Taking Stock &
Linear Regression in Experiments and Observational Studies

ICPSR 2022 Session 2

Jake Bowers & Tom Leavitt

July 26, 2022

Today

1. Agenda: Overview and context for where we are in the course, Statistical Inference and Causal Inference, Linear regresssion for estimation in randomized experiments, Linear regression for covariance adjustment in randomized experiments, Linear regression for covariance adjustment in observational studies ("controlling for")
2. Maybe open time for work on exercise with TAs and me around.
3. Questions arising from the reading or assignments or life.

Today

1. Agenda: Overview and context for where we are in the course, Statistical Inference and Causal Inference, Linear regresssion for estimation in randomized experiments, Linear regression for covariance adjustment in randomized experiments, Linear regression for covariance adjustment in observational studies ("controlling for")
2. Maybe open time for work on exercise with TAs and me around.
3. Questions arising from the reading or assignments or life.

Today

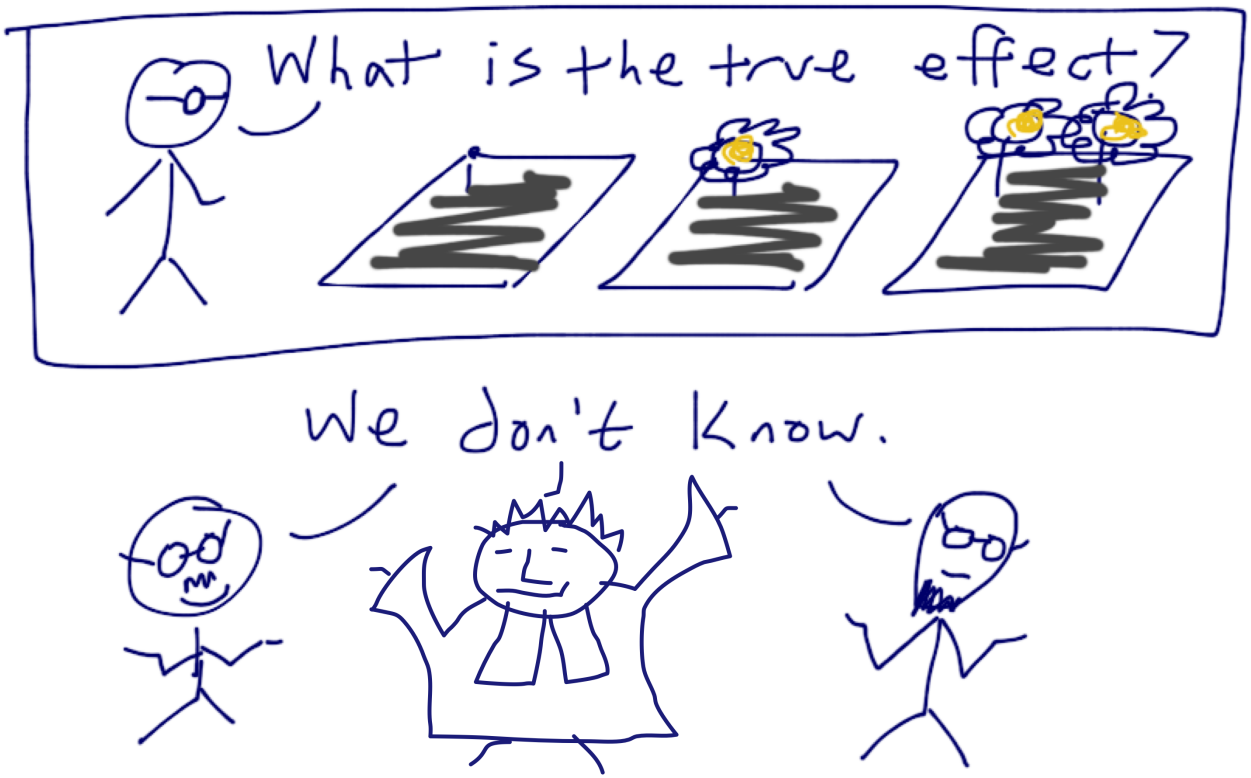
1. Agenda: Overview and context for where we are in the course, Statistical Inference and Causal Inference, Linear regresssion for estimation in randomized experiments, Linear regression for covariance adjustment in randomized experiments, Linear regression for covariance adjustment in observational studies ("controlling for")
2. Maybe open time for work on exercise with TAs and me around.
3. Questions arising from the reading or assignments or life.

1. An overview of approaches to statistical inference for causal quantities

2. Linear Regression in Experiments and Observational Studies

Three General Approaches To Learning About The Unobserved Using Data

Potential Outcomes



Imagine we would observe so many bushels of corn, y, if plot i were randomly assigned to new fertilizer, $y_{i,Z_i=1}$ (where $Z_i = 1$ means "assigned to new fertilizer" and $Z_i = 0$ means "assigned status quo fertilizer") and another amount of corn, $y_{i,Z_i=0}$, if the same plot were assigned the status quo fertilizer condition. These y are are potential or partially observed outcomes.

Notation

- Treatment $Z_i = 1$ for treatment and $Z_i = 0$ for control for units i
- In a two arm experiment each unit has at least a pair of potential outcomes $(y_{i,Z_i=1}, y_{i,Z_i=0})$ (also written $(y_{i,1}, y_{i,0})$ to indicate that $y_{1,Z_1=1,Z_2=1} = y_{1,Z_1=1,Z_2=0}$ — unit 1's outcome depends only on treatment assignment to unit 1 and not to unit 2.)
- Causal Effect for unit i is τ_i, $\tau_i = f(y_{i,1}, y_{i,0})$. For example, $\tau_i = y_{i,1} - y_{i,0}$.
- Fundamental Problem of (Counterfactual) Causality We only see one potential outcome $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ in our observed outcome, Y_i. Treatment reveals one potential outcome to us in a simple randomized experiment.

So how do we learn about τ_i if we cannot directly see it?

Notation

- Treatment $Z_i = 1$ for treatment and $Z_i = 0$ for control for units i
- In a two arm experiment each unit has at least a pair of potential outcomes $(y_{i,Z_i=1}, y_{i,Z_i=0})$ (also written $(y_{i,1}, y_{i,0})$ to indicate that $y_{1,Z_1=1,Z_2=1} = y_{1,Z_1=1,Z_2=0}$ — unit 1's outcome depends only on treatment assignment to unit 1 and not to unit 2.)
- Causal Effect for unit i is τ_i, $\tau_i = f(y_{i,1}, y_{i,0})$. For example, $\tau_i = y_{i,1} - y_{i,0}$.
- Fundamental Problem of (Counterfactual) Causality We only see one potential outcome $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ in our observed outcome, Y_i. Treatment reveals one potential outcome to us in a simple randomized experiment.

So how do we learn about τ_i if we cannot directly see it?

Notation

- Treatment $Z_i = 1$ for treatment and $Z_i = 0$ for control for units i
- In a two arm experiment each unit has at least a pair of potential outcomes $(y_{i,Z_i=1}, y_{i,Z_i=0})$ (also written $(y_{i,1}, y_{i,0})$ to indicate that $y_{1,Z_1=1,Z_2=1} = y_{1,Z_1=1,Z_2=0}$ — unit 1's outcome depends only on treatment assignment to unit 1 and not to unit 2.)
- Causal Effect for unit i is τ_i, $\tau_i = f(y_{i,1}, y_{i,0})$. For example, $\tau_i = y_{i,1} - y_{i,0}$.
- Fundamental Problem of (Counterfactual) Causality We only see one potential outcome $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ in our observed outcome, Y_i. Treatment reveals one potential outcome to us in a simple randomized experiment.

So how do we learn about τ_i if we cannot directly see it?

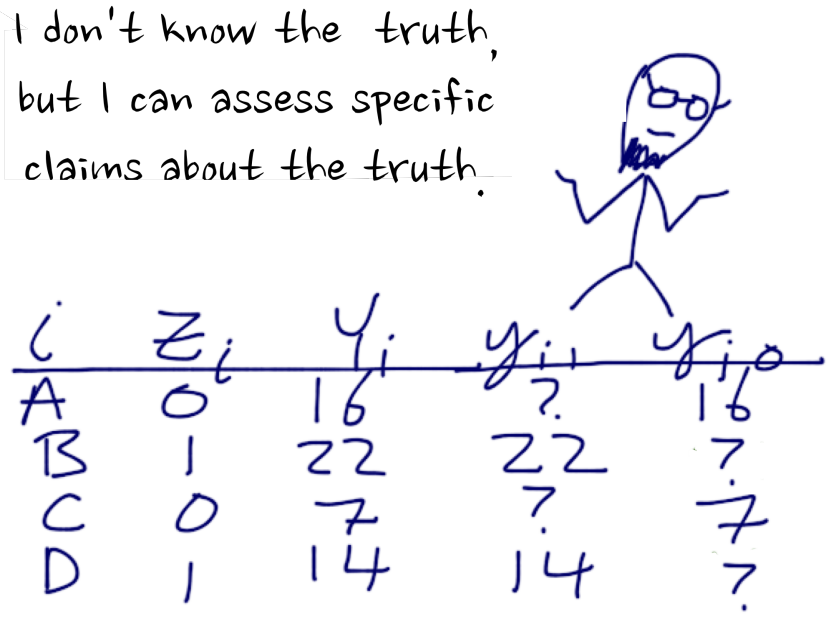
Notation

- Treatment $Z_i = 1$ for treatment and $Z_i = 0$ for control for units i
- In a two arm experiment each unit has at least a pair of potential outcomes $(y_{i,Z_i=1}, y_{i,Z_i=0})$ (also written $(y_{i,1}, y_{i,0})$ to indicate that $y_{1,Z_1=1,Z_2=1} = y_{1,Z_1=1,Z_2=0}$ — unit 1's outcome depends only on treatment assignment to unit 1 and not to unit 2.)
- Causal Effect for unit i is τ_i, $\tau_i = f(y_{i,1}, y_{i,0})$. For example, $\tau_i = y_{i,1} - y_{i,0}$.
- Fundamental Problem of (Counterfactual) Causality We only see one potential outcome $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ in our observed outcome, Y_i. Treatment reveals one potential outcome to us in a simple randomized experiment.

So how do we learn about τ_i if we cannot directly see it?

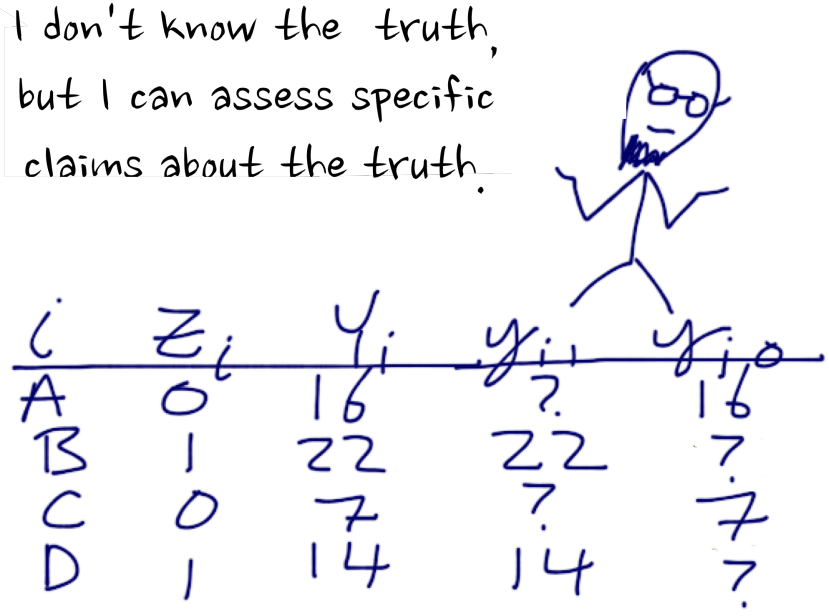
Design Based 1: Compare Models of Potential Outcomes to Data

1. Make a guess about (or model of) $\tau_i = f(y_{i,1}, y_{i,0})$. For example $H_0 : y_{i,1} = y_{i,0} + \tau_i$ and $\tau_i = 0$ is the sharp null hypothesis of no effects.
2. Measure consistency of the data with this model given the research design and choice of test statistic (summarizing the treatment-to-outcome relationship).



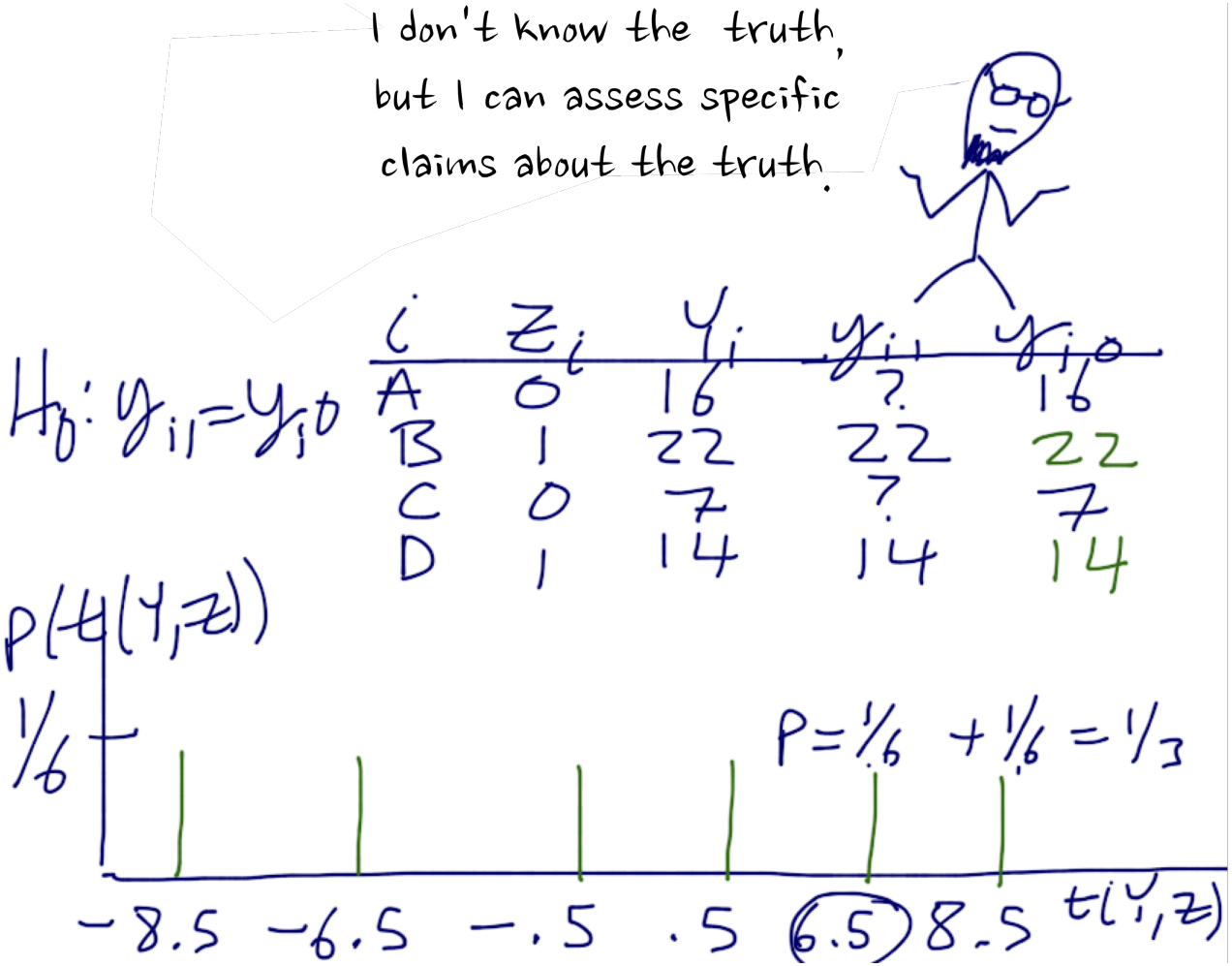
Design Based 1: Compare Models of Potential Outcomes to Data

1. Make a guess about (or model of) $\tau_i = f(y_{i,1}, y_{i,0})$. For example $H_0 : y_{i,1} = y_{i,0} + \tau_i$ and $\tau_i = 0$ is the sharp null hypothesis of no effects.
2. Measure consistency of the data with this model given the research design and choice of test statistic (summarizing the treatment-to-outcome relationship).



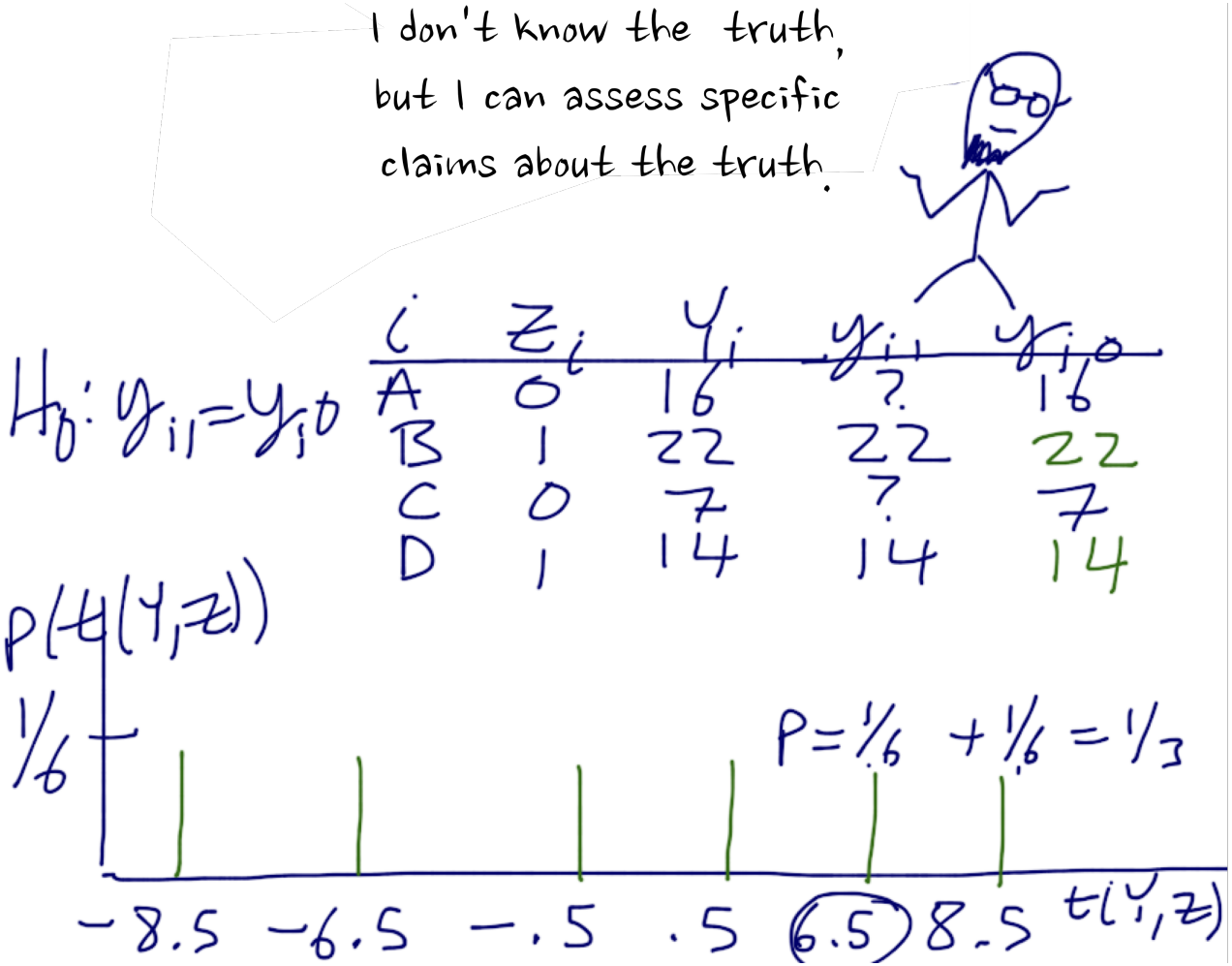
Design Based 1: Compare Models of Potential Outcomes to Data

1. Make a guess (or model of) about τ_i.
2. Measure consistency of data with this model given the design and test statistic.



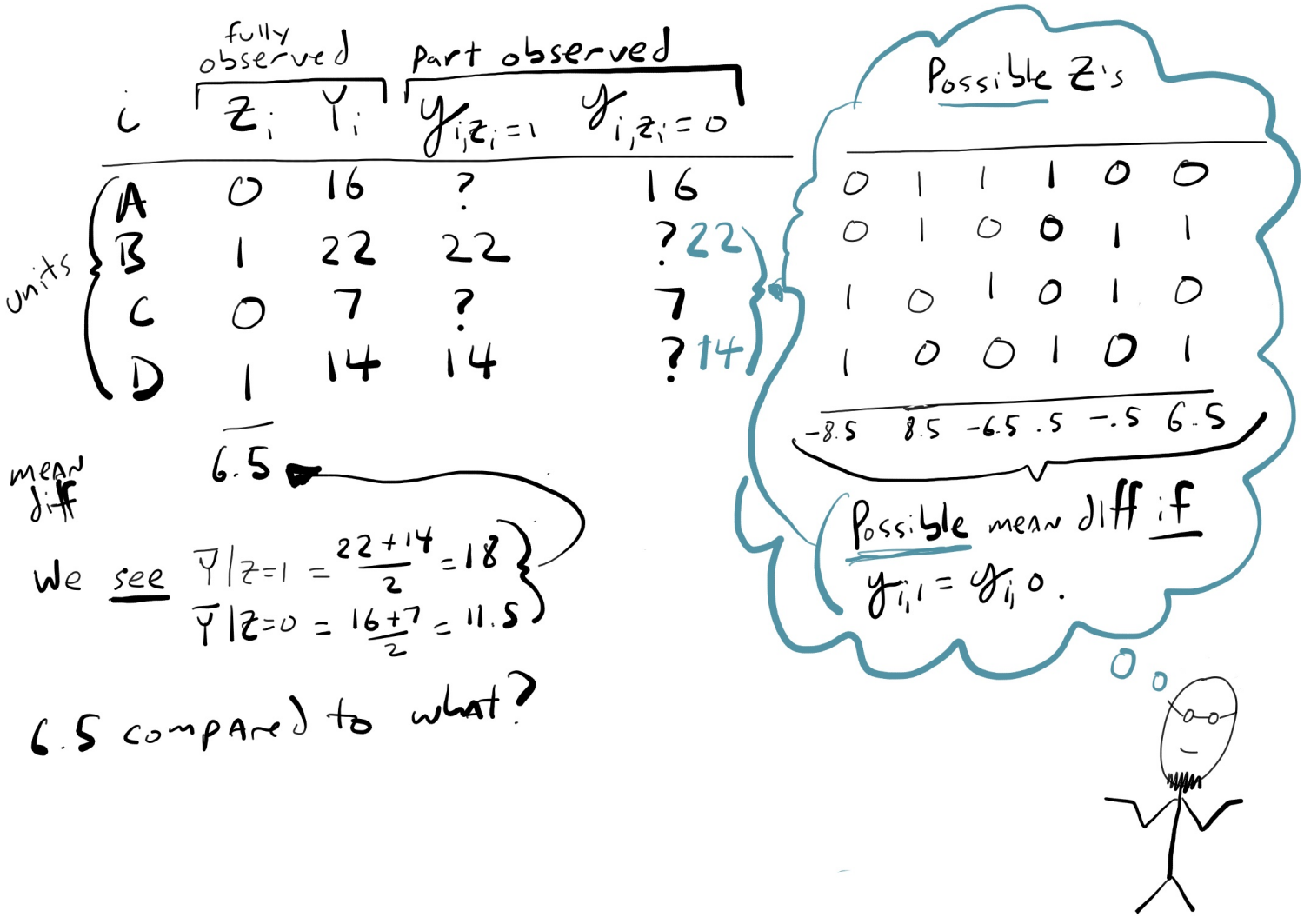
Design Based 1: Compare Models of Potential Outcomes to Data

1. Make a guess (or model of) about τ_i.
2. Measure consistency of data with this model given the design and test statistic.

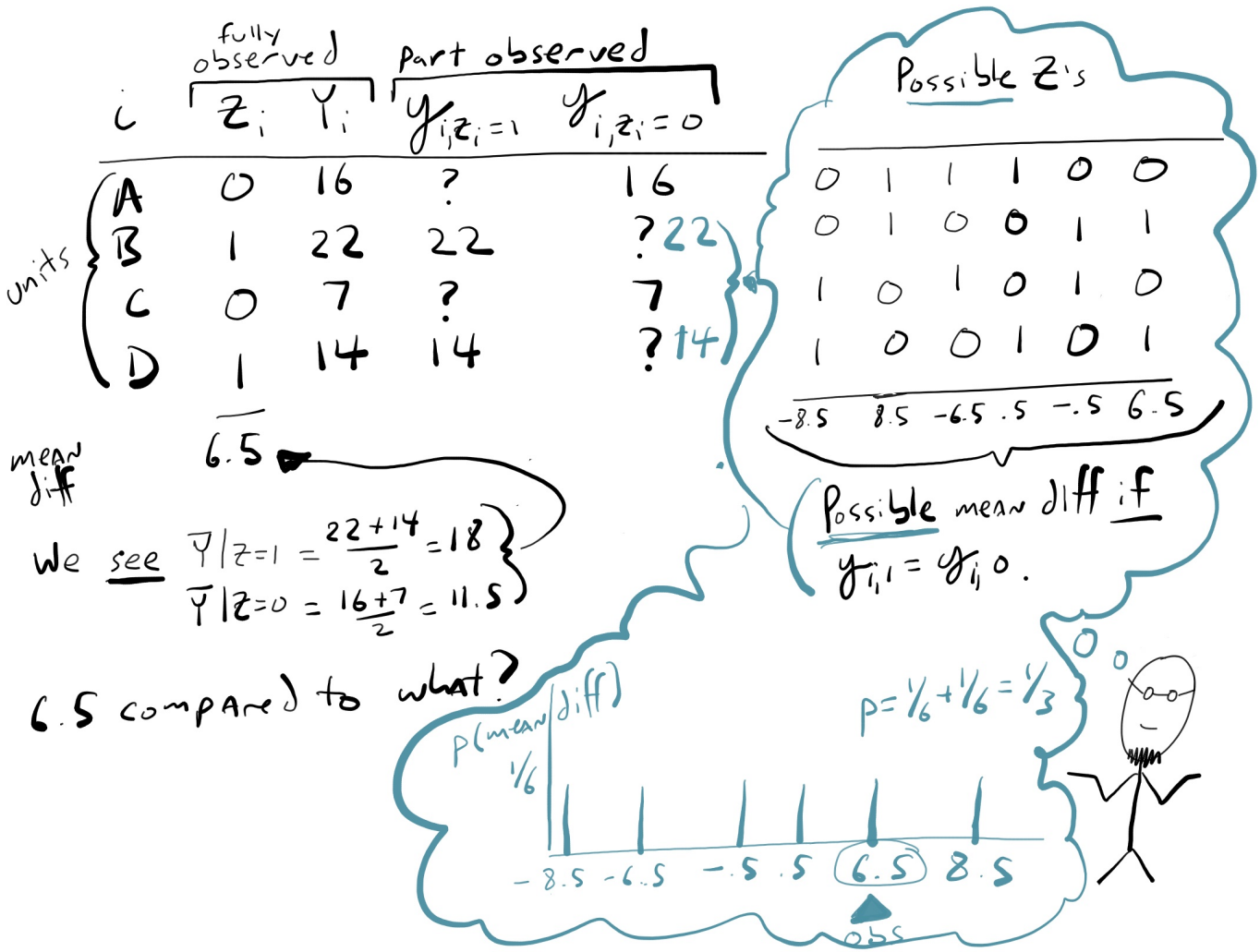


Design Based 1: Compare Models of Potential Outcomes to Data

Comparing the model, $H_0 : \tau_i = 0 \implies y_{i,1} = y_{i,0}$ to data:



Design Based 1: Compare Models of Potential Outcomes to Data



Design Based 1: Compare Models of Potential Outcomes to Data

Testing Models of No-Effects.

Here is some fake data from a tiny experiment with weird outcomes.

```
  Z    y0   y1    Y zF rY
1 0    16   16   16  0  2
2 1    22   24   24  1  3
3 0     7   10    7  0  1
4 1 3990 4000 4000  1  4
```

Next, define a function that compares treated to control outcomes:

```
## A mean difference test statistic
tz_mean_diff <- function(z, y) {
  mean(y[z == 1]) - mean(y[z == 0])
}
## A mean difference of ranks test statistic
tz_mean_rank_diff <- function(z, y) {
  ry <- rank(y)
  mean(ry[z == 1]) - mean(ry[z == 0])
}
```

And define a function to repeat the experimental randomization

```
## Function to repeat the experimental randomization
newexp <- function(z) {
  sample(z)
}
```

Design Based 1: Compare Models of Potential Outcomes to Data

Testing Models of No-Effects.

Here we **approximate the randomization distribution**:

```
set.seed(12345)
rand_dist_md <- with(smdat, replicate(1000, tz_mean_diff(z = newexp(Z), y = Y)))
rand_dist_rank_md <- with(smdat, replicate(1000, tz_mean_rank_diff(z = newexp(Z), y = Y)))
obs_md <- with(smdat, tz_mean_diff(z = Z, y = Y))
obs_rank_md <- with(smdat, tz_mean_rank_diff(z = Z, y = Y))
c(observed_mean_diff = obs_md, observed_mean_rank_diff = obs_rank_md)
```

```
     observed_mean_diff observed_mean_rank_diff
                   2000                       2
```

```
table(rand_dist_md) / 1000 ## Probability Distributions Under the Null of No Effects
```

```
rand_dist_md
-2000.5 -1992.5 -1983.5  1983.5  1992.5  2000.5
  0.188   0.163   0.161   0.172   0.155   0.161
```

```
table(rand_dist_rank_md) / 1000
```

```
rand_dist_rank_md
   -2    -1     0     1     2
0.172 0.197 0.318 0.161 0.152
```

```
p_md <- mean(rand_dist_md >= obs_md) ## P-Values
p_rank_md <- mean(rand_dist_rank_md >= obs_rank_md)
c(mean_diff_p = p_md, mean_rank_diff_p = p_rank_md)
```

Design Based 1: Compare Models of Potential Outcomes to Data

Testing Models of Effects.

To learn about whether the data are consistent with $\tau_i = 100$ for all i notice how treatment assignment reveals part of the unobserved outcomes:
$Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ and if $H_0 : \tau_i = 100$ or $H_0 : y_{i,1} = y_{i,0} + 100$ then:

$$
\begin{aligned}
Y_i &= Z_i(y_{i,0} + 100) + (1 - Z_i)y_{i,0} && (1)\\
&= Z_i y_{i,0} + Z_i 100 + y_{i,0} - Z_i y_{i,0} && (2)\\
&= Z_i 100 + y_{i,0} && (3)\\
y_{i,0} &= Y_i - Z_i 100 && (4)
\end{aligned}
$$

Design Based 1: Compare Models of Potential Outcomes to Data

Testing Models of Effects.

To test a model of causal effects we adjust the observed outcomes to be consistent with our hypothesis about unobserved outcomes and then repeat the experiment:

```
tz_mean_diff_effects <- function(z, y, tauvec) {
  adjy <- y - z * tauvec
  radjy <- rank(adjy)
  mean(radjy[z == 1]) - mean(radjy[z == 0])
}
rand_dist_md_tau_cae <- with(smdat, replicate(1000, tz_mean_diff_effects(z = newexp(Z), y = Y, ta
obs_md_tau_cae <- with(smdat, tz_mean_diff_effects(z = Z, y = Y, tauvec = c(100, 100, 100, 100)))
mean(rand_dist_md_tau_cae >= obs_md_tau_cae)
```

```
[1] 0.513
```

Design Based 1: Compare Models of Potential Outcomes to Data

Testing Models of Effects.

Now let's test a model with different effects for each unit — $H_0 : \boldsymbol{\tau} = \{0, 2, 3, 10\}$

```
rand_dist_md_taux <- with(smdat, replicate(1000, tz_mean_diff_effects(
  z = newexp(Z), y = Y,
  tauvec = c(0, 2, 3, 10)
)))
obs_md_taux <- with(smdat, tz_mean_diff_effects(z = Z, y = Y, tauvec = c(0, 2, 3, 10)))
mean(rand_dist_md_taux >= obs_md_taux)
```

```
[1] 0.158
```

So: We can learn about claims about unobserved potential outcomes by gauging the consistency our observed results with the distributions implied by the claims+test statistics+research design (including sample size and randomization scheme).

Questions about this first approach to using statistical inference to do counterfactual causal inference?

Design Based 2: Estimate Averages of Potential Outcomes

1. Notice that the observed Y_i are a sample from the (small, finite) population of unobserved potential outcomes $(y_{i,1}, y_{i,0})$.
2. Decide to focus on the average, $\bar{\tau}$, because sample averages, $\hat{\bar{\tau}}$ are unbiased and consistent estimators of population averages.
3. Estimate $\bar{\tau}$ with the observed difference in means as $\hat{\bar{\tau}}$.

I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	$y_{i,1}$	$y_{i,0}$
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i,1}$	$\bar{y}_{i,0}$

$$ATE = \bar{Y}_i|Z_i=1 - \bar{Y}_i|Z_i=0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

Design Based 2: Estimate Averages of Potential Outcomes

1. Notice that the observed Y_i are a sample from the (small, finite) population of unobserved potential outcomes $(y_{i,1}, y_{i,0})$.
2. Decide to focus on the average, $\bar{\tau}$, because sample averages, $\hat{\bar{\tau}}$ are unbiased and consistent estimators of population averages.
3. Estimate $\bar{\tau}$ with the observed difference in means as $\hat{\bar{\tau}}$.

I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	$y_{i,1}$	$y_{i,0}$
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i,1}$	$\bar{y}_{i,0}$

$$ATE = \bar{Y}_i|Z_i=1 - \bar{Y}_i|Z_i=0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

Design Based 2: Estimate Averages of Potential Outcomes

1. Notice that the observed Y_i are a sample from the (small, finite) population of unobserved potential outcomes $(y_{i,1}, y_{i,0})$.
2. Decide to focus on the average, $\bar{\tau}$, because sample averages, $\hat{\bar{\tau}}$ are unbiased and consistent estimators of population averages.
3. Estimate $\bar{\tau}$ with the observed difference in means as $\hat{\bar{\tau}}$.

I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	$y_{i,1}$	$y_{i,0}$
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i,1}$	$\bar{y}_{i,0}$

$$\widehat{ATE} = \bar{Y}_i|Z_i=1 - \bar{Y}_i|Z_i=0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

Design Based 2: Estimate Averages of Potential Outcomes

I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	y_{i1}	y_{i0}
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i1}$	$\bar{y}_{i0}$

$$ATE = \bar{Y}_i|Z_i=1 - \bar{Y}_i|Z_i=0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

Design Based 2: Estimate Averages of Potential Outcomes

Here using Neyman's standard errors (same as HC2 SEs) and Central Limit Theorem based p-values and 95% confidence intervals.
Notice that OLS/Linear Regression is just a difference of means calculator here.

```
with(smdat, mean(Y[Z == 1]) - mean(Y[Z == 0]))
```

```
[1] 2000
```

```
est_se_est_ate <- with(smdat, sqrt(var(Y[Z == 1]) / sum(Z) + var(Y[Z == 0]) / (sum(1 - Z))))
est_se_est_ate
```

```
[1] 1988
```

```
est1 <- difference_in_means(Y ~ Z, data = smdat)
tidy(est1)
```

```
  term estimate std.error statistic p.value conf.low conf.high df outcome
1    Z     2000      1988     1.006   0.498   -23259     27260  1       Y
```

```
lm1 <- lm_robust(Y ~ Z, data = smdat)
tidy(lm1)
```

```
         term estimate std.error statistic p.value  conf.low conf.high df outcome
1 (Intercept)     11.5       4.5     2.556  0.1250    -7.862     30.86  2       Y
2           Z   2000.5    1988.0     1.006  0.4202 -6553.196  10554.20  2       Y
```

Those p-values raise the next question: Where does the randomization distribution for these tests come from?

Design Based 2: Test hypotheses about averages.

A focus on the difference of average potential outcomes, on an average causal effect, also allows for testing hypotheses about these average causal efffects. This is called a test of the "weak null" hypothesis.
The challenge:

A hypothesis like $H_0 : \bar{\tau} = \tau_0$ is compatible with many different sharp null hypotheses: for example, $\tau = \{-5, 0, 0, 5\}$ and $\tau = \{0, 0, 0, 0\}$ are both compatible with $\bar{\tau} = 0$.

And a hypothesis test requires that we compare what we observe with what the hypothesis implies that we would observe. But a hypothesis about an average implies many different probability distributions: one for each sharp hypothesis.

Hypothesis tests of the weak null

- The finite population CLT tells us that

$$\frac{\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right) - \mathbb{E}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}{\sqrt{\mathrm{Var}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}} \xrightarrow{d} \mathcal{N}\left(0, 1\right)$$

- Diff-in-Means is unbiased, so write

$$\frac{\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right) - \tau}{\sqrt{\mathrm{Var}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}} \xrightarrow{d} \mathcal{N}\left(0, 1\right)$$

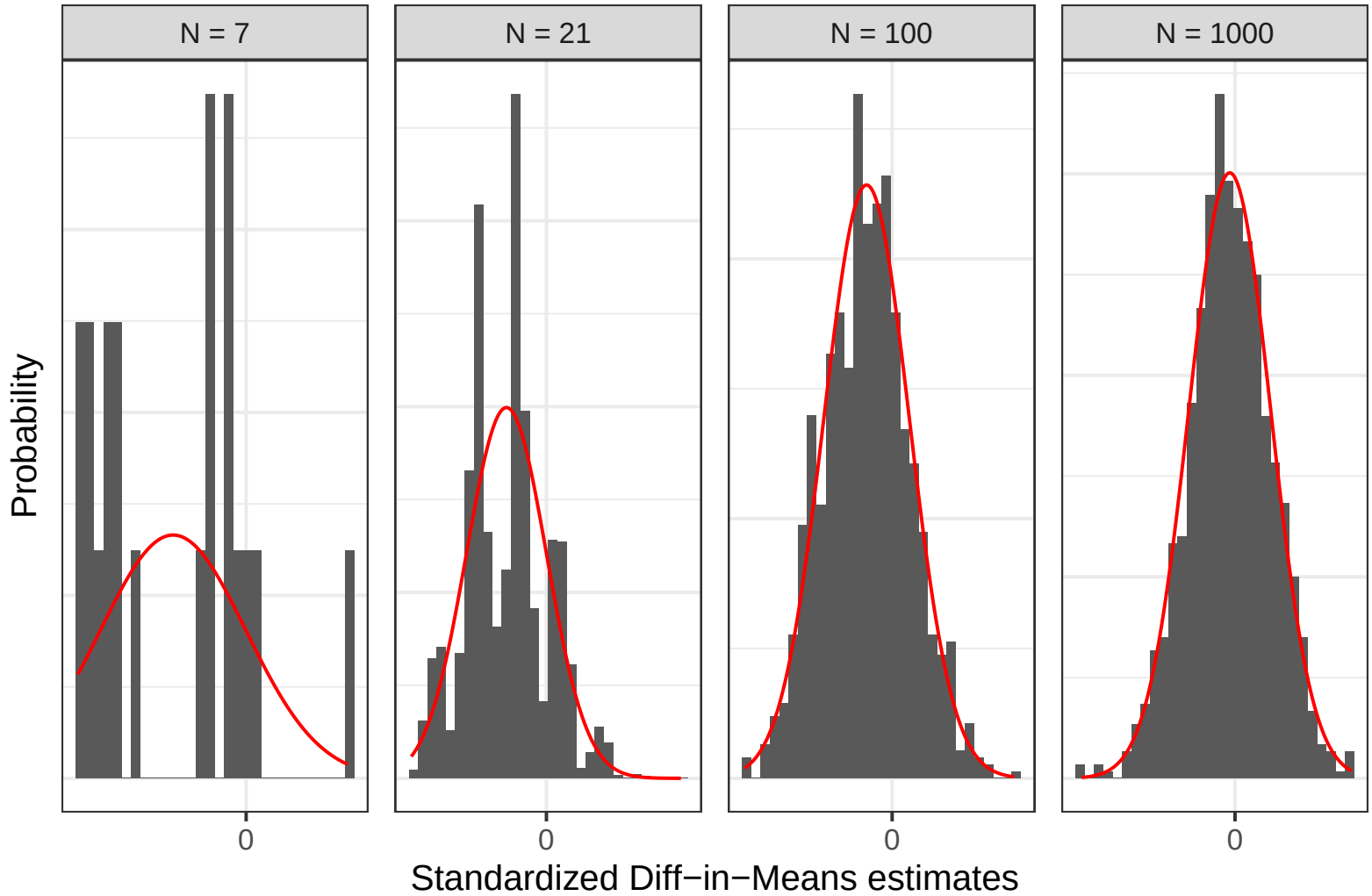
- The CLT is an asymptotic results as $N \to \infty$
- But we can bound error of Normal approximation for fixed N
- Thus, with experiments of at least moderate size and outcomes that aren't too skewed or have extreme outliers,

$$\frac{\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right) - \tau}{\sqrt{\mathrm{Var}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}} \overset{\text{approx.}}{\sim} \mathcal{N}\left(0, 1\right)$$

- This justifies use of standard Normal distribution for hypothesis tests

Hypothesis tests of the weak null

- "Village heads" example we calculate an estimate for each of the possible ways to randomize and subtract off the expected value (here "5") such that the distribution is now centered at 0 (a hypothesized value).



Hypothesis tests of the weak null

- To test null hypothesis relative to alternative

$$H_0 : \tau = \tau_0 \text{ versus either}$$
$$H_A : \tau > \tau_0,\ H_A : \tau < \tau_0 \text{ or } H_A : |\tau| > |\tau_0|$$

- Calculate upper(u), lower(l) or two-sided(t) p-value as

$$p_u = 1 - \Phi\left(\frac{\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right) - \tau_0}{\sqrt{\widehat{\mathrm{Var}}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}}\right)$$

$$p_l = \Phi\left(\frac{\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right) - \tau_0}{\sqrt{\widehat{\mathrm{Var}}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}}\right)$$

$$p_t = 2\left(1 - \Phi\left(\frac{|\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right) - \tau_0|}{\sqrt{\widehat{\mathrm{Var}}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]}}\right)\right)$$

- If p-value is less than size α-level of test, reject. Otherwise, don't
- Note that, since we don't know $\mathrm{Var}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]$, we have used its conservative estimator instead, $\widehat{\mathrm{Var}}\left[\hat{\tau}\left(\mathbf{Z}, \mathbf{Y}\right)\right]$

Hypothesis tests of the weak null

Hypothesis tests susceptible to two errors:

- Type I error: Rejecting null hypothesis when it is true
- Type II error: *Failing* to reject null hypothesis when it is false

A good test controls these errors:

1. Type I error probability is less than or equal to size (α-level) of test
2. Power (1 - type II error probability) is at least as great as α-level
3. Power tends to 1 as $N \to \infty$

Hypothesis tests of the weak null}

- We can prove that tests of weak null satisfy (1) – (3) as $N \to \infty$
- Thus, when experiments are large, we can often safely use such tests
- But (1) – (3) may not always be satisfied when experiments are small, have skewed outcome distributions or extreme outliers

Confidence intervals}

- Equivalence between hypothesis testing and confidence intervals
- Confidence interval is set of null hypotheses we fail to reject

Consider two-sided confidence interval, $\mathcal{C}_t$:

$$\mathcal{C}_t = \left\{ \tau_0 : \left| \frac{\hat{\tau}(\mathbf{Z}, \mathbf{Y}) - \tau_0}{\sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]}} \right| \leq z_{1-\alpha/2} \right\}$$

$$= \left\{ \tau_0 : -z_{1-\alpha/2} \leq \frac{\hat{\tau}(\mathbf{Z}, \mathbf{Y}) - \tau_0}{\sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]}} \leq z_{1-\alpha/2} \right\}$$

$$= \left\{ \tau_0 : -z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \leq \hat{\tau}(\mathbf{Z}, \mathbf{Y}) - \tau_0 \leq z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \right\}$$

$$= \left\{ \tau_0 : -\hat{\tau}(\mathbf{Z}, \mathbf{Y}) - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \leq -\tau_0 \leq -\hat{\tau}(\mathbf{Z}, \mathbf{Y}) + z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \right\}$$

$$= \left\{ \tau_0 : \hat{\tau}(\mathbf{Z}, \mathbf{Y}) + z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \geq \tau_0 \geq \hat{\tau}(\mathbf{Z}, \mathbf{Y}) - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \right\}$$

$$= \left\{ \tau_0 : \hat{\tau}(\mathbf{Z}, \mathbf{Y}) - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \leq \tau_0 \leq \hat{\tau}(\mathbf{Z}, \mathbf{Y}) + z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]} \right\}$$

Confidence intervals

Here is an example showing how a confidence interval is a collection of not-rejected hypotheses (here using the weak null hypothesis)

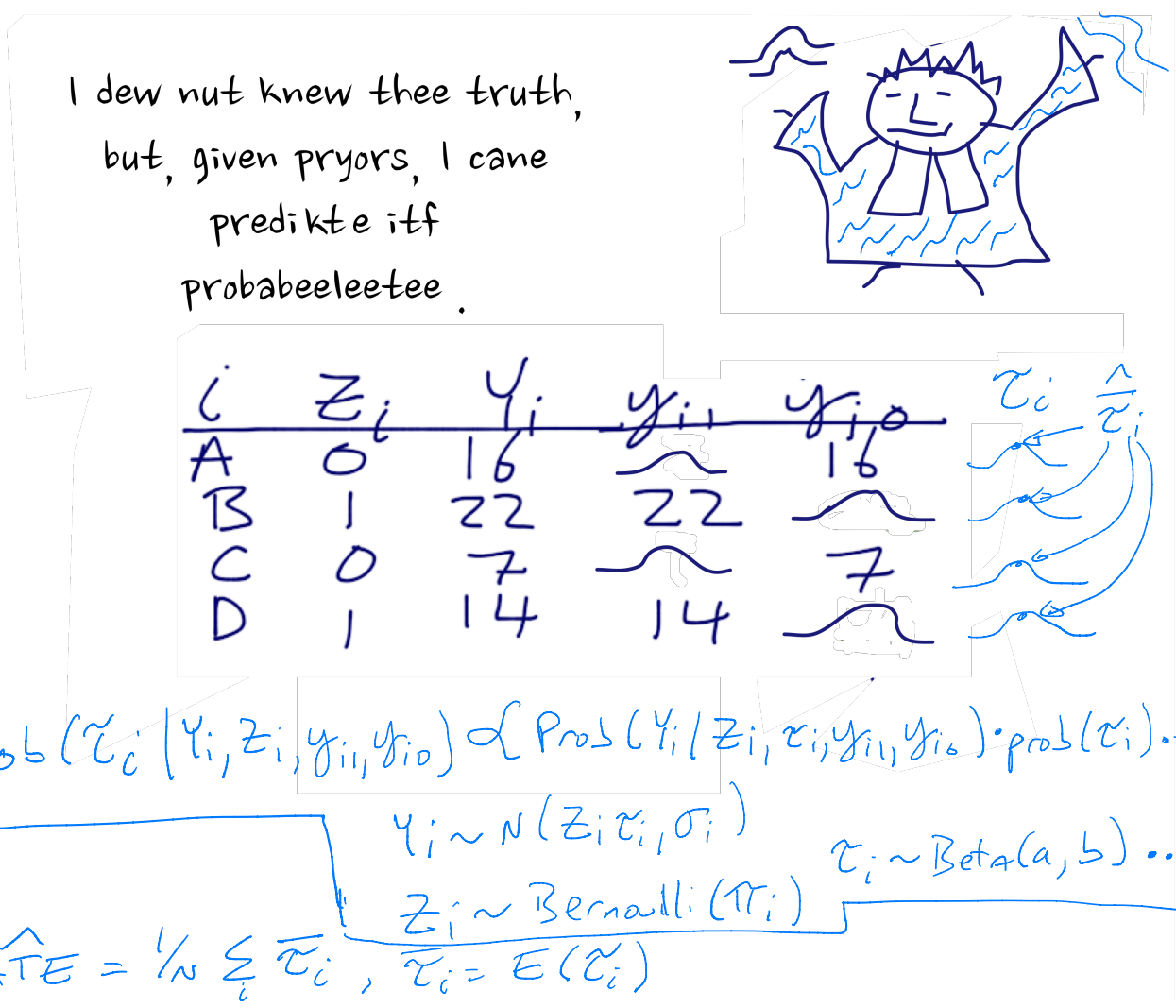
```
vh_dat <- data.frame(Y = Y, Z = Z, y_C = y_C, y_T = y_T)

p_val_fun <- function(h, thealpha = .05, return_p = FALSE) {
  ## Return a p-value for each weak null hypothesis
  test_res <- tidy(lm_robust(I(Y - h * Z) ~ Z, data = vh_dat))
  the_p <- test_res %>%
    filter(term == "Z") %>%
    select(p.value)
  if (return_p) {
    return(the_p[[1]])
  } else {
    return(thealpha - the_p[[1]])
  }
}

## First just test many hypotheses in a row. This is a "grid-search"
the_hyps <- seq(-20, 20, .1)
res <- sapply(the_hyps, function(h) {
  p_val_fun(h = h, return_p = TRUE)
})
hyp_res <- data.frame(h = the_hyps, p = res)
## Keep the largest and smallest hypotheses for which p>=.05
hyp_res %>%
  filter(abs(p) >= .05) %>%
  summarize(range(h))
```

```
  range(h)
```

Model Based 1: Predict Distributions of Potential Outcomes



Model Based 1: Predict Distributions of $(y_{i,1}, y_{i,0})$

1. Given a model of Y_i:[1]

$$\Pr(Y_i^{obs}|\mathrm{Z}, \theta) \sim \text{Normal}(Z_i \cdot \mu_1 + (1 - Z_i) \cdot \mu_0, Z_i\sigma_1^2 + (1 - Z_i) \cdot \sigma_0^2) \quad (5)$$

   where $\mu_0 = \alpha$ and $\mu_1 = \alpha + \tau$.

2. And a model of the pair $\{y_{i,0}, y_{i,1}\} \equiv \{Y_i(0), Y_i(1)\}$ but random not fixed as before (and so written as upper-case):

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Bigg| \theta \sim \text{Normal}\left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} \right) \quad (6)$$

3. And a model of Z_i is known because of randomization so we can write: $\Pr(\mathrm{Z}|\mathrm{Y}(0), \mathrm{Z}(1)) = \Pr(\mathrm{Z})$

4. And given priors on $\theta = \{\alpha, \tau, \sigma_c, \sigma_t\}$ (here make them all independent Normal(0,5))).

We can generate the posterior distribution of α, τ, σ_c, and σ_t and thus can impute $\{Y_i(0), Y_i(1)\}$ to generate a distribution for τ_i.

[1] see this website.

Model Based 1: Predict Distributions of Potential Outcomes

A snippet of `rctbayes.stan`:

```
model {
   // PRIORS
   alpha ~ normal(0, 5);
   tau ~ normal(0, 5);
   sigma_c ~ normal(0, 5);
   sigma_t ~ normal(0, 5);

   // LIKELIHOOD
   y ~ normal(alpha + tau*w, sigma_t*w + sigma_c*(1 - w));
}
```

Model Based 1: Predict Distributions of Potential Outcomes

```
## rho is correlation between the potential outcomes
stan_data <- list(N = 4, y = smdat$Y, w = smdat$Z, rho = 0)
# Compile and run the stan model
fit_simdat <- stan(file = "rctbayes.stan", data = stan_data, iter = 5000, warmup = 2500, chains =
res <- as.matrix(fit_simdat)
```

```
## Summary of the 2000 Predicted Treatment effects for units 1 and 4
t(apply(res[, c("tau_unit[1]", "tau_unit[4]")], 2, summary))
```

```
parameters      Min. 1st Qu.   Median     Mean 3rd Qu.   Max.
  tau_unit[1] -560.8   -98.6   -1.587   -2.516   92.39  474.6
  tau_unit[4] 3956.9  3986.8 3990.877 3991.262 3995.39 4029.3
```

```
## Probability that effect on unit 1 is greater than 0
mean(res[, "tau_unit[1]"] > 0)
```

```
[1] 0.496
```

```
## Overall mean of the effects:
mean_tau <- rowMeans(res[, c("tau_unit[1]", "tau_unit[2]", "tau_unit[3]", "tau_unit[4]")])
summary(mean_tau)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    821     970    1003    1003    1037    1179
```

Summmary: Modes of Statistical Inference for Causal Effects

We can infer about unobserved counterfactuals by:

1. assessing claims or models or hypotheses about relationships between unobserved potential outcomes (Fisher's sharp null testing approach via Rosenbaum) (this includes testing hypotheses of "no effects at all", "same effect for all", or different effects for each unit, or anything in between).
2. estimating averages (or other summaries) of unobserved potential outcomes (Neyman's estimation approach) (Unbiased estimators of effects, conservative variance estimators and the Central Limit Theorem allow us to use Normal Approximations to use the difference of means as a test statistic to test weak null hypotheses, too).
3. predicting individual level outcomes based on probability models of outcomes, interventions, etc. (Bayes's predictive approach via Rubin)

We can go from tests to intervals (even to quite narrow intervals) because a confidence interval is a collection of not-rejected hypothesis tests.

Summmary: Modes of Statistical Inference for Causal Effects

We can infer about unobserved counterfactuals by:

1. assessing claims or models or hypotheses about relationships between unobserved potential outcomes (Fisher's sharp null testing approach via Rosenbaum) (this includes testing hypotheses of "no effects at all", "same effect for all", or different effects for each unit, or anything in between).
2. estimating averages (or other summaries) of unobserved potential outcomes (Neyman's estimation approach) (Unbiased estimators of effects, conservative variance estimators and the Central Limit Theorem allow us to use Normal Approximations to use the difference of means as a test statistic to test weak null hypotheses, too).
3. predicting individual level outcomes based on probability models of outcomes, interventions, etc. (Bayes's predictive approach via Rubin)

We can go from tests to intervals (even to quite narrow intervals) because a confidence interval is a collection of not-rejected hypothesis tests.

Summmary: Modes of Statistical Inference for Causal Effects

We can infer about unobserved counterfactuals by:

1. assessing claims or models or hypotheses about relationships between unobserved potential outcomes (Fisher's sharp null testing approach via Rosenbaum) (this includes testing hypotheses of "no effects at all", "same effect for all", or different effects for each unit, or anything in between).
2. estimating averages (or other summaries) of unobserved potential outcomes (Neyman's estimation approach) (Unbiased estimators of effects, conservative variance estimators and the Central Limit Theorem allow us to use Normal Approximations to use the difference of means as a test statistic to test weak null hypotheses, too).
3. predicting individual level outcomes based on probability models of outcomes, interventions, etc. (Bayes's predictive approach via Rubin)

We can go from tests to intervals (even to quite narrow intervals) because a confidence interval is a collection of not-rejected hypothesis tests.

Summary: Modes of Statistical Inference for Causal Effects

Statistical inferences — formalized reasoning about "what if" statements ("What if I had randomly assigned other plots to treatment?") — and their properties (like bias, error rates, precision) arise from:

1. Repeating the design and using the hypothesis and test statistics to generate a reference distribution that describes the variation in the hypothetical world. Compare the observed to the hypothesized to measure consistency between hypothesis, or model, and observed outcomes (Fisher and Rosenbaum's randomization-based inference for individual causal effects).
2. Repeating the design and the estimation such that standard errors, p-values, and confidence intervals reflect design-based variability. Probability distributions (like the Normal or t-distribution) arise from Limit Theorems in large samples. (Neyman's randomization-based inference for average causal effects).
3. Repeatedly drawing from the probability distributions that generate the observed data (that represent the design) — the likelihood and the priors — to describe a posterior distribution for unit-level causal effects. Calculate posterior distributions for aggregated causal effects (like averages of individual level effects). (Bayes and Rubin's predictive model-based causal inference).

Summary: Modes of Statistical Inference for Causal Effects

Statistical inferences — formalized reasoning about "what if" statements ("What if I had randomly assigned other plots to treatment?") — and their properties (like bias, error rates, precision) arise from:

1. Repeating the design and using the hypothesis and test statistics to generate a reference distribution that describes the variation in the hypothetical world. Compare the observed to the hypothesized to measure consistency between hypothesis, or model, and observed outcomes (Fisher and Rosenbaum's randomization-based inference for individual causal effects).
2. Repeating the design and the estimation such that standard errors, p-values, and confidence intervals reflect design-based variability. Probability distributions (like the Normal or t-distribution) arise from Limit Theorems in large samples. (Neyman's randomization-based inference for average causal effects).
3. Repeatedly drawing from the probability distributions that generate the observed data (that represent the design) — the likelihood and the priors — to describe a posterior distribution for unit-level causal effects. Calculate posterior distributions for aggregated causal effects (like averages of individual level effects). (Bayes and Rubin's predictive model-based causal inference).

Summary: Modes of Statistical Inference for Causal Effects

Statistical inferences — formalized reasoning about "what if" statements ("What if I had randomly assigned other plots to treatment?") — and their properties (like bias, error rates, precision) arise from:

1. Repeating the design and using the hypothesis and test statistics to generate a reference distribution that describes the variation in the hypothetical world. Compare the observed to the hypothesized to measure consistency between hypothesis, or model, and observed outcomes (Fisher and Rosenbaum's randomization-based inference for individual causal effects).
2. Repeating the design and the estimation such that standard errors, p-values, and confidence intervals reflect design-based variability. Probability distributions (like the Normal or t-distribution) arise from Limit Theorems in large samples. (Neyman's randomization-based inference for average causal effects).
3. Repeatedly drawing from the probability distributions that generate the observed data (that represent the design) — the likelihood and the priors — to describe a posterior distribution for unit-level causal effects. Calculate posterior distributions for aggregated causal effects (like averages of individual level effects). (Bayes and Rubin's predictive model-based causal inference).

Summary: Applications of the Model-Based Prediction Approach

Examples of use of the model-based prediction approach:

- Estimating causal effects when we need to model processes of missing outcomes, missing treatment indicators, or complex non-compliance with treatment (Barnard et al. 2003)
- Searching for heterogeneity (subgroup differences) in how units react to treatment (ex. (Hahn, Murray, and Carvalho 2020) but see also literature on BART, Bayesian Machine Learning as applied to causal inference questions).

Summary: Applications of the Model-Based Prediction Approach

Examples of use of the model-based prediction approach:

- Estimating causal effects when we need to model processes of missing outcomes, missing treatment indicators, or complex non-compliance with treatment (Barnard et al. 2003)
- Searching for heterogeneity (subgroup differences) in how units react to treatment (ex. (Hahn, Murray, and Carvalho 2020) but see also literature on BART, Bayesian Machine Learning as applied to causal inference questions).

Summary: Applications of the Testing Approach

Examples of use of the testing approach:

- Assessing evidence of pareto optimal effects or no aberrant effect (i.e. no unit was made worse off by the treatment) (Caughey, Dafoe, and Miratrix 2016; P. Rosenbaum and Silber 2008).
- Assessing evidence that the treatment group was made better than the control group (but being agnostic about the precise nature of the difference) (ex. $p > .2$ with difference of means but $p < .001$ with difference of ranks in Office of Evaluation Sciences study of General Services Administration Auctions)
- Focusing on detection rather than on estimation (for example to identify promising sites for future research in experiments with many blocks or strata) (Bowers and Chen 2020 working paper).
- Assessing hypotheses of no effects in small samples, with rare outcomes, cluster randomization, or other designs where reference distributions may not be Normal (see for example, (Gerber and Green 2012)).
- Assessing structural models of causal effects (for example models of treatment effect propagation across networks) (Bowers, Desmarais, et al. 2018; Bowers, M. Fredrickson, and Aronow 2016; Bowers, M. M. Fredrickson, and Panagopoulos 2013).

Summary: Applications of the Testing Approach

Examples of use of the testing approach:

- Assessing evidence of pareto optimal effects or no aberrant effect (i.e. no unit was made worse off by the treatment) (Caughey, Dafoe, and Miratrix 2016; P. Rosenbaum and Silber 2008).
- Assessing evidence that the treatment group was made better than the control group (but being agnostic about the precise nature of the difference) (ex. $p > .2$ with difference of means but $p < .001$ with difference of ranks in Office of Evaluation Sciences study of General Services Administration Auctions)
- Focusing on detection rather than on estimation (for example to identify promising sites for future research in experiments with many blocks or strata) (Bowers and Chen 2020 working paper).
- Assessing hypotheses of no effects in small samples, with rare outcomes, cluster randomization, or other designs where reference distributions may not be Normal (see for example, (Gerber and Green 2012)).
- Assessing structural models of causal effects (for example models of treatment effect propagation across networks) (Bowers, Desmarais, et al. 2018; Bowers, M. Fredrickson, and Aronow 2016; Bowers, M. M. Fredrickson, and Panagopoulos 2013).

Summary: Applications of the Testing Approach

Examples of use of the testing approach:

- Assessing evidence of pareto optimal effects or no aberrant effect (i.e. no unit was made worse off by the treatment) (Caughey, Dafoe, and Miratrix 2016; P. Rosenbaum and Silber 2008).
- Assessing evidence that the treatment group was made better than the control group (but being agnostic about the precise nature of the difference) (ex. $p > .2$ with difference of means but $p < .001$ with difference of ranks in Office of Evaluation Sciences study of General Services Administration Auctions)
- Focusing on detection rather than on estimation (for example to identify promising sites for future research in experiments with many blocks or strata) (Bowers and Chen 2020 working paper).
- Assessing hypotheses of no effects in small samples, with rare outcomes, cluster randomization, or other designs where reference distributions may not be Normal (see for example, (Gerber and Green 2012)).
- Assessing structural models of causal effects (for example models of treatment effect propagation across networks) (Bowers, Desmarais, et al. 2018; Bowers, M. Fredrickson, and Aronow 2016; Bowers, M. M. Fredrickson, and Panagopoulos 2013).

Summary: Applications of the Testing Approach

Examples of use of the testing approach:

- Assessing evidence of pareto optimal effects or no aberrant effect (i.e. no unit was made worse off by the treatment) (Caughey, Dafoe, and Miratrix 2016; P. Rosenbaum and Silber 2008).
- Assessing evidence that the treatment group was made better than the control group (but being agnostic about the precise nature of the difference) (ex. $p > .2$ with difference of means but $p < .001$ with difference of ranks in Office of Evaluation Sciences study of General Services Administration Auctions)
- Focusing on detection rather than on estimation (for example to identify promising sites for future research in experiments with many blocks or strata) (Bowers and Chen 2020 working paper).
- Assessing hypotheses of no effects in small samples, with rare outcomes, cluster randomization, or other designs where reference distributions may not be Normal (see for example, (Gerber and Green 2012)).
- Assessing structural models of causal effects (for example models of treatment effect propagation across networks) (Bowers, Desmarais, et al. 2018; Bowers, M. Fredrickson, and Aronow 2016; Bowers, M. M. Fredrickson, and Panagopoulos 2013).

Summary: Applications of the Testing Approach

Examples of use of the testing approach:

- Assessing evidence of pareto optimal effects or no aberrant effect (i.e. no unit was made worse off by the treatment) (Caughey, Dafoe, and Miratrix 2016; P. Rosenbaum and Silber 2008).
- Assessing evidence that the treatment group was made better than the control group (but being agnostic about the precise nature of the difference) (ex. $p > .2$ with difference of means but $p < .001$ with difference of ranks in Office of Evaluation Sciences study of General Services Administration Auctions)
- Focusing on detection rather than on estimation (for example to identify promising sites for future research in experiments with many blocks or strata) (Bowers and Chen 2020 working paper).
- Assessing hypotheses of no effects in small samples, with rare outcomes, cluster randomization, or other designs where reference distributions may not be Normal (see for example, (Gerber and Green 2012)).
- Assessing structural models of causal effects (for example models of treatment effect propagation across networks) (Bowers, Desmarais, et al. 2018; Bowers, M. Fredrickson, and Aronow 2016; Bowers, M. M. Fredrickson, and Panagopoulos 2013).

1. An overview of approaches to statistical inference for causal quantities
2. Linear Regression in Experiments and Observational Studies

What does linear regression do in an experiment?

Let's make a slightly bigger dataset where the potential outcomes depend on a background covariate:

```
N <- 100
tau <- .2
dat <- data.frame(
  id = 1:N,
  x1 = rpois(n = N, lambda = 10)
)
dat <- mutate(dat,
  y0 = .5 * sd(x1) * x1 + runif(n = N, min = -2 * sd(x1), max = 2 * sd(x1)),
  y1 = y0 + tau * sd(y0) #+ runif(n = N,min=-.5*sd(y0),max=.5*sd(y0)),
)
set.seed(12345)
dat$Z <- complete_ra(N = N, m = floor(N / 2))
dat <- mutate(dat, Y = Z * y1 + (1 - Z) * y0)
dat$tau <- with(dat, y1 - y0)
head(dat)
```

```
  id x1     y0     y1 Z      Y   tau
1  1  8 13.436 14.594 0 13.436 1.159
2  2  7  9.182 10.340 1 10.340 1.159
3  3  6  3.747  4.905 1  4.905 1.159
4  4  9 10.311 11.469 1 11.469 1.159
5  5 16 23.106 24.264 1 24.264 1.159
6  6 14 22.721 23.879 0 22.721 1.159
```

```
## summary(lm(y0~x1,data=dat))$r.squared
## blah <- lm_robust(Y~Z,data=dat); blah$p.value["Z"]
## blah2 <- lm_robust(Y~Z+x1,data=dat); blah2$p.value["Z"]
```

What does linear regression do in an experiment?

Let's make a slightly bigger dataset where the potential outcomes depend on a background covariate:

Notice that background outcomes are now a function of a background covariate:

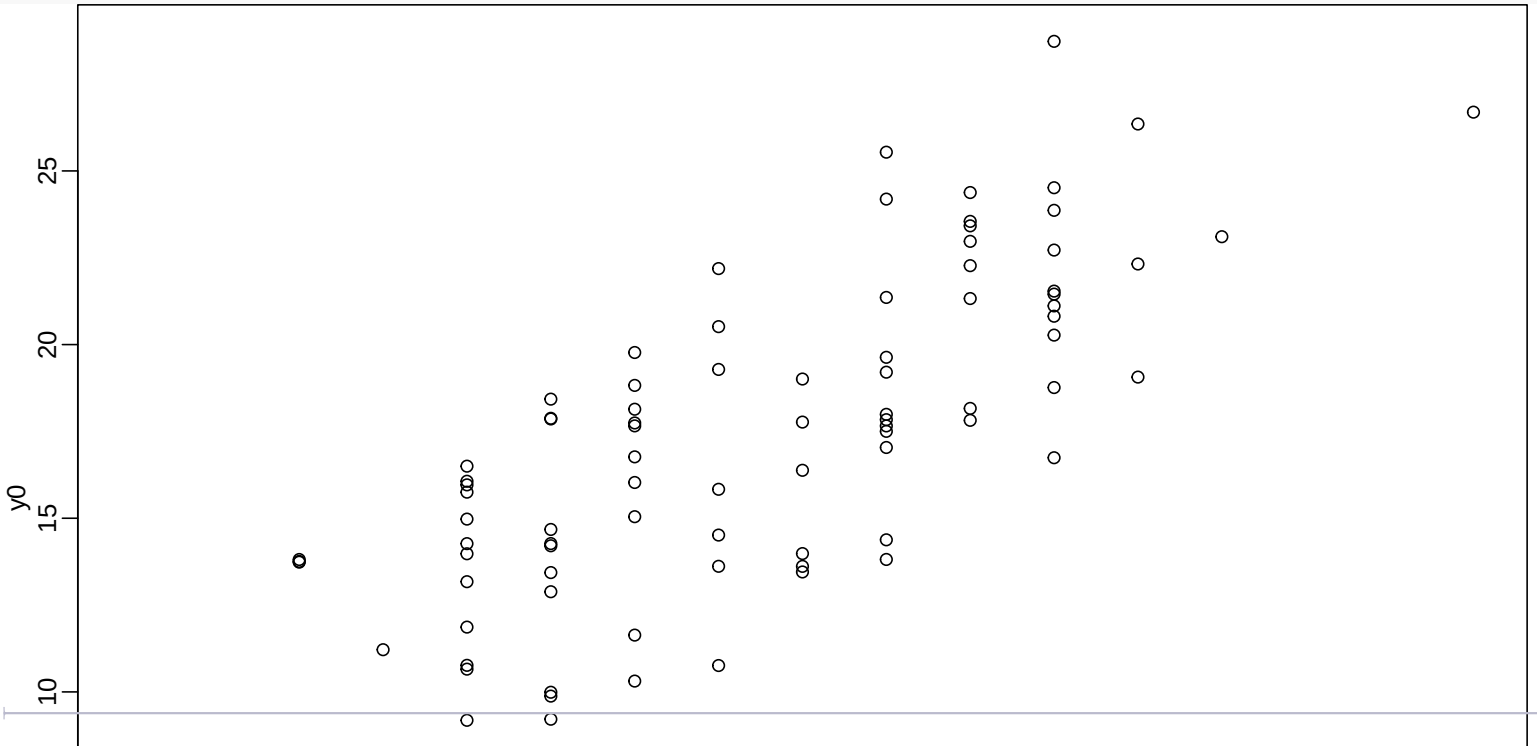
```
with(dat, cor(x1, y0))
```

```
[1] 0.8023
```

```
summary(lm(y0 ~ x1, data = dat))$r.squared
```

```
[1] 0.6437
```

```
with(dat, plot(x1, y0))
```



What does linear regression do in an experiment?

First, how does linear regression "adjust" for x1?

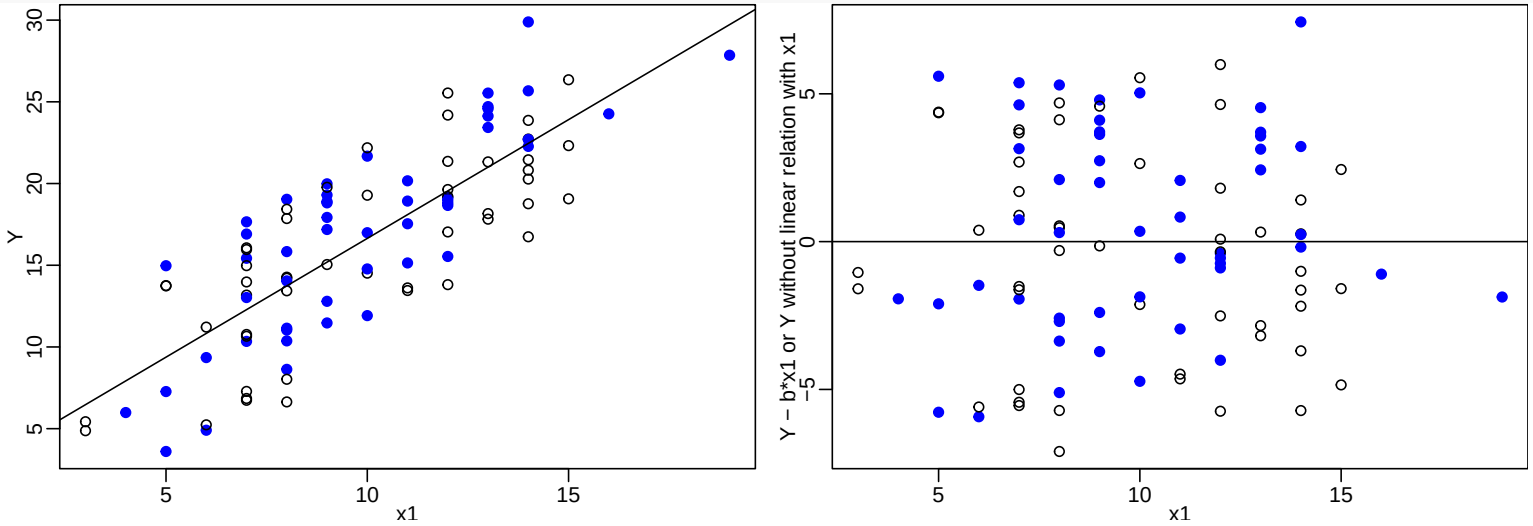
```
unadj_est <- lm_robust(Y ~ Z, data = dat)
adj_est <- lm_robust(Y ~ Z + x1, data = dat)
c(coef(unadj_est)[2], coef(adj_est)[2])
```

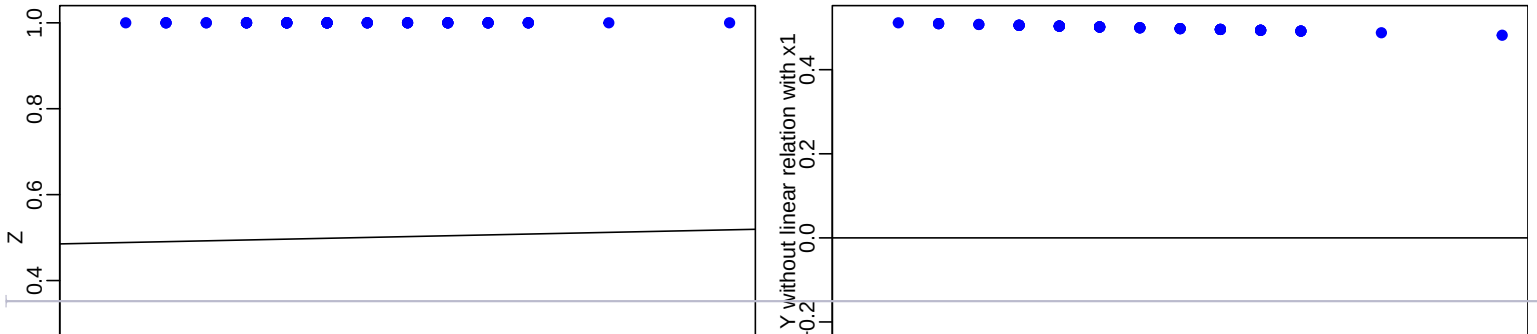
```
    Z     Z
1.148 1.032
```

What does linear regression do in an experiment?

First, how does linear regression "adjust" for x1? (Answer: by removing linear relationships between x1 and Z and between x1 and Y):

```
lm_Y_x1 <- lm(Y ~ x1, data = dat)
lm_Z_x1 <- lm(Z ~ x1, data = dat)
dat$resid_Y_x1 <- resid(lm_Y_x1)
dat$resid_Z_x1 <- resid(lm_Z_x1)
lm_resid_Y_x1 <- lm(resid_Y_x1 ~ x1, data = dat)
lm_resid_Z_x1 <- lm(resid_Z_x1 ~ x1, data = dat)
```





What does linear regression do in an experiment?

Linear regression "adjusts" the $Z \to Y$ relationship for a background covariate x by removing linear relationships: For example, here we show how it works after removing the linear relationships between `x1` and `Z` and between `x1` and `Y`):

```
coef(adj_est)[2]
```

```
    Z
1.032
```

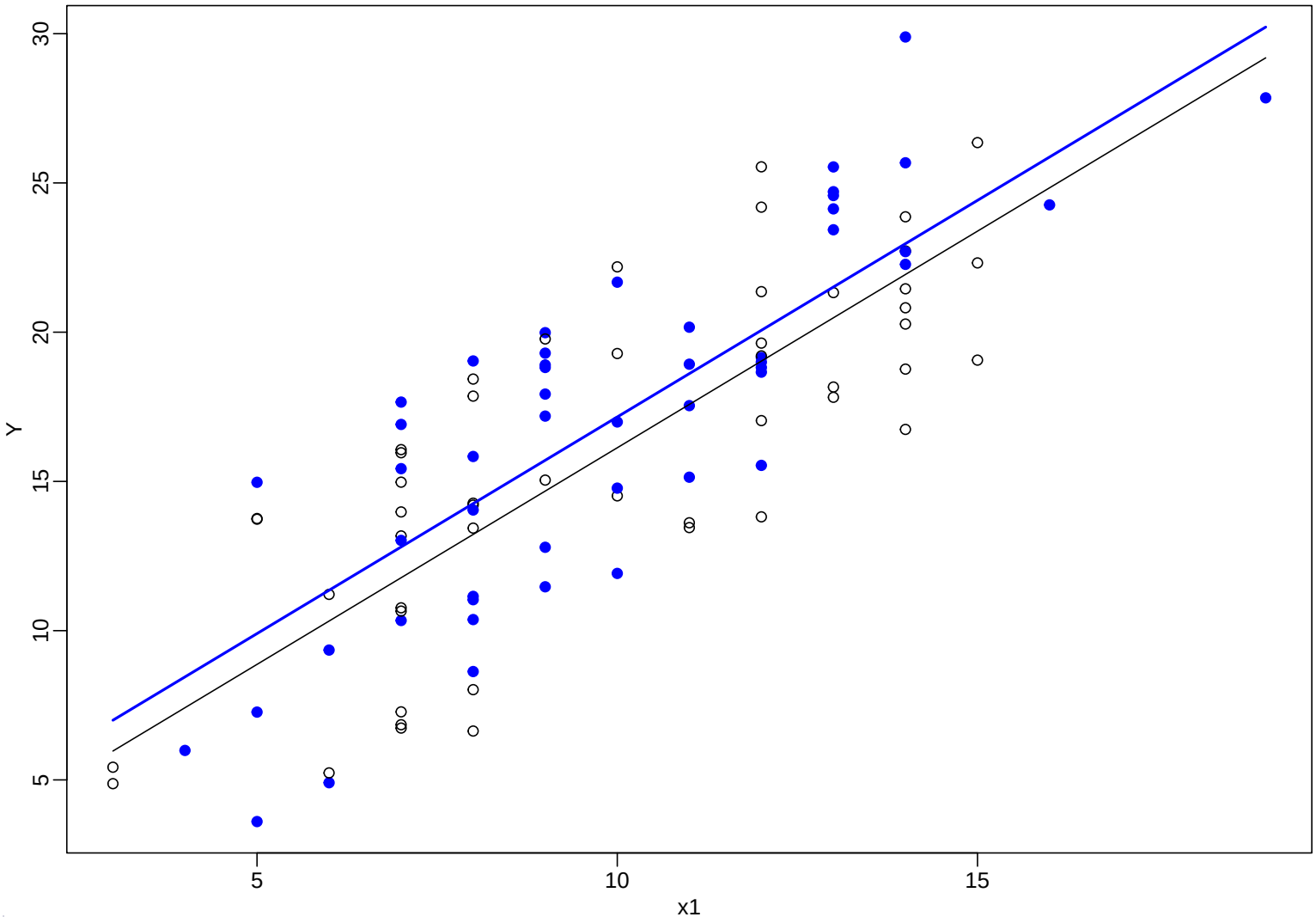
```
adj_est2 <- lm(resid_Y_x1 ~ resid_Z_x1, data = dat)
coef(adj_est2)[2]
```

```
resid_Z_x1
     1.032
```

```
stopifnot(all.equal(coef(adj_est)[[2]], coef(adj_est2)[[2]]))
```

What does linear regression do in an experiment?

Linear regression "adjusts" the $Z \to Y$ relationship for a background covariate x by removing linear relationships: For example, here we show how it works after removing the linear relationships between `x1` and `Z` and between `x1` and `Y`):



What does linear regression do in an experiment?

Why might we adjust in an experiment? The idea is to increase precision. Notice the smaller standard errors arising from the fact that we've removed noise independent of treatment from the outcome (recall that background covariates are indepedent of treatment assignment).

```
unadj_est$std.error["Z"]
```

```
    Z
1.164
```

```
adj_est$std.error["Z"]
```

```
     Z
0.6986
```

But what about bias? The point estimates differ here (because in finite samples, independence doesn't guarantee lack of correlation!)

```
coef(unadj_est)[["Z"]]
```

```
[1] 1.148
```

```
coef(adj_est)[["Z"]]
```

```
[1] 1.032
```

How would we know whether the bias is too big? I

Different may not mean worse. Let's see how these (and some other) estimators work. First, just specify a bunch of different estimators:

How would we know whether the bias is too big? II

```r
## The truth:
trueATE_estimand <- with(dat, mean(y1 - y0))

## A function to make a new experiment.
new_exp <- function(Z) {
  ## This next shuffles Z
  newZ <- sample(Z)
  return(newZ)
}

## The est1.. functions are estimators of the trueATE, each returning a single number --- the esti
est1 <- function(Z, Y, x = NULL) {
  mean(Y[Z == 1]) - mean(Y[Z == 0])
}
est2 <- function(Z, Y, x = NULL) {
  coef(lm(Y ~ Z))[["Z"]]
}
est3 <- function(Z, Y, x) {
  ## "controlling for" but not really because this is an experiment.
  coef(lm(Y ~ Z + x))[["Z"]]
}
est4 <- function(Z, Y, x) {
  quantY <- quantile(Y, .9)
  ## Recode highest 10% of outcome scores to 90% percentile to minimize effect of outliers.
  Y[Y > quantY] <- quantY
  coef(lm(Y ~ Z + x))[["Z"]]
}
est5 <- function(Z, Y, x) {
  ## Use the Lin method of covariance adjustment to minimize bias
```

How would we know whether the bias is too big? I

Different may not mean worse. Let's see how these (and some other) estimators work. First, just specify a bunch of different estimators:

```
set.seed(1235)
nsims <- 5000
dist_ests <- with(
  dat,
  replicate(nsims, compare_ests(Z = Z, y1 = y1, y0 = y0, x = x1))
)
apply(dist_ests, 1, summary)
```

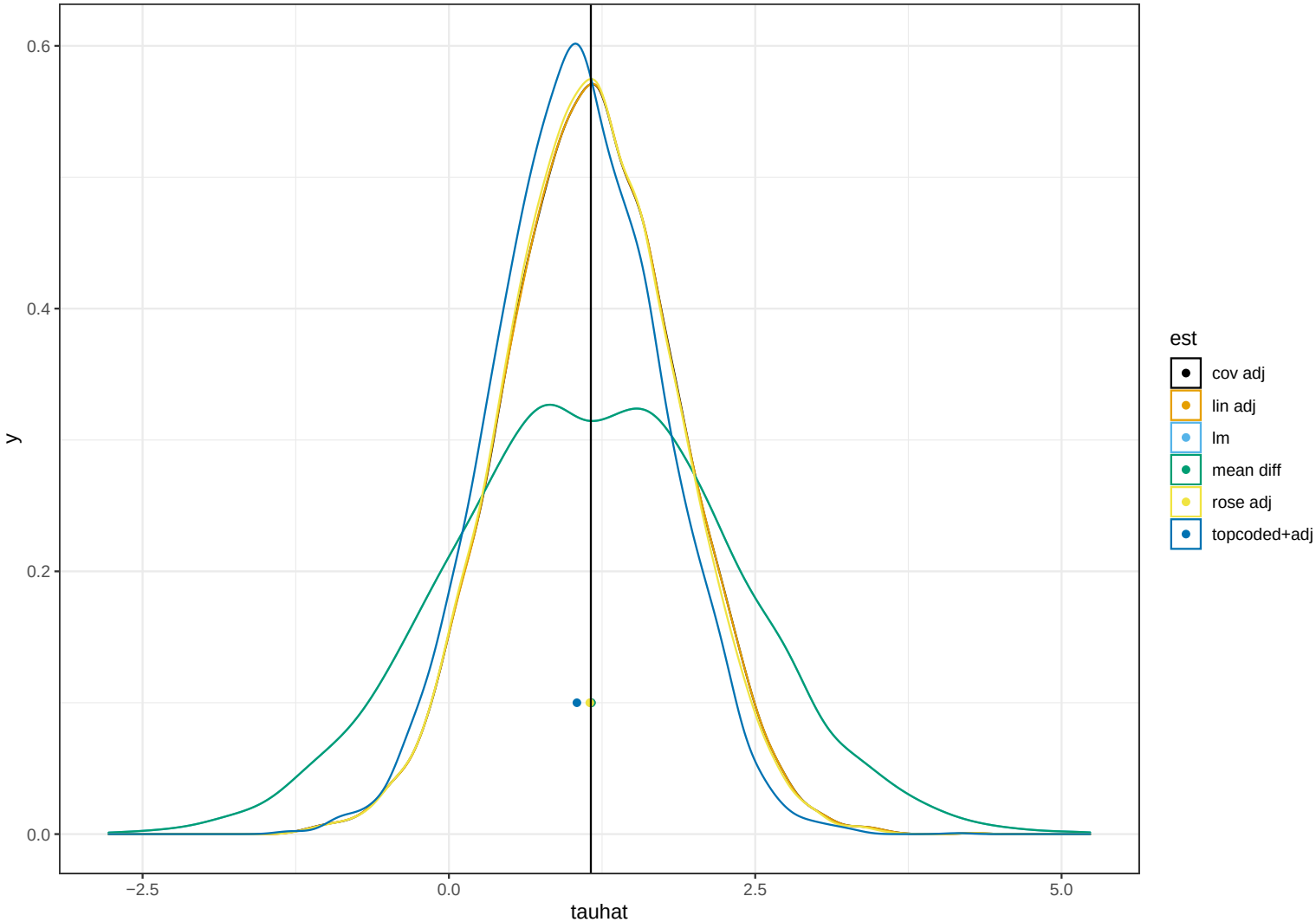
```
           est1    est2    est3    est4    est5    est6
Min.    -2.7775 -2.7775 -1.2075 -1.3458 -1.2058 -1.1921
1st Qu.  0.3732  0.3732  0.6849  0.5889  0.6871  0.6783
Median   1.1718  1.1718  1.1614  1.0393  1.1615  1.1479
Mean     1.1624  1.1624  1.1617  1.0446  1.1616  1.1499
3rd Qu.  1.9576  1.9576  1.6327  1.5006  1.6322  1.6159
Max.     5.2336  5.2336  4.2868  4.1551  4.2849  4.2862
```

```
apply(dist_ests, 1, function(x) {
  trueATE_estimand - mean(x)
})
```

```
     est1      est2      est3      est4      est5      est6
-0.003846 -0.003846 -0.003138  0.113950 -0.003071  0.008617
```

How would we know whether the bias is too big?

Different may not mean worse. Let's see how these estimators work (black line is true ATE):



More on the Rosenbaum 2002 approach:

The first approach worked because the treatment was independent of the covariate. Yet, in a finite sample, there is still some relationship — that was why the plot Z by x1 and the plot of the resid_Z_x1 by x1 were not the same.

```
summary(lm(Z ~ x1, dat))$r.squared
```

```
[1] 0.0001576
```

Another approach (from P. R. Rosenbaum (2002)) does not adjust Z at all — after all, we know that is randomized. We have no need to adjust Z we just want to remove non-treatment related noise in our outcome in order to increase precision. Notice that we also gain in precision here by not reducing the variation in Z.

```
rose_mod <- lm_robust(resid_Y_x1 ~ Z, data = dat)
rose_mod$p.value["Z"]
```

```
     Z
0.1409
```

```
unadj_est$p.value["Z"]
```

```
     Z
0.3267
```

```
adj_est$p.value["Z"]
```

```
    Z
0.143
```

More on the Lin 2013 approach:

Freedman (2008) showed the bias in simple covariance adjusted estimates of the ATE. Lin (2013) showed how to significantly reduce this bias.

```
## remove the mean from the covariate, center the covariate.
dat$x1_c <- with(dat, x1 - mean(x1))
## fit a model interacting the centered covariate with the treatment
lm_lin1 <- lm_robust(Y ~ Z + Z * x1_c, data = dat)
coef(lm_lin1)[["Z"]]
```

```
[1] 1.031
```

```
lm_lin2 <- lm_lin(Y ~ Z, covariates = ~x1, data = dat)
coef(lm_lin2)[["Z"]]
```

```
[1] 1.031
```

```
c(
  lm_lin2$std.error["Z"],
  lm_lin1$std.error["Z"],
  adj_est$std.error["Z"],
  rose_mod$std.error["Z"],
  unadj_est$std.error["Z"]
)
```

```
     Z      Z      Z      Z      Z
0.6953 0.6953 0.6986 0.6949 1.1644
```

Linear Regression for Covariance Adjustment in a RCT

Covariance adjustment is the removal of linear relationships. We can think of "removing a linear relationship" as "residualization".

- In randomized experiments:
  - Direct covariance adjustment can increase precision in the estimation of average causal effects.
  - The precision comes with some bias (perhaps a very small amount, perhaps worthwhile making the bias versus variance tradeoff in some cases, perhaps not in others).
  - We can reduce the bias following Lin (2013).
  - More precision can arise if we don't adjust the treatment following P. R. Rosenbaum (2002) (could involve some bias here too).

We don't say "controlling for" when we adjust for covariates in an experiment because we not removing their relationship with Z (after all, randomization is supposed to have already done that. If asked "What is the correlation between X and Z?" the answer is "On average, 0.")

Linear Regression for Covariance Adjustment in a RCT

Covariance adjustment is the removal of linear relationships. We can think of "removing a linear relationship" as "residualization".

- In randomized experiments:
  - Direct covariance adjustment can increase precision in the estimation of average causal effects.
  - The precision comes with some bias (perhaps a very small amount, perhaps worthwhile making the bias versus variance tradeoff in some cases, perhaps not in others).
  - We can reduce the bias following Lin (2013).
  - More precision can arise if we don't adjust the treatment following P. R. Rosenbaum (2002) (could involve some bias here too).

We don't say "controlling for" when we adjust for covariates in an experiment because we not removing their relationship with Z (after all, randomization is supposed to have already done that. If asked "What is the correlation between X and Z?" the answer is "On average, 0.")

Linear Regression for Covariance Adjustment in a RCT

Covariance adjustment is the removal of linear relationships. We can think of "removing a linear relationship" as "residualization".

- In randomized experiments:
  - Direct covariance adjustment can increase precision in the estimation of average causal effects.
  - The precision comes with some bias (perhaps a very small amount, perhaps worthwhile making the bias versus variance tradeoff in some cases, perhaps not in others).
  - We can reduce the bias following Lin (2013).
  - More precision can arise if we don't adjust the treatment following P. R. Rosenbaum (2002) (could involve some bias here too).

We don't say "controlling for" when we adjust for covariates in an experiment because we not removing their relationship with Z (after all, randomization is supposed to have already done that. If asked "What is the correlation between X and Z?" the answer is "On average, 0.")

Linear Regression for Covariance Adjustment in a RCT

Covariance adjustment is the removal of linear relationships. We can think of "removing a linear relationship" as "residualization".

- In randomized experiments:
  - Direct covariance adjustment can increase precision in the estimation of average causal effects.
  - The precision comes with some bias (perhaps a very small amount, perhaps worthwhile making the bias versus variance tradeoff in some cases, perhaps not in others).
  - We can reduce the bias following Lin (2013).
  - More precision can arise if we don't adjust the treatment following P. R. Rosenbaum (2002) (could involve some bias here too).

We don't say "controlling for" when we adjust for covariates in an experiment because we not removing their relationship with Z (after all, randomization is supposed to have already done that. If asked "What is the correlation between X and Z?" the answer is "On average, 0.")

Linear Regression for Covariance Adjustment in a RCT

Covariance adjustment is the removal of linear relationships. We can think of "removing a linear relationship" as "residualization".

- In randomized experiments:
  - Direct covariance adjustment can increase precision in the estimation of average causal effects.
  - The precision comes with some bias (perhaps a very small amount, perhaps worthwhile making the bias versus variance tradeoff in some cases, perhaps not in others).
  - We can reduce the bias following Lin (2013).
  - More precision can arise if we don't adjust the treatment following P. R. Rosenbaum (2002) (could involve some bias here too).

We don't say "controlling for" when we adjust for covariates in an experiment because we not removing their relationship with Z (after all, randomization is supposed to have already done that. If asked "What is the correlation between X and Z?" the answer is "On average, 0.")

Summary of the Day

- There is more than one way to use what we observe to reason about potential outcomes: testing, estimating, predicting.
- Linear regression is the most common way to "adjust for" background covariates in non-experimentable studies. We learned about what it means to adjust for covariates in experimental studies so that we can better evaluate linear regression in observational studies (tomorrow).
- Structure of the course. Where we are now. Where we are going: away from randomized experiments toward observational studies, but with the statistical basis developed from our understanding of randomized experiments.

Summary of the Day

- There is more than one way to use what we observe to reason about potential outcomes: testing, estimating, predicting.
- Linear regression is the most common way to "adjust for" background covariates in non-experimentable studies. We learned about what it means to adjust for covariates in experimental studies so that we can better evaluate linear regression in observational studies (tomorrow).
- Structure of the course. Where we are now. Where we are going: away from randomized experiments toward observational studies, but with the statistical basis developed from our understanding of randomized experiments.

Summary of the Day

- There is more than one way to use what we observe to reason about potential outcomes: testing, estimating, predicting.
- Linear regression is the most common way to "adjust for" background covariates in non-experimentable studies. We learned about what it means to adjust for covariates in experimental studies so that we can better evaluate linear regression in observational studies (tomorrow).
- Structure of the course. Where we are now. Where we are going: away from randomized experiments toward observational studies, but with the statistical basis developed from our understanding of randomized experiments.

References

- Barnard, J. et al. (2003). "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City.". In: Journal of the American Statistical Association 98.462, pp. 299–324.
- Bowers, Jake, Bruce A Desmarais, et al. (2018). "Models, methods and network topology: Experimental design for the study of interference". In: Social Networks 54, pp. 196–208.
- Bowers, Jake, Mark Fredrickson, and Peter M Aronow (2016). "Research Note: A more powerful test statistic for reasoning about interference between units". In: Political Analysis 24.3, pp. 395–403.
- Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos (2013). "Reasoning about Interference Between Units: A General Framework". In: Political Analysis 21.1, pp. 97–124.
- Caughey, Devin, Allan Dafoe, and Luke Miratrix (July 2016). "Beyond the Sharp Null: Permutation Tests Actually Test Heterogeneous Effects".
- Freedman, David A (2008). "On Regression Adjustments to Experimental Data". In: Advances in Applied Mathematics 40.2, pp. 180–193.


Gerber, Alan S and Donald P Green (2012). Field experiments: Design, analysis, and interpretation. WW Norton.


Hahn, P Richard, Jared S Murray, and Carlos M Carvalho (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects". In: Bayesian Analysis.


Lin, Winston (2013). "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique". In: The Annals of Applied Statistics 7.1, pp. 295–318.


Rosenbaum, Paul and Jeffrey H Silber (2008). "Aberrant effects of treatment". In: Journal of the American Statistical Association 103.481, pp. 240–247.


Rosenbaum, Paul R (2002). "Covariance adjustment in randomized experiments and observational studies". In: Statistical Science 17.3, pp. 286–327.