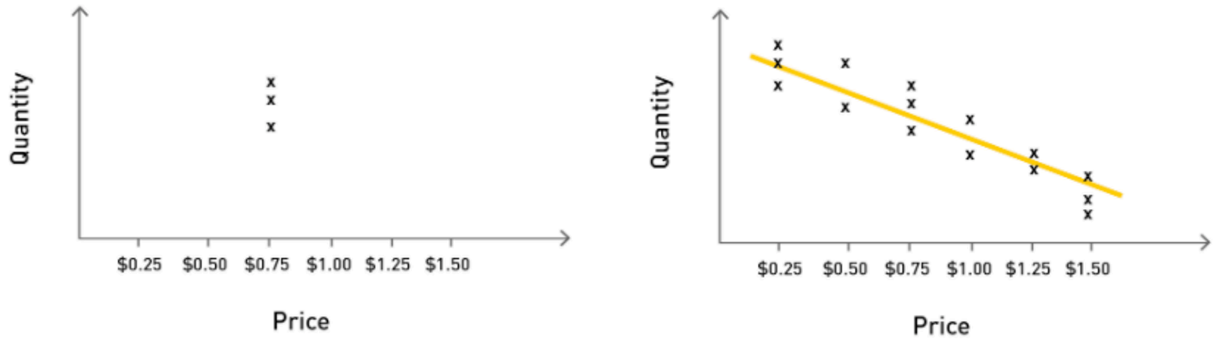
Field Experiments

1 Overview

- **Experiment**:
  - An intervention that creates variation in order to teach us about causal questions.

- Causal questions crucial in a variety of areas (e.g., business, public policy, individual decision making).
- Decisions concerned with counterfactuals.
  - The state of the world if we had done X? If we had done Y?
- Causal inference difficult because we can't observe both states of the world.
- Arguments based on intuition/anecdotes usually end in stalemates.
  - Extend unemployment benefits in a recession?
- Might settle with an experiment
  - Keep using Google ads?
- Causal questions settled with experiments in a way that avoids stalemates.
- Causal questions harder to get correct in social science than in physical science, because of heterogeneity.
  - All electrons the same.
  - No two human minds or bodies the same.
- Should be skeptical of causal inferences based on observational (i.e., nonexperimental) data because of the possibility of unobserved heterogeneity.
- Research without interventions leaves us uncertain about causality.
- Many causal questions are hard to answer because of a lack of naturally occurring data. For example:
  - Would doubling our advertising budget increase profits?
  - Would legalizing LSD cause bad health outcomes?
  - When the FCC sells spectrum rights, should it use a sealed-bid auction or an ascending-bid auction?
  - How much more likely is a campaign donor to get an appointment with a legislator, relative to a nondonating constituent?

1.1 Intervention vs Randomization



- We can't know anything about the demand curve if the price remains the same.
- To learn about the demand curve, we could deliberately vary the price.
    - Variation is crucial for making causal inferences.
    - This is **intervention**.
- We deliberately create variation when we run experiments.
- Experiments do not have to involve randomization.
- **Intervention is the key element of an experiment.**
- Randomization can be difficult to implement.
    - Example: randomly assigning price for each customer
        - Might take too much time
        - Might irritate customers
- We learn more with deliberate variation than without it.
- When is randomization useful?
    - We can control for confounds by repeating the experiment many times.
    - The treatment—price—should be independent of everything else that might influence sales (e.g., temperature, sun, holidays).
    - **Randomization guarantees this independence**.

1.2 Natural experiments

- A natural experiment is naturally occurring data that a researcher argues have the same properties as a true, controlled experiment.
- **Herschel's garden:** idea of viewing the night sky as a garden that features the same types of "plants" at different stages of development.
    - An idea that can be applied to the social sciences.

1.2.1 Observational Studies

- **Best**: Data from "naturally occurring experiments" (i.e., situations where variation was produced by something like random assignment)
    - Example: charter school lottery deciding applicant acceptance
        - Data about applicants possibly analyzed to study causal effects of charter school education
        - Same characteristics in both groups, those who attended the charter school and those who did not
    - Example: Vietnam draft lottery that caused some people to get more college education
        - Group that got more college education otherwise identical to group that did not
- **Less ideal**: decent reasons to believe that those who received an intervention are otherwise identical to those who did not
    - Example: snowplows clearing streets on alternate days—effect on business foot traffic?
        - Days of the week not necessarily identical in terms of foot traffic
    - Example: minimum wage increase in New Jersey but not in Pennsylvania—effect on fast-food employment?
        - Minimum wage in New Jersey not raised randomly
- **Bad**: no good reasons to believe that those who received an intervention are otherwise identical to those who did not

1.3 Online Auctions: Purpose and method

- **Auction formats**:
  - **English Auctions**: Ascending bids
    - People keep bidding higher till it ends.
  - **Sealed-bid auctions(first-price)**: One-time highest bids
    - Everyone submits a bid. Highest bid wins.
  - **Dutch Auctions**: descending price clock
    - Price starts at max and keeps dropping.
    - Person stopping clock before others wins.
  - **Vickrey's "second-price" auctions**:
    - Second-highest bid determines winner's price
    - If highest bid is 50 and second highest bid is 40, highest bidder wins and pays 41.
- Using field experiments to test equivalence between auction formats.
  - Which format makes more money?
    - According to theory, all formats should raise same amount on average.
  - 2 kinds of revenue equivalence
    - **Strategic equivalence**: strong prediction.
      - Dutch and first-price auctions are strategically equivalent.
        - Same amount of information
      - English and second-price auctions are strategically equivalent under "private values."
        - Dominant-strategy mechanisms for truthful revelation of valuations
        - i.e., regardless of others' strategies, my optimal strategy is the same: bid my maximum willingness to pay.
    - General revenue equivalence is weaker.
      - Expected revenue of all four formats should be the same if risk neutral.
  - **Violations to theory:** Actual results
    - First-price auctions raise more revenue than Dutch auctions.
    - With private values, subjects overbid in second-price auctions.
      - Yielded higher revenue than in English auctions
    - First/Dutch revenues higher than English/second.
      - Due to risk aversion?

2 Comparing Apples to Apples

- Experimentation delivers much more reliable causal inference than any observational method.
    - Allows us to compare two identical populations in which all that varies is treatment of interest.
- Conducting experiments correctly isn't easy.
- Concept of "potential outcomes" shows us what can go right and wrong.
    - If you can't imagine a manipulation that answers your question, it may not have a causal answer.
    - What would the same person do if in one treatment versus another?
    - Intervention is required to generate needed data, but sometimes imagining an intervention is impossible.
    - Example: What is the effect on mortality rates of being born in Africa?
        - What does this even mean for a particular person?
        - Yi(1) = outcome if person born in Africa
        - Yi(0) = outcome if same person born in the United States
        - Born in African hospital?
        - Lived entire life in Africa?
        - Question not posed well (FUQ'd – Fundamentally Unanswerable Question)
- **Potential outcome**:
    - Theoretical concepts useful for thinking about what an experiment could show.
    - Could never be derived from real data
    - Assumes an impossible amount of information
    - In practice: only treatment group observed in treatment, and only control group observed in control.
    - In theory: can imagine a group in two counterfactual states but can actually observe only one.
- **Notation**
    - $Y_i(1)$ = outcome if you were to be in treatment
    - $Y_i(0)$ = outcome if you were to be in control
    - $\tau_i = Y_i(1) - Y_i(0)$ = treatment effect

2.1 Experiments vs Observational data

- Key question in measuring causal effects of X on Y:
  - How does X vary?
  - Omitted variable—Christmas—causes increased advertising and increased sales.
  - Blindly running regression on observational data implicitly assumes advertising to be only variable responsible for increased sales.
  - Effects of advertising overestimated due to omitted-variable bias.
  - Using observational data; not comparing apples to apples.

- Examples of observational data providing inaccurate results
  - **Aggregate time-series data**
    - Advertising doesn't vary systematically over time.
    - **Reverse causality problem**.
      - Example: if advertising is 5% of sales, increase in sales means more advertising. (say during Christmas)
    - **Omitted-variable bias**.
      - Christmas example above.
  - **Individual cross-sectional data**
    - **Selection bias**:
      - Type of people who see ads not the same population as those who don't.
      - Example, people searching for online brokerage shown eTrade ads vs others not shown.
      - Even in absence of ads, shopping behavior might be different.

- **Key Points to note**:
  - Observational data can easily compare apples to oranges.
  - Selection bias:
    - Without a clean experiment, other factors can seem like treatment effects.
    - Those who select treatment often differ in other ways.
  - Experimentation more reliably estimates causal effects than observation.
    - Random assignment is gold standard.
  - Measuring effect of X on Y.
    - What are the potential outcomes for a given person?
    - What is the ideal experiment?
    - What causes the variation in X?

3 Quantifying Uncertainty

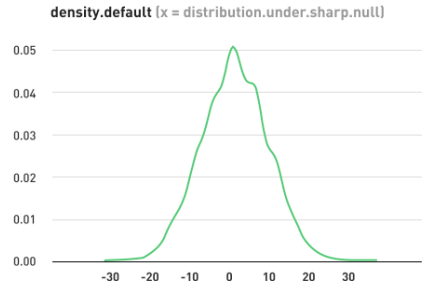
Sampling distribution: This is the frequency distribution of a statistic obtained from hypothetical replications of a randomized experiment.

Example: Does eating soybeans affect estrogen levels?

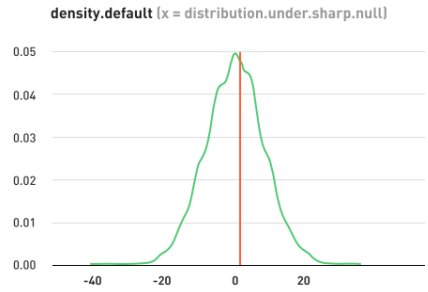
- We have 40 individuals: 20 men, 20 women.
  - `group <- c(rep("Man",20), rep("Woman",20))`
- Simulate the potential outcomes of the control group.
  - `po.control <- c(seq(from = 1, to = 20), seq(from = 51, to = 70))`
- Simulate the outcomes of the treatment group. (i.e no effect)
  - `po.treatment <- po.control`
- Now randomly pick 20 from treatment (i.e 1) and 20 from control (i.e 0)
  - `randomize <- function() sample(c(rep(0,20),rep(1,20)))`
  - `treatment <- randomize()`
- A single **realized outcome** of the experiment would be:
  - `outcomes <- po.treatment * treatment + po.control*(1-treatment)`
- Compute average treatment effect
  - `est.ate <- function(outcome, treat) mean(outcome[treat==1]) - mean(outcome[treat==0])`
  - `ate <- est.ate(outcomes, treatment)`
    - `OUTPUT: ate: [1] 1.3`
- We see a treatment effect even though we know **no** treatment was assigned.
- How do we that argue against a skeptic that a treatment has an effect.
  - We use the p-value.

Sharp null hypothesis: For every unit, there is no treatment effect.
Null Hypothesis: Average treatment effect is zero.

- Let's reshuffle the 40 people between treatment and control and compute ATE.
- Do these 5000 times and plot the distribution of the ATE.
    - `distribution.under.sharp.null <- replicate(5000, est.ate(outcomes, randomize()))`
    - `plot(density(distribution.under.sharp.null))`



- **Sampling distribution and P-value**:
    - How often did I get randomizations under the sharp null where the estimate was larger than my actual estimate?
    - For each, is it larger than the average treatment effect estimate?
    - How big is my estimate relative to the distribution of estimates?
        - `plot(density(distribution.under.sharp.null))`
        - `abline(v=ate) #Add a vertical line at our estimate`
        - `mean(ate < distribution.under.sharp.null)`
        - `Output: [1] 0.412`



- From the above we get the following:
    - 41.2% of values are greater than the ATE estimate of 1.3.
    - This means, there is a 41.2% chance we get a treatment effect greater than 1.3 even when one does not exist.
    - The p-value is thus 0.412
- Convention is to reject the null of no effect with p-value under 0.05.

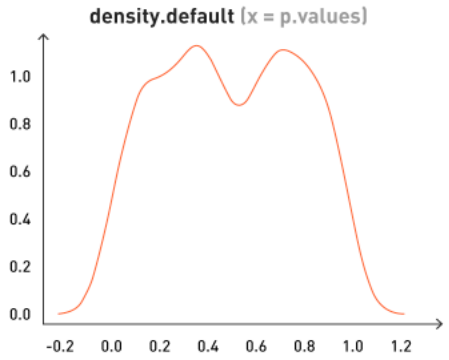
- p-values don't tell you for sure that the treatment has an effect.
- They just tell you ***how likely it is you would have gotten that result by chance***.
- The sampling distribution tells us how large the differences are we find by chance.
- **NOTE**: It is possible to find p-values < 0.05 even when the null hypothesis is correct.

Statistical Power

- Power is the probability of rejecting the null hypotheses, given that the null hypothesis is false; in other words, the probability of correctly rejecting H_0.
- When planning experiments, it is important to consider how large an effect you want to reliably detect.
    - If treatment effect is small, it is hard to detect because the p-value is insignificant.
    - How large a sample do you need to detect your desired effect size?
- The below example explains this.
    - We will replicate the soybean example from earlier with different treatment effects and compute the p-value.
    - The below R code is used

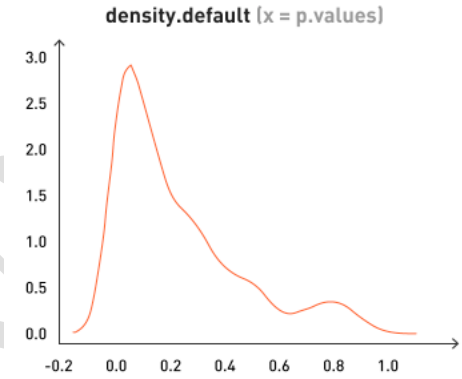
```
simulate.study <- function(treatment.effect.size){
+ po.control <- c(seq(from = 1, to = 20), seq(from = 51, to = 70))
+ po.treatment <- po.control + treatment.effect.size
+ treatment <- randomize()
+ outcomes <- po.treatment * treatment + po.control * (1-treatment)
+ ate <- est.ate(outcomes, treatment)
+ distribution.under.sharp.null <- replicate(1000, est.ate(outcomes,
randomize()))
+ return(mean(ate < distribution.under.sharp.null))
}
```

- **P-Value for no treatment effect**
    - `simulate.study(0)`
      `[1] 0.224`
    - The distribution of p-values for no treatment should be uniform.
        - `p.values <- replicate(10000, simulate.study(0))`
        - `plot(density(p.values)) #uniform distribution`
        - The below is not perfect square. But if we increase samples to say 100K, the plot smooths out further to be more uniform.



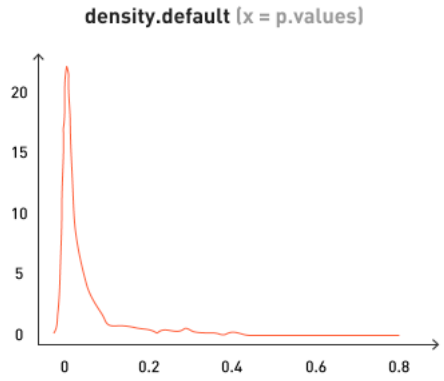
- How often can we expect to get a p-value less than 0.05 when there is no treatment?
  - `mean(p.values < 0.05)`
    `[1] 0.051`
  - As expected, since it's a uniform distribution, we can expect to see the value 5% of the time.
- Thus, **statistical power = 0.05. Unlikely to detect.**

- **P-Value when treatment effect is 10**
  - We now compute power for effect size of 10
  - `p.values <- replicate(500, simulate.study(10))`
  - `plot(density(p.values))`



  - `mean(p.values < 0.05)`
    `[1] 0.288`
  - We see **statistical power = 0.288** or somewhat likely to detect.

- **P-Value when treatment effect is 20**
  - We now compute power for effect size of 20
  - `p.values <- replicate(500, simulate.study(20))`
  - `plot(density(p.values))`



  - `mean(p.values < 0.05)`
    `[1] 0.764`
  - We see **statistical power = 0.764** or very likely to detect.

- **Increasing statistical power**. Power increases with:
  - Size of the effect. Larger effects easier to detect.
  - Increasing sample size. (power varies as square root of sample size)
  - Decreasing the standard error.
- **Decreasing statistical power**. Power decreases with:
  - Variation in outcome. More standard error, lower power.

Standard Error

- The spread of a sampling distribution is the standard error
- Estimated standard error for a sampling distribution is given by:

$$SE = \sqrt{\frac{var(Yi(0))}{N-m} + \frac{var(Yi(1))}{m}}$$

  - Where,
    - N → Sample size
    - m → Size of treatment group
    - $var(Yi(0))$ → Estimated variance of control
    - $var(Yi(1))$ → Estimated variance of treatment.
  - $NOTE$:
    - In the above formula, we assume treatment effect is same for all subjects which implies the correlation between Yi(0) and Yi(1) is 1.0.
    - If this assumption is not true, and we can compute the covariance, then the above formula cannot be used.

Confidence Interval

- The 95% CI is the interval that has a 0.95 probability of containing the true ATE.
- If α is the level of significance (typically 0.05)
  - 100(1- α) % CI = sample estimate +\- (multiplier).(standard error)
  - When $\alpha = 0.05$, multiplier = 1.96

Regression vs Permutation Inference

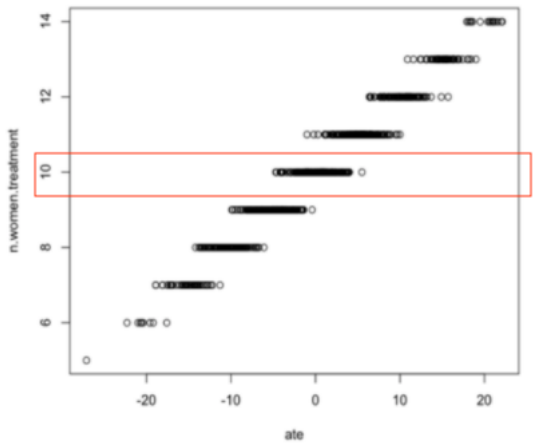
- The approach we used in previous example is called permutation inference.
- This is useful to understand the intuition behind the uncertainty estimates.
- If sample size > 50, regression gives the same answer much more quickly.

4 Blocking and Clustering

Blocking

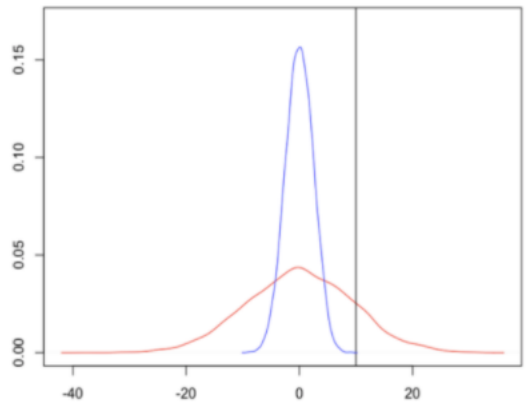
- When there are large differences between treatment and control, it is hard to know if it was due to chance
- We need to reduce the size of the differences that can arise by chance.
- Blocking is an approach to increase statistical power given an experiment with same sample and effect size.
- If some variables are related to the outcome, restrict ourselves to randomizations that keep treatment and control similar.
- **Example**:
    - In the soybean experiment discussed earlier, we randomize all 40 participants and place them either in control or treatment.
    - With the randomization we have, each permutation results in varying number of men and women in each group.
    - For example, we may get one sample having 9 men and 11 women in control and 11 men and 9 women in treatment. Another could have 13 and 7 split and so on.
    - If we conduct our earlier experiment that we used to compute statistical power and plot the results against the number of women in treatment, we see the following:

```
>sim.normal.study <- function(){
+ po.control <- c(seq(from = 1, to = 20), seq(from = 51, to =
70))
+ po.treatment <- po.control
+ treatment <- randomize()
+ outcomes <- po.treatment * treatment + po.control * (1-
treatment)
+ ate <- est.ate(outcomes, treatment)
+ n.women.treatment <- table(group, treatment)[2,2]
+ return(list(ate = ate, n.women.treatment = n.women.treatment))
+ }
>results <- t(replicate(1000, sim.normal.study()))
>plot(results)
```



- Clearly, we see the following:
  - Treatment effect is more when there are more women in treatment.
  - The confidence intervals on average show a treatment effect of zero, but we have a large variation (-20 to + 20) because the number of women vary.
  - We are really interested only in the band positioned where number of women=10.
  - So, if we randomize such that exactly 10 women are always in treatment and 10 in control, we significantly reduce the number of possible permutations and increase the power keeping sample size and effect size the same. This approach is called **blocking**.
- Let's compare the distribution of p-values under sharp null for the regular experiment and a blocked version of the experiment.

```
>plot(density(distribution.under.sharp.null), col="red",
ylim=c(0,.17)) #distribution without blocking
>abline(v=ate)
>distribution.under.sharp.null.blocked <- replicate(5000,
est.ate(outcomes.blocked, randomize.blocked()))
>lines(density(distribution.under.sharp.null.blocked),
col="blue") #distribution with blocking
```



- Clearly, blocking allows more precision by not conducting randomizations where covariates (e.g., gender) are very imbalanced.

- Blocking reduces probability of large differences that occur between treatment and control by chance by balancing presence of similar units across treatment and Control.
- Blocking can dramatically reduce the standard deviation of the sampling distribution.
- Power is chiefly determined by sample size and ratio of treatment effect to standard deviation in outcome.
- Standard deviation of outcome is often much smaller within groups of units identifiable ex-ante.
- It is successful to the extent blocks predict Y (especially good: pretreatment Y).
- **NOTE**:
  - The blocking behaviour can be achieved using regression with appropriate covariates without actually using blocking in the experiment.
  - For example, if we use the original randomization with regression, we see the following:

```
>po.control <- c(seq(from = 1, to = 20), seq(from = 51, to = 70))
>po.treatment <- po.control + 10 #simulate effect of 10
>treatment <- randomize()
>outcomes <- po.treatment * treatment + po.control * (1-
treatment)
```

  - **Example 1**: No blocking indicator

```
>summary(lm(outcomes ~ treatment)) #without block indicator
Coefficients:
                Estimate  Std.Error  t value   Pr(>|t|)
(Intercept)       34.650    5.883      5.890    8.06e-07***
treatment         11.700    8.319      1.406    0.168
```

- **Example 2**: Include gender as covariate (blocking indicator)

```
>summary(lm(outcomes ~ treatment + factor(group)))
Coefficients:
                  Estimate   Std.Error  t value    Pr(>|t|)
(Intercept)       12.000     1.524      7.873      2.01e-097***
treatment         6.667      1.825      3.653      0.000798***
factor
   (group)B       50.333     1.825      27.580     < 2e-16***
```

- We see above that regression gives results similar to blocking when we add the gender covariate.
- It also correctly shows the base difference between men and women in estrogen levels (50.33)
- With large samples, better to use regression. For small samples, we can use blocking.

Clustering

- Sometimes, we cannot individually assign treatment. In such cases, cluster assignment might be the only option.
- For example:
  - If we want to test whether lengthening school hours makes students more productive\improve test scores, we cannot assign treatment to each student. This would need to be done at a school level.
  - TV advertisements cannot individually show Ads, they need to cluster by geographic market.
  - Brick and mortar stores cannot assign prices randomly to customers.
    - Amazon had done this to estimate optimal pricing but got into trouble for doing this.

- **Clustering Example**
  - We will randomly assign teachers to 8 classrooms having 16 students each and analyze the treatment effect with clustering.
  - Setup the classrooms and students. Since we are clustering, individual students do not matter

```
>n.classrooms <- 8
>n.students <- 16
>classroom.ids <- unlist(lapply(1:n.classrooms, function(x)
rep(x,times=n.students)))
>classroom.ids
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 7
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
>all.classrooms <- unique(classroom.ids)
>all.classrooms
[1] 1 2 3 4 5 6 7 8
```

  - When treatment is assigned at a lever higher than the individual, the uncertainty at that level must be taken into account. (affects all individuals in a cluster)
  - In addition to variations at a classroom level, there could be variations at each student level too.
  - To account for the above, we add some noise both at classroom level as well as the individual level as follows

```
>classroom.level.noise <- rnorm(length(all.classrooms))
>classroom.level.noise
[1] -0.09677448 -0.92162355 -0.83917311 0.16667188 -1.55695401
-0.48004472 0.18754375 1.27884390
```

```
>student.outcomes.control <- rnorm(length(classroom.ids)) +
classroom.level.noise[classroom.ids]
>student.outcomes.control
[1] -0.26668159 -1.06631660 -1.24747184 -0.32880261 0.63158035
-0.38429133 -0.35639291 0.80059169 1.13219006 0.33447677 [……]
```

  - We simulate a treatment of 0.75

```
>student.outcomes.treat <- student.outcomes.control + 0.75
```

- We now compute the clustered ATE as follows:
  - Randomize at a cluster level – choose 4 out of 8 schools for treatment

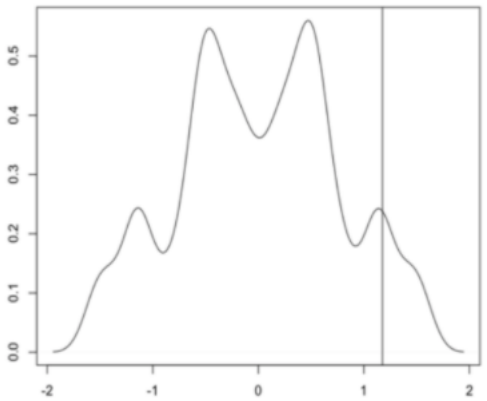
```
>randomize.clustered <- function(){
+ treat.classroom.ids <- sample(all.classrooms, n.classrooms/2)
+ return(
+ as.numeric(classroom.ids %in% treat.classroom.ids)
+ )
+ }

>treat <- randomize.clustered()
>treat
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

  - Compute ATE

```
>outcomes <- treat * student.outcomes.treat + (1-treat) *
student.outcomes.control
>ate <- est.ate(outcomes, treat)
>ate
[1] 0.7039964
```

  - Let's look at the distribution under sharp null and the p-value

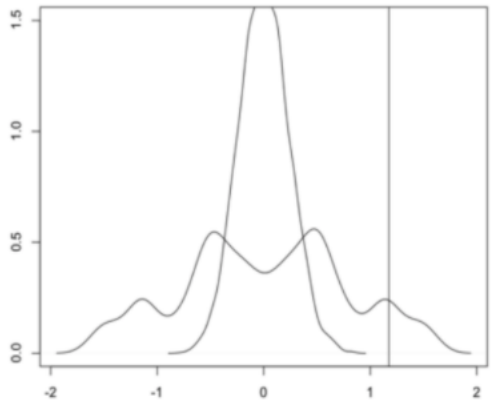
```
>distribution.under.sharp.null <- replicate(5000,
est.ate(outcomes, randomize.clustered()))
>plot(density(distribution.under.sharp.null))
>abline(v=ate) #pretty similar to one we'd get by chance
>mean(ate <= distribution.under.sharp.null) #p-value
[1] 0.0714
```



- From the above we see the following:
  - The ATE is pretty close to the actual treatment effect of 0.75.
  - However, the p-value is high indicating that an ATE of 0.70 could be reached by chance with really no treatment effect.
  - The distribution is not normal, but has multiple peaks because of the following reasons:
    - For central limit theorem to kick in, we need over 30 samples. Even though we have 128 students, we have only 8 clusters and our computation is at the cluster level. So, our sample size is very small. We may need at least 30 clusters.
    - Because of variations in clusters, its possible some classes generally perform higher and they also happen to get the better teacher (the treatment) or vice versa causing the peaks on the left and right of the distribution.

- **Example Ignoring clustering**
  - How do the estimates change if we ignore clustering.
    - This might happen if we unknowingly just compute the difference means of the 64 students in control vs treatment to compute ATE without considering the clusters.
  - We can compare the distributions as below:

```
>randomize.ignoring.clustering <- function()
sample(c(rep(0,n.classrooms*n.students/2),
rep(1,n.classrooms*n.students/2)))
>randomize.ignoring.clustering()
[1] 0 1 0 1 0 1 0 1 1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 1
1 1 0 1 0 1 1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 0 0
0 1 1 0 0 1 0 0 1 0 0 0 1 1 1 0 0 1 1 0 1 0 1 1 1 1 0 1 0 0 0 0 0
1 0 0 1 1 1 1 0 0 0 0 0 1 0 1 0 0 1 1 1 1 0 1 1 0 0 0 1 1 1 0
>distribution.under.sharp.null.wrong <- replicate(5000,
est.ate(outcomes, randomize.ignoring.clustering()))
>plot(density(distribution.under.sharp.null), ylim=c(0,1.5))
>lines(density(distribution.under.sharp.null.wrong))
>abline(v=ate)
```



    - As seen, we will see a much more sharper distribution because variations at cluster level are ignored. We get a much higher p-value and we overestimate the treatment effect.
- When units are assigned to treatment or control in clusters, larger differences between Treatment/Control outcomes will happen by chance.
- To account for this uncertainty, assign units together in these clusters when simulating the distribution under the null.
- Power worsens when:
    - Cluster-level noise is larger.
    - Number of clusters is smaller.
- Take into account the clustering process when estimating uncertainty, using randomization inference or regression.
- Cluster level noise: common shock that affects all individuals.

5 Covariates and Regression

Covariates

- Covariates are other variables we can measure besides the treatment variable and outcome variable. These can be used as follows:
    - Check for problems in experiment.
    - Improving precision of estimate (by using in regression)
- Covariates can indicate whether randomization worked and that we have an apples to apples comparison.
    - If there is a statistically significant difference in a covariate that could affect outcomes between treatment and control, then the randomization split didn't work\ experiment will produce inaccurate results.
    - **NOTE**: If randomization worked, each covariate should have a difference that is less than **2 standard deviations** from the mean. This is a good way to check if we are comparing apples to apples.
- If an experiment is complicated with many small treatment groups, it isn't possible to get precise measurements of the results for each group.
    - Confidence intervals can be too wide to say anything valuable.
    - In other words, we won't have enough statistical power in this design.
    - It's possible to fix an experiment where you spread your observations too thin.
    - **Note** that if you find you have low statistical power in your project, you may want to do fewer experimental treatments, or have a plan that will allow you to pool together data across multiple treatments in order to get more power.

- **Covariate imbalance**
    - Covariate imbalance in pre-treatment data is a good indicator that there is some error in the experiment.
    - These could be administrative errors or errors in randomization.
    - Suppose we notice that random assignment has produced an imbalance (say treatment group has higher pretest scores on average compared to control group), controlling for pre-test scores using regression will yield unbiased estimates.
        - Bias is a property of an estimation procedure, not a specific estimate.

Regression

- Regression provides a method to control for covariates
- Regression gives us a convenient way to summarize the results of an experiment, by specifying a single equation simultaneously containing both the treatment variable and the covariates we care about accounting for.
- A simple regression model could take on the following form:

$$Y_i = \beta_0 + \beta_1.X_i + \beta_2.A_i + \varepsilon_i$$

  - where,
    - The regression variables are:
      - Y is the outcome variable
      - X is the treatment variable
      - A is the control or covariate (0 or more depending on experiment)
    - The regression coefficients are:
      - β_0 is the intercept
      - β_1 is the treatment effect
      - β_2 is the effect of the covariate or control
    - ε is the residual\error

Difference in Differences Method

- Another approach to compute ATE using regression is to use difference in differences approach.
- The general assumption is that covariates remain fixed regardless of whether an observation is assigned to treatment or control. Depending on the experiment, this may not be true. So, this approach aims to account for these changes making the estimator for the ATE unbiased.
- Another advantage of this approach is that it reduces variability and thus the standard error (similar to blocking) and thus provides more precision and thus more statistical power.
- Let Y_i represent the outcome and X_i is the covariate we control for.
- In regular regression, we compare the average value of Y_i when d_i=1 (i.e dosage is given) to the average of Y_i where d_i=0. This is given by the coefficient of X_i.

$$Y_i = \beta_0 + \beta_1.X_i + \beta_2.A_i + \varepsilon_i$$

- In the ***difference in differences approach***,
  - we compare the average value of (Y_i - X_i) when d_i=1 to the average value of (Y_i-X_i) where d_i=0.
  - This approach constraints the slope to be 1. So, $\beta_0 = 1$.
  - Here, the experimental setup is as follows:
    - We denote a time period as t where t=0 indicates the period before randomization and t=1 indicates the period after treatment is provided.
    - The treatment variable is X having values 0 when treatment is not assigned and a value of 1 when treatment is assigned.
    - With this, we now get the following:
      - Pre-treatment covariates:
        - The values for the observations when t=0 (both in treatment and control)
      - Post treatment covariates:
        - The values for the observations when t=1(both in treatment and control)
    - The model is as follows:

$$Y_i = \beta_0 + \beta_1.X_i + \beta_2.t_i + \beta_3.X_i.t_i + \varepsilon_i$$

      - Where,
        - β_1 gives us selection effects i.e difference between treated and non-treated before treatment.
        - β_2 gives us change across time common to both treated and non-treated
        - β_3 gives us Treatment effect i.e difference in differences.
- **Issues to take care of**
  - Normally, post-treatment covariates are measured only after treatment is given.
  - In this approach, we need observations both pre and post treatment.
  - If we measure covariates prior to treatment, bias can be introduced if the pre-test provokes different reactions in control and treatment groups.

Omitted Variable Bias

- Instead of the previous equation (**long**), let's say we omit A and run a **short** regression.
- We would get:

$$Y_i = \beta_0 + \beta_1.X_i + \varepsilon_i$$

- Now, the effects of A can only be see using the above equation.
- This will introduce a bias in our estimates of β_1 the treatment effect.
  - If the omitted variable has no effect, we have no bias.
  - If the omitted variable is not correlated with the included variable, then there's no bias.
  - If the omitted variable is correlated with the included variable, then there is bias.
    - Positive correlation introduces a positive bias, so we overestimate the effects.
    - Negative correlation introduces a negative bias, so we underestimate the effects.

Measure Omitted Variable Bias

- Based on the previous definition we see that OVB is the difference between treatment effects when covariate is included and when it is excluded.
- Let the long form be:

$$Y_i = \beta_{01} + \beta_{11}.X_i + \beta_{21}.A_i + \varepsilon_{01_i}$$

- Let the short form be:

$$Y_i = \beta_{02} + \beta_{12}.X_i + \varepsilon_{02_i}$$

- This gives us:

$$OVB = \beta_{11} - \beta_{12}$$

- If X and A are corelated, this can be represented in regression form as:

$$A_i = \delta_0 + \delta_1.X_i + \delta_{2_i}$$ (regression anatomy)

- The **OVB Formula** is:

$$\boldsymbol{Effect\ of\ X_i\ in\ short} = \boldsymbol{Effect\ of\ X_i\ in\ long} + [(\boldsymbol{relation\ between\ A_i\ and\ X_i})\ \boldsymbol{x}\ (\boldsymbol{Effect\ of\ A_i\ in\ long})]$$

- The above gives us:

$$\beta_{12} = \beta_{11} + \delta_1.\beta_{21}$$

- Therefore, we get:

$$\mathbf{OVB} = \boldsymbol{\beta_{12} - \beta_{11} = \delta_1.\beta_{21}}$$

- This can be used to guess the consequences of omitting a variable.

- **Note**:
    - **Experimental Data**
        - With randomized treatment assignment, we know that treatment is uncorrelated with everything else: both observable covariates and un-observables we can't measure.
        - So, in an experiment, we don't have to worry about omitted-variable bias, because we should get approximately the same answer no matter how many covariates we include.
        - What including covariates does for us in an experiment is explain some of the residual variance in the regression, allowing us to shrink the standard error on the treatment effect. (i.e increase precision and statistical power)

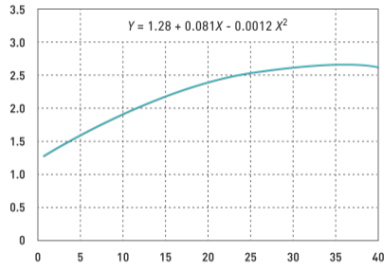
    - **Observational Data**
        - An observational study is where nothing changes and just record what you see, but an experimental study is where you have a control group and a testable group.
        - We always have the danger of OVB in observational data.
        - We cannot have causal inferences with observational studies\data.
        - When faced with observational data, we should always ask ourselves the following questions to better understand the potential for OVB:
            - What would be the ideal experiment for measuring the causal effect we're interested in?
            - Where does the variation in the observational data come from?
                - In other words, whom are you comparing to whom?
            - Can we tell stories about why the regression estimate might be biased, and what might be the direction of the bias?
                - In particular, why did people end up in the groups you are comparing in the first place?
                - And why might they differ for other reasons?

- **Examples for reference**
  - **Example 1**: Observational study to estimate returns to schooling i.e how much people earn based on number of years of schooling.
    - Below are the results

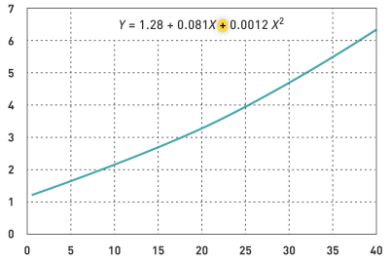
$$\ln Y_i = \underset{(.002)}{\alpha + .070\, S_i} + e_i$$

$$\ln Y_i = \underset{(.001)\quad(.001)\quad(.00002)}{\alpha + .107\, S_i + .081\, X_i - .0012\, X_i^2} + e_i$$

    - Omitted variable bias (OVB) causes us to underestimate the returns to schooling. Why?
      - This is observational data, so the two regressors can be correlated with each other.
      - Here, schooling and work experience are negatively correlated.
    - Both schooling and experience have positive effects on earnings.
    - When we omit experience, we force the effect of experience on earnings to become part of the schooling coefficient.
    - As people with more schooling have less experience, when we increase schooling, earnings don't increase as much as they would if we were holding experience constant as a covariate.
    - Hence, with OVB we measure the coefficient of interest to be 7% instead of 11%.
    - The quadratic specification in experience allows for the benefits of experience to accrue at a declining rate.
      - A positive coefficient on the linear term and a negative (smaller) coefficient on the squared term: **concave** function.



      - A positive coefficient on the linear term and a positive coefficient on the squared term: **convex** function.



- **Example 2**: Same observational study as example 1.
  - Griliches expected Mincer's estimates (in example 1) to be overstated due to the omission of ability from the regression. (he called **ability bias**)
  - Years of schooling are generally positively correlated with ability.
  - When Griliches included IQ as a covariate, the estimated returns to schooling fell from 6.8% per year of schooling to 5.9%.
  - Without IQ, OVB caused an overestimate of returns to schooling.
  - Have we now controlled for everything that might cause biased results?
    - No. There are more kinds of ability than IQ (e.g., emotional intelligence, curiosity, etc.).
    - We still have omitted variables that are likely to be correlated with years of schooling.
  - Thus, we cannot entirely trust the ~6% per year number.
- How could we be sure of the right result?
  - We would need an experiment that randomly sends some students to more years of school than others.
  - Then every possible omitted variable would be uncorrelated with years of schooling, eliminating OVB.

Attenuation Bias

- Angrist and Pischke mention a concept usually called attenuation bias.
- Imagine that we don't always correctly measure the treatment variable (years of schooling).
- With measurement error in the X variable, the ***resulting coefficient is biased toward zero***.
- For example,
    - If observations thought to be in the control group, were really in the treatment group (and vice versa), the effects would look smaller than actual.
    - This is an example of attenuation bias.
    - The X variable is treatment assignment, and it is measured with error, so the treatment effect will be biased toward zero.

Bad Controls

- When you analyze experimental results, **do not include other outcome variables as covariates** on the right-hand side of the regression.
- Generally, adding more covariates is good, but not when they are post treatment outcomes.
- **Example 1**:
    - Consider the study that wants to see the effects of schooling on income.
    - Including occupation as a covariate would be considered a bad control.
        - This is because in many cases, the schooling determines the occupation a person gets, and this makes it an outcome variable.
- **Example 2**:
    - Consider a study looking to see the effects of reputation on the revenue.
    - In such an analysis, including number of bids as a covariate could be considered a bad control.
        - A higher reputation could increase the number of bids and can thus be considered a post treatment outcome.
        - If we include number of bids, we will likely underestimate the overall effect of reputation on auction revenue.
            - Part of the effect of reputation is on attracting additional bidders, while part of the effect may come through getting existing bidders to bid more than they would have in the zero-reputation case
            - Including the "number of bids" covariate essentially makes us focus only on the latter effect, though the former effect may be really large

Regression Model Specifications

- The process of deciding what variables should be in a regression model and what form they should take is called specification.
- This changes the interpretation of the beta's. (coefficients)
- **Saturated Model**
    - Each value of the covariate gets a different dummy variable to represent it.
        - Example: If variable is years of schooling. Have separate dummy variables for 0 to 20.
    - With multiple covariates, we also have to include all possible interactions between these dummy variables (to estimate a two- or three-dimensional surface).
- Why might a researcher choose a model that is not fully saturated?
    - For example, use linear in a continuous variable (not a full set of dummies) or exclude interaction effects.
    - Answer:
        - Worry about too little data per cell to get identification and precision.
        - If you have only 500 observations, you can't estimate a full 10×10×10 model. But you could try a coarser model.
- What is good about using lots of covariates?
    - Increased statistical precision
- What is bad about using lots of covariates?
    - "Fishing expeditions" happen quite frequently with observational data.
    - Not much downside in an experiment, if we have made sure that X is independent of all covariates.

Questions We Should Ask in Any Causal-Effect Research

- What is the causal relationship of interest?
- What is the ideal experiment to measure this?
    - Ask this even if you're doing observational research.
    - If your question seems FUQ'd (Fundamentally Unanswerable Question), is there a different but related question that can be answered with a hypothetical experiment?
    - FUQ example: *What is the effect on earnings of being born in Africa instead of North America?*
- What is your identification strategy?
    - That is, where does your variation come from, and why do you think it is independent of potential outcomes?
- How are you computing your confidence intervals?

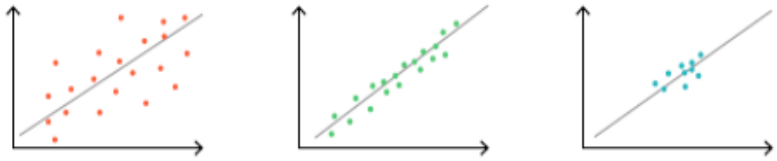
Standard Errors and Confidence Intervals

- Standard errors are (variance in estimate of slope coefficient):
  - Larger when variance in Y is larger
  - Smaller when variance in X is larger

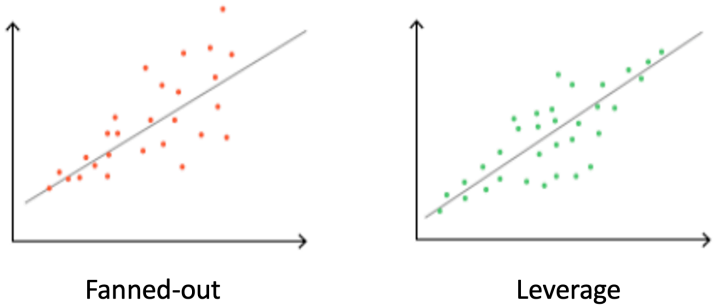
- **Treatment effect (with 95% CI) ≈ Slope coefficient ± 2 standard errors**

- **OLS Standard errors** assume that each observations' residual error (ε) is:
  - Independent: Randomization does this.
  - Identically distributed
    - That is, variance of error is constant. (Homoskedastic)

- **Homoskedasticity**
  - Homoskedasticity is the default assumption under which OLS standard errors are usually computed
  - Variance of Y around regression line does not change as X changes
  - Vertical error variance causes uncertainty about line's true slope



- **Heteroskedasticity**
  - Different observations with different error variances are heteroskedastic



  - Fanned-out data means the slope is very uncertain.
  - When data is most certain at endpoints, plot is more accurate; endpoints anchor slope.
  - **Leverage**: Data points nearer ends of regression line influence slope more.
  - If data is heteroskedastic or variance is uncertain, use **robust standard error** (instead of OLS standard error).

- **Robust Standard Error (RSE)**
  - RSEs give accurate confidence intervals when error variance varies with X
    - Also known as heteroskedasticity-robust standard errors or Eicher-White standard errors
  - No need to know shape of heteroskedasticity or which X variables correlate with variance
  - Very big/very small X values have more leverage on slope coefficient β
  - When estimating variance of β
    - Take weighted sum of squared residuals (i.e., squared vertical deviations from regression line)
      - Weights in weighted sum correspond to leverage of each observation
    - Divide by total variance in X i.e squared horizontal deviations from mean

$$\text{RSE} = \frac{\sum \varepsilon_i{}^2}{\frac{(Xi-\mu)^2}{2}}$$

- **Clustered Standard Errors** (CSE)
    - Precision of estimates in OLS assumes independence across observations.
    - If treatment assignment is clustered, use CSE to avoid unintentionally overstating precision
        - If everyone in a group gets same treatment (clustering), and if potential outcomes are similar within group but differ between groups, our information on treatment effect is less meaningful.
        - CSEs account for lack of independence, as RSEs account for heteroskedasticity
    - **Example**: Tennessee STAR Experiment: Randomized at classroom level
        - Each classroom's students had same teachers, similar backgrounds/experiences
        - More similarity within each classroom than between classrooms
        - Changing 20 similar students from control group to treatment group moves potential Y for all 20 at once
        - Result: more variance than if 20 randomly chosen people had been moved
        - In STAR experiment, CSEs are about three times bigger than OLS standard errors.

Multifactor Experiments (factorial design)

- These are experiments that have more than one treatment at a time.
- **Factorial Design**
    - A factorial design allocates subjects at random to every combination of experimental conditions.
    - So, if an experiment has factor 1 consisting of conditions A and B and another factor consisting of conditions C, D and E;
        - A factorial design would assign subjects to treatments {AC, AD, AE, BC, BD, BE}.
        - This would be a 2x3 design.
    - Factorial designs allow researchers to study treatment-by-treatment interactions.
    - Designs that omit some combinations are termed “fractional” factorial designs.

- **Treatment-by-treatment interaction**: Running regressions with interaction terms to tell us how much more one treatment matters when the other treatment is turned on.

- **Regression Analysis**
  - **Example**: An experiment to find out how state legislators would respond to request for help from a constituent seeking to find out about immigration rules.
  - **Factors**:
    - **Ethnicity**: Names Jose and Colin were used.
    - **Grammar**: Good grammar and bad grammar.
  - **Outcome**: Whether a response was received or not.
  - This experiment used a factorial design. The results are as depicted below:

	Colin	Jose
Good grammar	52%	37%
Bad grammar	29%	34%

  - The regression model can be analyzed in the following equivalent ways:
    - **Model 1**: Dummy variable for each factor:

$$Y_i = b_1 L_{\text{i Good Grammar}}^{\text{Non-Hispanic}} + b_2 L_{\text{i Good Grammar}}^{\text{Hispanic}} + b_3 L_{\text{i Bad Grammar}}^{\text{Non-Hispanic}} + b_4 L_{\text{i Bad Grammar}}^{\text{Hispanic}} + u_i$$

      - Each coefficient represents 1 cell in the above table.

    - **Model 2**: Using treatment interactions

$$Y_i = a + bJ_i + cG_i + d(J_i G_i) + u_i$$

      - Race of letter writer identified by J-indicator
        - J=1 for Jose
        - J=0 for Colin
      - Grammar identified by G-indicator
        - G=1 for bad grammar
        - G=0 for good grammar

- We get the following coefficients:

$$\widehat{Y}_i = 0.52 + (-0.15)J_i + (-0.23)G_i + (0.20)(J_iG_i)$$

  - The intercept a measures' the average response in the omitted category (Colin, no grammar errors).
    - As seen this matches the 52% in the table.
  - The coefficient b tells us how much the ethnicity matters, for the case where there are no grammar errors (G=0).
    - From that table, that would be, difference between J=1 and J=0 when G=0
    - That is 37-52 = -15
  - The coefficient c tells us how much the grammar errors matter, for the case of Colin (J=0).
    - From that table, that would be, difference between G=1 and G=0 when J=0
    - That is 29-52 = -23
  - The coefficient d (interaction coefficient) tells us how much more the grammar errors matter for Jose than for Colin.
    - Difference in differences in the table
    - **Option 1**: How much more does bad grammar (G=1) matter with J=1 vs. J=0?
      - Difference for J=1: $.34 - .37 = -.03$
      - Difference for J=0: $.29 - .52 = -.23$
      - DID = (- 0.03) - (- 0.23) = 0.20
    - **Option 2**: How much more does the Jose treatment (J=1) matter with G=1 vs. G=0?
      - Difference for G=1: $.34 - .29 = 0.05$
      - Difference for G=0: $.37 - .52 = -0.15$
      - DID = (0.05) - (-0.15) = 0.20

- **Benefits of using Regression analysis**
  - Regression output automatically includes standard errors, for easy hypothesis testing.
    - One can perform a t-test on the regression coefficient (coeff value divided by the standard error).
    - Easily compute confidence interval (coeff value +\- 2 times the standard error)
  - Regression also gracefully handles non-binary categorical variables.
  - Regression with dummy variables quickly summarizes treatment effects and interaction effects

6 Heterogeneous Treatment Effects (HTE)

- Heterogeneous treatment effects (HTEs):
  - same treatment having different effects on different subjects.
- Use of regression to measure HTEs.
  - Interaction between covariates of interest and treatment variable is important.

- **Estimating HTEs**: Example estimating student scores by teacher incentive with parent literacy as a factor.
  - **Approach 1**: Two-Sample Test (for 2 groups)
    - Split data into two separate samples.
    - Compute means and standard errors of the treatment effect in each group.
    - Do a two-sample test of means.
      - Group one: students whose parents have above-median literacy
      - Group two: students whose parents have below-median literacy

  - **Approach 2: Regression**

$$Y_i = \beta_0 + \beta_1 \cdot I_i + \beta_2 \cdot P_i + \beta_3 \cdot I_i \cdot P_i$$

    - Run a regression test with:
      - Dummy variable for treatment (I)
      - Dummy variable for covariate (P)
        - P stands for parents.
      - Interaction term (I·P)
    - Test if the interaction coefficient is zero.
      - If $\beta_3 \neq 0$, treatment impact varies with different levels of covariate.

  - With regression, we have two approaches:
    - **Treatment-covariate interactions**
      - Here, covariates partition subjects into subgroups
      - The ATE within a subgroup is called the **conditional average treatment effect (CATE)**.
      - Here, we should be careful about the multiple comparisons problem.
    - **Treatment-treatment interactions**
      - This is the factorial design discussed earlier.
      - Downside\risk to this approach is low.
      - Estimates could be imprecise if there is **non-compliance**.

        - A short regression model allows for heterogenous treatment effects but makes no attempt to explain why it arises.
            - By introducing treatment-covariate interaction or treatment-treatment interaction, the long model removes ***systematic heterogeneity*** from the pool of heterogeneity calling what remains as ***idiosyncratic heterogeneity***.
            - **Systematic heterogeneity**:
                - This is the variation in treatment that can be explained by observed covariates or other experimental factors.
            - **Idiosyncratic heterogeneity**:
                - This is the residual variation in treatment that cannot be explained by observed covariates or other experimental factors

- **Collinearity**
    - Collinearity should never be a problem between treatment variable and covariate.
    - Randomization guarantees independence between them.

- **Fishing Expedition (Multiple comparison problem)**
    - Trying out multiple hypotheses before picking a favorite.
        - More variables means more specification searching is possible.
        - Specifications can be changed until the coefficients are suitable.
        - All possible covariates can be tried until statistical significance is found.
        - **Example**: Testing each jellybean color and concluding green jellybean causes acne.
    - Violates assumptions that give valid confidence intervals.
    - Analyzing data in different ways is OK.
        - But the computed confidence intervals will be too narrow.
- **Solutions to fishing expeditions**
    - **Bonferroni Correction**:
        - helps avoid overstating statistical significance.
        - Critical values should be much higher (t=3 instead of t=2).
            - That is, 95% CI should be 3 times standard error instead of 2 times the standard error.
    - Consider findings to be interesting hypotheses.
        - Test them in another experiment to ensure replication.
- **Interpreting HTEs**
    - They explain how different subgroups respond to treatment.
    - They don't explain causal effects of reassigning people to new subgroups.

7 Non-Compliance

- Example of one-sided noncompliance
  - 200 subjects: 100 control, 100 treatment.
    - A week later, measure blood pressure for everyone.
    - 50 treatment subjects complied, 50 did not.
    - Mean BP of control group: 140.
    - Mean BP of compliers in treatment group: 120.
    - Mean BP of never-takers in treatment group: 180.
  - Our random assignment partitioned people only into treatment and control.
  - What comparison to make?
    - 50 treatment units who took pill and 50 treatment units who did not take pill are not comparable.
    - 100 control units not given pill and 50 treatment units who took pill are not comparable.
    - Ideal comparison: control units who would have taken the pill if given it.
    - Problem: can't know who would have taken pill if given it.
  - Assumptions for inference
    - Since assignments were randomized, the proportion of never takers in treatment and control are the same. So, assume 50 never takers in control too.
    - If a person does not comply, there is no treatment effect.
      - So assume same mean BP for never takers in control too.
    - Depicted below:

  - Computing treatment effect
    - First compute mean BP of control compliers (say X)
      - (180 + X)/2 = 140 ➔ X = 100
    - No comparing only the compliers we get
      - Treatment effect = 120-100 = 20

- From the last example we get the following:
  - Average Treatment effect (ATE) is really the **Intent to Treat** (ITT) effect
    - This is the overall treatment effect between both groups
    - ATE = ITT = 10
  - **Complier average causal effect** (CACE)
    - This is the actual treatment effect on the compliers.
    - Effect of treatment estimated for the kinds of units that take up the treatment.
    - This is what we really need to measure i.e. given a treatment what's its effect on people who ***received*** it.
    - CACE = 10

- **Formula for one-sided compliance**
  - Let α be the share of subjects who actually received treatment. (compliance rate)
  - Let ATE be the overall treatment effect = mean Treatment – mean control
  - Then,

    CACE = ATE / α

- **Exclusion Restriction**
  - Exclusion restriction assumes that potential outcomes respond only to treatments and not treatment assignments.
  - This assumption is important while computing CACE.
  - If this assumption is violated, we will need to design a new experiment.
  - Example
    - In the blood pressure example, we assumed there is no effect of being assigned to treatment or control.
    - Suppose being assigned to the control group generates anxiety. This ruins the experimental noncomparability.

- **Placebo Design**
  - Placebo design can tell us who are the control group compliers.
  - Can then compare the compliers in treatment and control directly.
  - This gives us **more statistical power** when computing CACE.
    - Here even though sample size reduces, effect is proportionally more
    - This increases precision.

- **Downstream Experiments**
    - Conceptualize a first experiment or its effects as an encouragement itself for a second analysis.
        - Example: What is the effect of smoking a first cigarette on the likelihood of being a smoker long term?
            - First experiment:
                - Create an incentive to smoke for some and not for others.
                - Who smokes that week?
            - Downstream experiment
                - Now for these compliers, the downstream experiment is to study long term effect of first cigarette.
            - We're finding the long-term effect on those who complied in the original study.
- **Attenuation Bias**
    - Treatment sometimes misapplied or mismeasured.
    - Common cause:
        - administrative or implementation errors.
        - Accidentally treating control group members
        - Accidentally categorizing treatment members as control members
    - Can lead to bias.
    - Biases estimate toward zero, unless treatment can cause mismeasurement to be more/less likely.
    - Can be conceptualized as noncompliance if application rate or error rate is known.

8 Spillovers

- Spillover is the effects of one person's treatment on the outcome of another person, regardless of whether the second person was treated.
    - ***Example***:
        - Person A in ad campaign treatment group discusses the ad with person B in the control group.
        - When the interaction with A causes B to make a purchase, she would not have made otherwise, this is spillover.
- With **positive** spillovers, we would expect a simple randomized experiment to **underestimate** the treatment effect.
    - Example: Word-of-mouth effect might increase purchases in control group reducing the treatment effect.

How to avoid spillover effects

- **Clustered Design**
    - Randomize at a group level instead of individuals
        - Example: Use geographic distance between groups to minimize chances of cross-group spillover
    - Advantages:
        - Reduced word-of-mouth spillovers
        - Easier administration
    - Disadvantages:
        - Less randomization
        - Less precision in estimated treatment effect
            - Due to possible correlation of outcomes within treatment.
            - Perhaps rain is correlated with certain employee or donor behaviors.
            - Clustered standard errors correctly estimate this uncertainty; regular standard errors are underestimated.

Spatial Spillovers

-
- Arbitrary assumptions about geographic extent of spillovers can radically change estimates.
    - What if range of spillover is 750 rather than 500 meters?
    - What if spillover effects are not constant out to some threshold distance?
        - Linear decline in treatment effect
        - Quadratic decline in treatment effect
- Contrast with usual virtue of experiments: Make measurements with minimal assumptions.
    - Assumption problems not unique to geographic spillovers.
    - Example: spillovers via social networks (Facebook, Twitter)

9 Problems and Diagnostics

End of Document