Code ▾

HW 3

Neal Kar (ink2105)

02/14/2020

Question 1

a)

Read in the SAT data. This data has information by state on the average verbal and math scores on the SAT for 2004 and 2005, and lists the participation percentage and region of each state. The SAT is an exam used for college applications, and higher scores imply better performance. Each section was scored from 200 to 800 in 2005.

Clean the variable names to be easier to work with. Assume that “MathSAT_2005…5” refers to the math scores in 2005 and “MathSAT_2005…7” refers to the math scores in 2004.

Code

b)

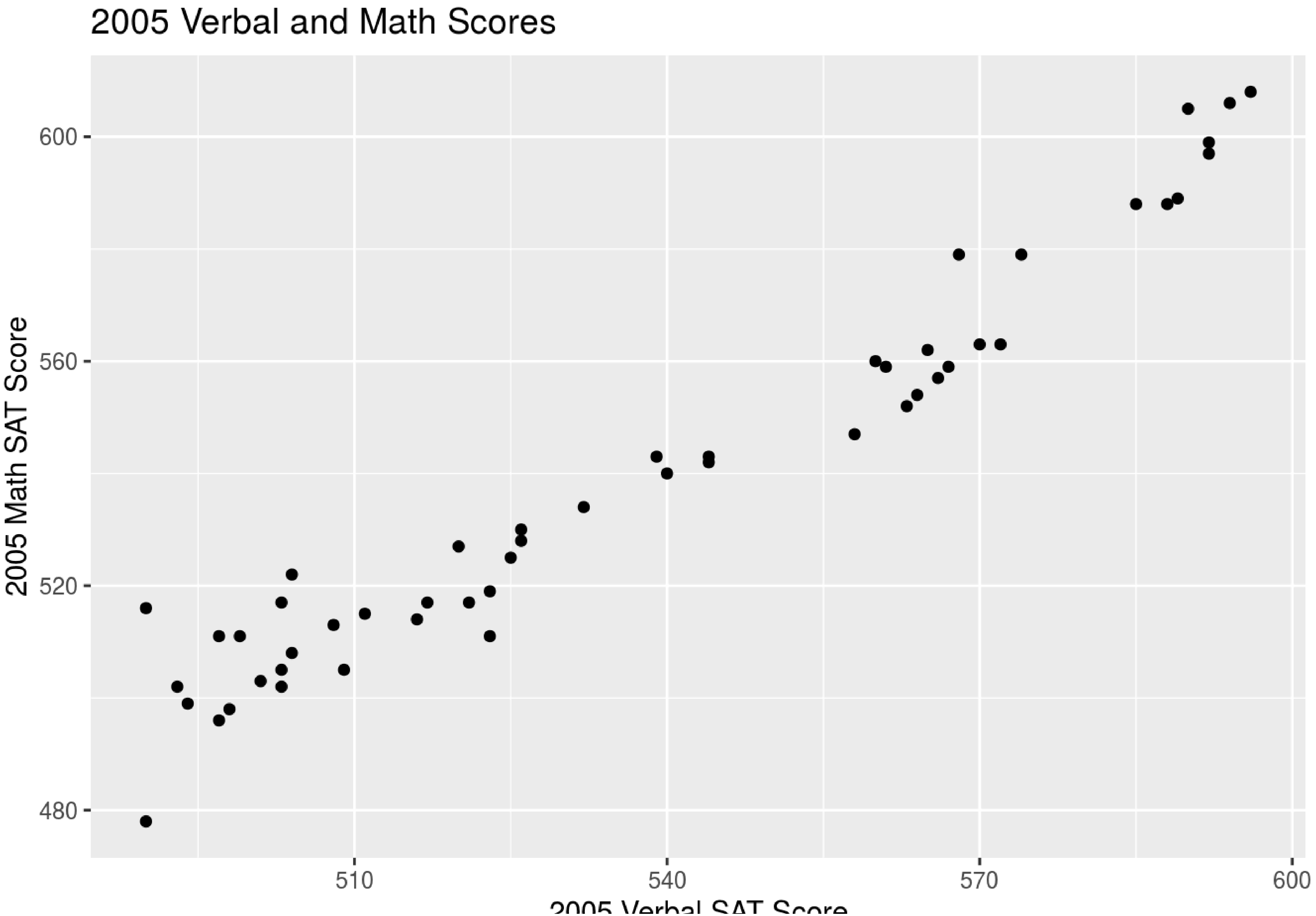
Once the data has been read in and cleaned, creating two new variables which calculate the total score for 2004 and 2005 for each state.

Code

c)

Was there a relationship between verbal and math SAT scores in 2005? Create a graph to evaluate this question, and be sure to include appropriate labels. Comment on the results.

Code


2005 Verbal and Math Scores
2005 Math SAT Score
600
560
520
480
510
540
570
600
2005 Verbal SAT Score

Yes, there appears to a relationship between verbal SAT score and math SAT score. Specifically, there seems to be a positive correlation between the scores since math SAT scores tended to increase as verbal SAT scores increased.

d)

Calculate the average total score by region in 2004 and 2005. Which region had the lowest performance in 2004? Which had the best performance in 2004? What about 2005? You can refer to the regions using the abbreviations from the dataset. You can use `kable()` to print out and refer to your results.

Code

region	mean_2004	mean_2005
WNC	1176.429	1181.143
ENC	1117.600	1126.400
ESC	1115.500	1124.750

region	mean_2004	mean_2005
WSC	1094.000	1092.500
MTN	1080.625	1084.750
PAC	1033.400	1038.800
NE	1025.500	1031.833
MA	1008.333	1010.667
SA	1003.222	1006.111

Code

region	mean_2004	mean_2005
WNC	1176.429	1181.143
ENC	1117.600	1126.400

region	mean_2004	mean_2005
ESC	1115.500	1124.750
WSC	1094.000	1092.500
MTN	1080.625	1084.750
PAC	1033.400	1038.800
NE	1025.500	1031.833
MA	1008.333	1010.667
SA	1003.222	1006.111

The SA region had the worst performance in 2004. The WNC region had the best performance in 2004. The SA region had the worst performance in 2005. The WNC region had the best performance in 2005.

e)

Use the `case_when` *function to create a new participation variable with three groups:*

- “low” for participation less than 50%

-“medium” for participation between 50% and 75%

- “high” for participation higher than 75%

Code

f)

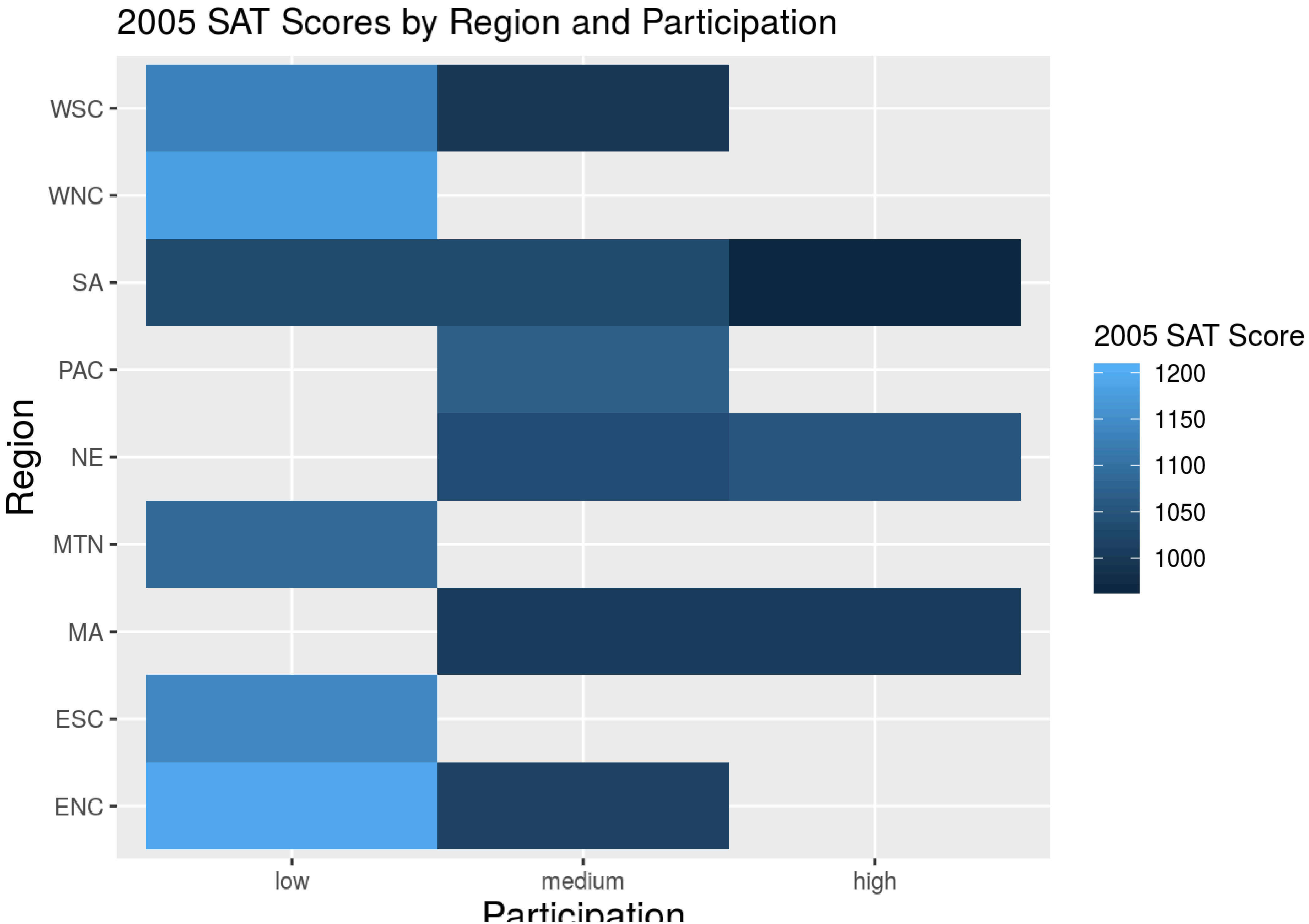
Create a heatmap showing average performance in 2005 by region and your new participation variable. To do this, make a `ggplot` *with* `geom_tile` *. The x aesthetic should be mapped to participation, the y aesthetic should be mapped to region, and the fill aesthetic should be mapped to performance. The geom_tile() geom will automatically calculate the average.*

Code

```
## # A tibble: 51 x 10
##    state region part_pct verbal_2005 math_2005 verbal_2004 ma
th_2004
##    <chr> <chr>     <dbl>       <dbl>     <dbl>       <dbl>
<dbl>
##  1 Alab… ESC          10         567       559         560
553
##  2 Ariz… MTN          33         526       530         523
524
##  3 Arka… WSC           6         563       552         569
555
##  4 Colo… MTN          26         560       560         554
553
##  5 Idaho MTN          21         544       542         540
539
##  6 Illi… ENC          10         594       606         585
597
##  7 Iowa  WNC           5         596       608         593
602
```

```
##  8 Kans… WNC           9         585       588         584
585
##  9 Kent… ESC          12         561       559         559
557
## 10 Loui… WSC           8         565       562         564
561
## # … with 41 more rows, and 3 more variables: sat_tot_2004 <db
l>,
## #   sat_tot_2005 <dbl>, part_cat <fct>
```

Code


2005 SAT Scores by Region and Participation
Region
WSC
WNC
SA
PAC
NE
MTN
MA
ESC
ENC
low
medium
high
Participation
2005 SAT Score
1200
1150
1100
1050
1000

Participation

g)

Improve the visualization of your heatmap by doing the following:

(1) Ensure that the plot has an appropriate title and that both axes and the legend are properly labeled. (2) Make sure the participation groups are presented in a logical order. (3) Use a `theme()` *statement to remove the grey panel background. (4) Use a* `theme()` *statement to increase the axis text size to 14.*

h)

Use your heatmap to answer the following questions:

(1) Does participation appear to be a good predictor for performance?

Participation does appear to be a good predictor of performance.

(2) If so, in what direction is the relationship?

The direction seems to be negative, meaning higher participation seems to be associated with lower scores.

(3) Does performance appear to change substantially region to region?

Yes, performance appears to change substantially region to region.

Question 2

You have received a dataset of treatment compliance for a 10-week study on physical activity and cognitive functioning. This data set can be found in the `activity_compliance.csv` *file.*

a)

Load the compliance data in from the `activity_compliance.csv` *file.*

Code

b)

The data set contains missing compliance observations. For the purpose of this exercise, we will treat missing data as non-compliant. Change all NA compliance values to 0 and use this updated dataset for the following problems.

Code

c)

Create a summary data frame of the proportion of compliant individuals per week and its 95% confidence interval. Report it using the `kable()` *function.*

Code

week	compliant	n	proportion	lower_bound_95_CI	upper_bound_95_CI
1	1	48	0.8000000	0.6868393	0.9131607
2	1	45	0.7500000	0.6234825	0.8765175

week	compliant	n	proportion	lower_bound_95_CI	upper_bound_95_CI
3	1	39	0.6500000	0.5003025	0.7996975
4	1	45	0.7500000	0.6234825	0.8765175
5	1	34	0.5666667	0.4000987	0.7332346
6	1	31	0.5166667	0.3407513	0.6925821
7	1	28	0.4666667	0.2818761	0.6514572
8	1	40	0.6666667	0.5205769	0.8127564
9	1	37	0.6166667	0.4600028	0.7733305
10	1	31	0.5166667	0.3407513	0.6925821

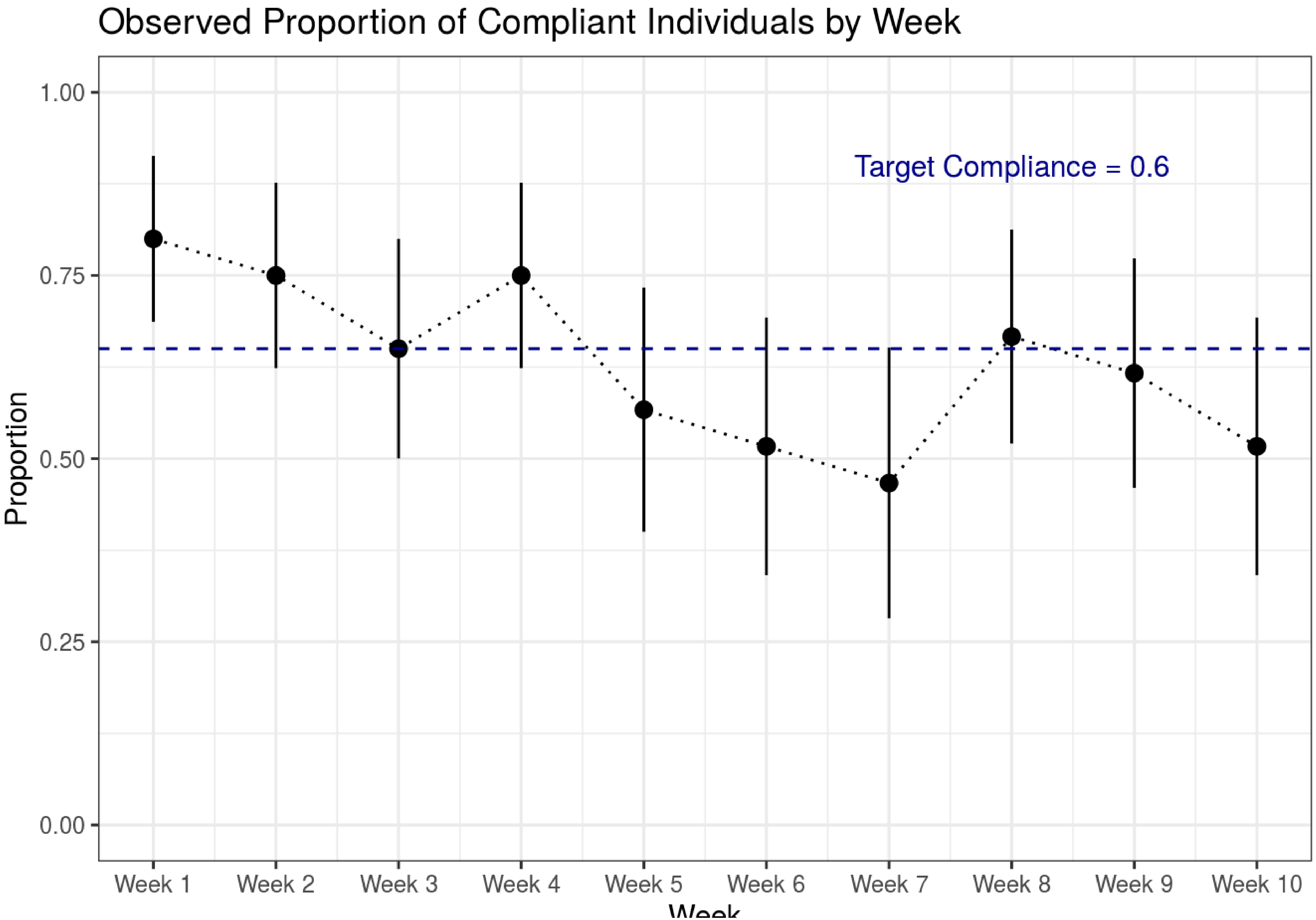
d)

The code skeleton below will create a graph that shows the proportion of compliant individuals and it’s 95% confidence interval per week once you plug in your data frame and the variables for x, y, ymax, and ymin. The ymax and ymin aesthetics should be mapped to the upper and lower bounds of your calculated 95% Confidence interval.

Once you have plugged in these values, you should see a plot of proportions with error bars marking the 95% confidence intervals. We have also added additional statements to improve our plot. In a series of sentences, please describe what each statement does. You need to describe the 8 statements starting with `geom_pointrange` *and ending with* `annotate` *. Hint: If you are stuck, try removing a statement and seeing what changes!*

When are done filling in the code, make sure to change the code chunk option below to eval = TRUE

Code


Observed Proportion of Compliant Individuals by Week
Target Compliance = 0.6
Proportion
1.00
0.75
0.50
0.25
0.00
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Week 9
Week 10
Week

Week

1. Line 211 graphs points with error bars and extends those errors bars to the upper and lower limits specified in the “ymax” and “ymin” statements, respectively.

2. Line 212 draws a dotted line connecting the dots created by line 211.

3. Line 213 creates a blue, dashed, horizontal line across the graph at the y-value 0.65.

4. Line 214 creates a white background behind the graph and makes the graph’s gridlines grey.

5. Line 215 labels the x-axis and y-axis and also creates a title that is set above the graph.

6. Line 216 specifies there should be labels along the x-axis at each point from x=1 to x=10. In addition, the labels should say “Week” followed by the week number (1-10).

7. Line 217 specifies the minmum value of the y-axis is 0 and the maximjm value of the y-axis is 1.

8. Line 218 embeds text in the graph, positions the center of the text at the coordinate (8, 0.9), and makes the text a dark blue color.

Question 3

In lecture and lab we have talked about power and how to calculate and visualize it using simulations. In all our work, we cheated a bit and used Standard Normal Distribution that can be used only if the population variance is known (We did this using the `rnorm()` *function and* `ggplot()` *.). Obviously, the variance that we observe is never a population variance but an observed variance. Because of that, all calculations and simulations should be done using t-distribution. In this excercise , you will compute and simulate power using the t-distribution instead of the z-distribution.*

We are managers at the Oxford Cereals plant. Boxes are expected to contain an average 368 grams of cereal and it was observed that the variance of the boxes weight was 225 grams. We know that cereal weights are normally distributed but we don’t know the population variance of the weights. We always take a sample of 31 cereal boxes to test whether the process (machine) significantly under-fills or over-fills the boxes compared to the label weight.

How likely are we to (rightfully) stop the production if the machine fills only 360 g on average?

a)

Simulate 100,000 observations using the null hypothesis distribution using the rt()` function. In order to do this, you will need to include the degrees of freedom for the t-distribution, which can be calculated using the formula $df = n - 1$.

Code

b)

Determine the rejection region of a t-distribution for a two-sided hypothesis test at a level of significance of 5%. Using qt function …. Hint (test what qnorm(.05, 0, 1) and qnorm(.025, 0, 1) gives you).

Code

The rejection regions are t < -2.0422725 and t > 2.0422725.

c)

Create a new categorical variable `rejection_region` based on whether a null hypothesis simulated observation is inside of the rejection region. What proportion of simulated observations using the null hypothesis fall inside the rejection region? Give precise number (upto 5 decimal places). Why is it not exactly .05?

Code

The proportion of observations that fell in the rejection region for the null hypothesis was 0.04976. It’s not exactly 0.05 because the variance of the distribution isn’t exactly one.

d)

Compute the corresponding t-score for a machine that fills boxes with 360 grams of cereals, on average. Note, that t-score and z-score formula don’t differ. Report the t-score and corresponding degrees of freedom.

Code

The t-statistic is -2.9694743, and the degrees of freedom is 30.

e)

Next, simulated 100,000 observations using the the alternative hypothesis using the `rt()` *function. In order to do this, you will need to include the degrees of freedom for the t-distribution, which can be calculated using the formula $df = n - 1$. Additionally, the t-distribution is always centered around 0. To center it around* *different number, we have to add/subtract the value we want the center (mean) to be. For instance to have the t-distribution centered around +2, we would add +2 to all the simulated observations.*

Code

f)

Create a new categorical variable `rejection_region` *based on whether a alternative hypothesis simulated observation is inside of the rejection region. What proportion of simulated observations using the alternative hypothesis fall inside the rejection region (this is a definition of power)? Give precise number (upto 5 decimal places).*

Code

The proportion of observations in the rejection region for the alternative hypothesis is 0.8176.

f)

Combine your alternative hypothesis simulated data with null hypothesis data into one dataset using `bind_rows()` *. Create a graph that displays the distributions under the null and alternative hypotheses that is just like the graph presented in lecture and lab.*

Key components your graph must have: (1) Custom colors for the rejection and non-rejection regions. (2) The graph must have facets for null hypothesis and alternative hypothesis observations. (3) The x-axis, y-axis, and legend must be properly labeled. (4) Your graph needs to have two vertical dashed lines that mark the boundary between rejection and non-rejection regions.

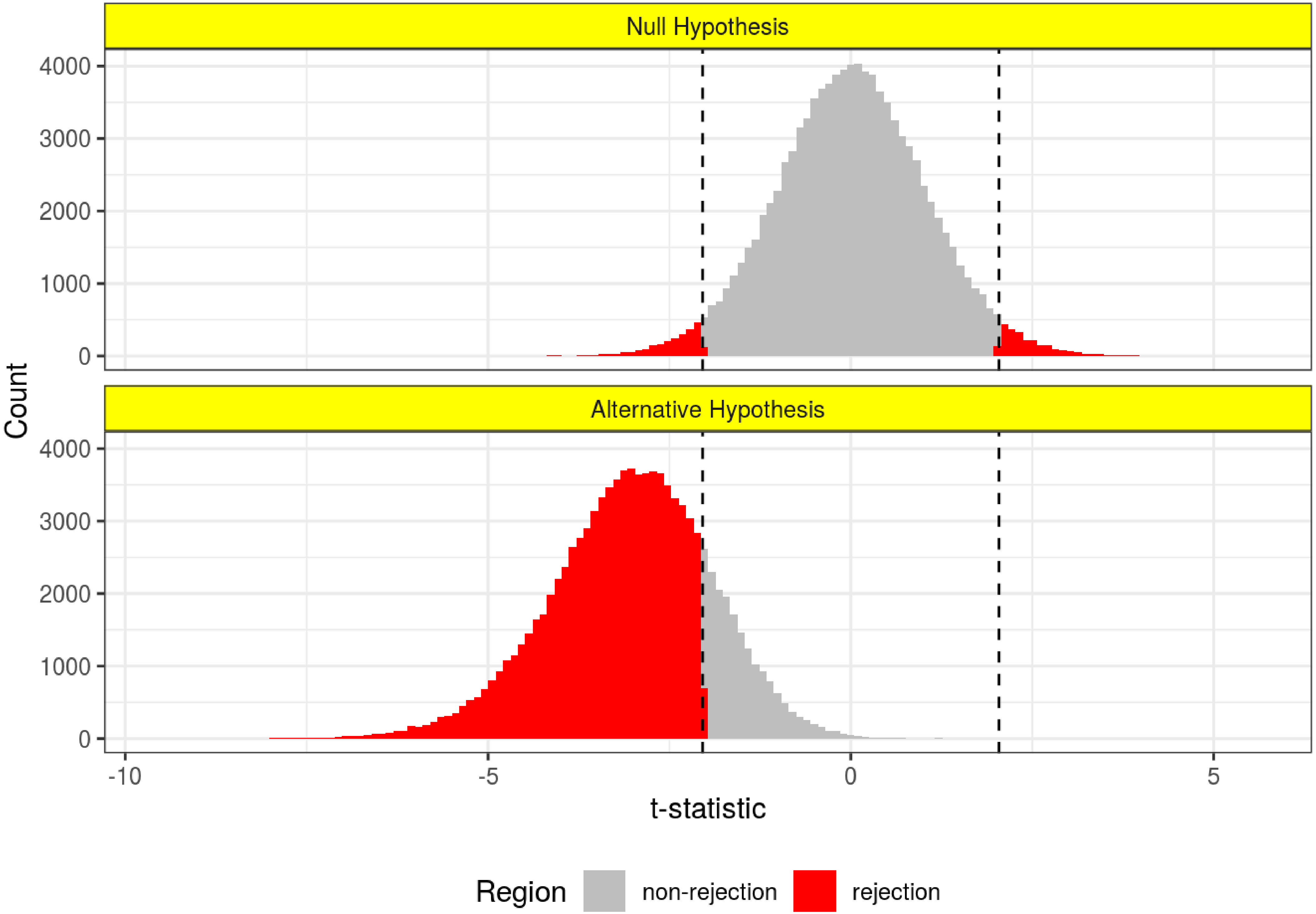
g)

In a sentence, describe how the calculated power using the t-distribution compares to the power we calculated in lecture and lab (use some of the following words "it is smaller", "it is larger", "it is the same"). What do you think about the difference between using the t-distribution and the z-distribution?

Code

```
## # A tibble: 200,000 x 3
##        obs scenario        region
##      <dbl> <fct>           <chr>
##  1 -0.289 Null Hypothesis non-rejection
##  2 -0.529 Null Hypothesis non-rejection
##  3  0.478 Null Hypothesis non-rejection
##  4 -0.281 Null Hypothesis non-rejection
##  5 -0.747 Null Hypothesis non-rejection
##  6 -0.333 Null Hypothesis non-rejection
##  7  0.538 Null Hypothesis non-rejection
##  8  0.395 Null Hypothesis non-rejection
##  9  0.287 Null Hypothesis non-rejection
## 10  2.49  Null Hypothesis rejection
## # … with 199,990 more rows
```

Code


Null Hypothesis
Alternative Hypothesis
Count
4000
3000
2000
1000
0
4000
3000
2000
1000
0
-10
-5
0
5
t-statistic
Region
non-rejection
rejection

The calculated power from the t-distribution (81.76%) is smaller than the calculated power (84.37%) from the z-distribution. The z-distribution assumes we know the population variance and this allows us to accurately maximize power, given a pre-set level of alpha. Since the t-distriibution implies uncertainty over the true population variance, maximum power (the probability of correctly rejecting the null) is lower, given the same pre-set level of alpha.

Question 4

Managers in rival cereal plant “No Fake News Cereals” learned about quality control used by managers at the Oxford Cereals plant and copied their system. “No Fake News Cereals” are expected to weigh on average 777 grams. The machines were observed to fill boxes with standard deviation of 17 grams. We know that cereal weights are normally distributed but we don’t know the population variance of the weights. The managers in “No Fake News Cereals” are not worried about overfilling the boxes, they only want to stop the process (machine) if the boxes are underfilled. In order to determine that, they always take a sample of 33 boxes.

a)

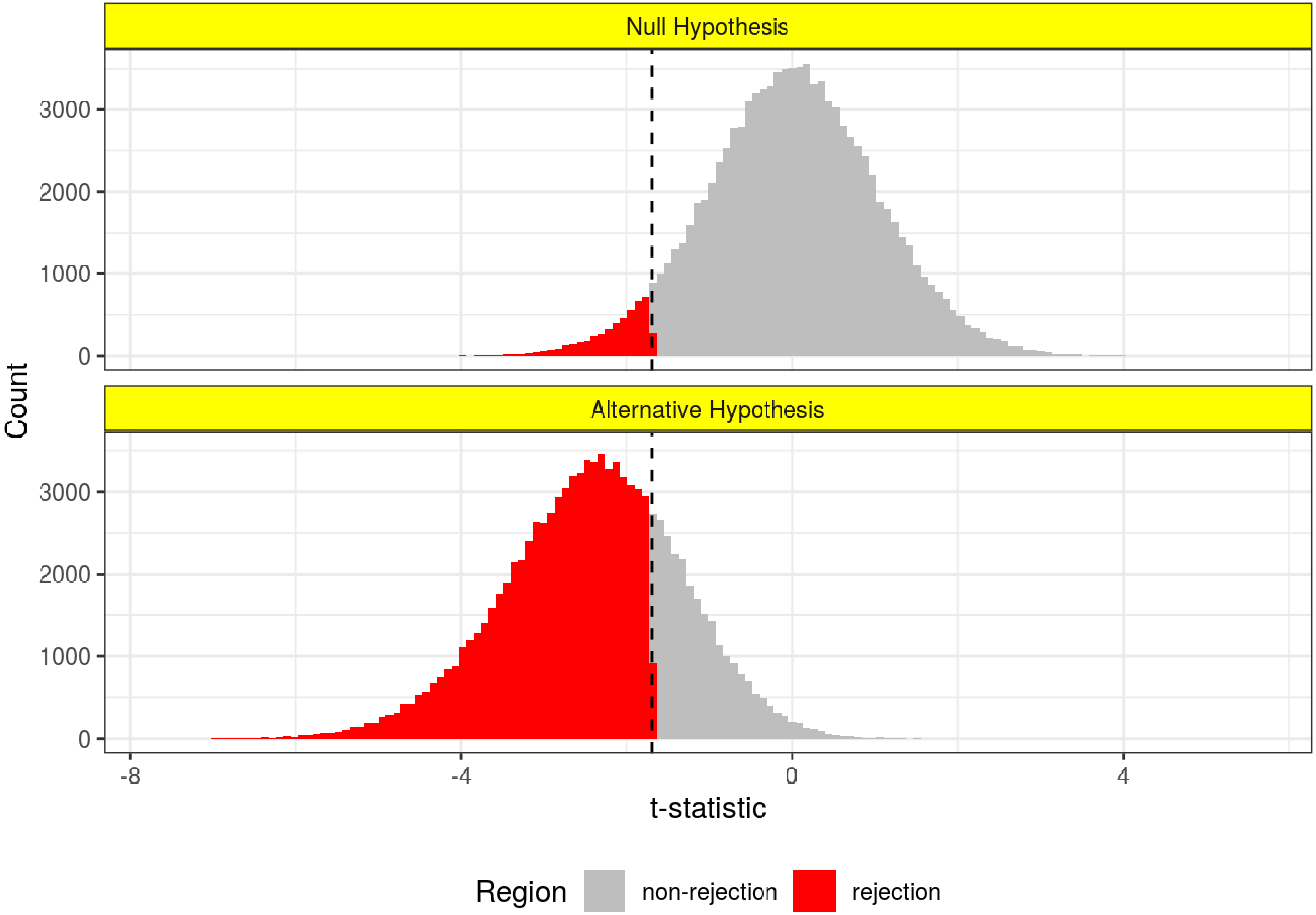
How likely are they to stop the production if the machine fills only 770 grams?

Compute and report power and create the relevant graphs. The graph must contain the same key components as your graph in part 4f.

Code

```
## # A tibble: 200,000 x 3
##        obs scenario        region
##      <dbl> <fct>           <chr>
##  1 -0.289 Null Hypothesis non-rejection
##  2 -0.529 Null Hypothesis non-rejection
##  3  0.478 Null Hypothesis non-rejection
##  4 -0.281 Null Hypothesis non-rejection
##  5 -0.747 Null Hypothesis non-rejection
##  6 -0.333 Null Hypothesis non-rejection
##  7  0.538 Null Hypothesis non-rejection
##  8  0.395 Null Hypothesis non-rejection
##  9  0.287 Null Hypothesis non-rejection
## 10  2.49  Null Hypothesis non-rejection
## # … with 199,990 more rows
```

Code


Null Hypothesis
Alternative Hypothesis
Count
3000
2000
1000
0
t-statistic
-8
-4
0
4
Region
non-rejection
rejection

Power would be 74.61%.

b)

What recommendations do you give to Managers in “No Fake News Cereals” plant in order for them to achieve 80% power.

I would recommend increasing the sample size of cereal boxes to 38 to achieve 80% power for stopping production if the machine fills only 770 grams.