Please Stop Calling

A Project 1 Analysis

I. Executive Summary

I seek to reduce the marginal cost of each customer acquisition by understanding the threshold where repeated call to the same individual no longer is reasonable. I fit models on a dataset of telemarketing calls for banks selling long-term deposits and determine the heterogeneity of the population. The resulting distribution of the number of calls before a contact was converted and the respective probabilities allows me to understand the likelihood the next call will result in a conversion. I also fit a model based on the specific attribute of education to uncover the propensities of different segments of our population. I use a Shifted Negative Binomial Distribution (NBD) to discover the parameters of the data. This model validated my hypothesis that the converted population has low heterogeneity, though higher than in the total population, and will need significantly fewer calls on average to convert than were dialed to the entire population.

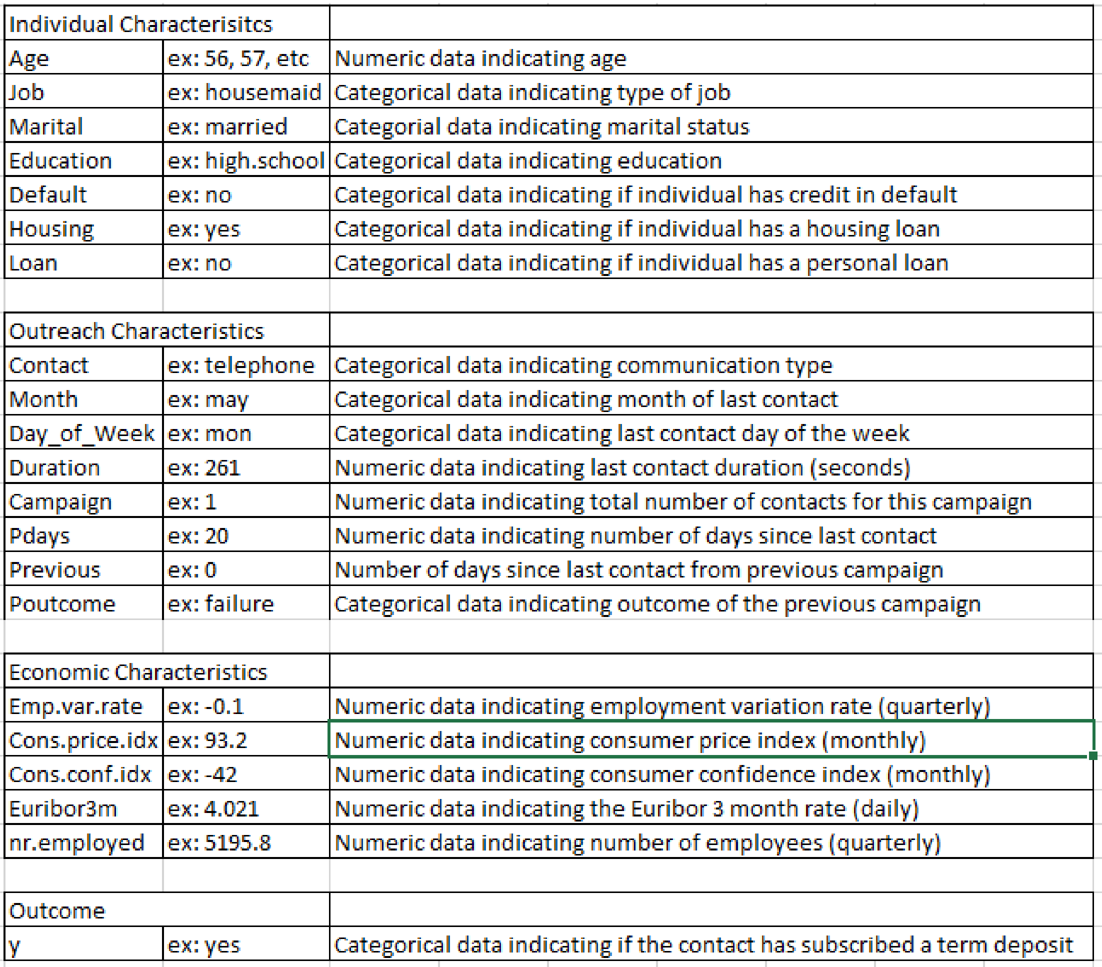
II. Managerial Problem

Any marketing campaign seeks to lower their customer acquisition cost, especially when their efforts constitute direct marketing utilizing a resource that does not scale well (i.e. people). By understanding the probability distribution of the likelihood of a conversion as the number of calls for a customer increases, firms will have valuable data to place an upper bound on their calls and no longer pursue low yield populations. If the probability that the 11th call to the same individual has an infinitesimally small chance of conversion, it likely is not reasonable to make that call. I examine the entire data population and a specific attribute of the population to determine how the probability distribution and propensities change across different populations.

III. Data Set Description

I use a data set collected from a Portuguese retail bank with 41,118 individuals with 20 attributes and the outcome of the outreach. Though the dataset only contains a minor number of successes (11.27%), I believe the number of converted individuals (4640) is significantly large enough to produce a sound analysis.

Each individual in the data set has attributes related to their own characteristics, outreach characteristics, and economic characteristics of the time. This is displayed in the following table:



Individual Characterisitcs		
Age	ex: 56, 57, etc	Numeric data indicating age
Job	ex: housemaid	Categorical data indicating type of job
Marital	ex: married	Categorial data indicating marital status
Education	ex: high.school	Categorical data indicating education
Default	ex: no	Categorical data indicating if individual has credit in default
Housing	ex: yes	Categorical data indicating if individual has a housing loan
Loan	ex: no	Categorical data indicating if individual has a personal loan
Outreach Characteristics		
Contact	ex: telephone	Categorical data indicating communication type
Month	ex: may	Categorical data indicating month of last contact
Day_of_Week	ex: mon	Categorical data indicating last contact day of the week
Duration	ex: 261	Numeric data indicating last contact duration (seconds)
Campaign	ex: 1	Numeric data indicating total number of contacts for this campaign
Pdays	ex: 20	Numeric data indicating number of days since last contact
Previous	ex: 0	Number of days since last contact from previous campaign
Poutcome	ex: failure	Categorical data indicating outcome of the previous campaign
Economic Characteristics		
Emp.var.rate	ex: -0.1	Numeric data indicating employment variation rate (quarterly)
Cons.price.idx	ex: 93.2	Numeric data indicating consumer price index (monthly)
Cons.conf.idx	ex: -42	Numeric data indicating consumer confidence index (monthly)
Euribor3m	ex: 4.021	Numeric data indicating the Euribor 3 month rate (daily)
nr.employed	ex: 5195.8	Numeric data indicating number of employees (quarterly)
Outcome		
y	ex: yes	Categorical data indicating if the contact has subscribed a term deposit

This data set was compiled by S. Moro, P. Cortez and P. Rita in their paper "A Data-Driven Approach to Predict the Success of Bank Telemarketing" and hosted on the University of California, Irvine Machine Learning Repository.

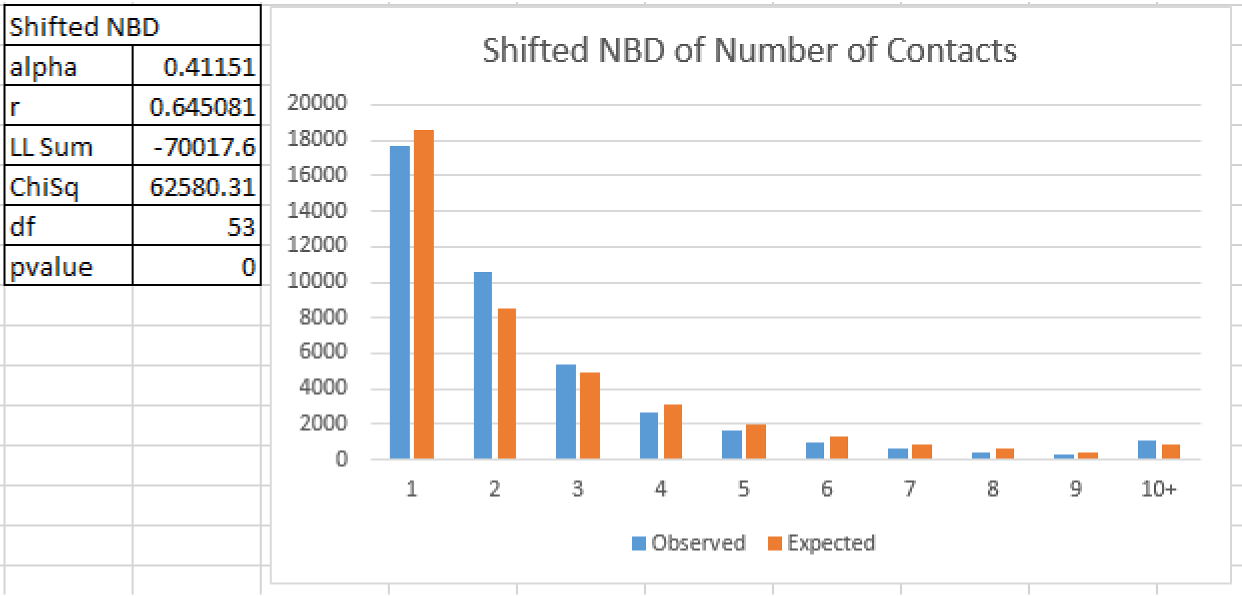
IV. Hypothesis

I believe converted contacts will have higher heterogeneity than the overall population and the heterogeneity of the contacts that refused. Intuitively, I believe most individuals will subscribe to a term deposit within the first few calls if they want a term deposit. Those seeking a term deposit will subscribe nearly immediately, and those uncertain about a term deposit with come to a conclusion within the next couple of calls. After a certain threshold of calls, the remaining population will mostly contain contacts who will never accept a term deposit, especially if they have been negatively impacted by the volume of calls. This will result in nearly all converted contacts distributed in the first few "bins" and extremely few converted contacts, if any, forming a long tail.

V. Analysis

I initially began modeling the entire data set to determine the overall heterogeneity of the population. As I had data that did not contain any "zeroes" (as we have no data on the contacts we didn’t contact) I could not use the standard NBD model and had to decide between the Shifted NBD Model and the Truncated NBD Model. Ultimately, I opted for the Shifted NBD Model as introducing and estimating the “zeroes” in the data was not relevant to the managerial problem I was solving. The quantity and probability of receiving zero calls would not be useful in determining the maximum number of calls one should dial each contact. We would only know the total population we had not reached.

The parameters of my Shifted NBD Model and my bar graph are below:

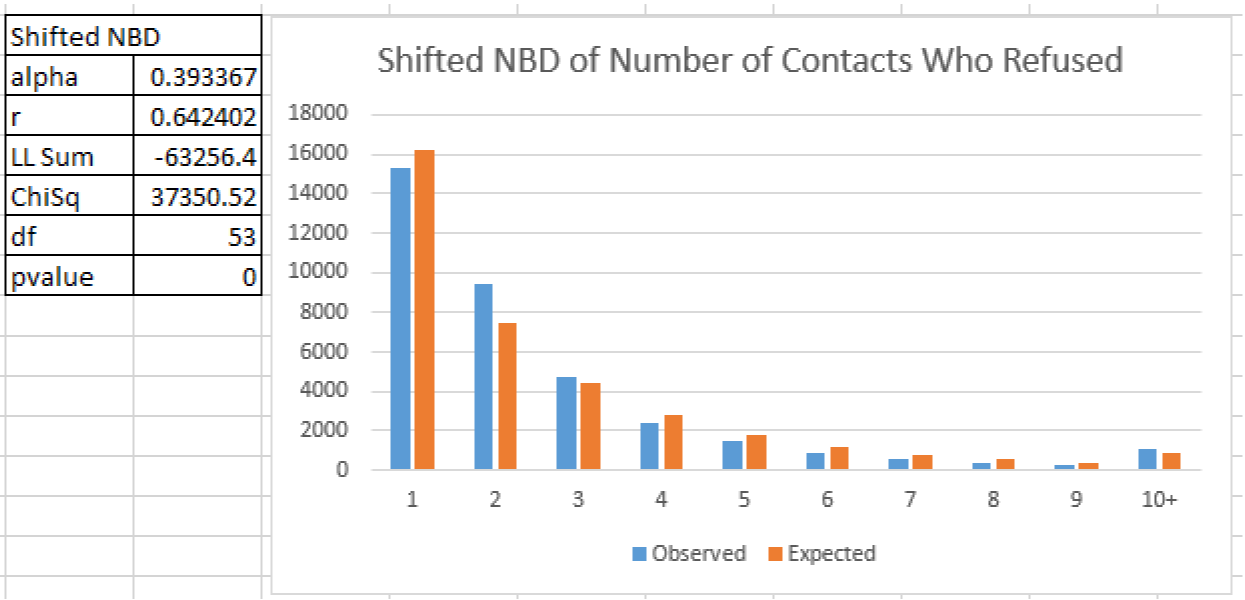


Shifted NBD	
alpha	0.41151
r	0.645081
LL Sum	-70017.6
ChiSq	62580.31
df	53
pvalue	0

*The buckets represent the number of calls each person received. I have binned the data set only graphically, not in my analysis, to provide a clearer picture of my results. The maximum number of calls a person received was 56. They still said no.

We see that we have an $r < 1$, and thus the distribution has a steep peak at 0 and a heavy tail and an *alpha* < 1, and thus a scale that stretches the distribution (as we observe in the bar graph). The implication of this is that the entire population has a relatively low heterogeneity, and most contacts are only called a few times. This intuitively makes sense, as most marketers will only be only be so persistent.

I then ran a Shifted NBD Model on the contacts that refused the term deposit to ensure my calculations were correct. The refusals should have roughly the same distribution as the whole population as they represent the majority of the population, and they do as shown below:

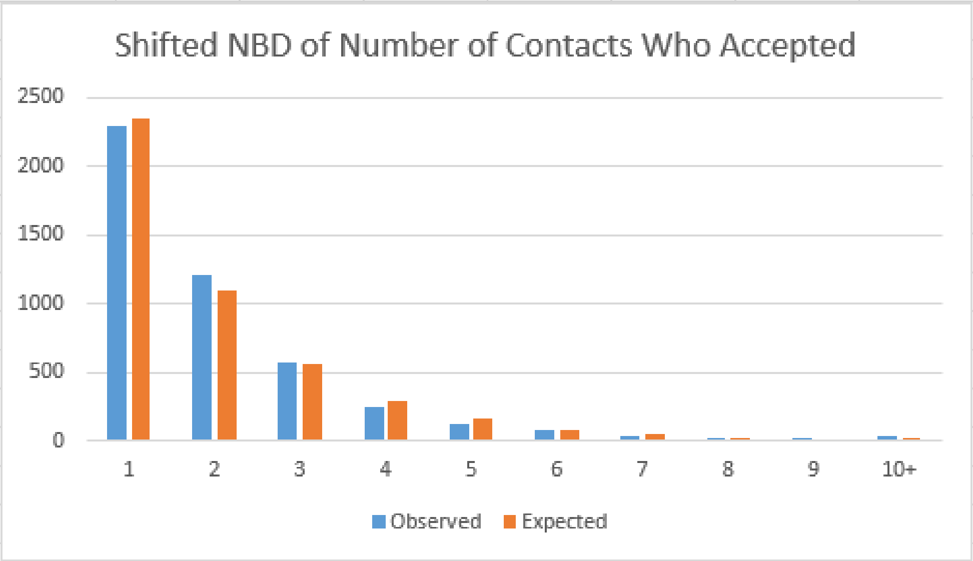


Shifted NBD	
alpha	0.393367
r	0.642402
LL Sum	-63256.4
ChiSq	37350.52
df	53
pvalue	0

*The data is binned again only for your visual benefit

To answer my managerial problem, I finally ran the Shifted NBD Model on the population who agreed to a term deposit:

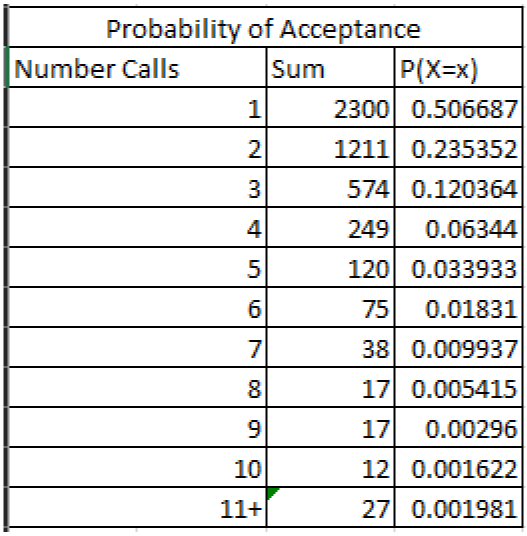
Shifted NBD	
alpha	0.790982
r	0.831896
LL Sum	-6587.81
ChiSq	495.6907
df	53
pvalue	3.67E-73



*The visualization is again binned for the reader's benefit. The most calls for an acceptance was 23.

We see that though our *r* and *alpha* are low, they are higher than observed in the Shifted NBD Model of both the entire population and the ones who refused. This validates my hypothesis. Since marketers will stop calling an individual if they accept and they are unlikely to accept after a high number of calls, the individuals in the "accept" population will be clustered towards the beginning of the x-axis and not have a long tail expressed by the whole population and the refusing population. This leads to the indication of a higher heterogeneity as the population will be more evenly distributed across the first few "bins". The scale *alpha* will also increase as the distribution is compressed more towards 0.

Now, diving into the probabilities of each "bin", we can answer our managerial problem:



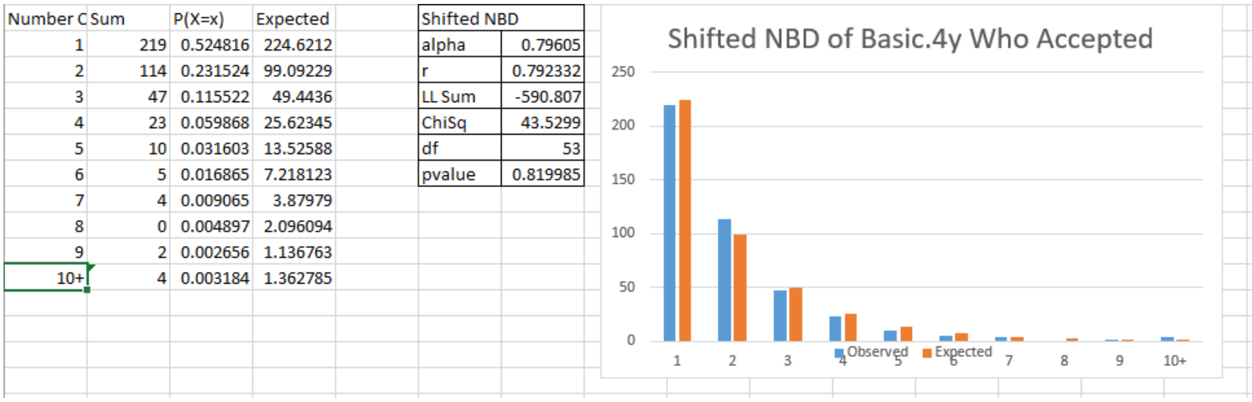
Probability of Acceptance		
Number Calls	Sum	P(X=x)
1	2300	0.506687
2	1211	0.235352
3	574	0.120364
4	249	0.06344
5	120	0.033933
6	75	0.01831
7	38	0.009937
8	17	0.005415
9	17	0.00296
10	12	0.001622
11+	27	0.001981

*The data is again binned for the reader's visual benefit

We can see the likelihood of each individual accepting the term deposit across the number of calls. Over 50% of individuals accepting the term deposit accept at the first call, confirming my initial hypothesis. We also see decreasing returns to calling, and see that the probability of an individual accepting the term deposit after any call after the 10th call is only 0.2%, scarcely higher than an individual accepting the term deposit on the 10th call.

We can use this information to inform our managerial problem through the lens of marginal revenue and marginal cost. We can calculate the marginal revenue of the nth call by multiplying the expected value of each individual accepting a term deposit with the probability of the individual accepting the term deposit on the nth call. If that is greater than the marginal cost, then it is sensible to keep calling.

The next part of my analysis focused on one of the attributes of the population, in this case, education. I created a Shifted NBD Model for each of the education attributes:, Basic 6 year education, Basic 9 year education, High School, Illiterate, Professional Course education, University Degree). I omitted unknown as the information we glean from that segment is unlikely to help us solve our managerial problem. Attached are the parameters and graphs of my findings:

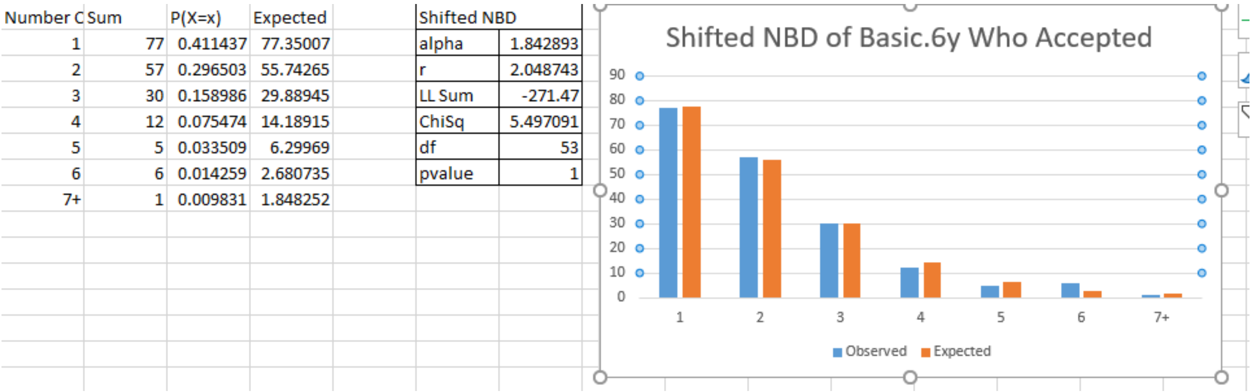


Number C	Sum	P(X=x)	Expected
1	219	0.524816	224.6212
2	114	0.231524	99.09229
3	47	0.115522	49.4436
4	23	0.059868	25.62345
5	10	0.031603	13.52588
6	5	0.016865	7.218123
7	4	0.009065	3.87979
8	0	0.004897	2.096094
9	2	0.002656	1.136763
10+	4	0.003184	1.362785

Shifted NBD	
alpha	0.79605
r	0.792332
LL Sum	-590.807
ChiSq	43.5299
df	53
pvalue	0.819985

*Binned again for your visual benefit

We see that if our contact only has a basic 4 year education, we have decreasing returns to scale and only a 0.3% chance of conversion for any call after the 9th call. The alpha and r are quite similar to the overall population of acceptances.

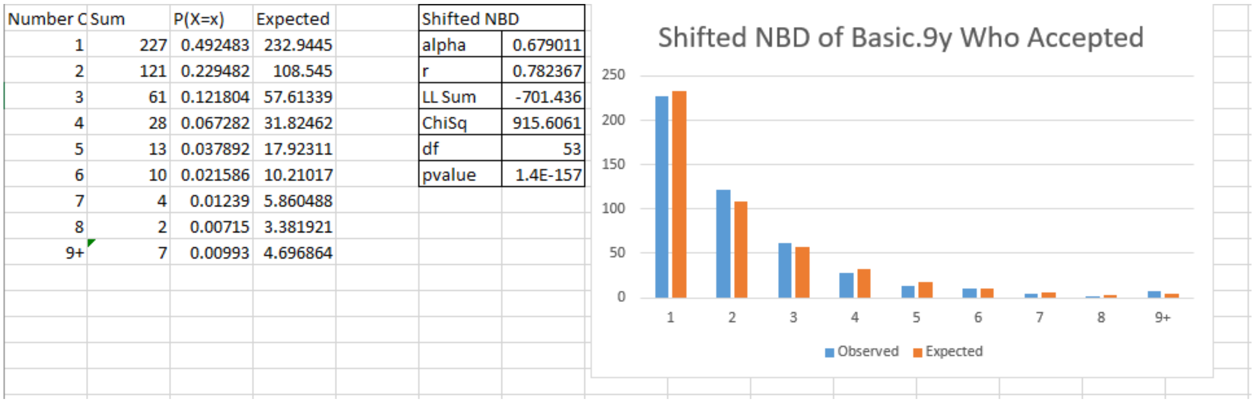


Number C	Sum	P(X=x)	Expected
1	77	0.411437	77.35007
2	57	0.296503	55.74265
3	30	0.158986	29.88945
4	12	0.075474	14.18915
5	5	0.033509	6.29969
6	6	0.014259	2.680735
7+	1	0.009831	1.848252

Shifted NBD	
alpha	1.842893
r	2.048743
LL Sum	-271.47
ChiSq	5.497091
df	53
pvalue	1

*Binned again for your visual benefit

We see that if our contact only has a basic 6 year education, we have decreasing returns to scale and only a 0.9% chance of conversion for any call after the 6th call. The alpha and r are higher than in the overall population of acceptances and this is due to the much smaller population in this segment that has an even shorter tail.

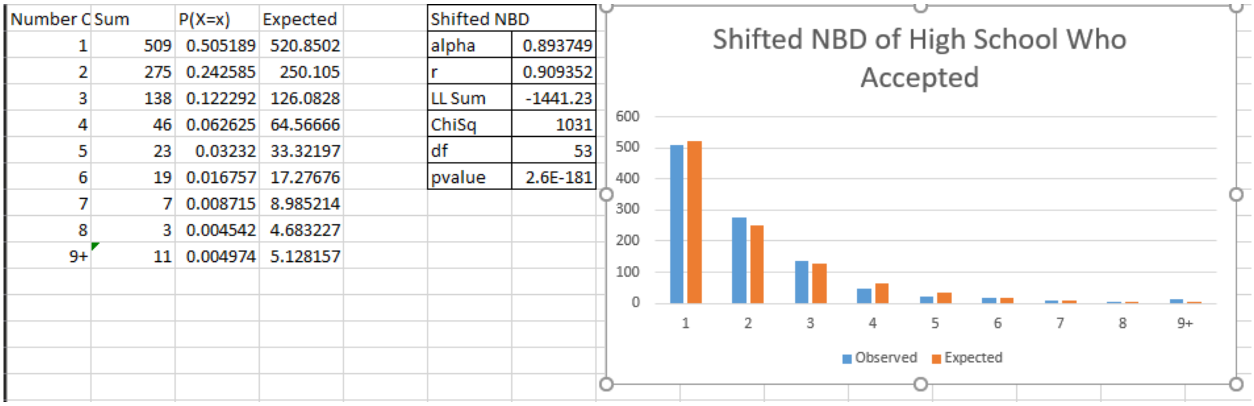


Number C	Sum	P(X=x)	Expected
1	227	0.492483	232.9445
2	121	0.229482	108.545
3	61	0.121804	57.61339
4	28	0.067282	31.82462
5	13	0.037892	17.92311
6	10	0.021586	10.21017
7	4	0.01239	5.860488
8	2	0.00715	3.381921
9+	7	0.00993	4.696864

Shifted NBD	
alpha	0.679011
r	0.782367
LL Sum	-701.436
ChiSq	915.6061
df	53
pvalue	1.4E-157

*Binned again for your visual benefit

We see that if our contact only has a basic 9 year education, we have decreasing returns to scale and only a 0.9% chance of conversion for any call after the 8th call. The alpha and r are about the same as in the overall population of acceptances, though the alpha scale is lower representing a longer tail.

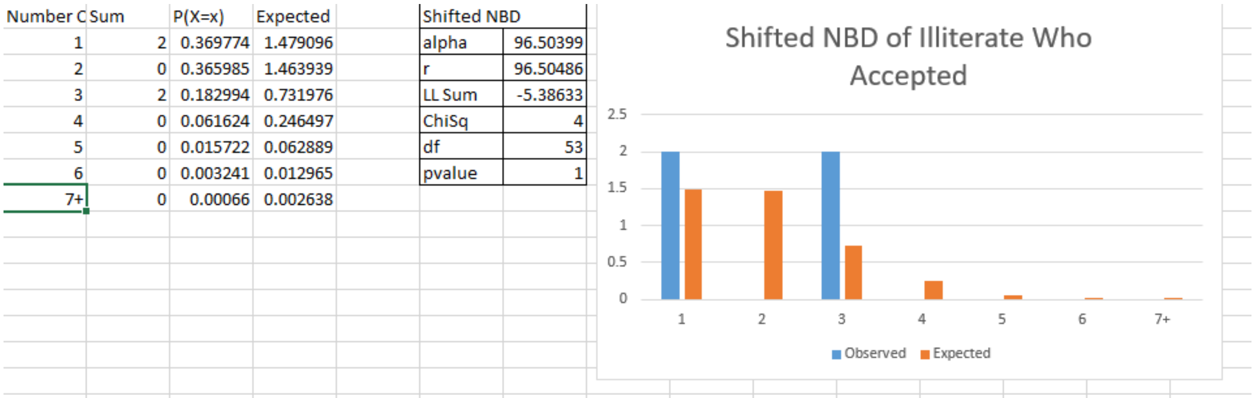


Number C	Sum	P(X=x)	Expected
1	509	0.505189	520.8502
2	275	0.242585	250.105
3	138	0.122292	126.0828
4	46	0.062625	64.56666
5	23	0.03232	33.32197
6	19	0.016757	17.27676
7	7	0.008715	8.985214
8	3	0.004542	4.683227
9+	11	0.004974	5.128157

Shifted NBD	
alpha	0.893749
r	0.909352
LL Sum	-1441.23
ChiSq	1031
df	53
pvalue	2.6E-181

*Binned again for your visual benefit

We see that if our contact only has a high school education, we have decreasing returns to scale and only a 0.4% chance of conversion for any call after the 8th call. The alpha and r are similar to the overall population of acceptances, though there is slightly more heterogeneity.

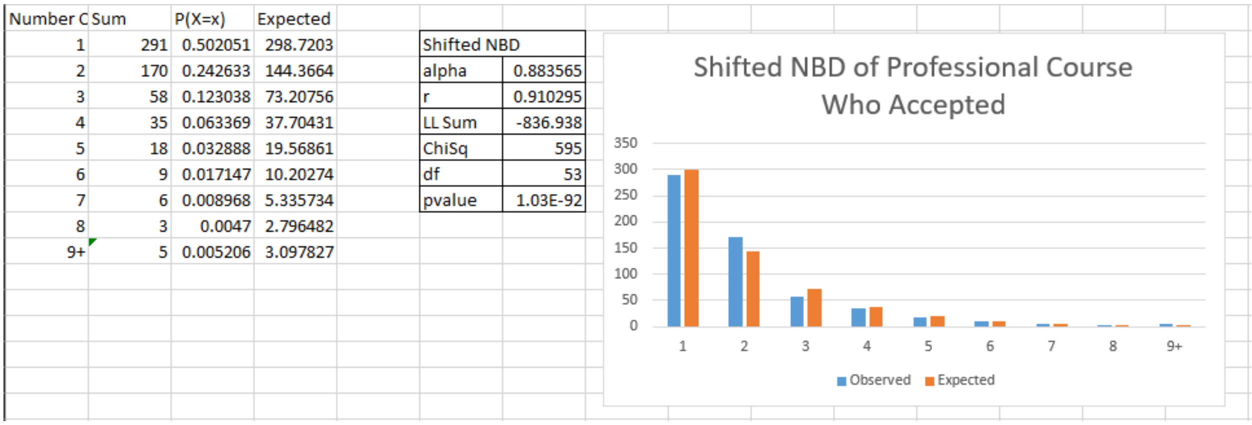


Number C	Sum	P(X=x)	Expected
1	2	0.369774	1.479096
2	0	0.365985	1.463939
3	2	0.182994	0.731976
4	0	0.061624	0.246497
5	0	0.015722	0.062889
6	0	0.003241	0.012965
7+	0	0.00066	0.002638

Shifted NBD	
alpha	96.50399
r	96.50486
LL Sum	-5.38633
ChiSq	4
df	53
pvalue	1

*Binned again for your visual benefit

We see that if our contact only has a high school education, we have decreasing returns to scale and only a 0.06% chance of conversion for any call after the 6th call. The alpha and r are vastly different than from the overall population, but these findings should be disregarded because the sample size(4) is too small.

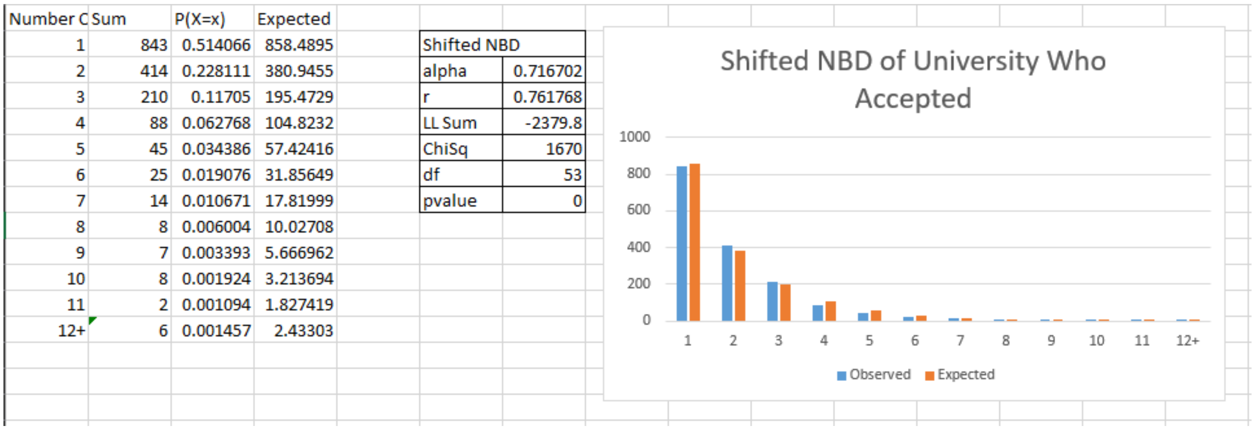


Number C	Sum	P(X=x)	Expected
1	291	0.502051	298.7203
2	170	0.242633	144.3664
3	58	0.123038	73.20756
4	35	0.063369	37.70431
5	18	0.032888	19.56861
6	9	0.017147	10.20274
7	6	0.008968	5.335734
8	3	0.0047	2.796482
9+	5	0.005206	3.097827

Shifted NBD	
alpha	0.883565
r	0.910295
LL Sum	-836.938
ChiSq	595
df	53
pvalue	1.03E-92

*Binned again for your visual benefit

We see that if our contact only has a high school education, we have decreasing returns to scale and only a 0.5% chance of conversion for any call after the 8th call. The alpha and r are about the same as the overall population of acceptances, though there is more heterogeneity.



Number C	Sum	P(X=x)	Expected
1	843	0.514066	858.4895
2	414	0.228111	380.9455
3	210	0.11705	195.4729
4	88	0.062768	104.8232
5	45	0.034386	57.42416
6	25	0.019076	31.85649
7	14	0.010671	17.81999
8	8	0.006004	10.02708
9	7	0.003393	5.666962
10	8	0.001924	3.213694
11	2	0.001094	1.827419
12+	6	0.001457	2.43303

Shifted NBD	
alpha	0.716702
r	0.761768
LL Sum	-2379.8
ChiSq	1670
df	53
pvalue	0

*Binned again for your visual benefit

We see that if our contact only has a high school education, we have decreasing returns to scale and only a 0.1% chance of conversion for any call after the 11th call. The alpha and r are about the same as the overall population acceptances, though there is less heterogeneity as the distribution is spread along the tail more.

VI. Conclusion

A basic framework in business is profit = revenue – cost. In this paper I have explored the methods for reducing the cost among marketers, using a dataset drawn from a Portuguese bank seeking to increase the number of consumers signed to a term deposit with them. We informally realize that their customers also follow the Pareto principle (11% of the total population accepted term deposits) and using my models, understand the marginal benefit of each additional call based on the probability distribution of acceptances on each call. We also are able to have a more granular view into one attribute of the population, and can see how the propensities and probability distributions of each population change based on those attributes.

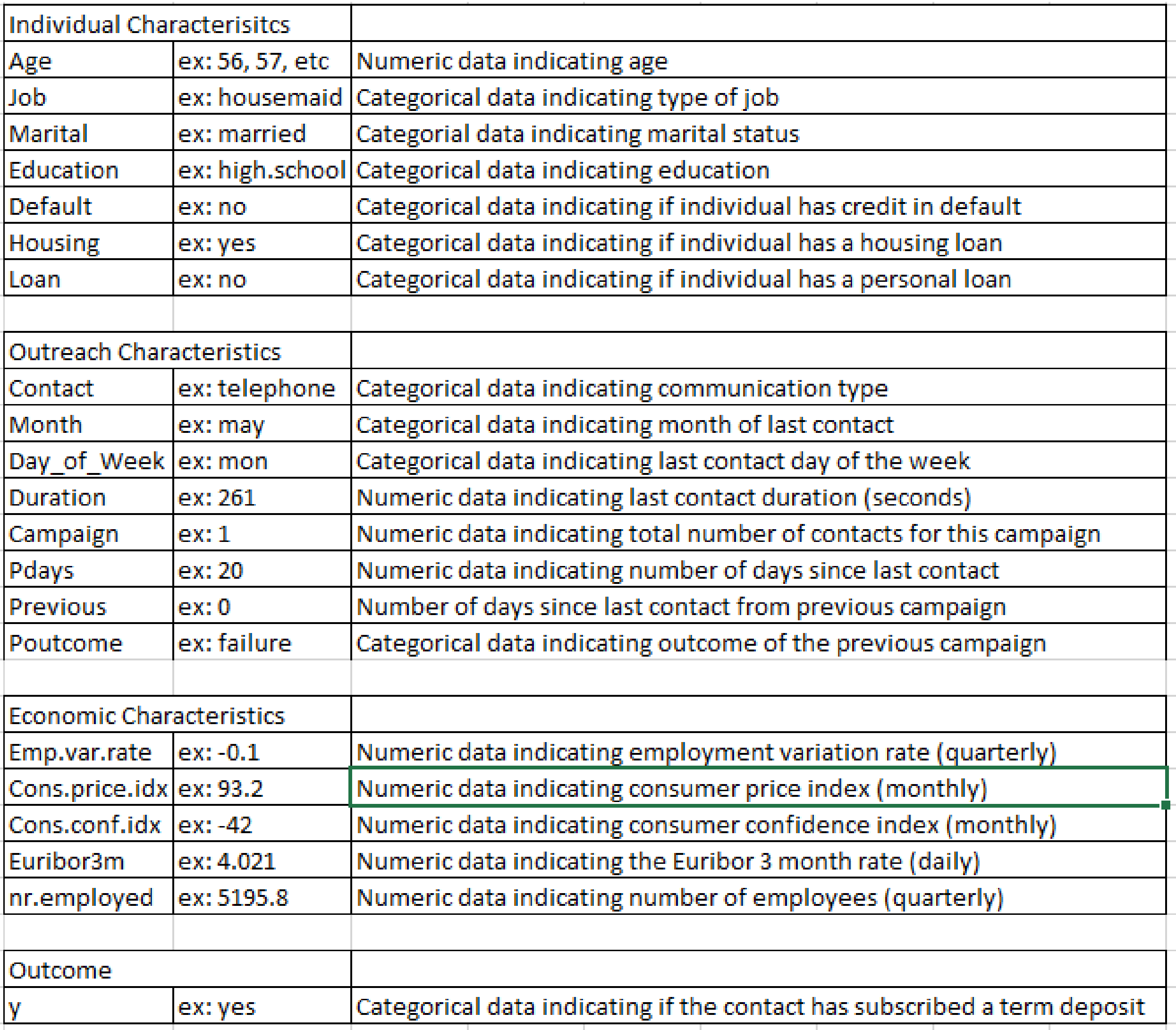
In this context, we can lower the cost of the Portuguese bank in this region by no longer reaching out to the same individual after the nth call, depending on the marginal cost and revenue. We see that on the 6th call there is a 1% probability of accepting and after the 8th call less than 1% no matter the number of calls. With this model, we can help the Portuguese bank lower costs and likely prevent instances of unnecessarily repeated calls. Like 56 calls to one individual.

The limitations to my analysis lie in the lack of space, constraints of the assignment, and data available to truly determine if other factors may shift the propensity of contacts to accept the term deposit within nth number of calls. I am also unaware of if and how these attributes interact with each other. In future work, I would dive deeper on how the other attributes affect acceptances on the number of calls, such as marital status or leasing information. I do not have the space here, however.

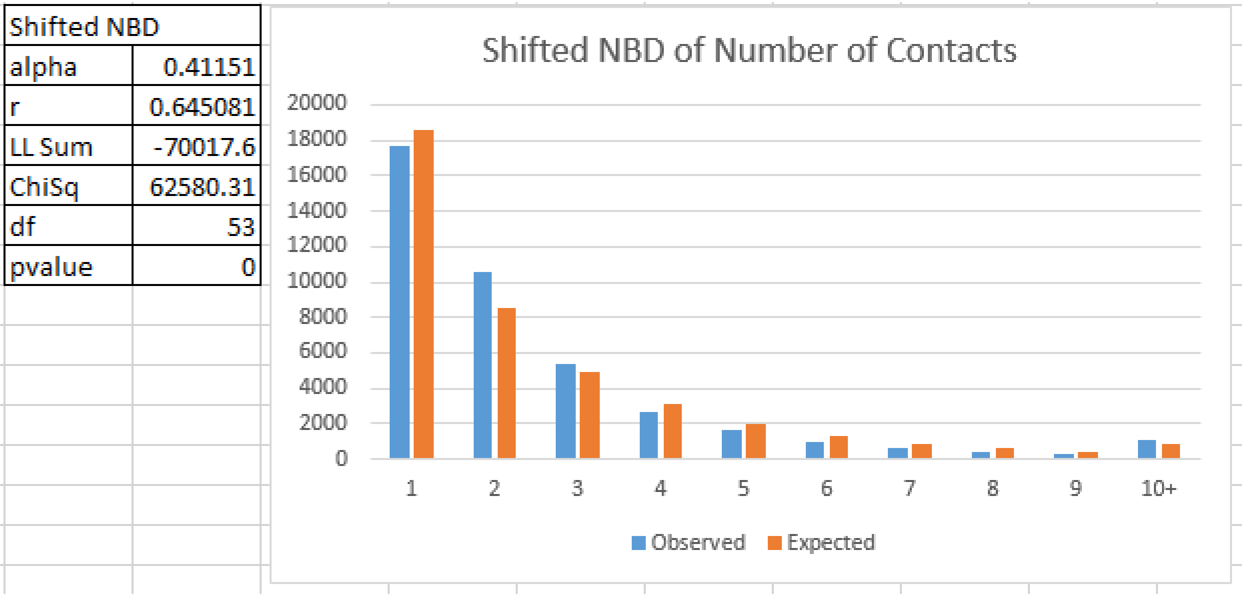
Further limitations lie in the sample data. It is unlikely the contacts were drawn from a randomized trial and the calling conducted in a scientific manner. The contacts may have been drawn from a sheet of individuals most likely to convert and different marketers may have different qualities of leads Perhaps some marketers have a higher propensity to sign contacts to a term deposit than others and the acceptances on nth calls represent their efforts. We are also unsure the year each call was performed, and the acceptances may lie only in the years of economic prosperity. Ultimately, while this data represents only one firms efforts in one area and the exact findings cannot be extrapolated to other populations, we can still use the same methods and modeling techniques to answer the same managerial problem in future populations.

VII. Appendix

a.

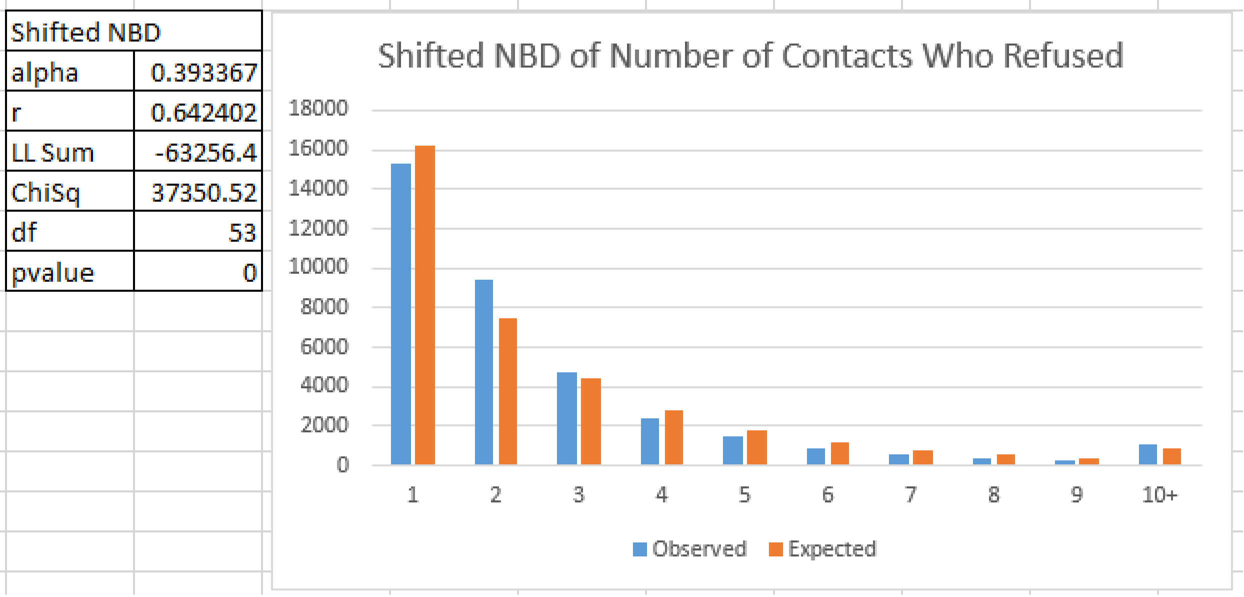


Individual Characterisitcs		
Age	ex: 56, 57, etc	Numeric data indicating age
Job	ex: housemaid	Categorical data indicating type of job
Marital	ex: married	Categorial data indicating marital status
Education	ex: high.school	Categorical data indicating education
Default	ex: no	Categorical data indicating if individual has credit in default
Housing	ex: yes	Categorical data indicating if individual has a housing loan
Loan	ex: no	Categorical data indicating if individual has a personal loan
Outreach Characteristics		
Contact	ex: telephone	Categorical data indicating communication type
Month	ex: may	Categorical data indicating month of last contact
Day_of_Week	ex: mon	Categorical data indicating last contact day of the week
Duration	ex: 261	Numeric data indicating last contact duration (seconds)
Campaign	ex: 1	Numeric data indicating total number of contacts for this campaign
Pdays	ex: 20	Numeric data indicating number of days since last contact
Previous	ex: 0	Number of days since last contact from previous campaign
Poutcome	ex: failure	Categorical data indicating outcome of the previous campaign
Economic Characteristics		
Emp.var.rate	ex: -0.1	Numeric data indicating employment variation rate (quarterly)
Cons.price.idx	ex: 93.2	Numeric data indicating consumer price index (monthly)
Cons.conf.idx	ex: -42	Numeric data indicating consumer confidence index (monthly)
Euribor3m	ex: 4.021	Numeric data indicating the Euribor 3 month rate (daily)
nr.employed	ex: 5195.8	Numeric data indicating number of employees (quarterly)
Outcome		
y	ex: yes	Categorical data indicating if the contact has subscribed a term deposit



Shifted NBD	
alpha	0.41151
r	0.645081
LL Sum	-70017.6
ChiSq	62580.31
df	53
pvalue	0

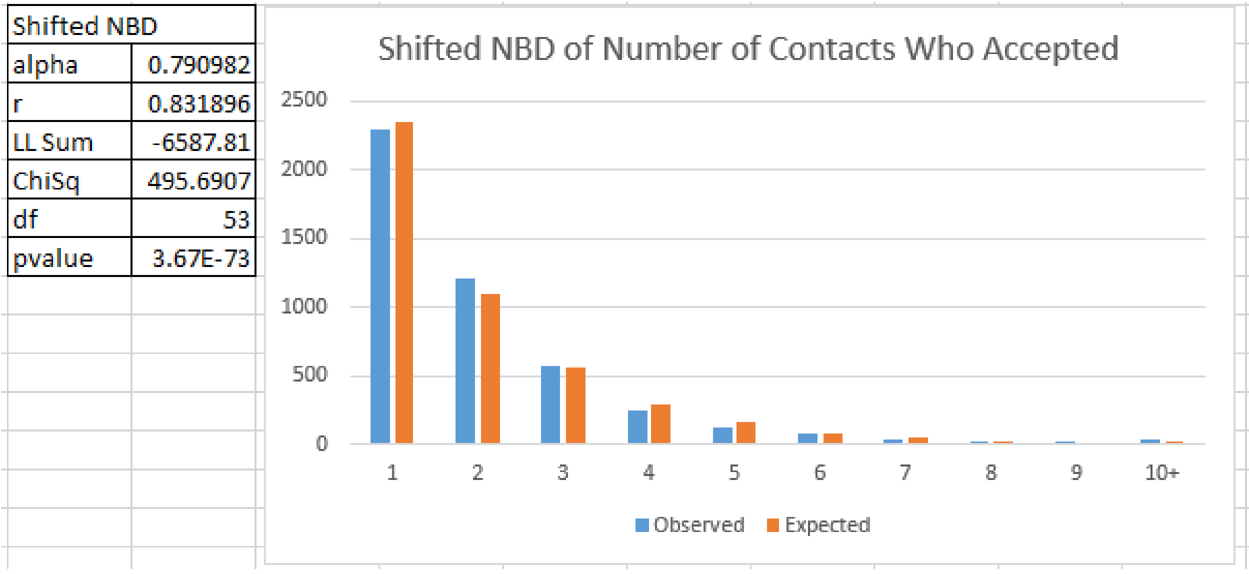
b.



Shifted NBD	
alpha	0.393367
r	0.642402
LL Sum	-63256.4
ChiSq	37350.52
df	53
pvalue	0

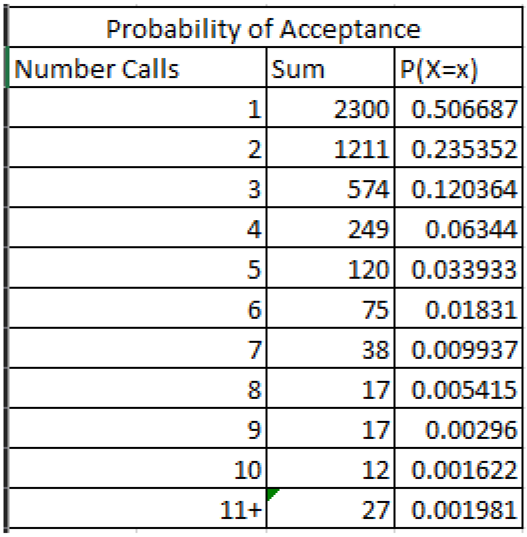
c.

d.



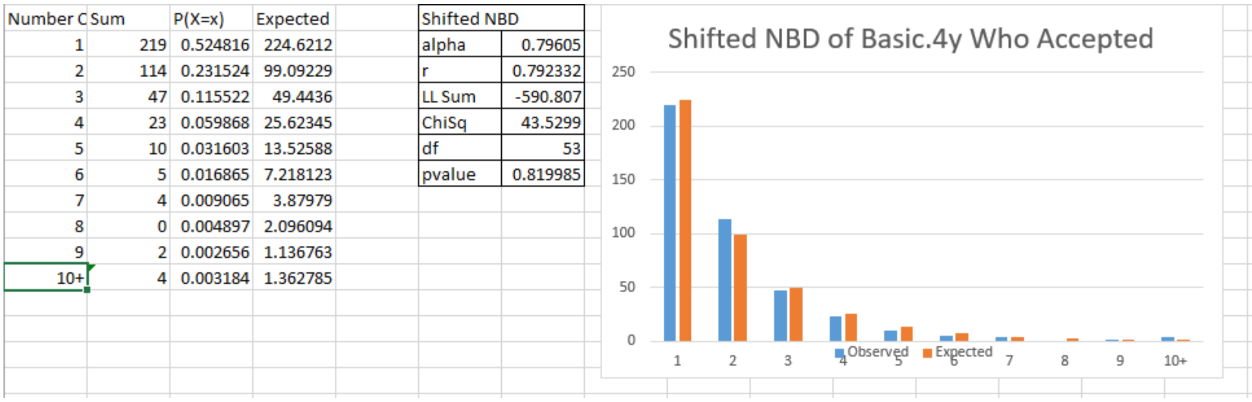
Shifted NBD	
alpha	0.790982
r	0.831896
LL Sum	-6587.81
ChiSq	495.6907
df	53
pvalue	3.67E-73

e.



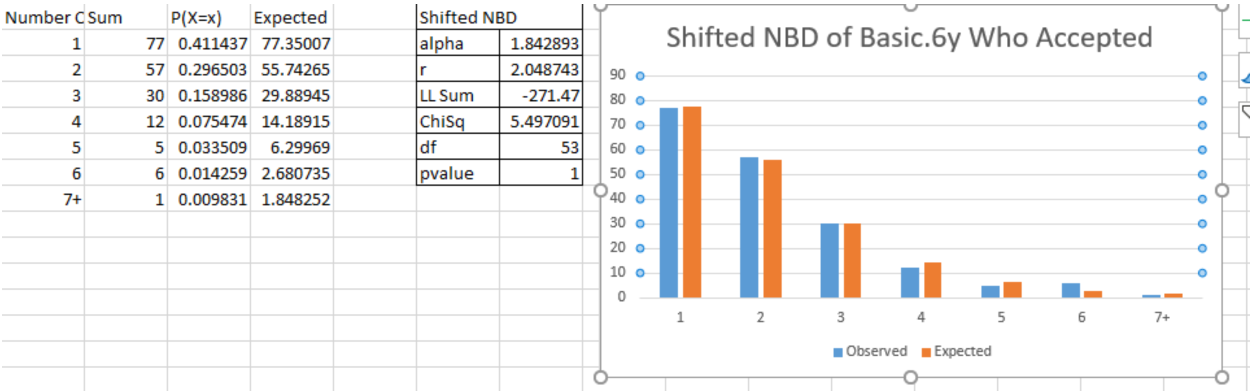
Probability of Acceptance		
Number Calls	Sum	P(X=x)
1	2300	0.506687
2	1211	0.235352
3	574	0.120364
4	249	0.06344
5	120	0.033933
6	75	0.01831
7	38	0.009937
8	17	0.005415
9	17	0.00296
10	12	0.001622
11+	27	0.001981

f.



Number C	Sum	P(X=x)	Expected
1	219	0.524816	224.6212
2	114	0.231524	99.09229
3	47	0.115522	49.4436
4	23	0.059868	25.62345
5	10	0.031603	13.52588
6	5	0.016865	7.218123
7	4	0.009065	3.87979
8	0	0.004897	2.096094
9	2	0.002656	1.136763
10+	4	0.003184	1.362785

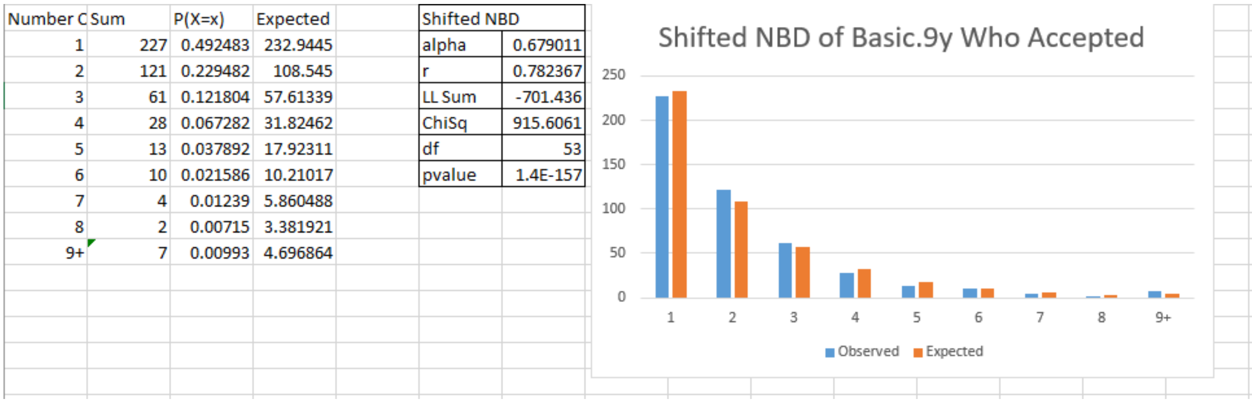
Shifted NBD	
alpha	0.79605
r	0.792332
LL Sum	-590.807
ChiSq	43.5299
df	53
pvalue	0.819985



Number C	Sum	P(X=x)	Expected
1	77	0.411437	77.35007
2	57	0.296503	55.74265
3	30	0.158986	29.88945
4	12	0.075474	14.18915
5	5	0.033509	6.29969
6	6	0.014259	2.680735
7+	1	0.009831	1.848252

Shifted NBD	
alpha	1.842893
r	2.048743
LL Sum	-271.47
ChiSq	5.497091
df	53
pvalue	1

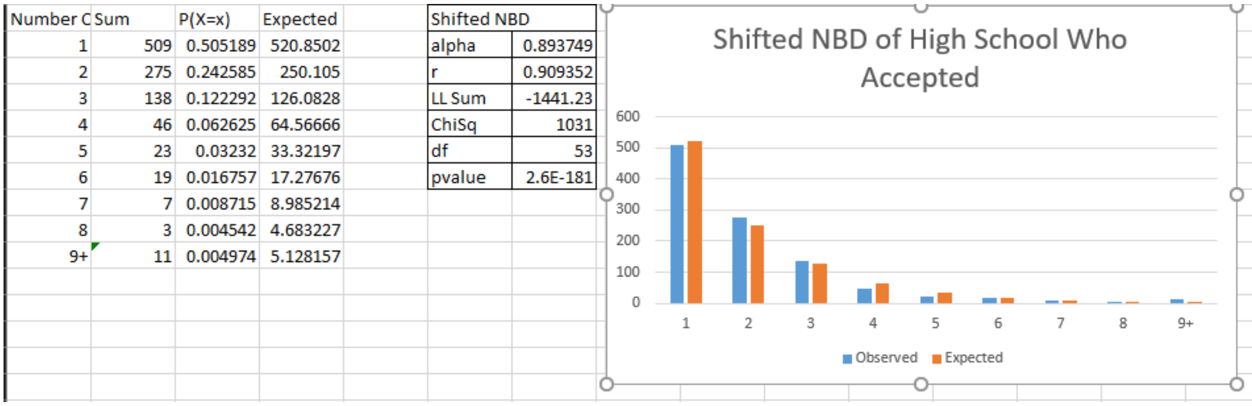
g.



Number C	Sum	P(X=x)	Expected
1	227	0.492483	232.9445
2	121	0.229482	108.545
3	61	0.121804	57.61339
4	28	0.067282	31.82462
5	13	0.037892	17.92311
6	10	0.021586	10.21017
7	4	0.01239	5.860488
8	2	0.00715	3.381921
9+	7	0.00993	4.696864

Shifted NBD	
alpha	0.679011
r	0.782367
LL Sum	-701.436
ChiSq	915.6061
df	53
pvalue	1.4E-157

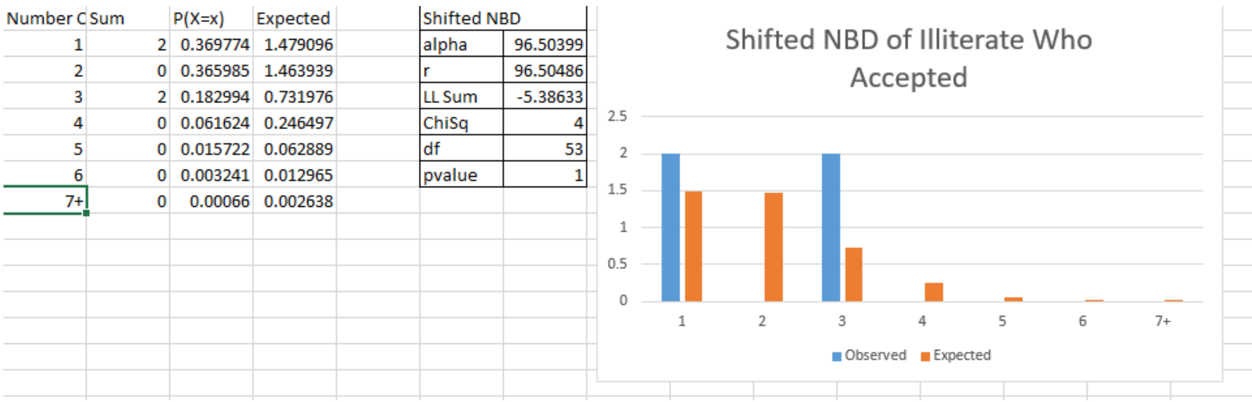
h.



Number C	Sum	P(X=x)	Expected
1	509	0.505189	520.8502
2	275	0.242585	250.105
3	138	0.122292	126.0828
4	46	0.062625	64.56666
5	23	0.03232	33.32197
6	19	0.016757	17.27676
7	7	0.008715	8.985214
8	3	0.004542	4.683227
9+	11	0.004974	5.128157

Shifted NBD	
alpha	0.893749
r	0.909352
LL Sum	-1441.23
ChiSq	1031
df	53
pvalue	2.6E-181

i.

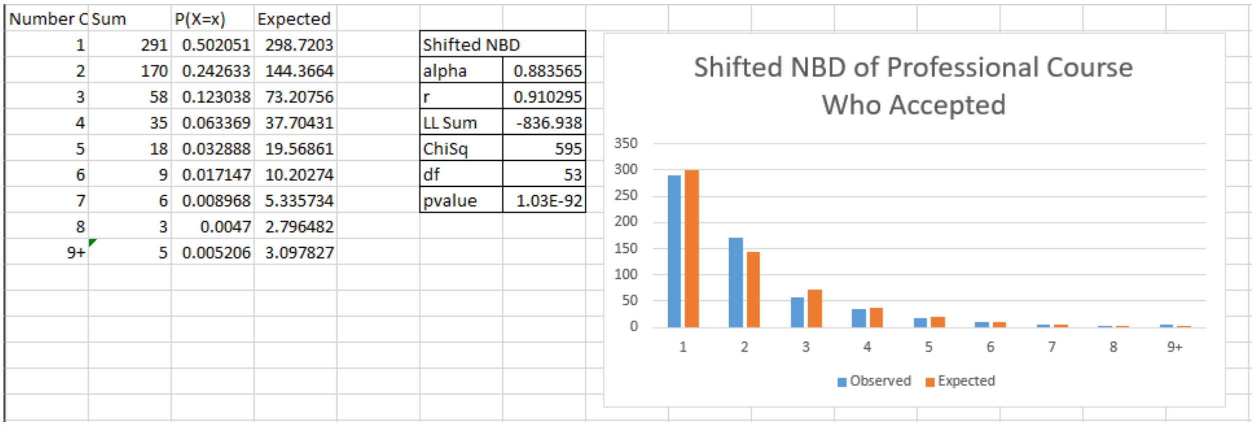


Number C	Sum	P(X=x)	Expected
1	2	0.369774	1.479096
2	0	0.365985	1.463939
3	2	0.182994	0.731976
4	0	0.061624	0.246497
5	0	0.015722	0.062889
6	0	0.003241	0.012965
7+	0	0.00066	0.002638

Shifted NBD	
alpha	96.50399
r	96.50486
LL Sum	-5.38633
ChiSq	4
df	53
pvalue	1

j.

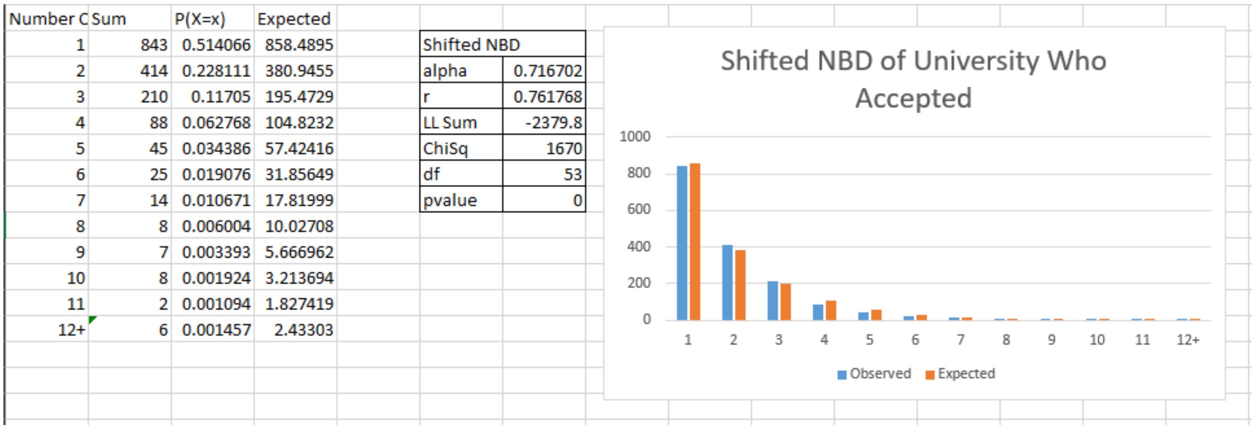
k.



Number C	Sum	P(X=x)	Expected
1	291	0.502051	298.7203
2	170	0.242633	144.3664
3	58	0.123038	73.20756
4	35	0.063369	37.70431
5	18	0.032888	19.56861
6	9	0.017147	10.20274
7	6	0.008968	5.335734
8	3	0.0047	2.796482
9+	5	0.005206	3.097827

Shifted NBD	
alpha	0.883565
r	0.910295
LL Sum	-836.938
ChiSq	595
df	53
pvalue	1.03E-92

l.



Number C	Sum	P(X=x)	Expected
1	843	0.514066	858.4895
2	414	0.228111	380.9455
3	210	0.11705	195.4729
4	88	0.062768	104.8232
5	45	0.034386	57.42416
6	25	0.019076	31.85649
7	14	0.010671	17.81999
8	8	0.006004	10.02708
9	7	0.003393	5.666962
10	8	0.001924	3.213694
11	2	0.001094	1.827419
12+	6	0.001457	2.43303

Shifted NBD	
alpha	0.716702
r	0.761768
LL Sum	-2379.8
ChiSq	1670
df	53
pvalue	0