Thanksgiving by Region

Max Wagner

December 7th, 2015

Part 1 - Introduction:

How do different regions in the United States celebrate Thanksgiving? This past Thanksgiving a survey was given by FiveThirtyEightLife with 64 questions on everything from the respondents main dish to how much they shop on Black Friday. The focus of this project will be mainly on whether the respondends celebrated, what their main dish was, and which side dishes they accompanied it with.

Part 2 - Data:

As said above, the data was collected by FiveThirtyEightLife via a SurveyMonkey on November 17th, 2015. The survey collected a total of 1,058 responses with varying amounts of response thoroughness. This project will be using a subset of the original data set found on FiveThirtyEightLife's Github page. In particular, the project will use the explanatory variable on the respondents' region, and study a range of response variables:

- Do you celebrate Thanksgiving?
- What is typically the main dish at your Thanksgiving dinner?
- Which of these side dishes are typically served at your Thanksgiving dinner?

A sample size of 1,058 is large enough to generalize to the public as a whole and to each individual region. A potential problem with the collection method is that it was collected only by visitors of that specific website. Causality can also not be determined due to it being a purely observational study.

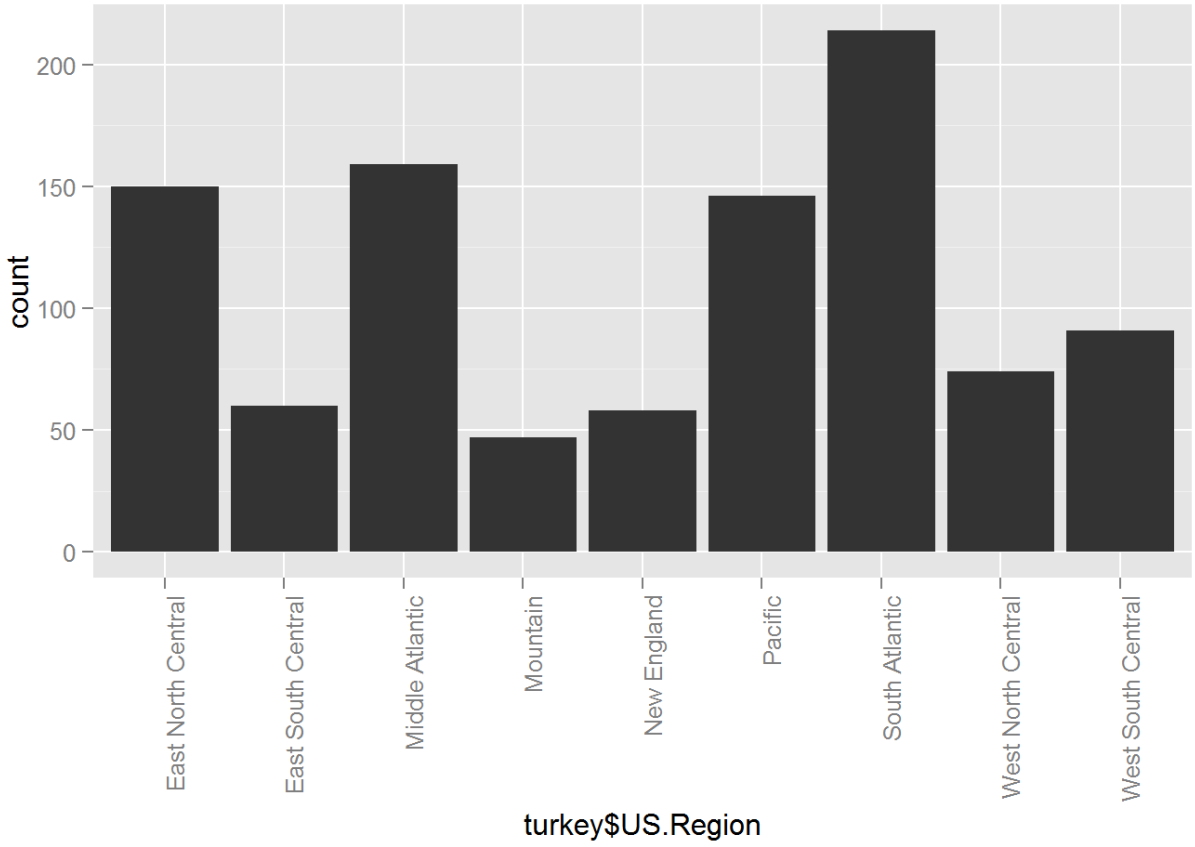
Part 3 - Exploratory data analysis:

Count of Regions First let's look at the spread of regions and number of responses from each of them. Out of 1,058 responses, 999 listed a region and 59 did not. As you can see below, the highest percentages of responses come from the South Atlantic and Middle Atlantic. The South Atlantic had over four times the responses as the Mountain, New England, and East South Central.

The following tables show first a raw count of responses per region, and secondly the frequency as a percentage. The plot below the two tables is a representation of the first raw count table.

	summary.turkey.US.Region.
EN Cntrl	150
ES Cntrl	60
Mid Atl	159
Mtn	47
New Eng	58
Pacif	146
S Atl	214
WN Cntrl	74
WS Cntrl	91

Var1	Freq
EN Cntrl	15.015015
ES Cntrl	6.006006
Mid Atl	15.915916
Mtn	4.704705
New Eng	5.805806
Pacif	14.614615
S Atl	21.421421
WN Cntrl	7.407407
WS Cntrl	9.109109

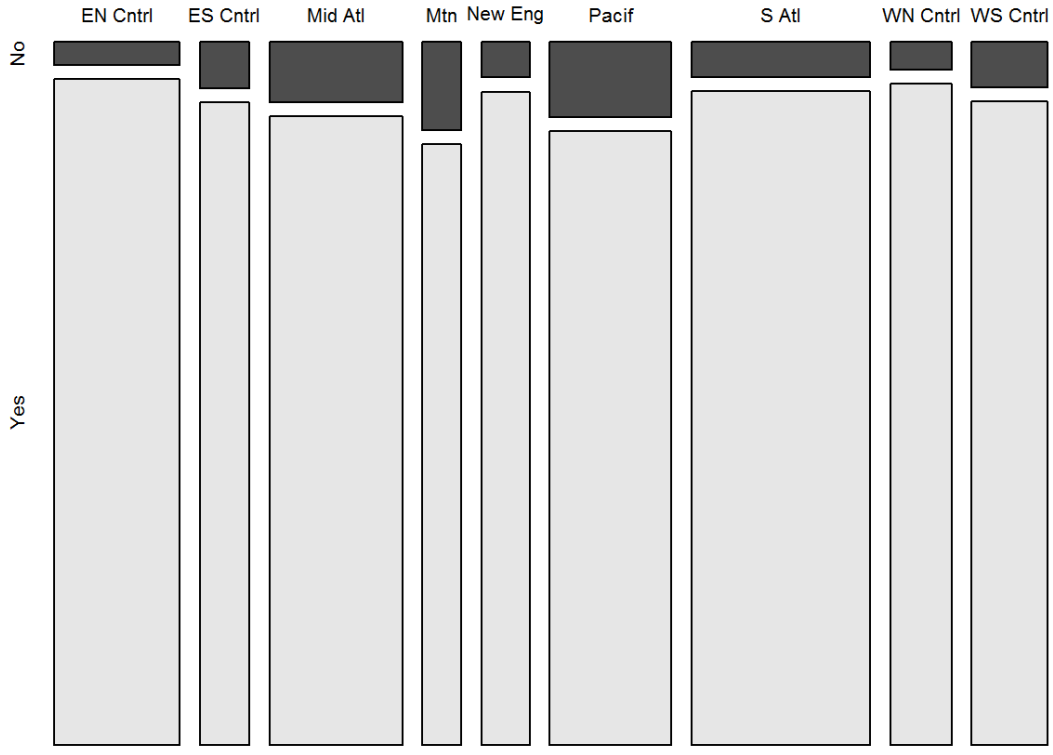


Celebration Percentage by Region The first response answered was whether or not the respondent celebrated Thanksgiving. The following tables and mosaic plot look at the counts of yes/no answers based on region, and the percentages. The Mountain and Pacific regions tend to be slightly lower than the others.

	No	Yes
EN Cntrl	5	145
ES Cntrl	4	56
Mid Atl	14	145

	No	Yes
Mtn	6	41
New Eng	3	55
Pacif	16	130
S Atl	11	203
WN Cntrl	3	71
WS Cntrl	6	85

	No	Yes
EN Cntrl	3.333333	96.66667
ES Cntrl	6.666667	93.33333
Mid Atl	8.805031	91.19497
Mtn	12.765957	87.23404
New Eng	5.172414	94.82759
Pacif	10.958904	89.04110
S Atl	5.140187	94.85981
WN Cntrl	4.054054	95.94595
WS Cntrl	6.593407	93.40659



Main Dish by Region After determining the respondents who did not celebrate Thanksgiving, the rows containing those values were removed as all other questions are blank. This leaves a remaining sample of 931 respondents. The obvious and overwhelming favorite is Turkey, with the Other category nearby.The differences between regions for the main dish is relatively small.

	Count
	0
Chicken	9
Ham/Pork	27
I don't know	2
Other	35
Roast beef	7
Tofurkey	20
Turducken	3
Turkey	828

	Freq
	0.0000000
Chicken	0.9667025
Ham/Pork	2.9001074
I don't know	0.2148228
Other	3.7593985
Roast beef	0.7518797
Tofurkey	2.1482277
Turducken	0.3222342
Turkey	88.9366273

		Chicken	Ham/Pork	I don't know	Other	Roast beef	Tofurkey	Turducken	Turkey
EN Cntrl	0	0	4	0	5	0	1	0	135
ES Cntrl	0	0	1	0	4	1	0	0	50
Mid Atl	0	1	2	0	4	2	5	1	130
Mtn	0	1	1	0	0	0	2	0	37
New Eng	0	2	0	0	1	0	1	0	51
Pacif	0	0	6	1	9	1	4	2	107
S Atl	0	3	7	0	6	3	3	0	181
WN Cntrl	0	1	4	1	3	0	2	0	60
WS Cntrl	0	1	2	0	3	0	2	0	77

		Chicken	Ham/Pork	I don't know	Other	Roast beef	Tofurkey	Turducken	Turkey
EN Cntrl	0	0.00	2.76	0.00	3.45	0.00	0.69	0.00	93.10
ES Cntrl	0	0.00	1.79	0.00	7.14	1.79	0.00	0.00	89.29
Mid Atl	0	0.69	1.38	0.00	2.76	1.38	3.45	0.69	89.66
Mtn	0	2.44	2.44	0.00	0.00	0.00	4.88	0.00	90.24
New Eng	0	3.64	0.00	0.00	1.82	0.00	1.82	0.00	92.73
Pacif	0	0.00	4.62	0.77	6.92	0.77	3.08	1.54	82.31
S Atl	0	1.48	3.45	0.00	2.96	1.48	1.48	0.00	89.16
WN Cntrl	0	1.41	5.63	1.41	4.23	0.00	2.82	0.00	84.51
WS Cntrl	0	1.18	2.35	0.00	3.53	0.00	2.35	0.00	90.59

Which Side Dishes are Preferred The first table shows the overall most eaten side dishes throughout the country. The three most common dishes were mashed potatoes, rolls/bisquets, and green beans. A notable omission from this list is stuffing. The question was not asked on the survey.

region	sprouts	carrts	cauli	corn	crnbrd	fruit	beans	pasta	potato	rolls	squash	veges	yam
All	16.22	25.24	9.24	48.98	24.49	21.91	71.75	21.48	85.5	79.91	17.72	21.27	66.38

The next table looks at each of the side dishes and their average prevalence in each regions' meals.

region	sprouts	carrts	cauli	corn	crnbrd	fruit	beans	pasta	potato	rolls	squash	veges	yam
EN Cntrl	15.17	19.31	8.97	52.41	15.86	13.79	71.03	14.48	87.59	84.83	10.34	17.93	62.07
ES Cntrl	12.50	26.79	8.93	55.36	28.57	35.71	87.50	37.50	80.36	87.50	21.43	14.29	78.57
Mid Atl	28.28	31.03	17.24	53.10	22.76	17.93	63.45	13.79	89.66	73.10	30.34	22.76	68.28
Mtn	14.63	26.83	9.76	41.46	24.39	26.83	75.61	7.32	92.68	80.49	12.20	29.27	63.41
New Eng	21.82	45.45	7.27	40.00	18.18	9.09	60.00	10.91	94.55	74.55	56.36	20.00	61.82
Pacif	22.31	23.08	13.85	42.31	28.46	26.92	64.62	13.85	86.15	76.15	13.08	30.00	65.38
S Atl	13.30	23.65	5.42	47.29	26.11	19.21	73.89	38.92	77.34	77.83	15.27	19.21	70.94
WN Cntrl	4.23	16.90	4.23	50.70	16.90	25.35	84.51	16.90	91.55	87.32	2.82	21.13	46.48
WS Cntrl	4.71	24.71	3.53	54.12	40.00	35.29	77.65	23.53	82.35	85.88	9.41	17.65	74.12

When we take the difference of each region and the overall average, we get the following table that shows which roughly equates to the residuals. The last row shows the mean difference for each side dish, and the last column shows the mean difference for each region. This allows a slightly different view on each side dish and region.

Region	sprouts	carrts	cauli	corn	crnbrd	fruit	beans
EN Cntrl	-1.05	-5.93	-0.27	3.43	-8.63	-8.12	-0.72
ES Cntrl	-3.72	1.55	-0.31	6.38	4.08	13.80	15.75
Mid Atl	12.06	5.79	8.00	4.12	-1.73	-3.98	-8.30
Mtn	-1.59	1.59	0.52	-7.52	-0.10	4.92	3.86
New Eng	5.60	20.21	-1.97	-8.98	-6.31	-12.82	-11.75
Pacif	6.09	-2.16	4.61	-6.67	3.97	5.01	-7.13
S Atl	-2.92	-1.59	-3.82	-1.69	1.62	-2.70	2.14
WN Cntrl	-11.99	-8.34	-5.01	1.72	-7.59	3.44	12.76
WS Cntrl	-11.51	-0.53	-5.71	5.14	15.51	13.38	5.90
means	-1.00	1.18	-0.44	-0.45	0.09	1.44	1.39

Region	pasta	potato	rolls	squash	veges	yam	means
EN Cntrl	-7.00	2.09	4.92	-7.38	-3.34	-4.31	-2.79
ES Cntrl	16.02	-5.14	7.59	3.71	-6.98	12.19	4.99
Mid Atl	-7.69	4.16	-6.81	12.62	1.49	1.90	1.66
Mtn	-14.16	7.18	0.58	-5.52	8.00	-2.97	-0.40
New Eng	-10.57	9.05	-5.36	38.64	-1.27	-4.56	0.76
Pacif	-7.63	0.65	-3.76	-4.64	8.73	-1.00	-0.30
S Atl	17.44	-8.16	-2.08	-2.45	-2.06	4.56	-0.13
WN Cntrl	-4.58	6.05	7.41	-14.90	-0.14	-19.90	-3.16

Region	pasta	potato	rolls	squash	veges	yam	means
WS Cntrl	2.05	-3.15	5.97	-8.31	-3.62	7.74	1.76
means	-1.79	1.41	0.94	1.31	0.09	-0.71	0.27

For instance it is simple to tell which region is the least or most likely to include squash in their meal. A slightly more interesting use of the table is understanding which region tends to have the most unique meal compared to other regions. This is possible by looking at the means column. East South Central tends to have the most unique meal which is likely due to their large residuals for fruit, beans, yams, and pasta. South Atlantic has the least unique as their high amount of pasta eaten is balanced by their relatively lower amounts of most other sides.

Part 4 - Inference:

Celebration The first step is figuring out whether the difference in celebration between regions is by chance, or if it statistically meaningful. Below is a quick table to see an overview of the responses, and another to view by region. Each is shown as a percentage.

	cele.total
No	6.81
Yes	93.19

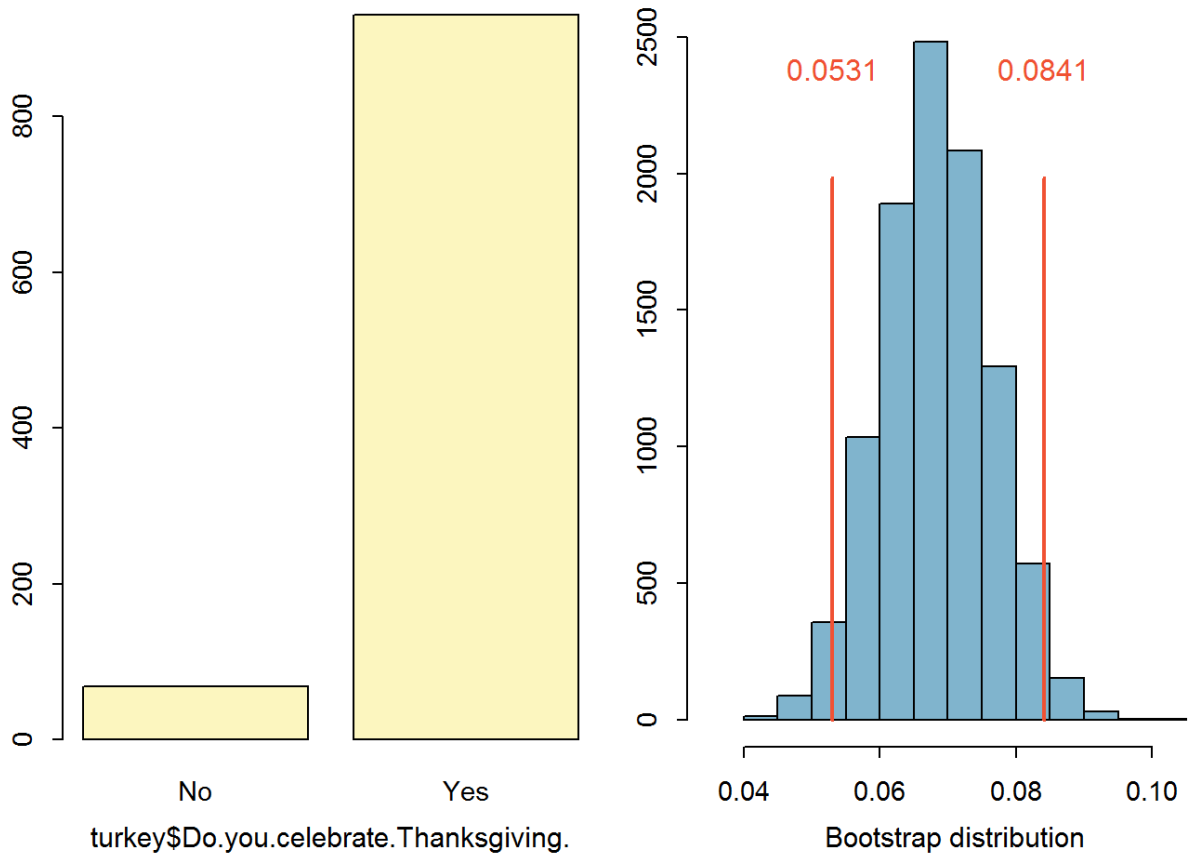
	No	Yes
EN Cntrl	3.33	96.67
ES Cntrl	6.67	93.33
Mid Atl	8.81	91.19
Mtn	12.77	87.23
New Eng	5.17	94.83
Pacif	10.96	89.04
S Atl	5.14	94.86
WN Cntrl	4.05	95.95
WS Cntrl	6.59	93.41

The inference function from a lab earlier this year tests the standard error and the interval, but first we need a check for the requirements. All observations seems to be independent, but the success failure check indicates that np = 5.508 and n(1 - p) = 62.492. Only one of the two needed passes the check so we will use a simulation for inference.

```
## Single proportion -- success: No
## Summary statistics:

## p_hat = 0.0681 ;  n = 999

## 95 % Bootstrap interval = ( 0.0531 , 0.0841 )
```



Side Dishes Instead of looking at all the regions individually, we can look at individual side dishes and if they are statistically different in each region compared to the overall average. For the sake of the length we will look at a limited number of side dishes. For each ANOVA test the hypothesis will be the same:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9$ where 1-9 represent each region

H_a : one or more means are different

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## sides.anova$region    8   5.19  0.6483   4.927 5.53e-06 ***
## Residuals           922 121.32  0.1316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taking a look at the statistics for sprouts, we can see that the p-value is leass than 0.05, so the null is rejected, which means that at least one region has as statistically different of sprouts eaten than the average.

The same can be done for all other side dishes. In the instance of the table for cauliflower below, the p-value is also below 0.05 which means we reject the null hypothesis and at least one region has a different amount of cauliflower eaten than the others. An important thing to note is that the p-value is much closer to 0.05 than the sprouts example.

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## sides.anova$region    8   1.98 0.24754       3 0.0025 **
## Residuals           922  76.08 0.08251
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table below is for pasta which had the largest difference in the section 3 table. Again we reject the null hypothesis. Interestingly the p-value is smaller than the value from the sprouts example. The other side dishes follow a similar pattern as the three example tables.

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## sides.anova$region    8  11.55  1.4443   9.154 3.57e-12 ***
## Residuals           922 145.48  0.1578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part 5 - Conclusion:

The brief look at Thanksgiving meals throughout the United States concludes a few important things. One is that each region certainly eats different things for Thanksgiving. The main dish of Turkey remains constant throughout the country, but individual side dishes differ greatly from region to region. The variation between regions was significantly more defined than I had expected going into the project. Being from New England myself, awareness of how others spend their holidays was an interesting topic. In the future, it would be beneficial to look at more of the personal aspects of Thanksgiving like how many family members came, how much they spent on Black Friday, or if they watched the Macy's Parade.

References:

- https://github.com/fivethirtyeight/data/tree/master/thanksgiving-2015