STA305/1004 - Week 4

January 27-30, 2020

Week 4 Outline

- Replication and Power: Case study on Power poses
- Power and Sample size formulae: Two-sample proportions
- Power via simulation
- Introduction to causal inference:

– The fundamental problem

– The assignment mechanism: Weight gain study

- SUTVA: Electric company example
- Observational studies

Replication and Power: Case Study on Power poses:

Professor Amy Cuddy

Can power poses significantly change outcomes in your life?

Cuddy's study methods:

- ▶ Randomly assigned 42 participants to the high-power pose or the low-power-pose condition.
- ▶ Participants believed that the study was about the science of physiological recordings and was focused on how placement of electrocardiography electrodes above and below the heart could influence data collection.
- ▶ Participants' bodies were posed by an experimenter into high-power or low-power poses. Each participant held two poses for 1 min each.
- ▶ Participants' risk taking was measured with a gambling task; feelings of power were measured with self-reports.
- ▶ Saliva samples, which were used to test cortisol and testosterone levels, were taken before and approximately 17 min after the power-pose manipulation.

(Carney, Cuddy, Yap, 2010)

Can power poses significantly change outcomes in your life?

Cuddy's study results:

As hypothesized, high-power poses caused an increase in testosterone compared with low-power poses, which caused a decrease in testosterone, $F(1, 39) = 4.29$, $p < .05$; $r = .34$. Also as hypothesized, high-power poses caused a decrease in cortisol compared with low-power poses, which caused an increase in cortisol, $F(1, 38) = 7.45$, $p < .02$; $r = .43$

Can power poses significantly change outcomes in your life?

- ▶ The study was replicated by Ranehill et al. (2015)
- ▶ An initial power analysis based on the effect sizes in Carney et al. (power $=$ 0.8, $\alpha = .05$) indicated that a sample size of 100 participants would be suitable.

```
library(pwr)
pwr.t.test(d=0.6,power = 0.8)
```

```
     Two-sample t test power calculation

              n = 44.58577
              d = 0.6
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Can power poses significantly change outcomes in your life?

- Ranehill et al. study used a sample of 200 participants to increase reliability.
- This study found none of the significant differences found in Cuddy's study.
- The replication study obtained very precise estimates of the effects.
- What happened?

Can power poses significantly change outcomes in your life?

- Sampling theory predicts that the variation between samples is proportional to $\frac{1}{\sqrt{n}}$.
- In small samples we can expect variability.
- Many researchers often expect that these samples will be more similar than sampling theory predicts.

Suppose that you have run an experiment on 20 subjects, and have obtained a significant result from a two-sided z-test ($H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$) which confirms your theory ($z = 2.23$, $p < 0.05$, two-tailed). The researcher is planning to run the same experiment on an additional 10 subjects. What is the probability that the results will be significant at the 5% level by a one-tailed test ($H_1 : \mu > 0$), separately for this group?

- Power and Sample size formulae: Two-sample proportions

Comparing Proportions for Binary Outcomes

- In many clinical trials, the primary endpoint is dichotomous, for example, whether a patient has responded to the treatment, or whether a patient has experienced toxicity.
- Consider a two-arm randomized trial with binary outcomes. Let p_1 denote the response rate of the experimental drug, p_2 as that of the standard drug, and the difference is $\theta = p_1 - p_2$.

Comparing Proportions for Binary Outcomes

Let Y_{ik} be the binary outcome for subject i in arm k; that is,

$$Y_{ik} = \begin{cases} 1 & \text{with probability } p_k \\ 0 & \text{with probability } 1 - p_k, \end{cases}$$

for $i = 1, ..., n_k$ and $k = 1, 2$. The sum of independent and identically distributed Bernoulli random variables has a binomial distribution,

$$\sum_{i=1}^{n_k} Y_{ik} \sim Bin(n_k, p_k), \ k = 1, 2.$$

(Yin, pg. 173-174)

Comparing Proportions for Binary Outcomes

The sample proportion for group k is

$$\hat{p}_k = \bar{Y}_k = (1/n_k) \sum_{i=1}^{n_k} Y_{ik}, \ k = 1, 2,$$

and $E\left(\bar{Y}_k\right) = p_k$ and $Var\left(\bar{Y}_k\right) = \frac{p_k(1-p_k)}{n_k}$.

The goal of the clinical trial is to determine if there is a difference between the two groups using a binary endpoint. That is, we want to test $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$.

The test statistic (assuming that H_0 is true) is:

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim N(0,1),$$

Comparing Proportions for Binary Outcomes

The test rejects at level α if and only if

$$|T| \geq z_{\alpha/2}.$$

Using the same argument as the case with continuous endpoints and ignoring terms smaller than $\alpha/2$ we can solve for β

$$\beta \approx \Phi\left(z_{\alpha/2} - \frac{|\theta_1|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right).$$

Comparing Proportions for Binary Outcomes

Using this formula to solve for sample size. If $n_1 = r \cdot n_2$ then

$$n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2}{\theta^2} \left(p_1(1 - p_1)/r + p_2(1 - p_2)\right).$$

Comparing Proportions for Binary Outcomes

▶ The built-in R function power.prop.test() can be used to calculate sample size or power.

▶ For example suppose that the standard treatment for a disease has a response rate of 20%, and an experimental treatment is anticipated to have a response rate of 28%.

▶ The researchers want both arms to have an equal number of subjects. How many patients should be enrolled if the study will conduct a two-sided test at the 5% level with 80% power?

```
power.prop.test(p1 = 0.2,p2 = 0.25,power = 0.8)
```

```
    Two-sample comparison of proportions power calculation

              n = 1093.739
             p1 = 0.2
             p2 = 0.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

- Power via simulation

- If the test statistic and distribution of the test statistic are known then the power of the test can be calculated via simulation.
- Consider a two-sample t-test with 30 subjects per group and the standard deviation of the clinical outcome is known to be 1.
- What is the power of the test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_0 : \mu_1 - \mu_2 = 0.5$, at the 5% significance level?
- The power is the proportion of times that the test correctly rejects the null hypothesis in repeated sampling.

Calculating Power by Simulation

We can simulate a single study using the rnorm() command. Let's assume that $n_1 = n_2 = 30, \mu_1 = 3.5, \mu_2 = 3, \sigma = 1, \alpha = 0.05$.

```
set.seed(2301)
t.test(rnorm(30,mean=3.5,sd=1),rnorm(30,mean=3,sd=1),var.equal = T)
```

```
    Two Sample t-test

data:  rnorm(30, mean = 3.5, sd = 1) and rnorm(30, mean = 3, sd = 1)
t = 2.1462, df = 58, p-value = 0.03605
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03458122 0.99248595
sample estimates:
mean of x mean of y
 3.339362  2.825828
```

Should you reject H_0?

Calculating Power by Simulation

- ▶ Suppose that 10 studies are simulated.
- ▶ What proportion of these 10 studies will reject the null hypothesis at the 5% level?
- ▶ To investigate how many times the two-sample t-test will reject at the 5% level the `replicate()` command will be used to generate 10 studies and calculate the p-value in each study.
- ▶ It will still be assumed that
  $n_1 = n_2 = 30, \mu_1 = 3.5, \mu_2 = 3, \sigma = 1, \alpha = 0.05$.

```
set.seed(2301)
pvals <- replicate(10,t.test(rnorm(30,mean=3.5,sd=1),
                              rnorm(30,mean=3,sd=1),
                              var.equal = T)$p.value)
pvals # print out 10 p-values
```

```
 [1] 0.03604893 0.15477655 0.01777959 0.40851999 0.34580930 0.11131007
 [7] 0.14788381 0.00317709 0.09452230 0.39173723
```

```
#power is the number of times the test rejects at the 5% level
sum(pvals<=0.05)/10
```

```
[1] 0.3
```

Calculating Power by Simulation

But, since we only simulated 10 studies the estimate of power will have a large standard error. So let's try simulating 10,000 studies so that we can obtain a more precise estimate of power.

```
set.seed(2301)
pvals <- replicate(10000,t.test(rnorm(30,mean=3.5,sd=1),
                                rnorm(30,mean=3,sd=1),
                           var.equal = T)$p.value)
sum(pvals<=0.05)/10000
```

```
[1] 0.4881
```

Calculating Power by Simulation

This is much closer to the theoretical power obtained from power.t.test().

```
power.t.test(n = 30,delta = 0.5,sd = 1,sig.level = 0.05)
```

```
     Two-sample t test power calculation

              n = 30
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.477841
    alternative = two.sided

NOTE: n is number in *each* group
```

- The built-in R functions `power.t.test()` and `power.prop.test()` don't have an option for calculating power where the there is unequal allocation of subjects between groups.
- These built-in functions don't have an option to investigate power if other assumptions don't hold (e.g., normality).
- One option is to simulate power for the scenarios that are of interest. Another option is to write your own function using the formula derived above.

Calculating Power by Simulation

- Suppose the standard treatment for a disease has a response rate of 20%, and an experimental treatment is anticipated to have a response rate of 28%.
- The researchers want both arms to have an equal number of subjects.
- A power calculation above revealed that the study will require 1094 patients for 80% power.
- What would happen to the power if the researchers put 1500 patients in the experimental arm and 500 patients in the control arm?

Calculating Power by Simulation

- The number of subjects in the experimental arm that have a positive response to treatment will be an observation from a $Bin(1500, 0.28)$.
- The number of subjects that have a positive response to the standard treatment will be an observation from a $Bin(500, 0.2)$.
- We can obtain simulated responses from these distributions using the `rbinom()` command in R.

```
set.seed(2301)
rbinom(1,1500,0.28)
```

```
[1] 403
```

```
rbinom(1,500,0.20)
```

```
[1] 89
```

Calculating Power by Simulation

▶ The p-value for this simulated study can be obtained using prop.test().

```
set.seed(2301)
prop.test(x=c(rbinom(1,1500,0.28),rbinom(1,500,0.20)),
          n=c(1500,500),correct = F)
```

```
    2-sample test for equality of proportions without continuity
    correction

data:  c(rbinom(1, 1500, 0.28), rbinom(1, 500, 0.2)) out of c(1500, 500
X-squared = 16.62, df = 1, p-value = 4.568e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.05032654 0.13100680
sample estimates:
   prop 1    prop 2
0.2686667 0.1780000
```

Calculating Power by Simulation

- A power simulation repeats this process a large number of times.
- In the example below we simulate 10,000 hypothetical studies to calculate power.

```
set.seed(2301)
pvals <- replicate(10000,
prop.test(x=c(rbinom(n = 1,size = 1500,prob = 0.25),
              rbinom(n=1,size=500,prob=0.20)),
              n=c(1500,500),correct=F)$p.value)
sum(pvals<=0.05)/10000
```

[1] 0.6231

If the researchers decide to have a 3:1 allocation ratio of patients in the treatment to control arm then the power will be _____?

- Introduction to Causal Inference

- Suppose Bob, at a particular point in time, is contemplating whether or not to take an aspirin for a headache.
- There are two treatment levels, taking an aspirin, and not taking an aspirin.
- If Bob takes the aspirin, his headache may be gone, or it may remain, say, an hour later; we denote this outcome, which can be either "Headache" or "No Headache," by Y(Aspirin).
- Similarly, if Bob does not take the aspirin, his headache may remain an hour later, or it may not; we denote this potential outcome by Y(No Aspirin), which also can be either "Headache," or "No Headache."
- There are therefore two potential outcomes, Y(Aspirin) and Y(No Aspirin), one for each level of the treatment. The causal effect of the treatment involves the comparison of these two potential outcomes.

Because in this example each potential outcome can take on only two values, the unit- level causal effect – the comparison of these two outcomes for the same unit – involves one of four (two by two) possibilities:

1. Headache gone only with aspirin: Y(Aspirin) = No Headache, Y(No Aspirin) = Headache
2. No effect of aspirin, with a headache in both cases: Y(Aspirin) = Headache, Y(No Aspirin) = Headache
3. No effect of aspirin, with the headache gone in both cases: Y(Aspirin) = No Headache, Y(No Aspirin) = No Headache
4. Headache gone only without aspirin: Y(Aspirin) = Headache, Y(No Aspirin) = No Headache

There are two important aspects of this definition of a causal effect.

1. The definition of the causal effect depends on the potential outcomes, but it does not depend on which outcome is actually observed.
2. The causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment.

- The causal effect is not defined in terms of comparisons of outcomes at different times, as in a before-and-after comparison of my headache before and after deciding to take or not to take the aspirin.

The fundamental problem of causal inference

"The fundamental problem of causal inference" (Holland, 1986, p. 947) is the problem that at most one of the potential outcomes can be realized and thus observed.

- If the action you take is Aspirin, you observe Y(Aspirin) and will never know the value of Y(No Aspirin) because you cannot go back in time.
- Similarly, if your action is No Aspirin, you observe Y(No Aspirin) but cannot know the value of Y(Aspirin).
- In general, therefore, even though the unit-level causal effect (the comparison of the two potential outcomes) may be well defined, by definition we cannot learn its value from just the single realized potential outcome.

The fundamental problem of causal inference

The outcomes that would be observed under control and treatment conditions are often called **counterfactuals** or **potential outcomes**.

- If Bob took aspirin for his headache then he would be assigned to the treatment condition so $T_i = 1$.
- Then $Y(\text{Aspirin})$ is observed and $Y(\text{No Aspirin})$ is the unobserved counterfactual outcome—it represents what would have happened to Bob if he had not taken aspirin.
- Conversely, if Bob had not taken aspirin then $Y(\text{No Aspirin})$ is observed and $Y(\text{Aspirin})$ is counterfactual.
- In either case, a simple treatment effect for Bob can be defined as

$$\text{treatment effect for Bob } = Y(\text{Aspirin}) - Y(\text{No Aspirin}).$$

- The problem is that we can only observe one outcome.

The assignment mechanism

- **Assignment mechanism:** The process for deciding which units receive treatment and which receive control.
- **Ignorable Assignment Mechanism:** The assignment of treatment or control for all units is independent of the unobserved potential outcomes ("nonignorable" means not ignorable)
- **Unconfounded Assignment Mechanism:** The assignment of treatment or control for all units is independent of all potential outcomes, observed or unobserved ("confounded" means not unconfounded)

The assignment mechanism

- Suppose that a doctor prescribes surgery (labeled 1) or drug (labeled 0) for a certain condition.
- The doctor knows enough about the potential outcomes of the patients so assigns each patient the treatment that is more beneficial to that patient.

unit	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
patient #1	1	7	6
patient #2	6	5	-1
patient #3	1	5	4
patient #4	8	7	-1
Average	4	6	2

Y is years of post-treatment survival.

The assignment mechanism

- Patients 1 and 3 will receive surgery and patients 2 and 4 will receive drug treatment.
- The observed treatments and outcomes are in this table.

unit	T_i	Y_i^{obs}	$Y_i(1)$	$Y_i(0)$
patient #1	1	7		
patient #2	0	6		
patient #3	1	5		
patient #4	0	8		
Average Drug		**7**		
Average Surg		**6**		

- This shows that we can reach invalid conclusions if we look at the observed values of potential outcomes without considering how the treatments were assigned.
- The assignment mechanism depended on the potential outcomes and was therefore nonignorable (implying that it was confounded).

The assignment mechanism

The observed difference in means is entirely misleading in this situation. The biggest problem when using the difference of sample means here is that we have effectively pretended that we had an unconfounded treatment assignment when in fact we did not. This example demonstrates the importance of finding a statistic that is appropriate for the actual assignment mechanism.

The assignment mechanism

Is the treatment assignment ignorable?

- The doctor knows enough about the potential outcomes of the patients so assigns each patient the treatment that is more beneficial to that patient.
- Suppose that a doctor prescribes surgery (labeled 1) or drug (labeled 0) for a certain condition by tossing a biased coin that depends on $Y(0)$ and $Y(1)$, where Y is years of post-treatment survival.
- If $Y(1) \geq Y(0)$ then $P(T_i = 1 | Y_i(0), Y_i(1)) = 0.8$.
- If $Y(1) < Y(0)$ then $P(T_i = 1 | Y_i(0), Y_i(1)) = 0.3$.

unit	$Y_i(0)$	$Y_i(1)$	p_1	p_0
patient #1	1	7	0.8	0.2
patient #2	6	5	0.3	0.7
patient #3	1	5	0.8	0.2
patient #4	8	7	0.3	0.7

where, $p_1 = P(T_i = 1 | Y_i(0), Y_i(1))$, and $p_0 = P(T_i = 0 | Y_i(0), Y_i(1))$.

Weight gain study

From Holland and Rubin (1983).

"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his [or her] arrival in September and his [or her] weight the following June are recorded."

- The average weight for Males was 180 in both September and June. Thus, the average weight gain for Males was zero.
- The average weight for Females was 130 in both September and June. Thus, the average weight gain for Females was zero.
- Question: What is the differential causal effect of the diet on male weights and on female weights?
- Statistician 1: Look at gain scores: No effect of diet on weight for either males or females, and no evidence of differential effect of the two sexes, because no group shows any systematic change.
- Statistician 2: Compare June weight for males and females with the same weight in September: On average, for a given September weight, men weigh more in June than women. Thus, the new diet leads to more weight gain for men.
- Is Statistician 1 correct? Statistician 2? Neither? Both?

Weight gain study

Questions:

1. What are the units?
2. What are the treatments?
3. What is the assignment mechanism?
4. Is the assignment mechanism useful for causal inference?
5. Would it have helped if all males received the dining hall diet and all females received the control diet?
6. Is Statistician 1 or Statistician 2 correct?

Getting around the fundamental problem by using close substitutes

- Are there situations where you can measure both Y_i^0 and Y_i^1 on the same unit?
- Drink tea one night and milk another night then measure the amount of sleep. What has been assumed?
- Divide a piece of plastic into two parts then expose each piece to a corrosive chemical. What has been assumed?
- Measure the effect of a new diet by comparing your weight before the diet and your weight after. What has been assumed?
- There are strong assumptions implicit in these types of strategies.

- The "statistical" idea of using the outcomes observed on a sample of units to learn about the distribution of outcomes in the population.
- The basic idea is that since we cannot compare treatment and control outcomes for the same units, we try to compare them on similar units.
- Similarity can be attained by using randomization to decide which units are assigned to the treatment group and which units are assigned to the control group.

- It is not always possible to achieve close similarity between the treated and control groups in a causal study.
- In observational studies, units often end up treated or not based on characteristics that are predictive of the outcome of interest (for example, men enter a job training program because they have low earnings and future earnings is the outcome of interest).
- **Randomized experiments can be impractical or unethical.**
- When treatment and control groups are not similar, modeling or other forms of statistical adjustment can be used to fill in the gap.

Fisherian Randomization Test

- The randomization test is related to a stochastic proof by contradiction giving the plausability of the null hypothesis of no treatment effect.
- The null hypothesis is $Y_i^0 = Y_i^1$, for all units.
- Under the null hypothesis all potential outcomes are known from Y^{obs} since $Y^{obs} = Y^1 = Y^0$.
- Under the null hypothesis, the observed value of any statistic, such as $\bar{y}^1 - \bar{y}^0$ is known for all possible treatment assignments.
- The randomization distribution of $\bar{y}^1 - \bar{y}^0$ can then be obtained.
- Unless the data suggest that the null hypothesis of no treatment effect is false then it's difficult to claim evidence that the treatments are different.

- Stable Unit Treatment Value Assumption (SUTVA)

To learn about causal effects through multiple experimental units assumptions are required.

The assumption. The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

Stable Unit Treatment Value Assumption (SUTVA)

- Involves assuming that treatments applied to one unit do not affect the outcome for another unit.
- For example, if Adam and Oliver are in different locations and have no contact with each other, it would appear reasonable to assume that if Oliver takes an aspirin for his headache then his behaviour has no effect on the status of Adam's headache.
- This assumption might not hold if Adam and Oliver are in the same location, and Adam's behavior, affects Oliver's behaviour.
- SUTVA incorporates the idea that Adam and Oliver do not interfere with one another and the idea that for each unit there is only a single version of each treatment level (e.g., there is only one dose of aspirin). (Imbens and Rubin, 2015)

The following example is from Gelman and Hill (2007).

An educational experiment performed on four elementary school classes (grades 1-4) was conducted.

- **Treatment:** Exposure to a new educational television show called The Electric Company.
- **Treatment assignment:** Classes were randomized into treated and control groups.
- **Primary study outcome:** At the end of the school year, students in all the classes were given a reading test, and the average test score within each class was recorded.

A pre-test was given at the beginning of the school year before the treatment. For grade 1 the pre-test was a subset of the longer post-test and in grades 2-4 the pre-test was the same as the post-test.
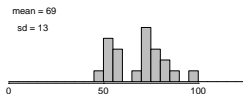
Electric company study

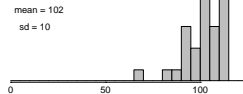The distribution of scores are shown for each treatment and control groups for each grade.
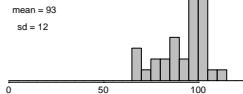

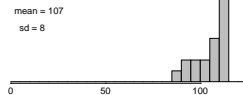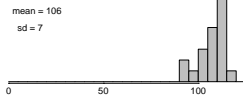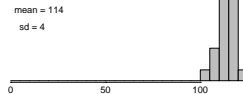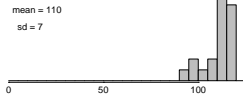**Test scores in control classes**   **Test scores in treated classes**

Grade 1   mean = 69   sd = 13   mean = 77   sd = 16

Grade 2   mean = 93   sd = 12   mean = 102   sd = 10

Grade 3   mean = 106   sd = 7   mean = 107   sd = 8

Grade 4   mean = 110   sd = 7   mean = 114   sd = 4
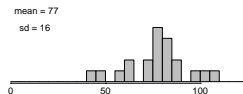
The average causal effect of the treatment is the coefficient in the regression,

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i,$$

for each grade.

Electric company study

```
Grade: 1

              2.5 %  97.5 %
(Intercept) 62.184754 75.3962
treatment   -1.041902 17.6419

Grade: 2

              2.5 %   97.5 %
(Intercept) 89.40453 97.01900
treatment    2.97458 13.74307

Grade: 3

               2.5 %    97.5 %
(Intercept) 102.809071 109.540929
treatment    -4.425143   5.095143

Grade: 4

                 2.5 %     97.5 %
(Intercept) 107.732487230 112.981798
treatment    -0.002299775   7.421347
```

What does the model

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i,$$

assume about pre-test scores?

Omitted variable bias in regression

Suppose that the following model is correct

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i,$$

where x_i is the ith pre-test score.

Assume that the confounding covariate, x_i, is ignored and we fit the model

$$y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i^*.$$

What is the relation between these models?

Electric company study

The pre-test can be controlled by fitting the model:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i,$$

where x_i is the pre-test.

```
attach(electric)
post.test <- c (treated.Posttest, control.Posttest)
pre.test <- c (treated.Pretest, control.Pretest)
grade <- rep (Grade, 2)
treatment <- rep (c(1,0), rep(length(treated.Posttest),2))

for (k in 1:4){
  cat("Grade:",k, "\n")
  print(summary((lm (post.test ~ treatment+pre.test, subset=(grade==k))
  print(confint(lm (post.test ~ treatment+pre.test, subset=(grade==k)))
}
```

Electric company study

```
Grade: 1
                2.5 %      97.5 %
(Intercept) -28.794124   6.748422
treatment     3.503741  14.069296
pre.test      3.996396   6.220489
Grade: 2
                2.5 %      97.5 %
(Intercept) 29.4582320  45.3993009
treatment    1.5517446   6.9797863
pre.test     0.6793567   0.8984661
Grade: 3
                2.5 %      97.5 %
(Intercept) 33.0310748  48.1374132
treatment    0.3370718   3.4823799
pre.test     0.6066683   0.7626586
Grade: 4
                2.5 %      97.5 %
(Intercept) 33.3343230  50.6551288
treatment    0.3151918   3.0876810
pre.test     0.5732627   0.7383963
```

- Observational Studies

- A solution to the fundamental problem of causal inference is, as we have described, to randomly sample a different set of units for each treatment group assignment from a common population, and then apply the appropriate treatments to each group; or
- Randomly assign the treatment conditions among a selected set of units. Both approaches are supposed to guarantee that, on average, the different treatment groups are balanced. In other words, $\bar{y}^0$ and $\bar{y}^1$ from the sample are estimating the average outcomes under control and treatment for the same population.

- In observational studies treatments are observed rather than assigned (e.g., comparison of lung cancer rates in smokers to nonsmokers), and the observed data under different treatments is usually not a random sample from a common population.
- In an observational study there might be systematic differences between groups of units, outside the control of the experimenter, that receive different treatments that can affect the outcome, y.