**Theodore Stetzel  
Baruch College – STA 9710**

**05/18/2015**



**Pizza Consumption in New York City**

A Dietary Intake Survey using Stratified Sampling

**Contents**

[Part One: Proposed Survey 1](#_Toc419119996)

[(a) Subject of Survey 1](#_Toc419119997)

[(b) Sampling Method 2](#_Toc419119998)

[(c) Estimation Method 4](#_Toc419119999)

[(d) Urn Model 6](#_Toc419120000)

[(e) Sample Size 6](#_Toc419120001)

[(f) Cost 7](#_Toc419120002)

[Part Two: Pilot Study 9](#_Toc419120003)

[(a) Description of Pilot Study 9](#_Toc419120004)

[(b) Data you collected 9](#_Toc419120005)

[(c) Results of Trial Run 10](#_Toc419120006)

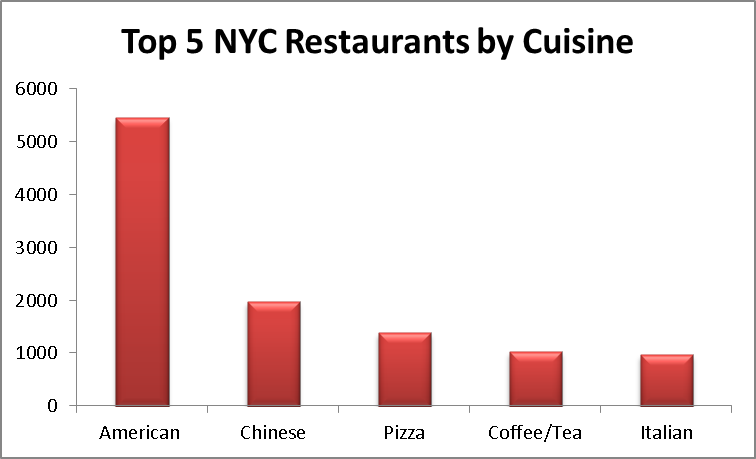
[(d) Sources of Bias 10](#_Toc419120007)

[Part Three: Data Appendix 12](#_Toc419120008)

*Cover image by Seattle Pi http://www.seattlepi.com/*

# Part One: Proposed Survey

(a) Subject of Survey

Pizza is one of the most popular foods in America. On any given day, an estimated average of 13% of Americans eat pizza for at least one meal. Nowhere is pizza more ubiquitous than in New York City where there are approximately 1353 pizzerias (4.7% of all NYC restaurants) operate. On most blocks in the city, one can find a pizzeria selling single slices, for approximately the price of a subway ride, a few steps away. As a result pizza has become an important part of the New York City experience.

My study seeks to estimate how many slices of pizza are sold in New York City in a single year. In addition to being an interesting fact about the city the study, this information could be useful to government agencies such as the USDA that study food consumption behavior. Its popularity has also made pizza as a major contributor of nutrients in the American diet and since pizza is more popular with children than it is adults[[1]](#footnote-1), it would be interesting to examine the overall health of students year over year and it’s correlation with the number of pizza slices sold in the city. Restaurateurs would also be likely interested in the results of this study in order to examine consumer behavioral trends in America’s largest restraint market.

In order to organize this undertaking, the parameters of the study will need to be established. The paper will first answer the “where” the study will be conducted aspects and then the “what” and “how” parts of the study. Establishing the “where” starts by creating what is called the sampling frame, or in this case, a list of every NYC pizzeria’s name and address. This frame will be used to conduct a sampling study and then to estimate the total number of slices sold after the sampling portion of the study is complete.

Most of the information for the survey’s sampling frame comes courtesy of the city of New York. The NYC department of Health and Mental Hygiene (DOHMH) inspects every restaurant in NYC before opening and conducts at least one unannounced inspection annual to ensure that proper standards for food preparation, hygiene and vermin control are followed. Since these inspections are required by law, the list provided by the DOH on the NYC open data website should be considered up to date.

The following criteria will need to be met in order to be included in the sampling frame:

* Must be on the NYC open data list of restaurants provided by the DOHMH
* Must have had a passing inspection conducted in the last 365 days
* Must have chosen either Pizza or Italian/Pizza as cuisine type

The resulting sampling contains 1353 restaurants and can be downloaded here in excel format: <https://www.dropbox.com/s/da5srjft7idkwpb/frame.xlsx?dl=0>

The population parameter that will be observed is the number of pizza slices that are sold at each restaurant. In order to collect data, someone will visit several pizzerias in New York City and sit in a location where he or she can inconspicuously overhear activity at the register and count the number of slices sold at each pizzeria.

(b) Sampling Method

Since it is not practical to have someone sit at each pizzeria in NYC for a whole year and count the number of slices that are sold every day, sampling from the will be utilized to estimate the total number of slices sold. There are three possible methods that could be used to sample. The pros and cons of each are detailed below:

**Simple Random Sampling (SRS)**

SRS could be used to determine sample from the population of pizzerias in order to estimate the total number of slices ordered in NYC per year (tau). First we would establish a percentage of restaurants we would want to audit, let’s say use a naive 5% for this example (sample size will be discussed further in section e of this chapter), and then randomly select 68 restaurants from the total 1343 (n=68 N=1343) to which we would conduct our study. We would then send out our team to these 68 different restaurants on a given day and they would count the number of slices sold at each restaurant.

To determine the total amount of combinations of restaurants that we can sample from, taking into consideration that we don’t want to take into consideration the order in which they are selected, the following formula can be used (n=68 N=1343):

NPn =  =  3.67\*10115

Which is an extremely large number of combinations. We’ll revisit how this number is used to infer the total number of slices sold in section d of this chapter.

While this is the simplest way to sample it also the most expensive due to the wide range of pizzerias that occur in a single selection. Since SRS requires going to 68 restaurants which could be very far apart. If we wanted to conduct these surveys all on the same day, we would need to coordinate 68 people to arriving at each the specified locations. A very challenging task from a managerial standpoint considering each person would need to get their own and it would be difficult to provide managerial oversight.

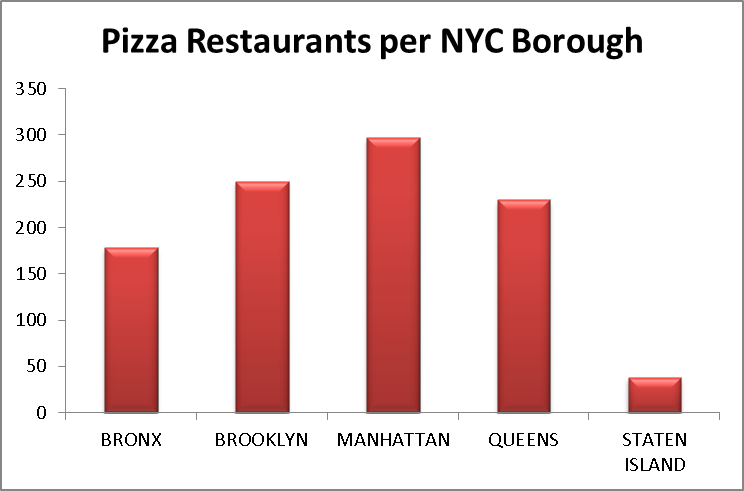
**Cluster Sampling**

In order to save money we could use the method of sampling known as cluster sampling. This method would involve sampling from “clusters” or larger groups of data points that share some commonality (typically geographic). For this study we could perform stage-one cluster sampling by picking a group of restaurants to sample using the NYC community districts that have been predetermined by the U.S. Census Bureau (map of CDs is included in the appendix of this document). Our team of samplers would then go out and sample at every pizzeria within that district. Another approach would be to perform two-stage cluster sampling which would involve sampling from smaller clusters such as a five square block radius, from within each of the NYC Community Districts. Either of the cluster approaches would be advantageous from a managerial standpoint since a single manager could check on all of the surveyors and a system could be set up to give them a break if needed.

However using this method could introduce bias if a selected cluster neighborhood is not representative of the total population or if the clusters chosen are not representative of the population as a whole. This type of sampling is also best used when the sampling frame is not well established and since we are afforded extra information form the DOH website and studies by the USDA, which can be considered fairly accurate, there are better more effective ways to way to sample than Cluster, which is prone to error, or SRS, which is difficult to manages and is expensive.

**Stratified Sampling**

Since we can know some of the characters of when and where people eat pizza we can use stratified sampling which can be considered a hybrid solution. The DOHMH website information can be used to determine the number pizza restaurants in each borough. We can this by-borough create pre-determined groups called strata.

Additionally, strata could be broken out by time of day as a means of reducing the amount of time each pizzeria would need to be observed. Since it is highly likely that the majority of pizza slices are served during traditional lunch and dinner times, our study could focus more of our survey in during these times.

However this adds another layer of complexity and the potential to introduce bias since outer boroughs will likely have lower lunch time consumption compared to Manhattan since it is the city’s center of business. In order to introduce borough/time strata we’d need to know the proportions of how much pizza is eaten in which boroughs at different times of the data to set up our strata. Since this information is not available, the study will require that those collecting data stay at each pizzeria for an entire day to observe during busy and non-busy times.

For this study, the stratified random sampling methodwill be utilized and each borough will represent a strata. This will be executed by randomly picking the appropriate number of observations from each borough (based on the proportion of pizzerias there) and then sending agents to count the number of slices sold at each of these pizzerias.

(c) Estimation Method

Since this survey will only collect data from a small portion of the total number of pizzerias in New York City, a methodology needs to be established to determine how the sampled data will be used to estimate the total number of slices sold by all restaurants (i.e. the population total ). There are three popular methods that are used to estimate characteristics of a total population based on a smaller sampling of data.

**Indirect Method**

Indirect and Indirect ratio methods could be utilized if the study did not directly count the number of slices ordered at each restaurant. For instance an indirect method such as counting the number of plates or the amount of flour or cheese or flour could that each pizzeria order could be performed. Instead of watching activity at each register, we could ask pizzerias to tell us the amount of each of these ingredients and supplies.

**Ratio Method**

Another possible indirect method that could be used is to construct a ratio and then use this as a multiplier to estimate the total number of slices sold in New York City in a single. In the case of this study would be to calculate the time between slices sold and then use the total time each pizzeria is open to calculate the pizzeria’s throughput.

The ratio method has pros and cons just like the other methods. The ratio method can help to stabilize variance by reducing heteroskedasticity, where the levels of variance change at different parts of the model. This would be a useful in this case considering that pizza sales likely vary greatly at different times of the day. However this method is not considered an unbiased estimator (which will be discussed further in second d). Although both approaches would save both time and money it’s not likely that this would be as accurate since these same ingredients and supplies could be used

**Direct Method**

The direct method is the simplest of the three estimation methods. It simply takes the average of the all samples taken and then multiplies that average by total number of restaurants in the population. It can be expressed as the following equation: 

Where N is equal to the total number of restaurants and hours they are open if we collect less than a full day’s worth of data. And  is average of all samples collected, expressed as= (1/n), where n is the number of samples collected from the population of pizzerias expressed as N.

This approach can is reasonable if the restaurants chosen are done so at random. In the case where stratified sampling is used, the restaurants inside each strata must be selected at random.

(d) Urn Model

In order to explain why the direct method is an effective unbiased estimator, the Urn model and the story of many possible samples can be used a means of explaining this rather complex concept.

Imagine that there is a large urn and that this urn could be filled with marbles in order to simulate random selection process, such as a lottery or a BINGO game. However, instead of each marble representing a square on a bingo card, each marble marble in this urn represents a pizzeria in Brooklyn. To be more specific, we’ll take 368 brand new marbles and write the name and address of each pizzeria onto the marble and then place it in the urn. This will represent the sampling space for the Brooklyn in this study.

Later in this paper, it will be determined that we need to selected 25 pizzerias from Brooklyn to get an accurate sample for now, take this number at its face value. In order to select these pizzerias, we’ll pick 25 marbles out from the urn (with replacement for simplicity) and write these results down on to another marble. Taking into account that we do not care which order the pizzerias are selected, we can calculate the total number of possible combinations using the formula below:

NPn =  =  1.345\*1041

This large number represents the number of possible combinations or picking 25 pizzerias from the total 368 in Brooklyn or the total number of different combinations we can pull out of the urn. In order to conduct a study, we bring in a second, “child” urn that will hold these the results of experiment. If someone kept pulling marbles in theory he or she could pull each of the possible 1.345\*1041 combination or marbles and then writing the results onto a second marble and then putting these new marbles into our second urn.

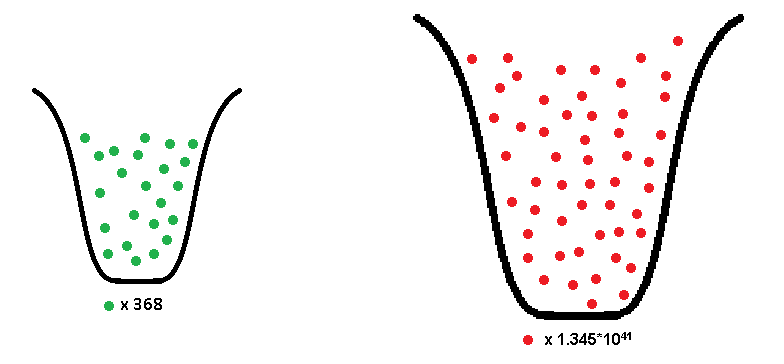
So far what has been conducted is known as a random experiment which meets the following criteria:

1. The exact outcome of each time the experiment is run cannot be known in advance
2. All possible outcomes of the experiment are known in advance
3. The experiment can be repeated under the same conditions

However, in order to quantify the results of each experiment we need to add a fourth requirement to qualify our marble pulls as a what is known as **random variable**: each marble that we place into the new urn needs to have a numeric value assigned. In the case of this study that number will be the average number of slices eaten at each of the 25 pizzerias from the pulled from the first urn. Now to recap what we have:

**Urn 1:** A “parent” urn that has the name and address of each pizzeria in Brooklyn written on it

**Urn 2**: The average number of slices sold at each possible combination of the pizzerias.



While it isn’t inherently apparent, these two urns have something in common other than that they are both full of marbles. If someone were to go to every one of the 368 pizzerias listed in the first urn on the same day and calculate the average number of slices sold at those pizzerias, the grand average would be equal to the grand average of all of the values in urn #2. This relationship is what constitutes the sampling method as a form *unbiased estimation*.

The direct projection method is an extension of this approach and thus it is unbiased under the condition the restaurants in each strata are selected at random, just as our pull of 25 marbles are random sample. Since each strata will contain a different number of observations, this proportion can be adjusted when the total population is calculated by weighing each strata by the proportion of the total represented by each strata.

The urn model is also useful for introducing another concept that will be introduced later in this paper, the confidence interval. Using the formula  we can create a confidence interval based on a single sample of 25 restaurants. The result of the formula is an interval with a max number and a min number that constitutes our interval. If the data points used to create this interval are reflective of the population, and this population is distributed somewhat normally, then we can be confident (or a different percentage can be assigned depending on the application of the study) that 95% of the total possible combinations that we can pull out of urn #1 will have an average between the max and min of our confidence interval.

(e) Sample Size

In order to determine the sample size the amount of uncertainty that is acceptable in this study needs to be explored and defined. The amount of uncertainty is subjective as it needs to be amount that would be acceptable to people who will utilize the results of this study. While getting rid of all uncertainty seems like an attractive proposal, this will be extremely expensive and likely impossible. To determine this number, let’s use the example of asking a pizzeria how many slices they have sold in a day. If the manager were able to tell you that number, give or take a 2 pizzas (16 slices) then that would be acceptable to most people to assume that this is a fairly well-run business that keeps acceptably accurate records. While this seems like a small number for a day, it is important to consider that the study seeks to estimate the number slices sold at every pizzeria, for every day over the course of an entire whole year. If each of the 1353 pizzerias in the city were off by four whole pies every day for a whole year then that would be almost 8,000,000 slices, almost enough to give every resident of the city a free slice of pizza.

While this seems like a large number we can compare it to the estimated amount of pizza that the average American eats multiplied by the number of people who live New Yorkers to get a better feel for the magnitude of this error. According to one source the average American eats 46 slices in a year[[2]](#footnote-2). While I suspect that the average New Yorker’s pizza consumption might be higher, we’ll multiply this number times the 2010 census population of New York 8.19 million to get 376 million. Now 8,000,000 slices compares to this number seems rather small when compared to 376 million.

We can now use the following equation to determine the sample size, which will be distributed across our strata.



* Where E is our acceptable error for a day = 16
* Sigma our standard deviation based on the sample trail described in part two of the study = 77
* And http://www.isixsigma.com/wp-content/uploads/images/stories/migrated/graphics/zalphaover2.gif?4b7c21 is equal to our critical value of 1.96

We can now calculate that

= 89

This means that 89 pizzerias will need to be sampled across the five boroughs to be 95 percent confident that the sample average will be within 16 slices of the true population average of slices sold at each pizzeria over the course of a day.

(f) Cost

The cost of conducting this survey assumes the following:

* All those sampling will be paid $15 an hour and will be given breaks and a stipend for lunch. The estimated rate will then be $17 per hour.
* Each Pizzeria will be observed for 12 hours on a given day.
* Managerial time will be estimated to be $2000. This assumes that two managers will be paid $25 per hour and will each work 5 full work days

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| n |  | **# of Pizzerias** | **% of total** | **% of sample** | **Cost to Sample** |
| n1 | **BRONX** | 216 | 16.77% | 15 | $2,040.00 |
| n2 | **BROOKLYN** | 368 | 28.57% | 25 | $3,400.00 |
| n3 | **MANHATTAN** | 394 | 30.59% | 27 | $3,672.00 |
| n4 | **QUEENS** | 272 | 21.12% | 19 | $2,584.00 |
| n5 | **STATEN ISLAND** | 38 | 2.95% | 3 | $408.00 |
|  | **Total Sampler costs** | 1288 | 100.00% | 89 | **$12,104.00** |
|  | **Management cost** |  |  |  | **$2,000.00** |
|  | **Total Costs** |  |  |  | **$14,104.00** |

The distribution of strata can be expressed as:

n = n1 + n2 + n3 + n4 + n3

n = N1(.1677) + N2(.2857) + N3(.3059) + N4(.2112) + N3(.295)

n = 216(.1677) + 368(.1677) + 394(.1677) + 272(.1677) + 38(.1677)

n = 15 + 25+ 27 + 19+ 3

n = 89

# Part Two: Pilot Study

(a) Description of Pilot Study

In order to determine information that could be used for planning a large scale study a pilot study was conducted during April and May of 2015. The goal of the pilot study is to collect enough information to establish a sample standard deviation that could be used to conduct a sample and a set of standard operational producers/definitions that could be used for a large scale study. In order to collect data I went to a series of pizzerias in Brooklyn and Manhattan to collect data about how to observe the number of pizza slices sold in New York City pizzerias, (the results of this study are posted in part b of this section).

The biggest difference between the pilot and the proposed study is that I was only able to sit for 15 minutes at each pizzeria due to availability constraints. Since the rate at which the number of pizza slices varies greatly over the course of a day, collecting only a 15 minutes sample is not nearly enough to get a large enough sample that is not influenced by random variation introduces by short term spikes or lulls in sales.

However the experience did afford me the opportunity to develop a set of directions that could be given to those conducting a full study. The following procedures are followed during the sampling procedure:

* Sit in a spot in the restaurant where you can observe the register and count the number of slices ordered.
* If a person orders 2 or more slices, all of these slices are counted
* Any pizza ordered by the person observing sales will not be counted.
* The pizza has to come across the counter to be counted. An order placed but not given to the customer will not be counted.

(b) Data you collected

I performed an initial cluster sample around Baruch and my apartment located in the Park Slope Section of Brooklyn

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Start Time** | **End Time** | **Restaurant** | **Address** | **Borough** | **Slices** |
| 4/16 | 6:58 PM | 7:13 PM | Frank’s Pizza | 127 E 23rd St | Manhattan | 8 |
| 4/17 | 12:14 PM | 12:29 PM | Numero 28 | 137 7th Ave | Brooklyn | 11 |
| 4/18 | 6:00 PM | 6:15 PM | Roma Pizza | 85 7th Ave | Brooklyn | 13 |
| 4/18 | 6:30 PM | 6:45 PM | Pino’s Pizza | 181 7th Ave | Brooklyn | 25 |
| 4/29 | 9:30 PM | 9:45 PM | Antonio’s Pizza | 318 Flatbush | Brooklyn | 14 |
| 5/03 | 11:30 AM | 11:45AM | Corner Pizza | 226 Church Ave | Brooklyn | 8 |
| 5/08 | 2:00 PM | 2:15 PM | 2 Bros Pizza | 395 Flatbush | Brooklyn | 11 |
| 5/12 | 4:45 PM | 5:00 PM | Pizza Plus | 359 7th Ave | Brooklyn | 5 |
| 5/14 | 12:30 PM | 12:45 PM | 99 Cent Fresh | 51 Willoughby St | Brooklyn | 24 |

(c) Results of Trial Run

The data from above can be used to calculate a sample stratum for Brooklyn, which has represents 28.5% of the total number of pizzerias in New York City. The sample of eight pizzerias for 15 minute intervals is only .66% of the total amount of observation time for this stratum in the proposed full study. Additionally, these observations will need to be multiplied by 48 to estimate the total number of slices sold over the course of a day since 15 minutes is too short. Later, the estimated total number of slices will be multiplied by 365 to determine the number sold per year.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Restaurant** | **Borough** | **Slices** | **Est. of slices sold p/d (x48)** |
| 1 | Numero 28 | Brooklyn | 11 | 528 |
| 2 | Roma Pizza | Brooklyn | 13 | 624 |
| 3 | Pino’s Pizza | Brooklyn | 25 | 1200 |
| 4 | Antonio’s Pizza | Brooklyn | 14 | 672 |
| 5 | Corner Pizza | Brooklyn | 8 | 384 |
| 6 | 2 Bros Pizza | Brooklyn | 11 | 528 |
| 7 | Pizza Plus | Brooklyn | 5 | 240 |
| 8 | 99 Cent Fresh | Brooklyn | 23 | 1104 |

In order to calculate strata average, variance, and a 95% confidence interval, we can use the following equations:

First find the average of the samples taken:



Using the direct projection method, the sample average will be multiplied by the total number of pizzerias in Brooklyn (368) to estimate the total number of slices sold:







To extend this projection on to the entire year, we can multiply this number by the number of days in a year:





Next we can calculate the standard deviation of 

can be found using the chart below. The sample have been

adjusted for pizza sales on an annual basis:

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | ^2 |
| 1 | | 192720 | 2537136900 |
| 2 | | 227760 | 235008900 |
| 3 | | 438000 | 37989908100 |
| 4 | | 245280 | 4796100 |
| 5 | | 140160 | 10594584900 |
| 6 | | 192720 | 2537136900 |
| 7 | | 87600 | 24177140100 |
| 8 | | 420480 | 31467212100 |
|  | = | | 1.09543E+11 |

125,095

Next we can determine the variance for the entire population of Brooklyn Pizzerias using the following formula:



 =

Now that variance has been determined, the 95% confidence interval can be created using the following equation:

,

Where  is a t-interval for n-L (8-1=7) degrees of freedom = 2.365



**Lower Confidence Interval:**

 = 50,585,637

**Upper Confidence Interval:**

= 126,716,763

To Summarize:

|  |  |
| --- | --- |
| ybar | 240900 |
| stdev | 125095 |
| Ni | 368 |
| Tauhat | 88651200 |
| Sample size | 8 |
| EstVar(Tauhats) | 259143449454000.00 |
| Low Conf | 50,585,637 |
| Upper Conf | 126,716,762 |

**Conclusion**

Our sample concludes that by using direct method its estimated 89 million slices of pizza are sold in Brooklyn in a year. Given the level of sampling we are ready to perform, we can be 95% confident that the true total is between 50 million and 120 million.

To give a rough number to the goal of answering the question of how many slices of pizza are sold in a year in New York City, we can extend our estimate for Brooklyn using the direct projection method. This approach makes the very big assumptions that pizza consumption in Brooklyn is representative of the city as a whole, which has yet to be determined.

By going back the proportions of pizzeria distribution established in part section f or part 1, we know that Brooklyn contains 28.57% of all of NYC’s pizzerias. Assuming that 89 million slices represents 28.57% of we can set up the following proportion to estimate that New York City eats 331 million slices of pizza per year

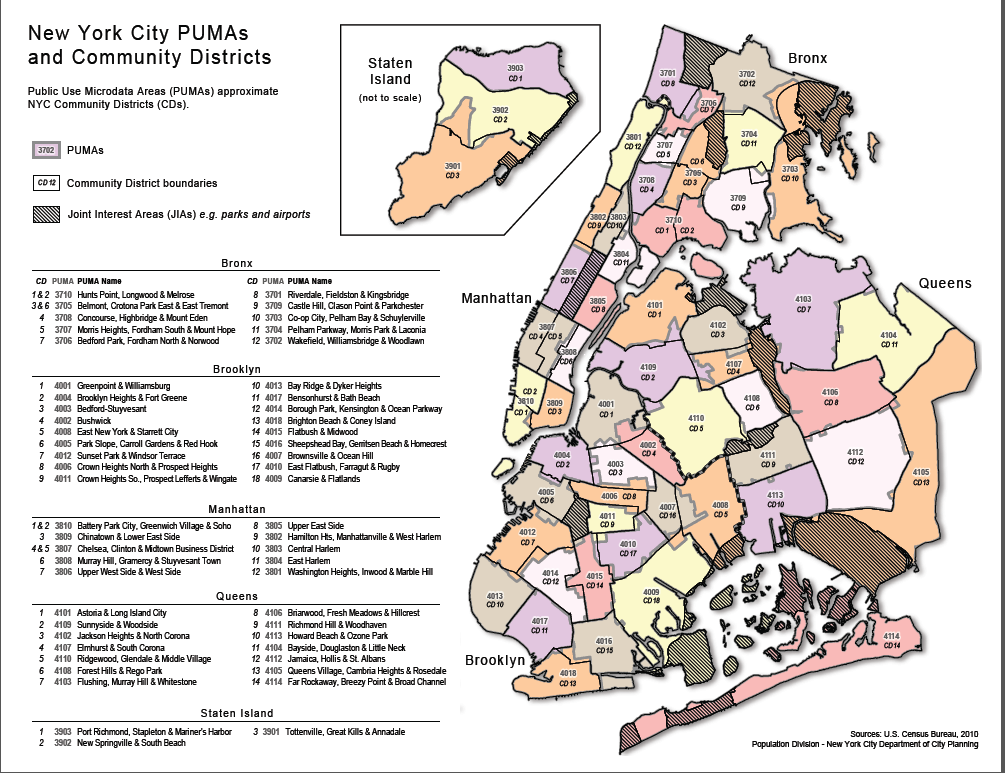
(d) Sources of Bias

Weekend vs weekday consumption and seasonality are concerns about the sampling process that could potentially introduce bias. Without conducting a large study beforehand it is difficult to know the balance between how much pizza is eaten in Manhattan, where people from all over the tri-state area commute in for work, vs the weekend when there are fewer people in the borough. Because of this characteristic, using the direct estimation method to calculate total consumption with data that has been collected on weekdays may artificially inflate the estimated total. Also if the data is collected during summer holidays or during summer vacation months the estimated amount may be lower than the actual average.

Location of the samples taken within Manhattan may also introduce bias. Commercially zoned areas that experience little tourist traffic during the weekend likely get very little business over the weekend if they are even open at all. Areas such as the Finical District, Metro Tech in Brooklyn and commercially zoned parts of midtown will likely contain higher numbers of pizzerias that are only open during lunch hours.

Finally, pizza is severed a number of different ways in restaurants across New York City and these discrepancies represent a possible source of bias to this study. For instance, pizza slices like nearly all non-prepackaged foods, come in a variety of different sizes. In addition to serving triangular slices that are larger or smaller than most establishments, some pizzerias serve pizza in squares or rectangular shapes. Additionally, some pizzerias serve pizza by the entire pie instead of by the slice and it is difficult to determine the exact number of restaurants that follow this practice. For instance, pizzerias such as Lombardi’s in Little Italy, John’s on Bleeker Street, and Grimaldi’s in Brooklyn are extremely popular tourist destinations with long lines that only serve pizza by the entire pie. Currently this study only seeks to understand consumption by the slice so these establishments introduce bias.

# Part Three: Data Appendix

****

1. U.S. Department of Agriculture. Consumption of Pizza. 2014 http://www.ars.usda.gov/SP2UserFiles/Place/80400530/pdf/DBrief/11\_consumption\_of\_pizza\_0710.pdf [↑](#footnote-ref-1)
2. U.S. Department of Agriculture. Consumption of Pizza. 2014 http://www.ars.usda.gov/SP2UserFiles/Place/80400530/pdf/DBrief/11\_consumption\_of\_pizza\_0710.pdf [↑](#footnote-ref-2)