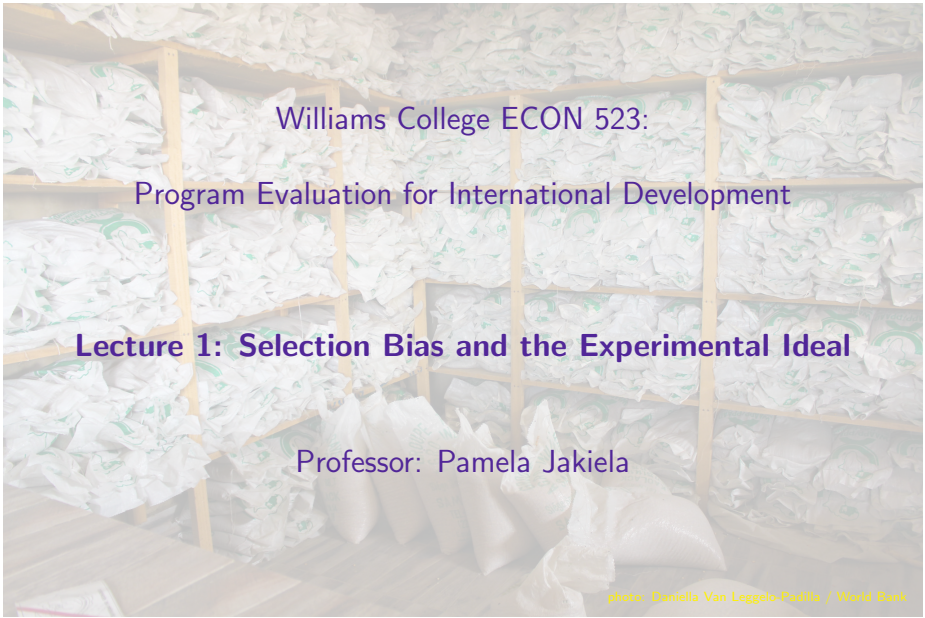


Williams College ECON 523:

Program Evaluation for International Development

Lecture 1: Selection Bias and the Experimental Ideal

Professor: Pamela Jakiela

photo: Daniella Van Leggelo-Padilla / World Bank

Potential Outcomes

Do Hospitals Make People Healthier?

Your health status is: excellent, very good, good, fair, or poor?

	Hospital	**No Hospital**	**Difference**
Health status	3.21	3.93	-0.72^{***}
	(0.014)	(0.003)	
Observations	7,774	90,049	

source: 2005 National Health Interview Survey (Angrist & Pischke 2009)

A comparison of means suggests hospitals make people worse off: those with a hospital stay in last 6 months are, on average, less healthy than those that were not admitted to the hospital

- What's wrong with this picture?

The Causal Impact of Treatment

We are interested in the relationship between some "**treatment**" (e.g. going to the hospital) and some outcome that may be impacted by the treatment (eg. self-assessed health status)

Each individual is either treated or not:

- D_i = is a **treatment dummy** equal to 1 if i is treated and 0 otherwise

Outcome of interest:

- Y = outcome we are interested in studying (e.g. health)
- Y_i = value of outcome of interest *for individual i*

Potential Outcomes

For each individual, there are two **potential outcomes**:

- $Y_{0,i} = i$'s outcome if she **doesn't** receive treatment
- $Y_{1,i} = i$'s outcome if she **does** receive treatment

The **causal impact** of treatment on individual i is: $Y_{1,i} - Y_{0,i}$

- How much does treatment change outcome of interest for i?
- We are interested in **average treatment effect** – average of $Y_{1,i} - Y_{0,i}$ across people

Potential Outcomes: Example

Alejandro has a broken leg.

- $Y_{0,a}$ = If he doesn't go to the hospital, his leg won't heal properly
- $Y_{1,a}$ = If he goes to the hospital, his leg heals completely

Benicio doesn't have any broken bones. His health is fine.

- $Y_{0,b}$ = If he doesn't go to the hospital, his health is still fine
- $Y_{1,b}$ = If he goes to the hospital, his health is still fine

Potential Outcomes: Example

	Yes Hospital	**No Hospital**
Alejandro	$Y_{1,a}$	$Y_{0,a}$
Benicio	$Y_{1,b}$	$Y_{0,b}$

The Fundamental Problem of Causal Inference

The fundamental problem of causal inference:

We never observe both potential outcomes for the same individual

$\Rightarrow$ Creates a missing data problem whenever we try to compare treated to untreated

For any individual, we can only observe one potential outcome:

$$Y_i = \begin{cases} Y_{0,i} & \text{if } D_i = 0 \\ Y_{1,i} & \text{if } D_i = 1 \end{cases}$$

Potential outcomes without treatment (i.e. values of $Y_{0,i}$) may differ between those who choose to take-up treatment (Alejandro with a broken leg) and those who do not (healthy Benicio)

Selection Bias: Example

	$Y_{1,i}$	$Y_{0,i}$
Alice	6	4
Betty	7	5
Carol	3	1
Diana	4	2

Selection Bias: Example

	$Y_{1,i}$	$Y_{0,i}$
Alice	6	4
Betty	7	5
Carol	3	1
Diana	4	2

Alice and Betty take up treatment

Selection Bias: Example

	$Y_{1,i}$	$Y_{0,i}$
Alice	6	
Betty	7	
Carol		1
Diana		2

Alice and Betty take up treatment

$$\Rightarrow \bar{Y}_{treatment} = 6.5$$

Carol and Diana do not participate

$$\Rightarrow \bar{Y}_{comparison} = 1.5$$

$$\bar{Y}_{treatment} - \bar{Y}_{comparison} = 6.5 - 1.5 = 5$$

Selection Bias

Comparing the mean outcome among program participants to the mean outcome among those who don't choose to participate doesn't normally provide an unbiased estimate of causal impact

- Treated, untreated likely different in absence of program
- Difference in potential outcomes without treatment leads to **selection bias**
- The difference in outcome means, $\bar{Y}_T - \bar{Y}_C$, is a biased estimator of program impacts
- $\bar{Y}_T - \bar{Y}_C$ could be biased up or down, relative to true average causal effect of treatment
- Bias does not disappear in large samples, even large numbers of controls may not help

Notation: Mathematical Expectations

The **expected value** or mathematical expectation of Y_i, $E[Y_i]$:

- Equivalent to population mean or sample average in an infinite population
  - Example: probability coin flipped lands heads
  - Equivalent to fraction heads after a (very) large number of flips

Law of Large Numbers:

- In small samples, realized average of Y_i might be far from the true mean of Y_i
- Average of Y_i gets very close to $E[Y_i]$ as number of observations gets large

Notation: Conditional Expectations

Conditional expectation:

$E[Y_i|X_i = x]$

Conditional expectation of Y_i given $X_i = x$ is average of Y_i in infinite population where $X_i = x$

Example:
Let Y_i be height, and let $X_i \in \{0, 1\}$ be an "economics professor dummy"

- $E[Y_i|X_i = 1]$ is the average height among (infinitely many) economics professors
- $E[Y_i|X_i = 0]$ is the population mean of height among everybody else

Notation: Average Treatment Effect (ATE)

The quantity of interest is the **average treatment effect** (ATE), or average causal effect, or conditional average treatment effect, or average impact, or treatment effect...

$$E[Y_{1,i} - Y_{0,i}|D_i = 1] = E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]$$

- ATE is average difference in potential outcomes (usually) across treated population
- Fundamental problem of causal inference: we never observe $Y_{0,i}$ for treatment group
  - $\bar{Y}_T$ is an unbiased estimator of $E[Y_i|D_i = 1] = E[Y_{1,i}|D_i = 1]$
  - We need an unbiased estimator of $E[Y_{0,i}|D_i = 1]$

Notation: Selection Bias

When we compare (many) participants to (many) non-participants:

$$E[\bar{Y}_T - \bar{Y}_C] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$
$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

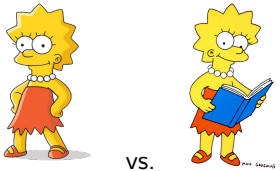
Adding in $\underbrace{-E[Y_{0,i}|D_i = 1] + E[Y_{0,i}|D_i = 1]}_{=0}$, we get:

Difference in group means

$$= \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{average causal effect on participants}} + \underbrace{E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]}_{\text{selection bias}}$$

Summary

We would like to calculate average treatment effect by comparing potential outcomes for i both with and without treatment, but for each i we can only observe one potential outcome

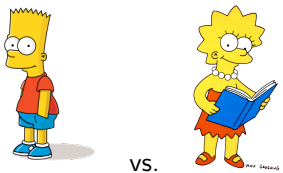


In the real world, we either observe Lisa Simpson with a textbook or without a textbook

- Can't observe the **counterfactual** (i.e. the other potential outcome)
- We need to find a comparison group to approximate Lisa's outcome without a textbook

Summary

To estimate causal impacts on the set of people who choose to take up treatment, we must identify a comparison group that is similar to the treatment group in the absence of treatment



This is hard – typically impossible in observational data

- An **identification strategy** is a research design specifying treatment, comparison groups
- A good identification strategy: variation in treatment status that is good-as-random

The Experimental Ideal

Random Assignment Eliminates Selection Bias

Experimental approach:

- **Random assignment to treatment:** eligibility for program is determined at random, e.g. via pulling names out of a hat, or using a computer pseudo-random number generator

When treatment status is randomly assigned,
treatment, control groups are random samples of a single population
(e.g. the population of all eligible applicants for the program)

$$\Rightarrow E[Y_{0,i}|D_i = 1] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Expected outcomes are equal in the absence of the program

Random Assignment Eliminates Selection Bias

$\bar{Y}_T - \bar{Y}_C$ provides an unbiased estimate of the (casual) average treatment effect (or ATE):

$$= E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1] + E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

$$= \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{average treatment effect on participants}} + \underbrace{E[Y_{0,i}] - E[Y_{0,i}]}_{=0}$$

$$= \underbrace{E[Y_{1,i}] - E[Y_{0,i}]}_{\text{ATE}}$$

Random Assignment Eliminates Selection Bias: Assumptions

Excellent news: random assignment eliminates selection bias*

*Some restrictions apply

Random assignment is not (quite) magic:

- Relies on Law of Large Numbers, which only makes sense for large(ish) samples
- Stable Unit Treatment Value Assumption (SUTVA): individual outcomes depend on one's own treatment status, but not on anyone else's treatment status (i.e. no spillovers)

Sample Size Matters: Example

Example: imagine that I want to evaluate the impact of fancy new software Stata 138, so I randomly choose which of my two research assistants (below) should receive a copy

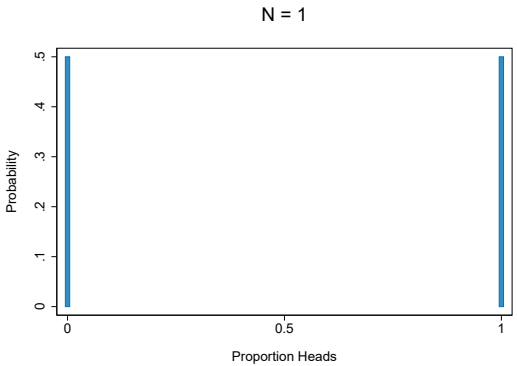
They're different! Omitted variables likely to matter – by chance – in small samples

> *"Randomization works not by eliminating individual difference but rather by ensuring that the mix of individuals being compared is the same. Think of this as comparing barrels that include equal proportions of apples and oranges."*
>
> – Angrist and Pischke (2009)

The Law of Large Numbers in Practice

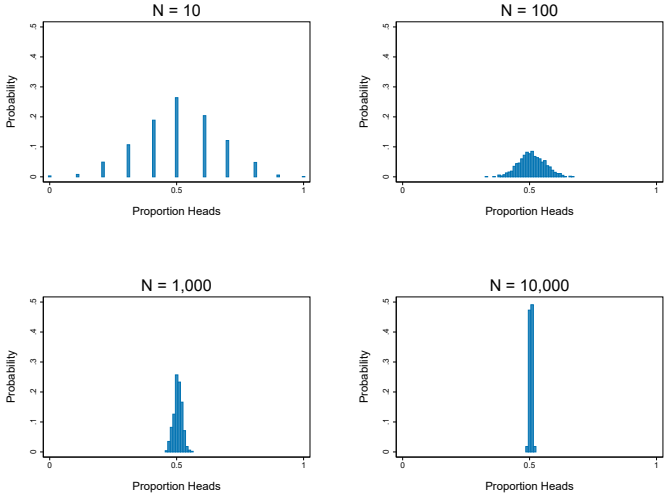
The probability a fair coin lands "heads" is 0.5, but the observed average proportion heads after a single coin flip is either 0 or 1

The Law of Large Numbers in Practice



Law of Large Numbers: sample average can be brought as close as we like to population mean (i.e. probability that average is far from population mean can be made as low as we like)

The Law of Large Numbers in Practice



Stable Unit Treatment Value Assumption (SUTVA)

The **Stable Unit Treatment Value Assumption (SUTVA)**:

- Imbens and Rubin (2015):
  "potential outcomes for any unit do not vary with the treatments assigned to other units"
- Remember: binary treatment, two potential outcomes is only a model

When is SUTVA likely to be violated?

- When there are spillovers (so i's treatment impacts j)
- Examples: vaccination/health, network externalities, equilibrium effects
  - This is why we have "cluster-randomized" trials

Summary

When treatment is randomly assigned (at an appropriate level), difference in outcomes between treatment and control groups provides an unbiased estimate of the causal impact of treatment

Randomly assigning treatment status eliminates selection bias (at least in expectation) because treatment, control groups are random samples of same underlying population

Randomization: A Short History

The Idea of Randomization

Petrarch (1364):

"If a hundred thousand men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one half followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on nature's instincts, I have no doubt as to which half would escape."

van Helmont (who died in 1644):

"Let us take out of the Hospitals, pit of the Camps, or from elsewhere, 200 or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one half of them may fall to my share, and the other to yours; I will cure them without bloodletting...
we shall see how many Funerals both of us shall have."

Source: Jamison (2019)

Randomization: A Timeline (Part I)

1885 Psychologists Charles Pierce and Joseph Jastrow use randomization in a psychology experiment, varying the order in which stimuli are presented to subjects (not to estimate treatment effects)

1898 Johannes Fibiger conducts a trial of an anti-diphtheria serum in which every other subject was assigned to treatment (or control), considered to be the first controlled clinical trial

1923 Jerzy Neyman suggests the idea of potential outcomes

1925 **Ronald Fisher suggests explicit randomization of treatments (in agricultural experiments)**

1926 J.B. Amberson et al. study of sanocrysin treatments for tuberculosis: flipped a coin to determine which group received sanocrysin treatment, which group served as controls

1948 Randomized trial of streptomycin treatment for tuberculosis conducted by the Medical Research Council of Great Britain, first medical trial where treatment randomized at individual level

$\Rightarrow$ Randomized evaluations become the norm in medicine

The Lady Tasting Tea

Chapter II of Fisher's *The Design of Experiments* begins:

> *"A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup."*

The Lady Tasting Tea

Null hypothesis (aka H_0):

- Fisher believes that Dr. Bristol cannot taste the difference

A test of the hypothesis:

- *"Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in random order."*

Research design:

- Treatment: an indicator for having the milk poured in first
- Outcome of interest: a dummy for Muriel Bristol believing the milk was poured in first

The Lady Tasting Tea: Experimental Design

Rule #1: do not confound your own treatment

- Critical assumption: if Dr. Bristol is unable to detect whether the milk was poured in first, she will choose four cups at random (probability of selection equal for treatment, control)
  - Allows us to calculate probability four correct cups chosen by chance "under the null"
- Fisher points out that the experimenter could screw this up:

> *"If all those cups made with the milk first had sugar added,*
> *while those made with the tea first had none,*
> *a very obvious difference in flavour would have been introduced*
> *which might well ensure that all those made with sugar should be classed alike."*

- Gerber and Green (2012) refer to this as **excludability**

The Lady Tasting Tea: Experimental Design

Rule #1B: do not accidentally confound your own treatment

- Fisher, in (perhaps) the earliest known scientific subtweet:

> *"It is not sufficient remedy to insist that*
> *'all the cups must be exactly alike' in every respect except that to be tested.*
> *For this is a totally impossible requirement."*

- To minimize likelihood of accidentally confounding your treatment, it's best is to constrain yourself by randomizing treatment assignments (à la Pierce and Jastrow, British TB trial)
  - Minimizes the likelihood of unfortunate coincidences (in some circumstances)
  - Highly controversial position at the time, and is still debated in some circles; alternative is to force balance on observables (and then just hope that unobservables don't matter too much)

The Lady Tasting Tea: A Hypothesis Test

How should we interpret data from this experiment?

Suppose Dr. Bristol correctly identified all 4 "treated" cups

- How likely is it that this could have occurred by chance?
  - There are $\binom{8}{4} = 70$ possible ways to choose 4 of 8 cups, and only one is correct
  - A subject with no ability to tell treated from untreated cups has a $1/70$ chance of success
  - The **p-value** is the probability that an outcome at least as extreme as the one observed would occur under the null (i.e. if the null hypothesis of no treatment effect were true)
  - The p-value associated with this outcome is $1/70 \approx 0.014$, less than the cutoff for the "standard level of significance" of 0.05 (as characterized by Fisher himself)

Fisher's Exact Test

	Identified by Dr. Bristol?	
	Yes	**No**
Milk poured first	a	b
Tea poured first	c	d

Is Dr. Bristol more likely to select cups where the milk was poured first?

- She chooses a of $a + b$ treated cups correctly, and c of $c + d$ untreated cups incorrectly
- How likely was such an outcome to have occurred at random (under the null)?

Fisher's Exact Test

Is Dr. Bristol more likely to select cups where the milk was poured first?

$$\text{probability} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{\frac{(a+b)!}{a!b!}\frac{(c+d)!}{c!d!}}{\frac{(a+b+c+d)!}{(a+c)!(b+d)!}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

The p-value is the sum of the probabilities of observing outcomes that are at least as extreme (i.e. at least as unlikely under the null hypothesis that Muriel Bristol chooses cups at random)

The Lady Tasting Tea: Testing Alternative Hypotheses

Suppose Dr. Bristol correctly identified 3 "treated" cups

- How likely is it that this could have occurred by chance?
  - There are $\binom{4}{3} \times \binom{4}{1} = 16$ possible ways to choose 3 of 8 cups
    - There are 17 ways to choose **at least** 3 correct cups
  - The p-value associated with this outcome is $17/70 \approx 0.243$
  - If our cutoff for significance is 0.05, we would not reject the null hypothesis

Implication: we should only reject H_0 if Dr. Bristol identified all 4 treated cups

- In the actual experiment, Dr. Bristol identified all four cups correctly

The Lady Tasting Tea: Size and Power

The size of a test is the likelihood of rejecting a true null (finding an impact when there is none)

- Fisher asserts that a test size 0.05 is typical

Alternative experiment: what if we had treated 3 out of 6 cups?

- There are $\binom{6}{3} = 20$ possible ways to choose 3 of 6 cups
- Best possible p-value is therefore 0.05

Alternative experiment: what if we had treated 3 out of 8 cups?

- There are $\binom{8}{3} = 56$ possible ways to choose 3 of 8 cups
- Best possible p-value is therefore 0.017, which is less than 0.014

$\Rightarrow$ **More trials increases power, best to have equal numbers of treated, untreated cups**

The Lady Tasting Tea: Size and Power

An alternate experiment: an unknown number of treated cups

- Under the null, the probability of getting 8 right is 1 in 2^8
- Probability of getting 7 right is $8/256 = 0.03125$

Design would achieve higher power with the same number of trials

- Research design would make it possible to reject the null hypothesis (that Dr. Bristol cannot tell when the milk is added first) even when her ability to discriminate is imperfect

As Fisher discusses, uncertainty about number of treated cups has other unattractive properties

- Would subjects believe that either 0 or 8 cups might be treated?

Ronald Fisher's Contributions to Statistics

Key lesson to take away from "lady tasting tea" anecdote:
caffeine breaks with colleagues critical to advancement of science

Other contributions:

1. Introduced the modern randomized trial
2. Introduced the idea of permutation tests
3. Fixed "standard" test size at 0.05

Fisher is also a clear example of a not-so-nice man who made a strong contribution to science

Randomization: A Timeline (Part II)

1942 Launch of Cambridge-Somerville Youth Study of at-risk boys

1962 Perry Preschool Project (in Ypsilanti, MI) and Early Training Project (in Murfreesboro, TN) experiments randomized assignment of at-risk, low-income children to high-quality preschools

1967 New Jersey Income Maintenance Experiment (proposed by graduate student Heather Ross), four other negative income tax experiments in the US between 1971 and 1982

1972 Abecedarian Project (in Orange County, NC) randomized intervention for at-risk infants

1974 Rubin introduces the concept of potential outcomes (as we know it)

1994 National Job Corps Study (by Mathematica/US Dept. of Labor)

1995 PROGRESA evaluation launched by Mexican government, evaluated by researchers at IFPRI

1998 Dutch NGO ICS Africa begins randomized trial of "deworming" in Kenyan primary schools... in partnership with Michael Kremer, an Assistant Professor of Economics at Harvard University

RCTs in Development Economics: Mexico's Progresa

photo: Curt Carnemark / World Bank

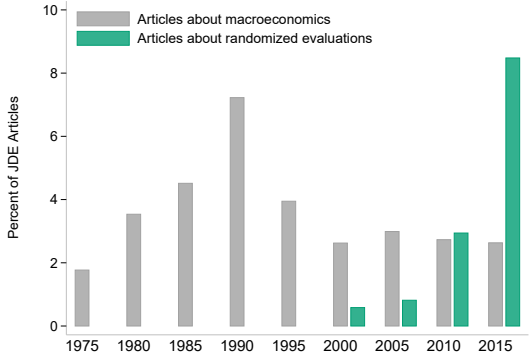
- Mexican government piloted conditional cash transfers (CCTs) in the mid-1990s
- Economists within president's office pushed for randomized roll out of pilot
- IFPRI researchers published initial findings in late 1990s

RCTs in Development Economics: Busia, Kenya

photo: Stephanie Skinner / Deworm the World

- Michael Kremer convinces NGO ICS Africa to randomize interventions in Kenyan schools
- Study of deworming (w/ Edward Miguel) effectively launches RCT movement

RCTs in Development Economics: Trends



Abstracts of 2,695 *Journal of Development Economics* articles
(all articles published prior to 2019, starting form Volume 1 in 1974)

RCTs in Development Economics

In 2019, Michael Kremer, Esther Duflo, and Abhijit Banerjee won the Nobel Prize in economics for their promotion of RCTs and their "experimental approach to alleviating global poverty"

Regression Analysis of RCTs

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i|D_i = 0] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

equal to population mean because control group is a random sample

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i|D_i = 0] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Expected value of treatment group mean:

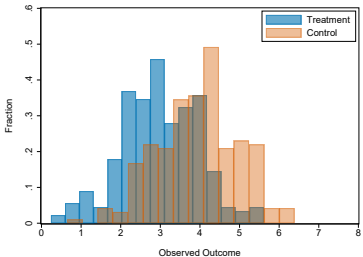
$$\begin{aligned} E[\bar{Y}_T] = E[Y_i|D_i = 1] &= E[Y_{1,i}|D_i = 1] \\ &= E[\delta_i + Y_{0,i}|D_i = 1] \\ &= E[\delta_i|D_i = 1] + E[Y_{0,i}|D_i = 1] \\ &= E[\delta_i] + E[Y_{0,i}] \end{aligned}$$

H_0: $ATE = 0$

Null hypothesis (H_0):

The average treatment effect is zero: $ATE = 0$

Or, equivalently: $\bar{Y}_T = \bar{Y}_C$



In Stata:

```
ttest y, by(t)
```

Testing the Equality of Means in Stata

Stata: `ttest y, by(t)`

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	50	3.00936	.1324447	.9365253	2.743202	3.275517
1	50	4.096623	.1474163	1.042391	3.800379	4.392868
combined	100	3.552992	.1127134	1.127134	3.329344	3.77664
diff		-1.087263	.1981745		-1.480534	-.6939925

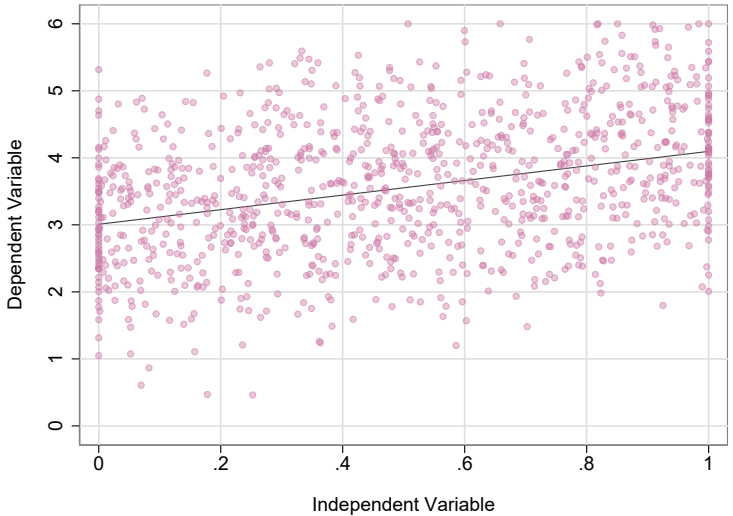
diff = mean(0) - mean(1) t = -5.4864

Ho: diff = 0 degrees of freedom = 98

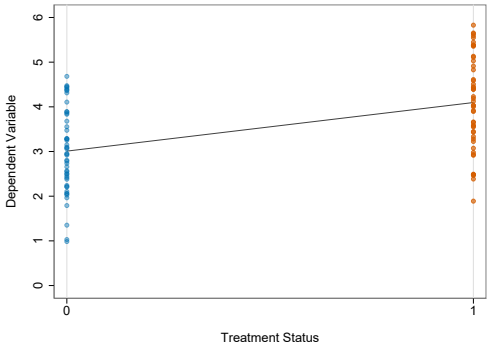
Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

OLS Regression



OLS Regression on a Binary Independent Variable



Simple regression framework for analyzing RCTs: $Y_i = \alpha + \beta D_i + \varepsilon_i$

Treatment indicator $D_i = 0, 1 \Rightarrow$ only two sensible values of $\hat{Y}_i$

OLS Regression on a Binary Independent Variable

Stata: `reg y t`

Source	SS	df	MS	Number of obs =	100
				F(1, 98) =	30.10
Model	29.5535457	1	29.5535457	Prob > F =	0.0000
Residual	96.2192216	98	.981828792	R-squared =	0.2350
				Adj R-squared =	0.2272
Total	125.772767	99	1.27043199	Root MSE =	.99087

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
t	1.087263	.1981745	5.49	0.000	.6939925	1.480534
_cons	3.00936	.1401306	21.48	0.000	2.731275	3.287445

Comparing the Approaches

`ttest y, by(t)`

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	50	3.00936	.1324447	.9365253	2.743202	3.275517
1	50	4.096623	.1474163	1.042391	3.800379	4.392868
combined	100	3.552992	.1127134	1.127134	3.329344	3.77664
diff		-1.087263	.1981745		-1.480534	-.6939925

diff = mean(0) - mean(1) t = -5.4864
Ho: diff = 0 degrees of freedom = 98

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

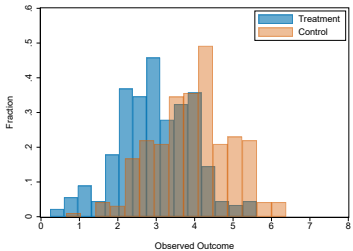
`reg y t`

Source	SS	df	MS		
Model	29.5535457	1	29.5535457	Prob > F	= 0.0000
Residual	96.2192216	98	.981828792	R-squared	= 0.2350
				Adj R-squared	= 0.2272
Total	125.772767	99	1.27043199	Root MSE	= .99087

Number of obs	= 100
F(1, 98)	= 30.10

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
t	1.087263	.1981745	5.49	0.000	.6939925	1.480534
_cons	3.00936	.1401306	21.48	0.000	2.731275	3.287445

The Standard Error of a Difference in Means



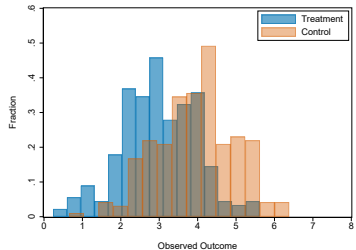
When $\bar{Y}_T$ and $\bar{Y}_C$ are independent:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$$SE_{\bar{Y}_T} = \sqrt{\frac{s_T^2}{n_T}} = \sqrt{\frac{\sum_{i \in T}\left(Y_i - \bar{Y}\right)^2}{n_T(n_T - 1)}}$$

where n_T is treatment observations, and $\sum_{i \in T}$ sums over treated i

The Standard Error of a Difference in Means

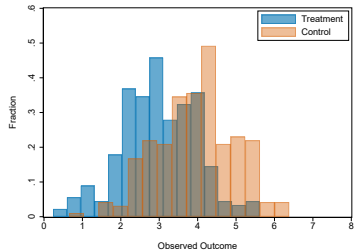


When $\bar{Y}_T$ and $\bar{Y}_C$ are independent:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$$\Rightarrow t = \left(\bar{Y}_T - \bar{Y}_C\right) / \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

The Standard Error of a Difference in Means



When $\bar{Y}_T$ and $\bar{Y}_C$ have variance s^2:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{s^2/n_T + s^2/n_C}$$

where: $s^2 = \frac{\sum_i \left(Y_i - \bar{Y}\right)^2}{(N-2)}$

Empirical Exercise

Subsidizing Malaria Treatment in Kenya

American Economic Review 2015, 105(2): 609–645
http://dx.doi.org/10.1257/aer.20130267

Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial†

By Jessica Cohen, Pascaline Dupas, and Simone Schaner*

Both under- and over-treatment of communicable diseases are public bads. But efforts to decrease one run the risk of increasing the other. Using rich experimental data on household treatment-seeking behavior in Kenya, we study the implications of this trade-off for subsidizing life-saving antimalarials sold over-the-counter at retail drug outlets. We show that a very high subsidy (such as the one under consideration by the international community) dramatically increases access, but nearly one-half of subsidized pills go to patients without malaria. We study two ways to better target subsidized drugs: reducing the subsidy level, and introducing rapid malaria tests over-the-counter. (*JEL* D12, D82, I12, O12, O15)

Subsidizing Malaria Treatment in Kenya

Comparison Group

No subsidy. Households received vouchers to purchase unsubsidized ACTs at the pre-AMFm retail price in Kenya: KSh 500 (approximately US$6.25, using a 2009 exchange rate of KSh 80/ US$1).

ACT Subsidy

Households were randomly selected to receive vouchers for ACTs at one of three subsidy levels:

- **92 percent** (US$0.50 per adult dose, corresponds to the Kenyan government's target retail price of KSh 40 under the AMFm)
- **88 percent** (US$0.75 per adult dose)
- **80 percent** (US$1.25 per adult dose)

ACT & RDT Subsidy

Households received one of the three ACT subsidy levels above and were also randomly assigned to receive vouchers for rapid diagnostic tests (RDTs) either for free or at an 85 percent subsidy (US$0.20).

PHOTO BY AUDE GUERRUCCI

Subsidizing Malaria Treatment in Kenya

TABLE 2—IMPACT OF ACT SUBSIDY ON TREATMENT SEEKING AND ACT ACCESS

	Took ACT (1)	Took ACT from drug shop (2)	Took ACT from health center (3)	Visited drug shop (4)	Visited health center (5)	Sought no care (6)	Took malaria test (7)	Took antibiotic (8)
Panel A. Pooled impact								
Any ACT subsidy	0.187***	0.222***	−0.038	0.167***	−0.079*	−0.096***	−0.014	−0.072**
	(0.038)	(0.031)	(0.030)	(0.046)	(0.042)	(0.036)	(0.038)	(0.034)
Panel B. Impact by subsidy level								
B1. ACT subsidy = 92 percent	0.225***	0.249***	−0.024	0.159***	−0.055	−0.110***	−0.031	−0.046
	(0.053)	(0.046)	(0.037)	(0.058)	(0.053)	(0.042)	(0.048)	(0.043)
B2. ACT subsidy = 88 percent	0.161***	0.217***	−0.056	0.167***	−0.070	−0.097**	−0.042	−0.062
	(0.050)	(0.043)	(0.037)	(0.058)	(0.052)	(0.042)	(0.047)	(0.040)
B3. ACT subsidy = 80 percent	0.178***	0.206***	−0.035	0.173***	−0.106**	−0.085*	0.023	−0.100***
	(0.048)	(0.042)	(0.035)	(0.054)	(0.047)	(0.045)	(0.046)	(0.038)
p-value: B1 = B2 = B3 = 0	0.000***	0.000***	0.498	0.004***	0.164	0.048**	0.533	0.066
p-value: B1 = B2 = B3	0.531	0.723	0.660	0.968	0.535	0.846	0.362	0.304
DV mean (control group)	0.190	0.071	0.119	0.488	0.286	0.226	0.214	0.185
Observations	631	631	631	631	631	631	631	631

Notes: "Substandard" malaria treatment includes non-ACT antimalarials and antipyretics. Sample excludes all households selected for a surprise or subsidized RDT. The unit of observation is the first illness episode with at least one malaria-like symptom that the household experienced following the baseline. A few households have multiple observations if multiple household members were ill simultaneously. Robust standard errors clustered at the household level in parentheses. All regressions control for household head age and a full set of strata dummies.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Subsidizing Malaria Treatment in Kenya

POLICY LESSONS

PHOTO BY IPA KENYA

Understanding the Context

The data collected during this evaluation suggest that households in the study area:

- Tend to bypass the public health care system if they are poor, likely because they live far from health centers, making travel costs too high. Instead they rely on local drug shops that do not offer diagnostic services.
- Experience illnesses suspected to be malaria very often. These illness episodes are generally not formally diagnosed and are typically presumptively treated with less effective antimalarials procured from a drug shop.

Subsidizing ACTs provides measurable benefits, especially for vulnerable children and the poorest households. Many households effectively miss out on the existing free treatment at public facilities and either do not seek care for malaria at all or take less effective medicines. For these families, a retail-sector ACT subsidy substantially improves access to proper treatment.

A slightly lower subsidy can improve targeting without compromising access for children. Moving from the AMFm target subsidy level (roughly 92 percent) to a somewhat lower subsidy (80 percent) reduced overtreatment among adults, while keeping access constant for children. These results suggest that an ACT subsidy is clearly needed, but that a slightly lower subsidy may achieve similar benefits at a lower cost.

Rapid diagnostic tests may be a promising means to improve targeting. People were very willing to try out rapid diagnostic testing, including sharing the cost of the test. More than half of adults who suspected malaria but got a negative test result decided not to purchase the subsidized ACT. Imperfect compliance with malaria test results is also common among public health workers, and thus it may take some time for people with malaria to become familiar with and trust RDTs.

Empirical Exercise: Takeaways

1. You should be able to open the Cohen, Dupas, and Schaner (2015) data set in Stata
2. In a bivariate regression on a (single) dummy variable, the estimated OLS coefficient $\hat{\beta}$ is the difference in means between the treatment group and the comparison group, which can also be recovered from a t-test of the equality of means in the two groups
3. Same logic applies when we include separate dummies for multiple (randomly-assigned) treatments, with no interaction terms and no additional covariates (or strata dummies)
4. When the treatment dummy aggregates multiple distinct treatment intensities, each treated observation weighted equally in calculating the estimated treatment effect

The End!

Regression Analysis of RCTs in R and Python

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i|D_i = 0] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

equal to population mean because control group is a random sample

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i|D_i = 0] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Expected value of treatment group mean:

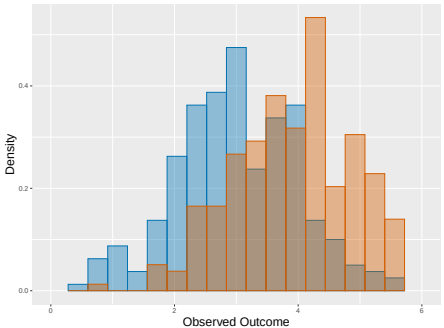
$$\begin{aligned} E[\bar{Y}_T] = E[Y_i|D_i = 1] &= E[Y_{1,i}|D_i = 1] \\ &= E[\delta_i + Y_{0,i}|D_i = 1] \\ &= E[\delta_i|D_i = 1] + E[Y_{0,i}|D_i = 1] \\ &= E[\delta_i] + E[Y_{0,i}] \end{aligned}$$

H_0: $ATE = 0$

Null hypothesis (H_0):

The average treatment effect is zero: $ATE = 0$

Or, equivalently: $\bar{Y}_T = \bar{Y}_C$



In Stata: `ttest y, by(t)`

In R: `t.test(y ~ t, data = df)`

In Python:

```
scipy.stats.ttest_ind(
    df['y'][df['t'] == 1],
    df['y'][df['t'] == 0],
    alternative="two-sided")
```

Testing the Equality of Means in R and Python

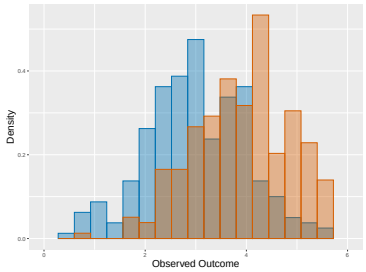
R:

```
t.test(y ~ t, data = df)

Welch Two Sample t-test

data: y by t
t = -5.7048,
df = 85.77,
p-value = 1.626e-07
alternative hypothesis: true ...
95 percent confidence interval:
 -1.2873994 -0.6220043
sample estimates:
mean in group 0 mean in group 1
 3.085472 4.040173
```

Python:

```
import scipy
scipy.stats.ttest_ind(
    df['y'][df['t'] == 1],
    df['y'][df['t'] == 0],
    nan_policy = "omit",
    alternative="two-sided")

TtestResult(statistic=5.6380327702693,
pvalue=5.8686215171853224e-08,
df=198.0)
```

The Standard Error of a Difference in Means



When $\bar{Y}_T$ and $\bar{Y}_C$ are independent:

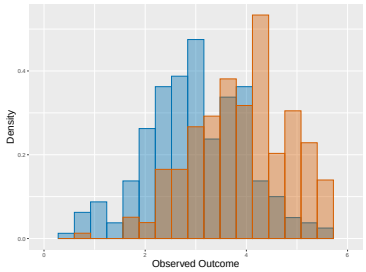
$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$$SE_{\bar{Y}_T} = \sqrt{\frac{s_T^2}{n_T}}$$

$$= \sqrt{\frac{\sum_{i \in T}\left(Y_i - \bar{Y}\right)^2}{n_T(n_T - 1)}}$$

where n_T is treatment observations, and $\sum_{i \in T}$ sums over treated i

The Standard Error of a Difference in Means

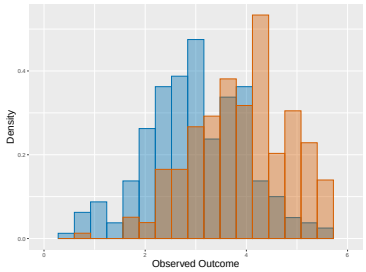


When $\bar{Y}_T$ and $\bar{Y}_C$ are independent:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$$\Rightarrow t = \left(\bar{Y}_T - \bar{Y}_C\right) / \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

The Standard Error of a Difference in Means

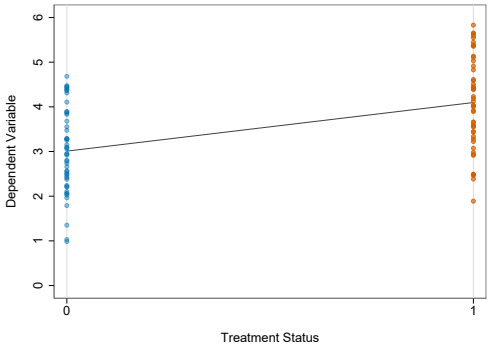


When $\bar{Y}_T$ and $\bar{Y}_C$ have variance s^2:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{s^2/n_T + s^2/n_C}$$

where: $s^2 = \frac{\sum_i \left(Y_i - \bar{Y}\right)^2}{(N-2)}$

OLS Regression on a Binary Independent Variable



Simple regression framework for analyzing RCTs: $Y_i = \alpha + \beta D_i + \varepsilon_i$

Treatment indicator $D_i = 0, 1 \Rightarrow$ only two sensible values of $\hat{Y}_i$

OLS in R and Python

R:

```
ols <- lm(y ~ t, data = df)
summary(ols)
```

```
Call:
lm(formula = y ~ t, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.8871 -0.5451 -0.0462  0.6731  2.5591

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0855     0.1466  21.040  < 2e-16 ***
t             0.9547     0.1693   5.638 5.87e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.037 on 198 degrees of freedom
Multiple R-squared:  0.1383,    Adjusted R-squared:  0.134
F-statistic: 31.79 on 1 and 198 DF,  p-value: 5.869e-08
```

Python:

```
import statsmodels.formula.api as smf
model = smf.ols('y t', data = df).fit()
print(model.summary())
```

```
In [18]: model = smf.ols('y~t', data = df).fit()
    ...: print(model.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.138
Model:                            OLS   Adj. R-squared:                  0.134
Method:                 Least Squares   F-statistic:                     31.79
Date:                Sun, 09 Feb 2025   Prob (F-statistic):           5.87e-08
Time:                        16:58:41   Log-Likelihood:                -290.04
No. Observations:                 200   AIC:                             584.1
Df Residuals:                     198   BIC:                             590.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.0855      0.147     21.040      0.000       2.796       3.375
t              0.9547      0.169      5.638      0.000       0.621       1.289
==============================================================================
Omnibus:                        0.464   Durbin-Watson:                   1.896
Prob(Omnibus):                  0.793   Jarque-Bera (JB):                0.247
Skew:                          -0.067   Prob(JB):                        0.884
Kurtosis:                       3.108   Cond. No.                         3.78
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Empirical Exercise

Subsidizing Malaria Treatment in Kenya

American Economic Review 2015, 105(2): 609–645
http://dx.doi.org/10.1257/aer.20130267

Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial†

By Jessica Cohen, Pascaline Dupas, and Simone Schaner*

Both under- and over-treatment of communicable diseases are public bads. But efforts to decrease one run the risk of increasing the other. Using rich experimental data on household treatment-seeking behavior in Kenya, we study the implications of this trade-off for subsidizing life-saving antimalarials sold over-the-counter at retail drug outlets. We show that a very high subsidy (such as the one under consideration by the international community) dramatically increases access, but nearly one-half of subsidized pills go to patients without malaria. We study two ways to better target subsidized drugs: reducing the subsidy level, and introducing rapid malaria tests over-the-counter. (*JEL* D12, D82, I12, O12, O15)

Subsidizing Malaria Treatment in Kenya

Comparison Group

No subsidy. Households received vouchers to purchase unsubsidized ACTs at the pre-AMFm retail price in Kenya: KSh 500 (approximately US$6.25, using a 2009 exchange rate of KSh 80/ US$1).

ACT Subsidy

Households were randomly selected to receive vouchers for ACTs at one of three subsidy levels:

- **92 percent** (US$0.50 per adult dose, corresponds to the Kenyan government's target retail price of KSh 40 under the AMFm)
- **88 percent** (US$0.75 per adult dose)
- **80 percent** (US$1.25 per adult dose)

ACT & RDT Subsidy

Households received one of the three ACT subsidy levels above and were also randomly assigned to receive vouchers for rapid diagnostic tests (RDTs) either for free or at an 85 percent subsidy (US$0.20).

PHOTO BY AUDE GUERRUCCI

Subsidizing Malaria Treatment in Kenya

TABLE 2—IMPACT OF ACT SUBSIDY ON TREATMENT SEEKING AND ACT ACCESS

	Took ACT (1)	Took ACT from drug shop (2)	Took ACT from health center (3)	Visited drug shop (4)	Visited health center (5)	Sought no care (6)	Took malaria test (7)	Took antibiotic (8)
Panel A. Pooled impact								
Any ACT subsidy	0.187***	0.222***	−0.038	0.167***	−0.079*	−0.096***	−0.014	−0.072**
	(0.038)	(0.031)	(0.030)	(0.046)	(0.042)	(0.036)	(0.038)	(0.034)
Panel B. Impact by subsidy level								
B1. ACT subsidy = 92 percent	0.225***	0.249***	−0.024	0.159***	−0.055	−0.110***	−0.031	−0.046
	(0.053)	(0.046)	(0.037)	(0.058)	(0.053)	(0.042)	(0.048)	(0.043)
B2. ACT subsidy = 88 percent	0.161***	0.217***	−0.056	0.167***	−0.070	−0.097**	−0.042	−0.062
	(0.050)	(0.043)	(0.037)	(0.058)	(0.052)	(0.042)	(0.047)	(0.040)
B3. ACT subsidy = 80 percent	0.178***	0.206***	−0.035	0.173***	−0.106**	−0.085*	0.023	−0.100***
	(0.048)	(0.042)	(0.035)	(0.054)	(0.047)	(0.045)	(0.046)	(0.038)
p-value: B1 = B2 = B3 = 0	0.000***	0.000***	0.498	0.004***	0.164	0.048**	0.533	0.066
p-value: B1 = B2 = B3	0.531	0.723	0.660	0.968	0.535	0.846	0.362	0.304
DV mean (control group)	0.190	0.071	0.119	0.488	0.286	0.226	0.214	0.185
Observations	631	631	631	631	631	631	631	631

Notes: "Substandard" malaria treatment includes non-ACT antimalarials and antipyretics. Sample excludes all households selected for a surprise or subsidized RDT. The unit of observation is the first illness episode with at least one malaria-like symptom that the household experienced following the baseline. A few households have multiple observations if multiple household members were ill simultaneously. Robust standard errors clustered at the household level in parentheses. All regressions control for household head age and a full set of strata dummies.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Subsidizing Malaria Treatment in Kenya

POLICY LESSONS

PHOTO BY IPA KENYA

Understanding the Context

The data collected during this evaluation suggest that households in the study area:

- Tend to bypass the public health care system if they are poor, likely because they live far from health centers, making travel costs too high. Instead they rely on local drug shops that do not offer diagnostic services.
- Experience illnesses suspected to be malaria very often. These illness episodes are generally not formally diagnosed and are typically presumptively treated with less effective antimalarials procured from a drug shop.

Subsidizing ACTs provides measurable benefits, especially for vulnerable children and the poorest households. Many households effectively miss out on the existing free treatment at public facilities and either do not seek care for malaria at all or take less effective medicines. For these families, a retail-sector ACT subsidy substantially improves access to proper treatment.

A slightly lower subsidy can improve targeting without compromising access for children. Moving from the AMFm target subsidy level (roughly 92 percent) to a somewhat lower subsidy (80 percent) reduced overtreatment among adults, while keeping access constant for children. These results suggest that an ACT subsidy is clearly needed, but that a slightly lower subsidy may achieve similar benefits at a lower cost.

Rapid diagnostic tests may be a promising means to improve targeting. People were very willing to try out rapid diagnostic testing, including sharing the cost of the test. More than half of adults who suspected malaria but got a negative test result decided not to purchase the subsidized ACT. Imperfect compliance with malaria test results is also common among public health workers, and thus it may take some time for people with malaria to become familiar with and trust RDTs.

Empirical Exercise: Takeaways

1. You should be able to open the Cohen, Dupas, and Schaner (2015) data set in Stata
2. In a bivariate regression on a (single) dummy variable, the estimated OLS coefficient $\hat{\beta}$ is the difference in means between the treatment group and the comparison group, which can also be recovered from a t-test of the equality of means in the two groups
3. Same logic applies when we include separate dummies for multiple (randomly-assigned) treatments, with no interaction terms and no additional covariates (or strata dummies)
4. When the treatment dummy aggregates multiple distinct treatment intensities, each treated observation weighted equally in calculating the estimated treatment effect