Pizza Consumption in New York City

A Dietary Survey using Stratified Sampling

Theodore Stetzel

Baruch College – STA 9710

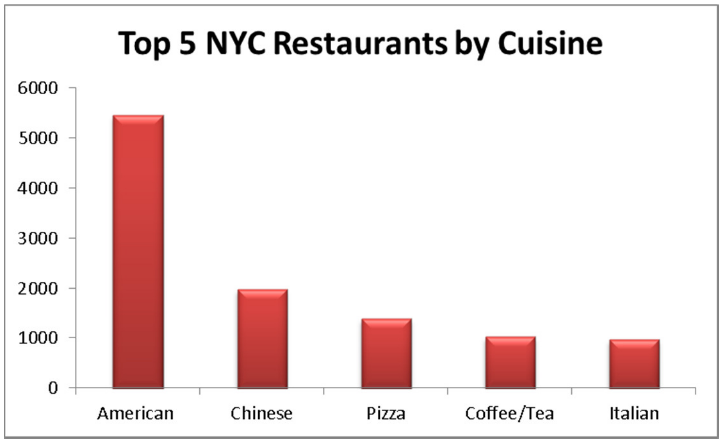
Spring 2015

Contents

Cover image by Seattle Pi http://www.seattlepi.com/ - staff photographer

Part One: Proposed Survey

(a) Subject of Survey



Pizza is one of the most popular foods in America. On any given day an average of about 13% of Americans eat pizza and as a result pizza is recognized as a major contributor of nutrients in the American diet[1]. 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.

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 could be useful to government agencies that study food consumption such as the USDA. Since pizza is more popular with children than it is adults, 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 likely be interested in the results of the study in order to examine consumer behavioral trends in America's largest restaurant market.

This paper examines sampling methods and then establishes a framework that can be used to execute a study large enough to estimate all pizza by slice consumption in New York City. To begin to build this framework, the paper will first answer the "where" aspects of a potential study and then examine the "what" and "how" of the study.

[1] U.S. Department Of Agriculture. What We Eat: Pizza. 2014. http://www.ars.usda.gov/SP2UserFiles/Place/80400530/pdf/DBrief/11_consumption_of_pizza_0710.pdf

The “where” part starts by creating what is called the **sampling frame** or in the case of this study a list of every pizzeria’s name and address. The sampling 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 it can begin operations and then conducts at least one unannounced inspection annually to ensure that proper practices for food preparation and vermin control are followed. Since these inspections are required by law, the DOHMH list on the NYC open data website can be considered reliable and up to date. Restaurants that meet 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
- Must have had a passing inspection conducted in the last 365 days
- Must have chosen either Pizza or Italian/Pizza as the cuisine type

Using these rules results in a population of 1353 restaurants and can be downloaded here in excel format: https://www.dropbox.com/s/da5srjft7idkwpb/frame.xlsx?dl=0

The “what” and “how” aspects of the study will now be defined. The population parameter that will be observed is the number of pizza slices that are sold at each restaurant. This will be then used to estimate the total number of slices sold in the city in a single year. 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 every pizzeria in NYC for a whole year and count the number of slices that are sold every day, sampling methods 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 is a fairly simple method that could be used to collect sample information from the population of pizzerias in order to estimate the total number of slices ordered in NYC per year. The approach simply takes a list of all pizzerias in New York City and selects restaurants to sample from this list at random. In order to use SRS, first we would establish a percentage of restaurants we would want to sample, let's say use a simple 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) from which we would conduct our sampling study. We would then send out samplers to these pizzerias who would observe pizza sales 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 number of combinations of restaurants that we can sample from, taking into consideration that the order in which they are selected is not of concern to the study, the following formula can be used (where n=68 N=1343):

$$\text{nPn} = \frac{\text{N!}}{(\text{N}-\text{n})!} = \frac{{}_{1343}\text{P}_{68}}{68!} = \frac{1343\cdot 1342\cdot \ldots \cdot 1277\cdot 1276}{68\cdot 67\cdot \ldots \cdot 2\cdot 1} = 3.67*10^{115},$$

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 since this method requires going to 68 restaurants which could be very far apart. If we wanted to conduct the entire study on the same day, we would need to coordinate 68 people to arriving at each the specified locations. This would be a very challenging task from a managerial standpoint considering each person would need to get their own and there would be little to no management support (e.g. coordinate breaks, answer questions).

Cluster Sampling

In order to save money, a method of sampling known as cluster sampling could be utilized. Cluster sampling involves sampling from "clusters" of data points and is done by classifying what we will sample in larger groups of data points that share some commonality (typically geographic). For this study we could possibly 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 NYC Community Districts is included in the appendix of this document). Our team of samplers would then go out and sample at every pizzeria within that community 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 community districts. Either of the cluster approaches would be advantageous from a managerial standpoint since a single supervisor could check on all of the surveyors and a system could be set up to give them a break if needed.

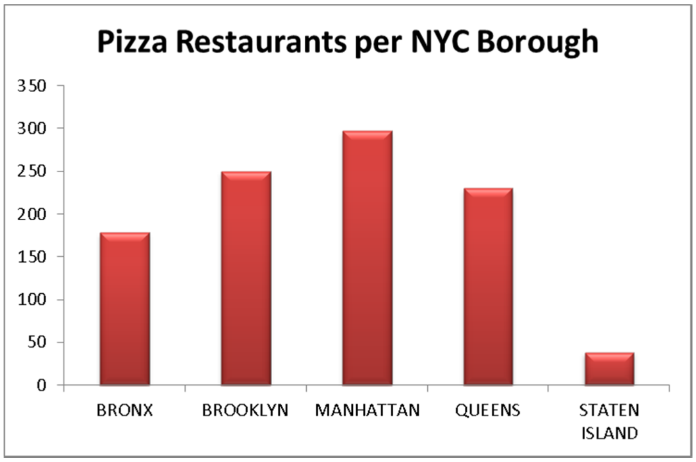
However this method is not without a drawback. Cluster sampling could introduce bias if the clusters chosen are not representative of the population as a whole. This type of sampling is also best used when the sampling frame is not well established and since we are afforded extra information form the 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 know some of the characteristics of where people eat pizza, we can use stratified sampling which can be considered a "best of both" hybrid solution. The DOHMH website information discussed earlier in this paper can be used to determine the number pizza restaurants in each borough. We can this by-borough create pre-determined groups called strata.

Additionally, if we wanted to get more specific in hopes to reducing variance, our stratum could be broken out by time of day. 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 the survey during these times. However this approach 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 stratified *random sampling will be utilized* and *each borough will represent a strata.* This is done 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 (the population total a.k.a. $\hat{\tau}$). There are three popular methods that can be used to estimate characteristics of a total population based on a smaller sampling of data.

Indirect Method

The indirect method 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 estimation 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, an example of this method 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 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 could be useful in this particular study considering that pizza sales likely vary greatly at different times of the day. This approach excluded this method from being called an unbiased estimator (which will be discussed further in second d) but it is still usefull in certain circumstances. Although both the indirect and ratio approaches could potentially 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. This method 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: $\hat{\tau} = N\ \bar{y}$, 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 $\bar{y}$ is average of all samples collected, expressed as $\bar{y} = (1/n)\sum_{j=1}^{n} y_i$, where n is the number of samples collected. This approach is considered unbiased if the restaurants chosen are done so in a random fashion. In the case where stratified sampling is used, the restaurants inside each stratum 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 explain a complex concept. For a moment, use your imagination to picture a large urn in an empty room and that this urn could be filled with marbles in order to simulate the random selection process used in sampling.

For this study, pizzerias in each borough of New York City area will be sampled. In order to sample using the stratified method, we'll start with Brooklyn and place a marble in this urn for each pizzeria in the borough. To do so, we'll take 368 brand new marbles and write the name and address of each pizzeria onto the marble (using a very small pen) and then place it in the urn.

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

$$nPn = \frac{N!}{(N-n)!} = \frac{{}_{368}P_{25}}{25!} = \frac{368*367*...*344*343}{25*24*...*2*1} = 1.345*10^{41}$$

This large number represents the total number of possible combinations of picking 25 from the total number of 368 pizzerias in Brooklyn or the total number of different combinations we can pull out of the urn for this stratum. So far what we have conducted with our imaginary urns is known as a **random experiment** because it meets the following criteria:

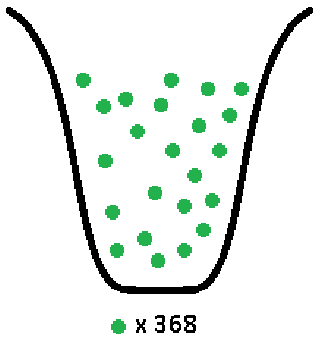
1. The exact outcome of each time the experiment is run cannot be known in advance
2. All possible outcomes are known in advance
3. The experiment can be repeated under the same conditions

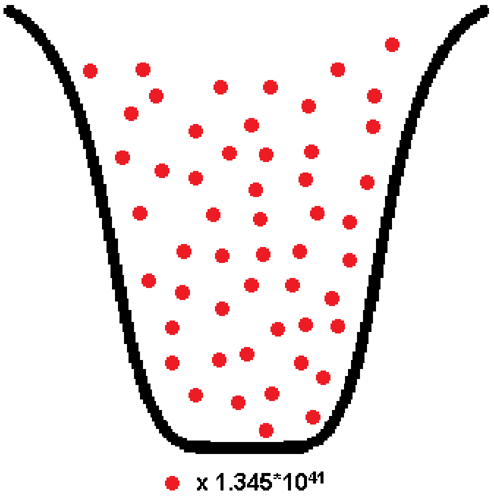
However, in order to make is so that these samples can be used to perform estimates of a larger population, we need to add a fourth requirement in order to qualify our marble pulls as a **random variable**. Each marble that we place into the new urn needs to have a numeric value assigned. In the case of this study, the grade average number of slices eaten at each of the pizzerias from the pulled from the first urn. In order do this, we could bring in a second urn (called a daughter or son urn) that will hold these the results of this experiment. The process of filling the second urn will consist of pulling each of the possible $1.345*10^{41}$ combinations of marbles from the first urn and then writing the results onto another marble which is then put into our second urn.

Now to recap what we have in terms of our imaginary urns:

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, there 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 grand average of slices sold at those pizzerias, that number 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 if the restaurants in each strata are selected at random, (just as marbles are pulled from urn #1). 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 (this break down is provided in section f)

The urn model is also useful for getting acquainted with another concept that will be introduced later in this paper, the idea of a **confidence interval**. Using the formula $\hat{\tau} \pm t_{n-1}\hat{\sigma}_{\hat{\tau}}$ 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. 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 (of a different value depending on the application of our 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

Sample size is important because it determines the cost and the usefulness of the study. Since there is no way to hit the exact number of slices sold in New York City right on the nose, we'll need to provide a range that our estimated total falls into. If this range is extremely large, our study will not seem very useful to our intended audience. In order to get this range to be very small, an extortionate number of samples will need to be taken and will likely make this study too expensive to conduct. In order to determine the sample size, first the amount of uncertainty that is acceptable in this study needs to be defined.

To determine a practical value for this range, 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 2 whole pizzas (16 slices), then that would be probably be acceptable to most people to assume that this is a fairly well-run business that keeps reasonably accurate records of their sales given that most pizzerias make over 100 pies a day. While this seems like a small error, 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 a whole year. If each of the 1353 pizzerias in the city were off by 2 whole pies every day for a whole year then that would be almost 8,000,000 slices lost.

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] 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. Given that the error range of 8 million is just over 2% of estimated total of 376 million our estimate seems to be an acceptable amount of error in scale and believability from an operational standpoint.

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

$$n = \left[\frac{z_{\alpha/2}\sigma}{\mathrm{E}}\right]^2$$

- 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 $z_{\alpha/2}$ is equal to our critical value of 1.96

We can now calculate that

$$\left\{\frac{1.96*77}{16}\right\}\text{\textasciicircum}2 = 89$$

[2] U.S. Department Of Agriculture. What We Eat: Pizza. 2014. http://www.ars.usda.gov/SP2UserFiles/Place/80400530/pdf/DBrief/11_consumption_of_pizza_0710.pdf

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	pizzerias	Cost
n_1	**BRONX**	216	16.77%	15	$2,040.00
n_2	**BROOKLYN**	368	28.57%	25	$3,400.00
n_3	**MANHATTAN**	394	30.59%	27	$3,672.00
n_4	**QUEENS**	272	21.12%	19	$2,584.00
n_5	**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:

$n = n_1 + n_2 + n_3 + n_4 + n_3$

$n = N_1(.1677) + N_2(.2857) + N_3(.3059) + N_4(.2112) + N_3(.295)$

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

$n = 15 + 25 + 27 + 19 + 3$

$\mathbf{n = 89}$

Part Two: Pilot Study

(a) Description of Pilot Study

In order to determine standard operational producers and definitions and to also determine metrics such as a sample standard deviation that could be used for a large scale study, a pilot study was conducted during the months of April and May in 2015. In order to collect data, I went to pizzerias in Brooklyn and Manhattan to collect data about how to also observe the number of pizza slices sold in New York City pizzerias. The results of this experiment are described in section b.

Over the course of watching the pizza selling process, I noticed some aspects of the observation process that I did not initially expect to see when I conceptualized the study. For instance pizza is sold in different ways. For instance, some pizzerias offer square slices in addition to traditional slices. Many also sell non-pizza items such as garlic knots, sandwiches and pastas. The experience afforded me the opportunity to develop a set of directions that could be given to those conducting a full study.

These standard operational producers include:

- 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 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.

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 my own availability constraints. Since the rate at which the number of pizza slices sell 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 somewhat influenced by random variation introduced by short term spikes or lulls in sales at each restaurant.

(b) Data Collected

I performed an initial sample around Baruch College 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	11:45 AM	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 the table above can be used to calculate a stratum for Brooklyn, which represents 28.5% of the total number of pizzerias. It should be noted that this is an extremely small data set as the eight pizzerias were observed for 15 minute intervals is only .66% of the total amount of observation time for this stratum in the proposed full study. Despite the incompleteness of the sample, we'll use this data set to conduct some initial estimates of total sales. To make up for the fact that these observations are less than a full day's worth of data, these numbers will need to be multiplied by 48 to estimate the total number of slices sold over a 12 hour day. Later, the estimated total number of slices will be multiplied by 365 to determine the number sold per year.

Index	Restaurant	Borough	Slices	Est. sold per day (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 the strata average, strata variance, and a 95% confidence interval, we can use the following equations:

First find the average of the samples taken in Brooklyn:

$$\bar{y}_{BK} = \frac{528 + 624 + 1200 + 672 + 384 + 528 + 240 + 1104}{8} \approx 660$$

This means that based on this sample it can be can estimated that each pizzeria in Brooklyn sells about 660 slices every day. 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:

$$\hat{\tau}_{BK} = N_{BK}\bar{y}_{BK}$$

$$\hat{\tau}_{BK} = 368(660)$$

$$\hat{\tau}_{BK} = 242{,}880$$

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

$$\hat{\tau}_{BK} = 242{,}880(365)$$

$$\hat{\tau}_{BK} = 88{,}651{,}200$$

This calculation concludes that it can be estimated that around 88 million slices are sold in Brooklyn in a single year.

Next we can calculate the standard deviation of $s_{BK}{}^2$ using a standard deviation formula

$$s_{BK}{}^2 = \sqrt{\frac{\sum_{j1}^{n1}(y_{lj}-\bar{y})^2}{n-1}}$$

$\sum_{j=1}^{n_1}\left(y_{1j}-\bar{y}_{BK}\right)^2$ can be found using the chart below. The sample have been adjusted for pizza sales on an annual basis:

	y_{BKj}	$\left(y_{BKj}-\bar{y}_{BK}\right)$^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
$\sum_{j=1}^{n_1}\left(y_{1j}-\bar{y}_1\right)^2 =$		1.09543E+11

We can then determine that:

$$s_{BK}{}^2 = \sqrt{\frac{1.09543\text{E}+11}{8-1}} = 125{,}095$$

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

$$\hat{V}(\hat{\tau}_{BK}) = N_{BK}{}^2\left(\frac{N_{BK}-n_{BK}}{N_{BK}}\right)\frac{s_{BK}{}^2}{n_{BK}}$$

$$\hat{V}(\hat{\tau}_{BK}) = 368^2\left(\frac{368-8}{368}\right)\frac{125{,}095}{8} = 259{,}143{,}449{,}454{,}000$$

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

$$\hat{\tau}_{BK} \pm t_{n-L}\sqrt{\hat{V}(\hat{\tau}_{BK})},$$

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

$$88651200 \pm 2.365\sqrt{259{,}143{,}449{,}454{,}000}$$

Lower Confidence Interval:

$$88651200 - 2.365\sqrt{259,143,449,454,000} = 50,585,637$$

Upper Confidence Interval:

$$88651200 + 2.365\sqrt{259,143,449,454,000} = 126,716,763$$

To Summarize:

ybar	240900
stdev	125095
Ni	368
Tauhat	88,651,200
Sample size	8
EstVar(Tauhats)	259143449454000.00
Low Conf	50585637
Upper Conf	1267167623

Conclusion:

Based on the calculations performed above, it can be estimated that the total number of slices sold in Brooklyn in a single year is around 89 million and it can be stated that we are 95% confident that the true total number of slices sold is between 50 and 127 million based on our confidence interval.

We could also use this number to estimate the total number of slices sold in New York City. Referring back to our break down of pizzerias by borough, we know that 28.6% of all pizzerias in New York City are located in Brooklyn. Assuming that pizza consumption is the same across the city (which has not ben proven), we can estimate the other 71.4% of the city's pizza sales using the Brooklyn's estimated total of 89 million. To demonstrate this I've set up the following proportion:

$$\frac{89}{.286} = \frac{331}{1}$$

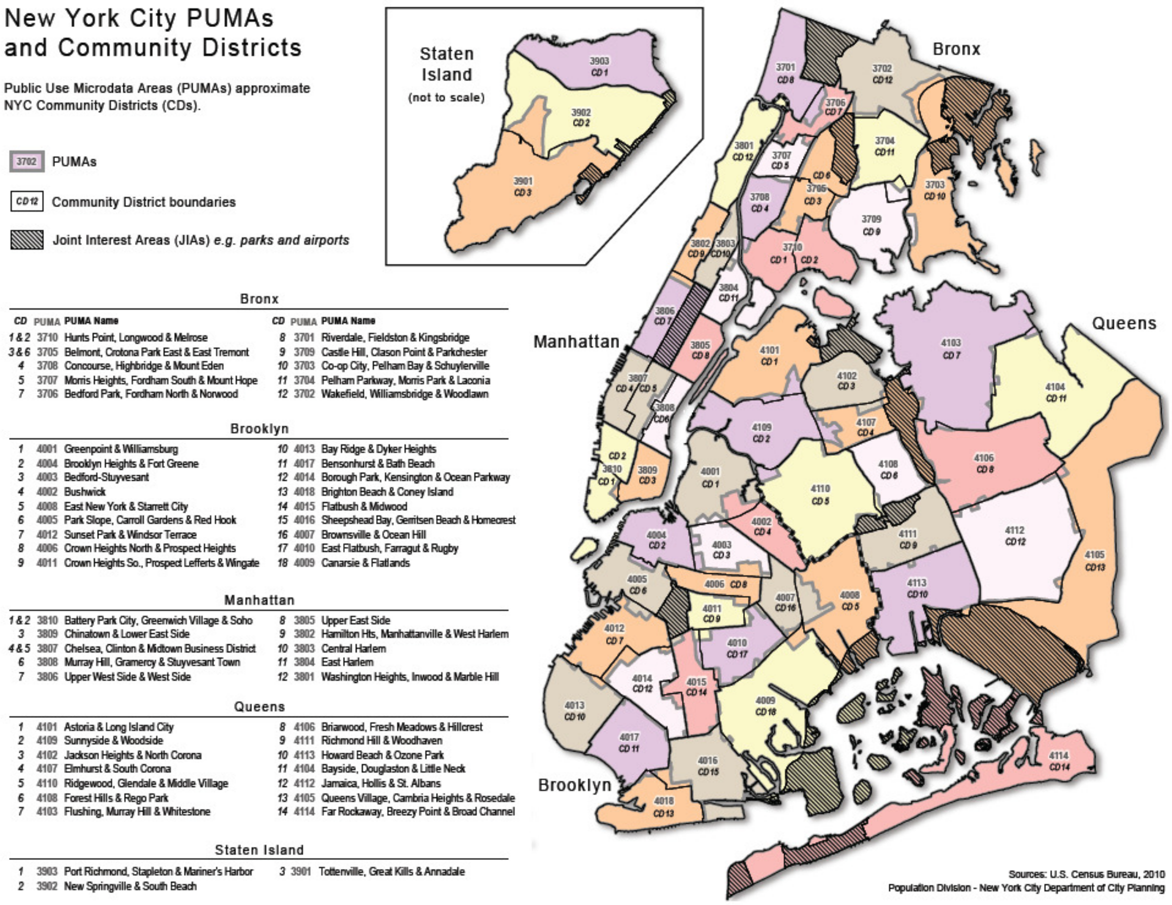
to estimate that approximately 331 million slices are sold in New York City every year. This seems to be in the same general area as the estimate of 376 million slices that I put together using the information provided by the USDA in section e of part 1 of this study. While this may be totally by chance, it is of interest for future studies.

(d) Sources of Bias

Given that the number of slices sold can vary greatly depending on the time of day, week or year, the timeframe in which study is conducted could introduce unwanted bias. Weekend vs weekday consumption and seasonality are of the most concern. Without conducting a large study beforehand it is difficult to know the balance between how much pizza is eaten in Manhattan on a weekday, when people from all over the tri-state area commute in for work and leave in the evening, vs the weekend, when there are fewer people in the borough. Because of this characteristic of the borough, 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 holidays or during summer vacation months the estimated amount may be lower than the actual amount. Location of the samples taken within Manhattan may also introduce bias due to the daily influx of commuters on weekdays. Commercially zoned areas that experience little tourist traffic during the weekend are likely to get very little business over the weekend if they are even open at all. These areas include neighborhoods such as the Finical District, Metro Tech in Brooklyn and office heavy parts of midtown likely contain higher numbers of pizzerias that are only open during lunch hours on weekdays.

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 food served in restaurants, come in a variety of different sizes and shapes. In addition to serving tradational triangular slices that are larger or smaller than most establishments, some pizzerias serve pizza in squares or rectangular shapes. Additionally, some pizzerias incuded in the sampling frame only serve pizza by the entire pie instead of by the slice and thus 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 not having a way to exclude these establishments from the study introduces a potential source of bias.

Part Three: Data Appendix



New York City PUMAs and Community Districts

Public Use Microdata Areas (PUMAs) approximate NYC Community Districts (CDs).

3702	PUMAs
CD 12	Community District boundaries
▨	Joint Interest Areas (JIAs) *e.g. parks and airports*

Bronx

CD	PUMA	PUMA Name	CD	PUMA	PUMA Name
1 & 2	3710	Hunts Point, Longwood & Melrose	*8*	3701	Riverdale, Fieldston & Kingsbridge
3 & 6	3705	Belmont, Crotona Park East & East Tremont	*9*	3709	Castle Hill, Clason Point & Parkchester
4	3708	Concourse, Highbridge & Mount Eden	*10*	3703	Co-op City, Pelham Bay & Schuylerville
5	3707	Morris Heights, Fordham South & Mount Hope	*11*	3704	Pelham Parkway, Morris Park & Laconia
7	3706	Bedford Park, Fordham North & Norwood	*12*	3702	Wakefield, Williamsbridge & Woodlawn

Brooklyn

1	4001	Greenpoint & Williamsburg	*10*	4013	Bay Ridge & Dyker Heights
2	4004	Brooklyn Heights & Fort Greene	*11*	4017	Bensonhurst & Bath Beach
3	4003	Bedford-Stuyvesant	*12*	4014	Borough Park, Kensington & Ocean Parkway
4	4002	Bushwick	*13*	4018	Brighton Beach & Coney Island
5	4008	East New York & Starrett City	*14*	4015	Flatbush & Midwood
6	4005	Park Slope, Carroll Gardens & Red Hook	*15*	4016	Sheepshead Bay, Gerritsen Beach & Homecrest
7	4012	Sunset Park & Windsor Terrace	*16*	4007	Brownsville & Ocean Hill
8	4006	Crown Heights North & Prospect Heights	*17*	4010	East Flatbush, Farragut & Rugby
9	4011	Crown Heights So., Prospect Lefferts & Wingate	*18*	4009	Canarsie & Flatlands

Manhattan

1 & 2	3810	Battery Park City, Greenwich Village & Soho	*8*	3805	Upper East Side
3	3809	Chinatown & Lower East Side	*9*	3802	Hamilton Hts, Manhattanville & West Harlem
4 & 5	3807	Chelsea, Clinton & Midtown Business District	*10*	3803	Central Harlem
6	3808	Murray Hill, Gramercy & Stuyvesant Town	*11*	3804	East Harlem
7	3806	Upper West Side & West Side	*12*	3801	Washington Heights, Inwood & Marble Hill

Queens

1	4101	Astoria & Long Island City	*8*	4106	Briarwood, Fresh Meadows & Hillcrest
2	4109	Sunnyside & Woodside	*9*	4111	Richmond Hill & Woodhaven
3	4102	Jackson Heights & North Corona	*10*	4113	Howard Beach & Ozone Park
4	4107	Elmhurst & South Corona	*11*	4104	Bayside, Douglaston & Little Neck
5	4110	Ridgewood, Glendale & Middle Village	*12*	4112	Jamaica, Hollis & St. Albans
6	4108	Forest Hills & Rego Park	*13*	4105	Queens Village, Cambria Heights & Rosedale
7	4103	Flushing, Murray Hill & Whitestone	*14*	4114	Far Rockaway, Breezy Point & Broad Channel

Staten Island

1	3903	Port Richmond, Stapleton & Mariner's Harbor	*3*	3901	Tottenville, Great Kills & Annadale
2	3902	New Springville & South Beach			

Sources: U.S. Census Bureau, 2010
Population Division - New York City Department of City Planning