PS0700
Basic Statistical Methods: Hypothesis Testing and T-Tests

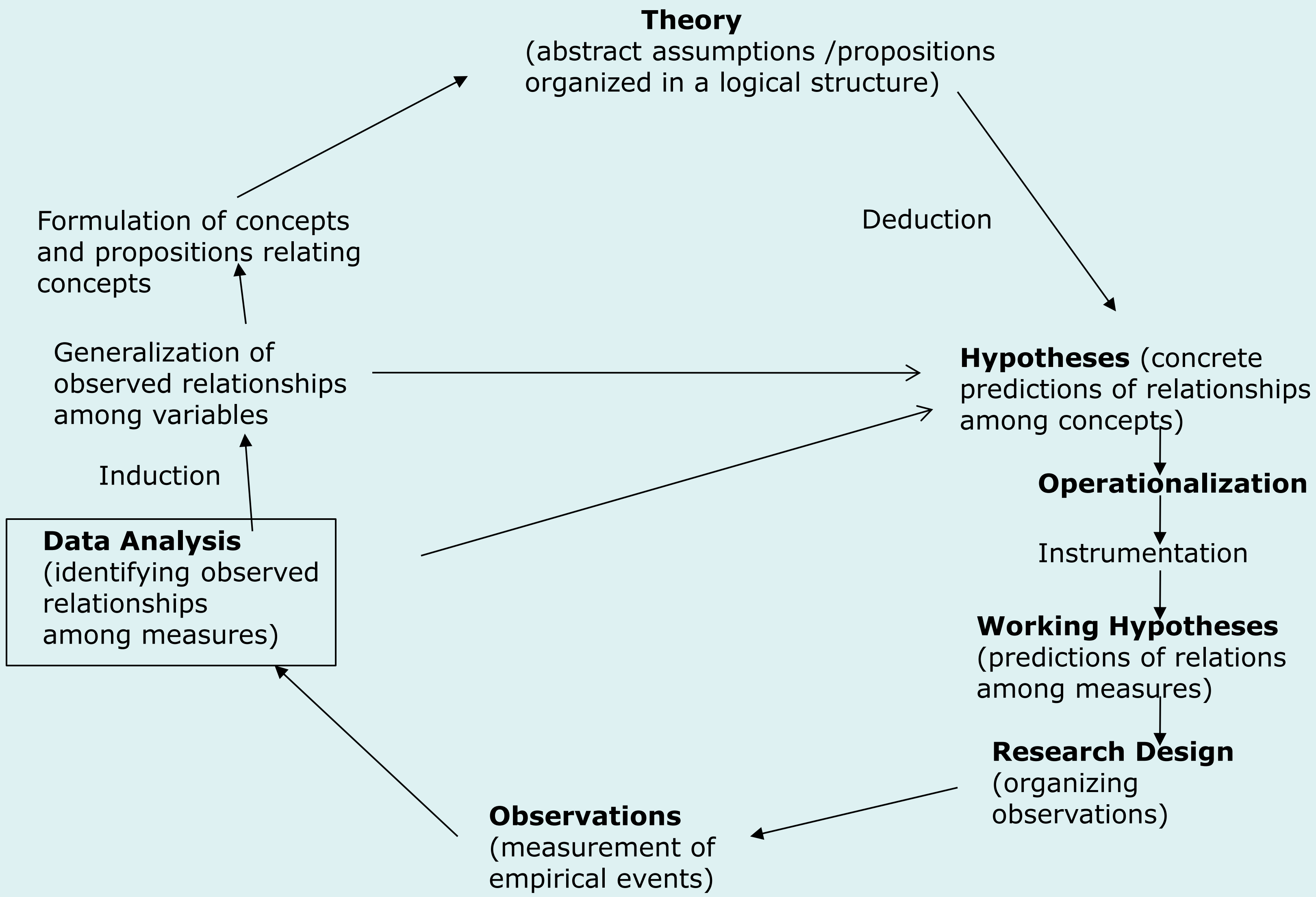
Political Science Research Methods

Professor Steven Finkel

Fall Semester 2022

Week 10-11

A Model of the Research Process



Theory
(abstract assumptions /propositions organized in a logical structure)

Deduction

Hypotheses (concrete predictions of relationships among concepts)

Operationalization

Instrumentation

Working Hypotheses
(predictions of relations among measures)

Research Design
(organizing observations)

Observations
(measurement of empirical events)

Data Analysis
(identifying observed relationships among measures)

Induction

Generalization of observed relationships among variables

Formulation of concepts and propositions relating concepts

Testing Statistical Hypotheses

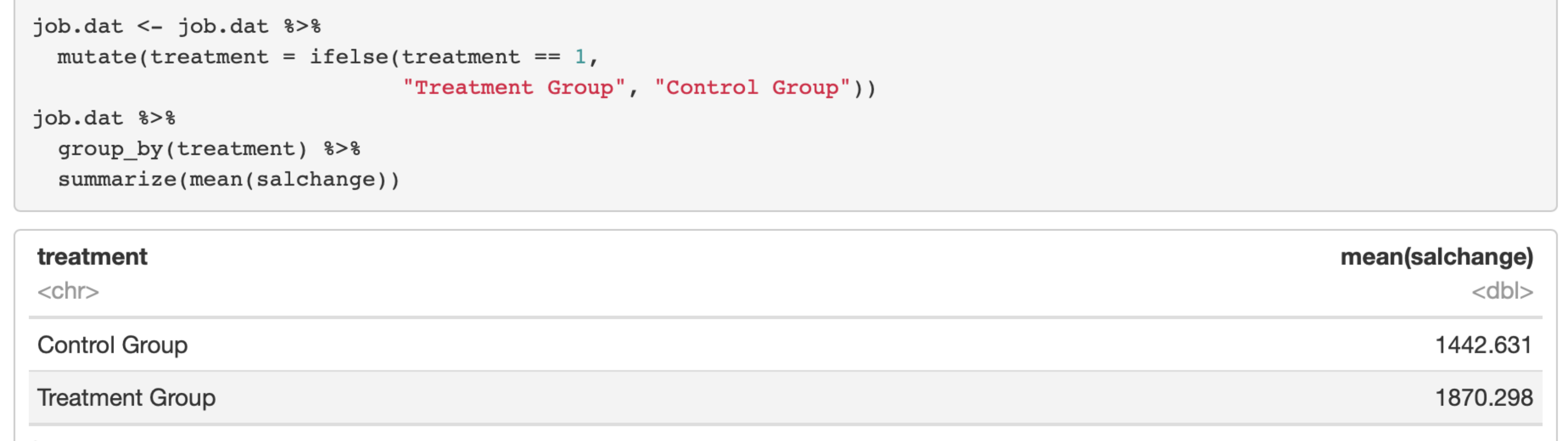
- Are the relationships between variables that we see in our *sample* indicative of a "true" relationship between those variables in the overall *population*?

- We can never be *absolutely sure* that they are (because we don't observe the whole population), but the set of procedures outlined here will allow us to draw conclusions with a known degree of uncertainty, *or probability of being wrong*. So we will make statements like ***"there is a statistically significant relationship between education and income at the .05 level"***, which will indicate (informally) that there is a 5% chance of error.

- These procedures draw on the ideas of sampling distributions and confidence intervals we have considered so far, along with a dose of the philosophy of social science from earlier in the course!

Example: Does Participation in a Job Training Program Increase a Person's Salary?

Output in our sample: We see a $428 "effect" of participating in the program on changes in salary [$1870 – $1442=$428]



```
job.dat <- job.dat %>%
  mutate(treatment = ifelse(treatment == 1,
                             "Treatment Group", "Control Group"))
job.dat %>%
  group_by(treatment) %>%
  summarize(mean(salchange))
```

treatment	mean(salchange)
<chr>	<dbl>
Control Group	1442.631
Treatment Group	1870.298

Questions: Is this effect "statistically significant" in the overall population? Or did it come about in our sample through random chance, or sampling error?

Steps in Hypothesis Testing

- State the hypothesis in statistical language
- Establish criteria for rejecting the hypothesis, i.e. how different must your sample estimates be from “no relationship” for you to reject the hypothesis that the sample came from a population with “no relationship”?
- Compute a “test statistic”--this will tell you, e.g., how many standard errors away from a hypothesized population mean your sample estimate is. Based on this information, you make a:
- Decision--reject or do not reject the hypothesis

Step 1: The "Null" Hypothesis

- The first step in trying to prove that a relationship between variables exists is to construct what is called the "null hypothesis" which is usually the opposite of what we are trying to prove. We then try to "reject the null hypothesis" and therefore show that "not nothing" is going on in our population.
- Our Example:
  - The **Research, or Substantive, or Alternative Hypothesis**: "Participation in Job Training Programs Positively Increases the Individual's Subsequent Salary"
  - The **Null Hypothesis**: "There is no effect of participation in job training programs on the individual's subsequent salary"

- In Statistical Symbols:
  - H_0: μ (treatment) - μ (control) = 0
  - H_A: μ (treatment) - μ (control) ≠ 0

The Null Hypothesis (H_0) says that the difference between the salary change for the treatment group and the salary change for the control group *in the population* is zero; i.e., no relationship between treatment and salary

The Research Hypothesis (H_A) says that there is some difference between the salary change for the two groups *in the population*

Why Do We Test the Null Hypothesis and Not the Research Hypothesis Directly?

- Social science is inherently conservative: we want to be extremely cautious in asserting that variables truly are related, or that there is really a process other than chance operating in a given situation. So we construct a "hypothetical world" where there is no relationship, or no differences, and we make it very hard to reject the hypothesis that our data reflect a different "world“
- The triumph of the “falsificationist” idea in the philosophy of social science. We can never “prove” the research hypothesis is true – there may be other cases, other situations, other variables, etc. that still need to be considered. So we try instead to falsify nulls, saying that we reject hypotheses of “no relationship”. As we do this more and more often, we get more support for the substantive or research hypothesis, but we never “prove” it.

Step 2: Establish Criteria for Rejecting or Not Rejecting the Null Hypothesis

- At what point will we say that our sample estimates are *so far away* from the null hypothesis that they could not have come about by chance?
- We see a $428 salary difference between the treatment and control? Is that 'big enough" of a difference? How should we conceive of what "big enough" means?
- Answer: We usually reject the null if our sample estimates are ***more than 1.96 standard errors away from zero***, or more generally, at the point where our sample estimates ***could have come from a population with "no relationship" 5% of the time or less.***

- Logic: Imagine a population where the null hypothesis was true, i.e., no difference between treatment and control groups, i.e.: μ (treatment) - μ (control) = 0
- **Let's call this the "null hypothesis world"**
- If we took repeated random samples of a certain size of treatment and control groups from the **null hypothesis world** and plotted the difference between salary increases between the groups in each of those random samples, we would have a *sampling distribution of mean differences.* Following the Central Limit Theorem, this sampling distribution would be *normally distributed*, centered around 0 with a standard deviation or standard error that we could calculate, based on the standard deviations in the population of the treatment and control groups

- We then want to know: Where is our sample difference of $428 on this distribution? Is it so far away from 0 that we reject the idea that our samples reflect a population with no differences between treatment and control groups, i.e., can we "reject" the "null hypothesis word"?
- If our difference is more than 1.96 standard errors away from 0, it would mean that the chances of observing a sample difference of $428, if the "true" difference were 0, was less than .05, i.e., it would happen less than 5% of the time.
- If so, we will *reject the null hypothesis* at the .05 level, and we say that there is a "statistically significant effect of treatment on salary at the .05 level"

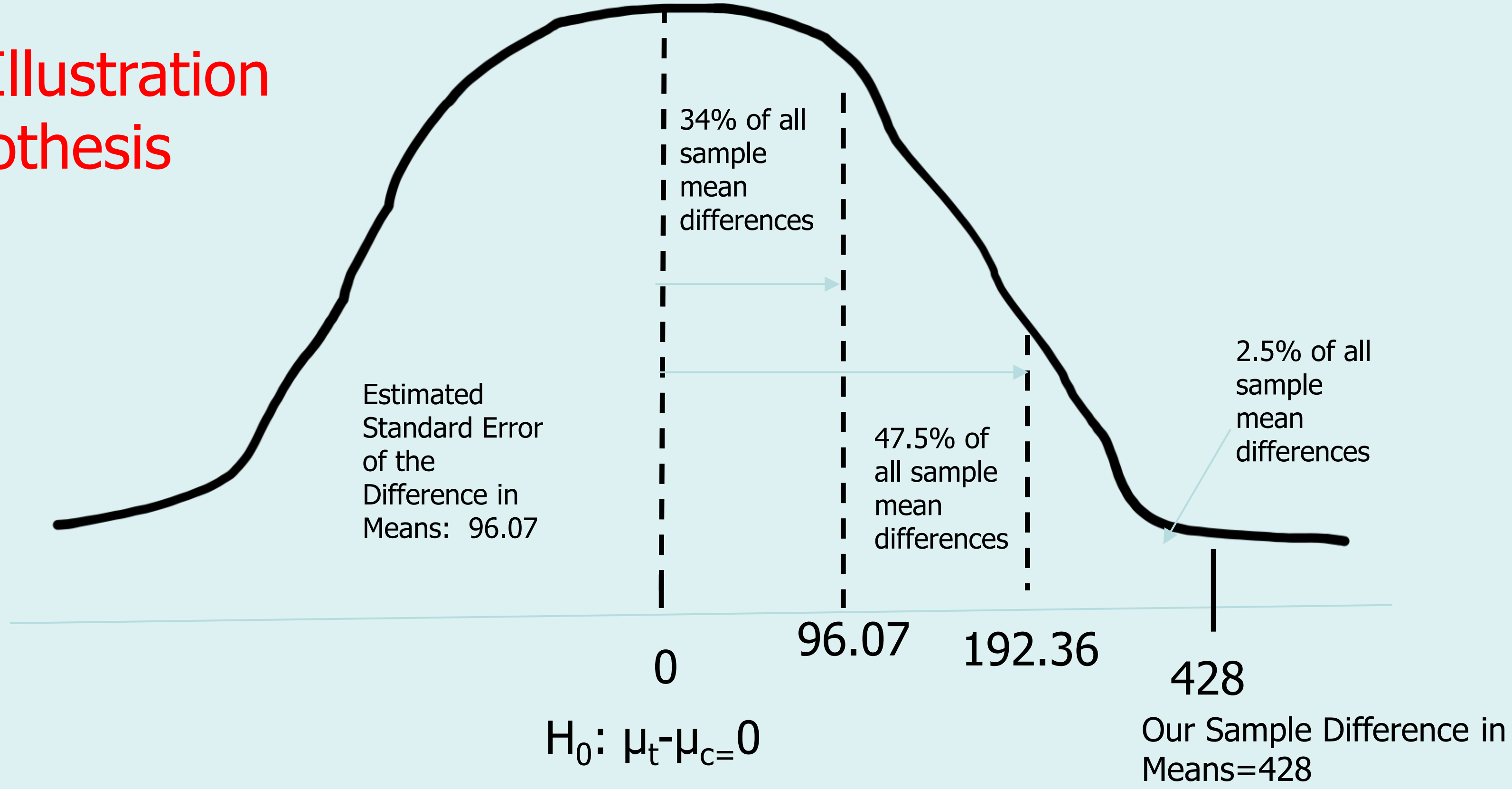
- If our sample difference *could have been observed* from a population with "no difference" *more than 5% of the* time, then we will ***fail to reject*** *the null hypothesis*, and that means we cannot reject the idea that there are no differences between the groups in the population (at the .05 level)

- Why do we choose .05 as the significance level? Social science convention, for the most part. We make it really hard but not impossible to reject null hypothesis, and the .05 level seems to satisfy a balance between conservatism and risk. This means, though, that we always have a 5% chance of being wrong! We will reject H_o but it is really true 5% of the time.

Steps 3: Construct the Test Statistic

- The statistic for testing differences between means is called a “**t-test**”. It gives you the same information as a “z” score, i.e., how many standard errors away from 0 is our sample difference of $428
- If the value of “t” associated with $428 is greater than 1.96, we will reject H_0 at the .05 level
- R output provides the *exact probability* of observing a value of “t” of the size (or greater) that we did observe if the null hypothesis value of 0 were true. If this probability (*p*) is less than .05, we reject the null
- So reject H_0 if “t” is greater than 1.96 (or if “t” is less than -1.96). Reject H_o if *p* is less than .05. Same thing!
- **Calculation of “t”:** $\boldsymbol{t = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}}}$
- t= (Difference in Means)/(Standard Error of Difference in Means)

Visual Illustration of Hypothesis Testing



34% of all sample mean differences

Estimated Standard Error of the Difference in Means: 96.07

47.5% of all sample mean differences

2.5% of all sample mean differences

0

96.07

192.36

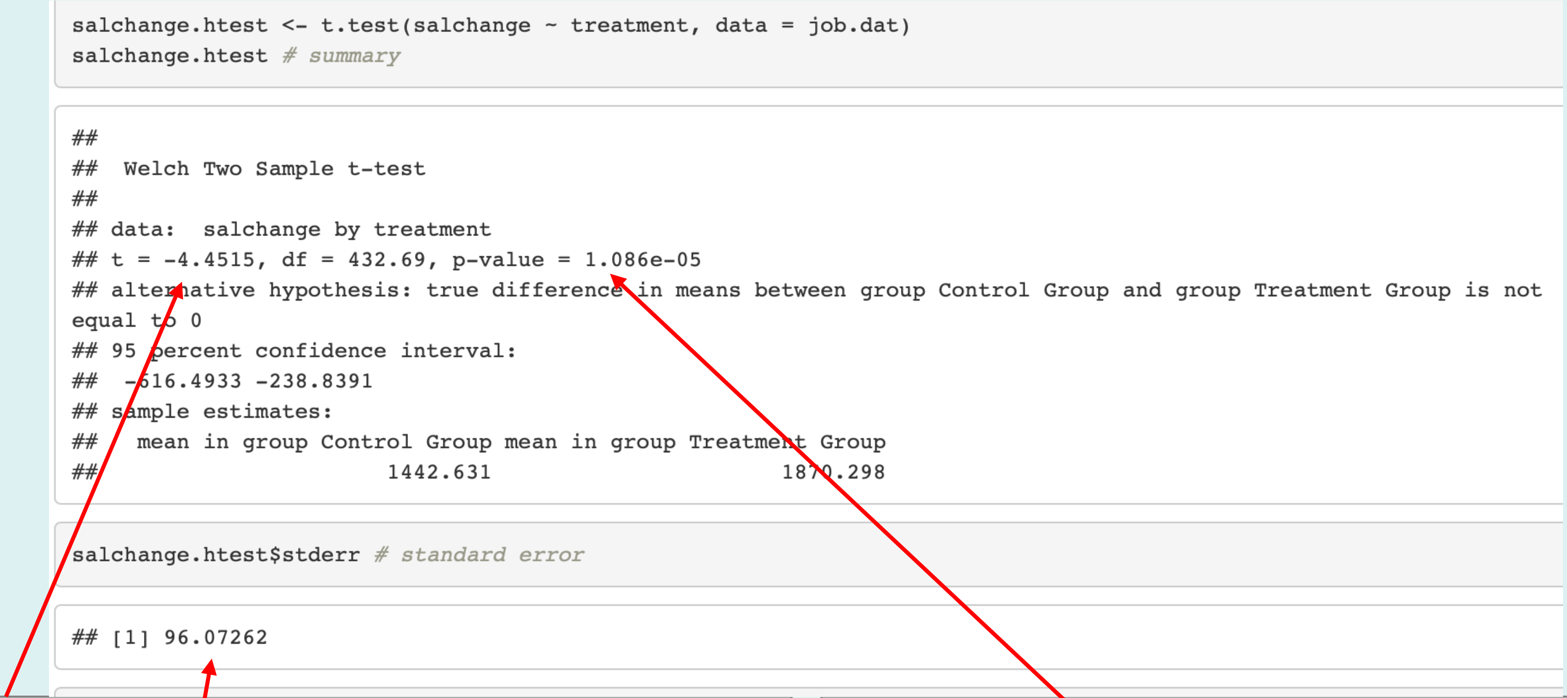
428

$H_0: \mu_t\text{-}\mu_{c=}0$

Our Sample Difference in Means=428

- How likely would it have been to obtain a sample difference between treatment and control groups of 428, if the *TRUE* population difference was 0 (i.e., no difference between the groups?
- How many “standard errors” away from 0 is 428?
- T=428-0/96.07 = 4.45   (see slide 15 for R output of the standard error)
- Our sample mean difference of 428 is 4.45 standard errors away from 0; this difference occurs less than 1 time in 10,000 samples from a population where the difference was 0. This is *well beyond” the .05 significance level, so we say that the difference between treatment and control groups is “statistically significant” at the .05 significance level. We conclude that the differences are “real” and not the result of random chance (but we could be wrong – 5% of the time)!

R Results: T-test



```
salchange.htest <- t.test(salchange ~ treatment, data = job.dat)
salchange.htest # summary
```

```
##
##  Welch Two Sample t-test
##
## data:  salchange by treatment
## t = -4.4515, df = 432.69, p-value = 1.086e-05
## alternative hypothesis: true difference in means between group Control Group and group Treatment Group is not
equal to 0
## 95 percent confidence interval:
##  -616.4933 -238.8391
## sample estimates:
##   mean in group Control Group mean in group Treatment Group
##                      1442.631                      1870.298
```

```
salchange.htest$stderr # standard error
```

```
## [1] 96.07262
```

This is the obtained value of “t” : -4.452.

It is calculated as the difference between the two groups (-428) divided by the standard error of the difference between the groups (96.07)

This is the probability of obtaining a value of “t” of this size or greater *if the null hypothesis of no difference were true*. It is less than 1 in 1000.

Step 4: Decision

- We **reject H_0** since "t" is less than -1.96. The probability of obtaining this size "t" if the null were true is less than 1 in 1000. **We say that the relationship between treatment and salary is *statistically significant at the .05 level*.**
- Note: The "t" is negative only because R treated the control group as the first group, subtracted the average salary increase of the treatment group from that of the control group. This gave -428 as the difference between the group means. If the treatment group were coded as "0" and the control group as "1", it would have been a difference of +428, and a positive "t" of 4.451. Same substantive interpretation.
- **But always pay attention to the coding of the variables when interpreting the results of any statistical test.**

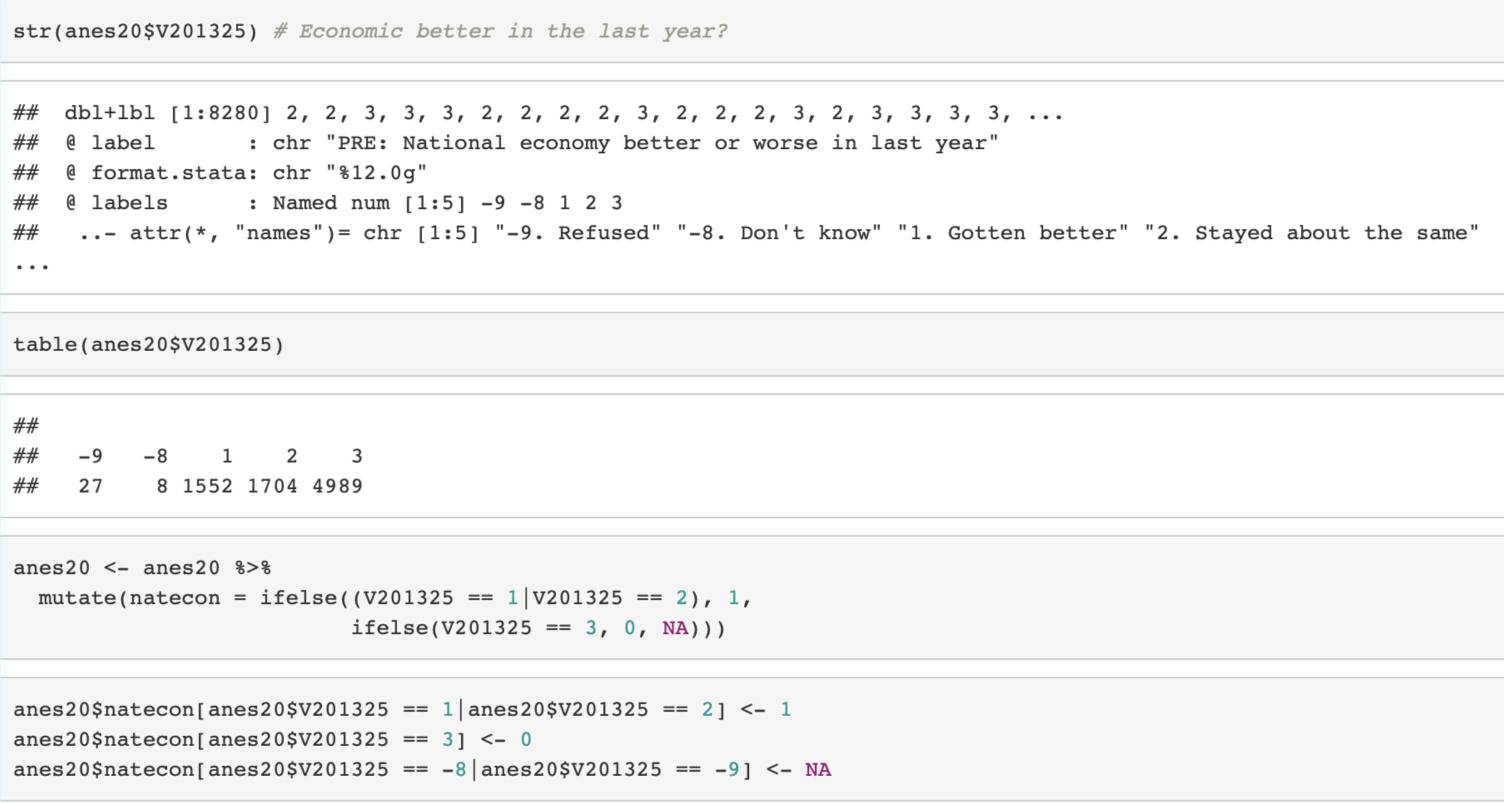
Notes on T-Tests

- When sample size is large enough (>100 or so), the "t" test converges to a perfectly normal distribution and it is identical then to a "z" test. R and other packages always call it a "t" test, however.
- There are variants of the t-test depending on whether the standard deviations of the two groups are the same or not, whether you have equal or unequal number of cases in the groups, etc.
- Statistical significance depends on how large the difference in your samples are, as well as the number of cases you have observed. This means that even very small "substantive" differences may be "statistically significant," given enough observations. So be sure to keep substantive and statistical significance separate and evaluate both
- Statistical significance does not prove **causality** either! (But you certainly knew this already!!!)

R Example: 2020 Election Data

- Question: Do attitudes about Donald Trump and Joe Biden depend on individuals’ perceptions of the national economy?
- Steps:

1. Get 2020 American National Data, and read the data into R
2. Determine the independent and dependent variables from the codebook
   - Independent Variable: Perceptions of National Economy (V20135)
   - Dependent Variable: Difference in Feeling Thermometer Ratings, Trump Minus Biden (V201151, V201152)
3. Run frequencies on each variable to see the range of valid responses, and whether you need to assign “missing values” and/or recode the variable to make the analysis meaningful
4. Recode and/or generate new variables to prepare for analysis
5. Generate t-test and interpret the results



```
str(anes20$V201325) # Economic better in the last year?
```

```
## dbl+lbl [1:8280] 2, 2, 3, 3, 3, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2, 3, 3, 3, 3, ...
## @ label       : chr "PRE: National economy better or worse in last year"
## @ format.stata: chr "%12.0g"
## @ labels      : Named num [1:5] -9 -8 1 2 3
##  ..- attr(*, "names")= chr [1:5] "-9. Refused" "-8. Don't know" "1. Gotten better" "2. Stayed about the same"
...
```

```
table(anes20$V201325)
```

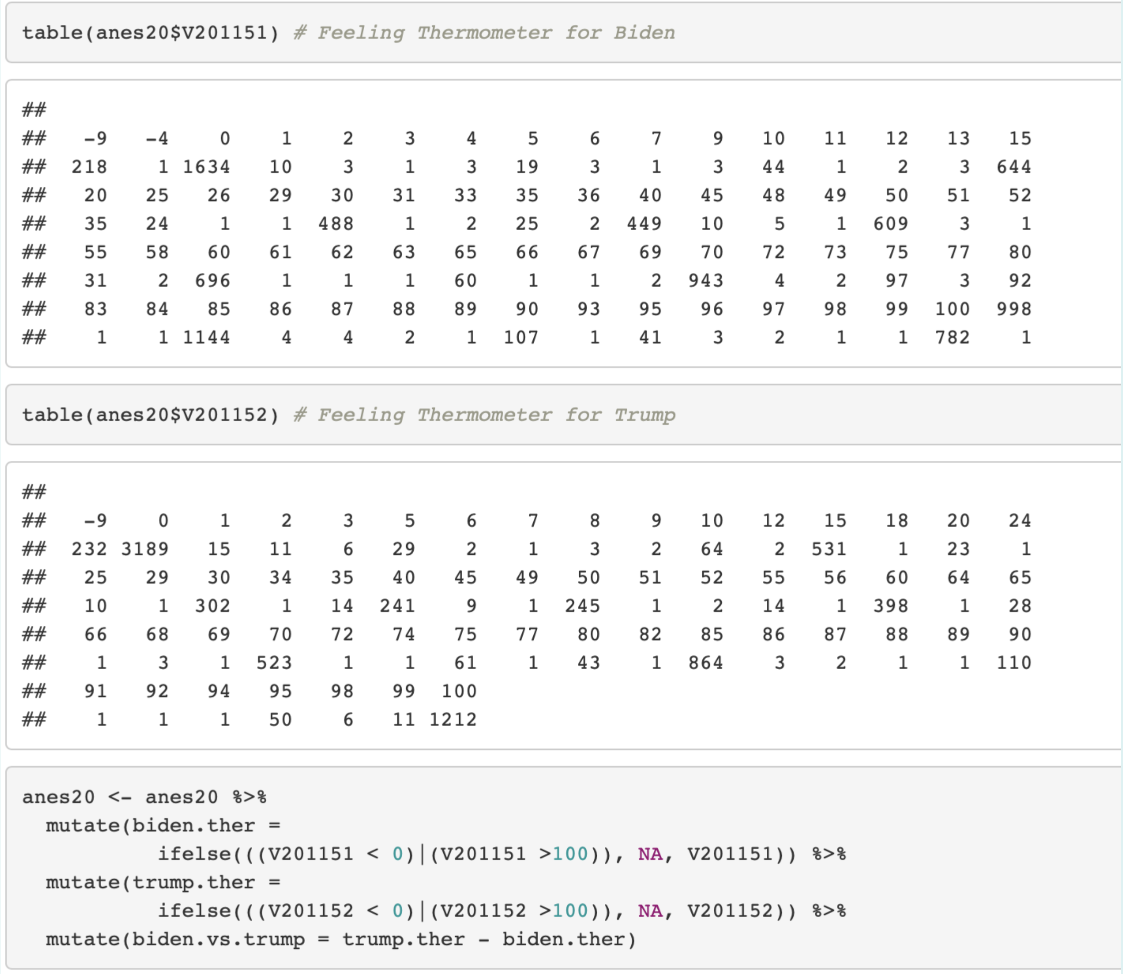
```
##
##   -9   -8    1    2    3
##   27    8 1552 1704 4989
```

```
anes20 <- anes20 %>%
  mutate(natecon = ifelse((V201325 == 1|V201325 == 2), 1,
                          ifelse(V201325 == 3, 0, NA)))
```

```
anes20$natecon[anes20$V201325 == 1|anes20$V201325 == 2] <- 1
anes20$natecon[anes20$V201325 == 3] <- 0
anes20$natecon[anes20$V201325 == -8|anes20$V201325 == -9] <- NA
```

This is the perceptions of national economy variable

I created a two-category variable from this: value 1 will correspond to perceptions of the economy that are “better in the last year” or “stayed about the same”, and value 0 will correspond to perceptions that are “worse”. Values -9 and -8 will be excluded from the analysis. I’ll call this new variable **“natecon”**



```
table(anes20$V201151) # Feeling Thermometer for Biden
```

```
##
##   -9   -4    0    1    2    3    4    5    6    7    9   10   11   12   13   15
##  218    1 1634   10    3    1    3   19    3    1    3   44    1    2    3  644
##   20   25   26   29   30   31   33   35   36   40   45   48   49   50   51   52
##   35   24    1    1  488    1    2   25    2  449   10    5    1  609    3    1
##   55   58   60   61   62   63   65   66   67   69   70   72   73   75   77   80
##   31    2  696    1    1    1   60    1    1    2  943    4    2   97    3   92
##   83   84   85   86   87   88   89   90   93   95   96   97   98   99  100  998
##    1    1 1144    4    4    2    1  107    1   41    3    2    1    1  782    1
```

```
table(anes20$V201152) # Feeling Thermometer for Trump
```

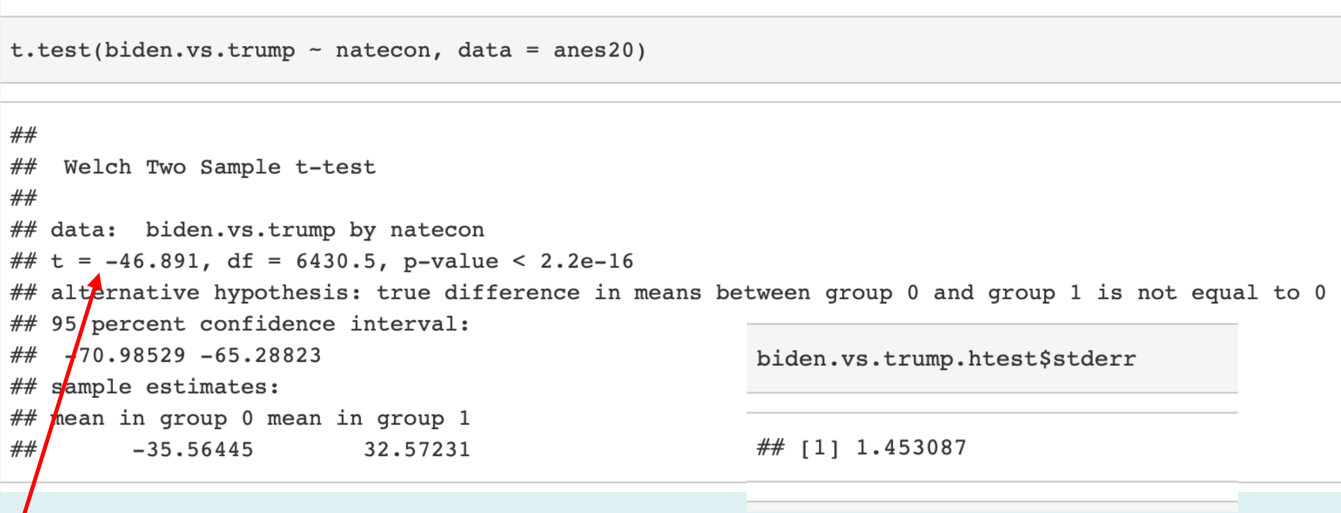
```
##
##   -9    0    1    2    3    5    6    7    8    9   10   12   15   18   20   24
##  232 3189   15   11    6   29    2    1    3    2   64    2  531    1   23    1
##   25   29   30   34   35   40   45   49   50   51   52   55   56   60   64   65
##   10    1  302    1   14  241    9    1  245    1    2   14    1  398    1   28
##   66   68   69   70   72   74   75   77   80   82   85   86   87   88   89   90
##    1    3    1  523    1    1   61    1   43    1  864    3    2    1    1  110
##   91   92   94   95   98   99  100
##    1    1    1   50    6   11 1212
```

```
anes20 <- anes20 %>%
  mutate(biden.ther =
           ifelse(((V201151 < 0)|(V201151 >100)), NA, V201151)) %>%
  mutate(trump.ther =
           ifelse(((V201152 < 0)|(V201152 >100)), NA, V201152)) %>%
  mutate(biden.vs.trump = trump.ther - biden.ther)
```

These are the Biden and Trump Thermometer scores. Values less the 0 or greater than 100 are not valid; otherwise the scores run from 0 to 100.

So I'll tell R to consider only those values between 0 and 100, and then create a new variable from the difference between the Trump and Biden scores.

This will give the **NET THERMOMETER RATING** between the two candidates. Now we want to test whether differences in perception of the national economy determine the relative difference in thermometer ratings between Trump and Biden



```
t.test(biden.vs.trump ~ natecon, data = anes20)
```

```
##
##  Welch Two Sample t-test
##
## data:  biden.vs.trump by natecon
## t = -46.891, df = 6430.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -70.98529 -65.28823
## sample estimates:
## mean in group 0 mean in group 1
##       -35.56445       32.57231
```

```
biden.vs.trump.htest$stderr
```

```
## [1] 1.453087
```

State the null and research hypotheses in statistical language, set the significance level, then conduct the t-test.

This is the probability of obtaining a value of "t" of this size or greater *if the null hypothesis of no difference were true*. It is less than 1 in 1000000

Result: People who think the economy is worse (Group 0) rate Biden higher than Trump by 35.56 points; People who think the economy is better or same (Group 1) rate Trump higher than Biden by 32.57 points. That's a big difference (68.14 total points)!

Conclusion: There is a statistically significant difference between Net Thermometer Ratings of Trump and Biden and the individuals' perception of the state of the national economy at the .05 significance level. And it is a big effect substantively.

Next steps: Compare to other variables' effects, and control for possible Z variables in multivariate analyses. Think: what could those Z variables be???