detailed solutions will also be provided via Canvas

# Summary: workshop

## Additional insights from workshop:

1. We can test randomization by evaluating whether X-variables can predict treatment
  - If random, treatment should be uncorrelated with EVERYTHING else and hence coefficients on X-variables should be insignificant
  
2. What are good controls to improve precision?
  - Information on past behavior tends to be very powerful
  - This is a general finding in many settings
  - E.g. in consumer packaged goods markets, knowing which brand a consumer purchased is very predictive of future purchases (demographics less so)