Probability Sampling: An Application

Yuhang Liu
Amy Nussbaum

Introduction

- General Probability Sampling
  - Definition of a Probability Sample
  - Cost Function and Constraints
  - Sample Design
  - Objective Function
  - Tools
- Application
  - Background Information
  - Development of Cost Function, Constraints
  - Development of Sample Design
  - Development of Objective Function
  - Implementation
- Morals of the Story

Probability Sample

- Definition: A sample selected in such a way that every member of the sample has a known chance of selection (and every member of the population has a nonzero chance of selection)
- This definition does not require EQUAL probability of selection, as long as we use some sort of randomness to make our decisions.
- Sampling Units
  - Primary Sampling Unit (PSU)
  - Secondary Sampling Unit (SSU)

Probability Sample

- Advantages
  - Reduces risk of selection bias (allows for unbiased estimators of population parameters)
  - Produces results that are generalizable to the U.S. population (allow for known margins of error from population parameters)

Cost Function and Constraints

- Cost Function Formula for a Balanced Survey:

$$c_{total} = c_1 n + c_2 mn$$

n is the total number of PSUs

m is the total number of SSUs

- When the survey is unbalanced, replace m with $\bar{m}$
- Constraints can be
  - Financial
  - Time
  - Resources

Sample Design

- Depends on what information you have and what information you want…
  - Simple Random Sample
  - Cluster Sample
- Requires the use of auxiliary variable:
  - Stratified Sample
  - Probability Proportional to Size Sampling

Objective Function

- Helps us choose optimal sample size with two goals:
  - Minimize the cost with respect to a certain variance
  - Minimize the variance with respect to a certain cost
  - Both reduce to the same answer!
- Either way, we need to know how to calculate the variance of our statistic of interest.
  - Total
  - Mean
  - Proportion
  - Regression

Tools for Design

- Excel
  - Solver
    - Allows for customizable objective function
    - Allows for additional constraints
    - Extremely useful for choosing sample sizes
    - Can solve problems numerically
  - "Using Spreadsheet Solvers in Sample Design"
    - (Stokes & Plummer, 2004).

Tools for Analysis

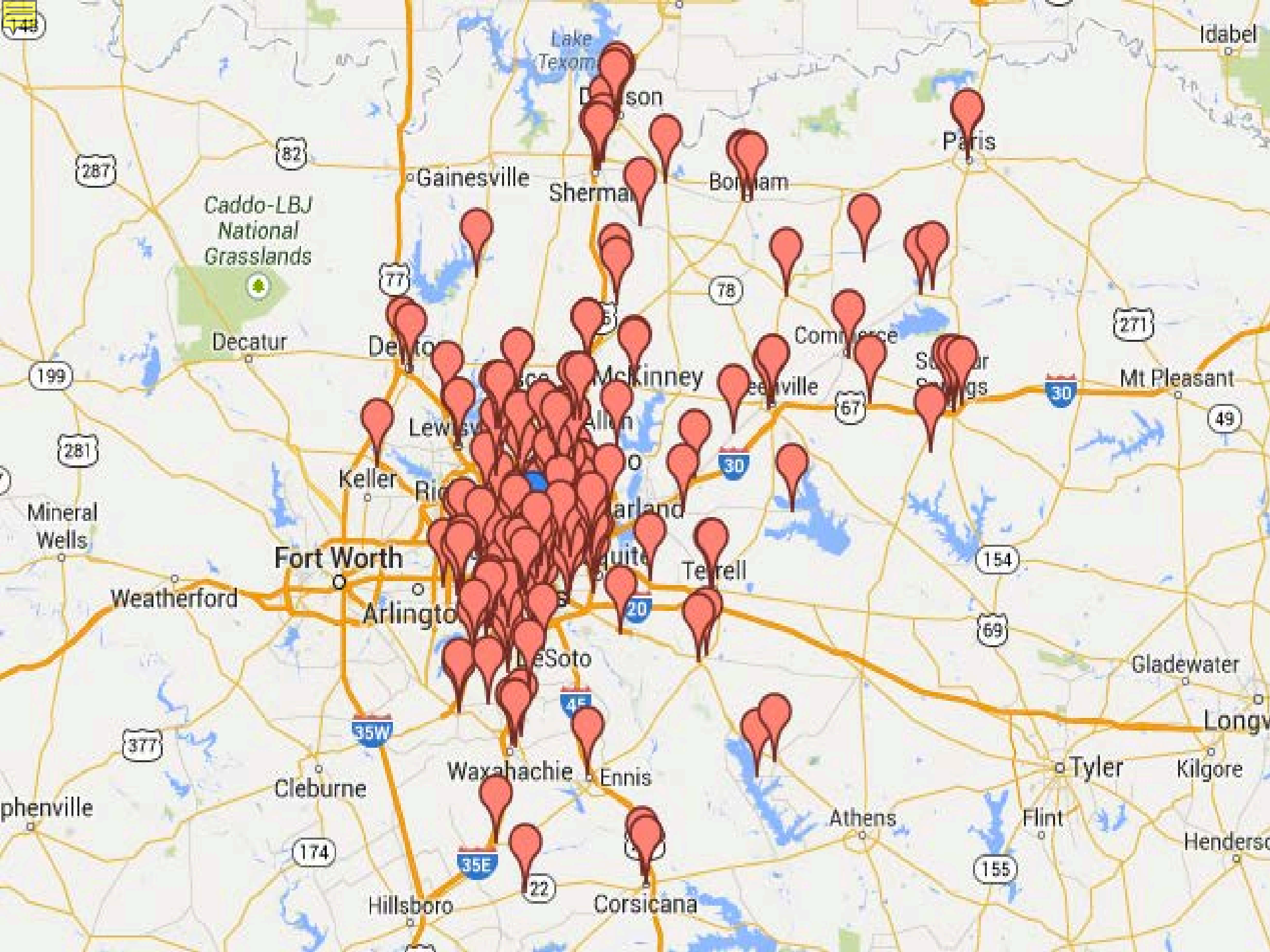
- SAS
  - Selection: SURVEYSELECT
  - Analysis of Means and Totals: SURVEYMEANS
  - Analysis of Regression: SURVEYREG
  - Analysis of Generalized Regression: SURVEYLOGISTIC
  - Analysis of Frequency Tables: SURVEYFREQ
  - Easily allows use of sampling unit weights
- Do not ignore weights when analyzing data!

Application

- Background Information:
  - Predict food-insecurity in the Greater North Texas Area
  - The grant, from the Hunger Center of Dallas, Subsidiary of North Texas Food Bank, focuses on
    - Financial Literacy
    - Social Network
    - Educational Level
  - Reaches beyond the unidirectional definition of poverty
- Sampling Units
  - 172 Food Banks (PSU)
  - Approximately 1000 Clients (SSU)

Application

- Additional Information on Food Banks
  - ID Number
  - Name
  - Number of clients served in the last three months
  - Mileage
  - Housing in America Survey Info
  - County
  - City
  - Zip Code
  - IBSCO Map
  - Unemployment Rate by County – Proxy Variable


Lake Texoma
Idabel
Paris
Gainesville
Sherman
Bonham
Caddo-LBJ National Grasslands
Decatur
McKinney
Commerce
Mt Pleasant
Keller
Mineral Wells
Fort Worth
Weatherford
Arlington
Terrell
DeSoto
Gladewater
Waxahachie
Ennis
Cleburne
Tyler
Kilgore
Athens
Flint
Hillsboro
Corsicana
82
287
77
78
271
199
281
67
30
49
154
69
20
35W
45
377
35E
22
174
155

Development of Cost Function

- Constraints:
  - Financial
  - Cost per survey: $50 per respondent
    - $20 incentive, given to the respondent
    - $15 paycheck, given to the interviewer
    - $10 paycheck, given for data entry
    - $5 reimbursement for gas and travel time
  - For N=1000, total cost is $50,000

Development of Cost Function

- Constraints
  - Time
    - Time per survey: 40 minutes
    - 35 minutes for survey
    - 5 minutes for travel time
  - For N=1000, total time is 660 hours
- Interviewers
  - 4 surveyors, contributing 165 hours each
  - If they spend 10 hours per week interviewing, the survey will take 16 weeks to complete.
  - Plan: conduct surveys November – February (4 months total)

Development of Cost Function

- Other practical issues:
  - Interviewers could work for a maximum in one day.
  - Interviews could potentially take longer than 30 minutes.
  - Interviewers were sent in pairs for safety reasons.

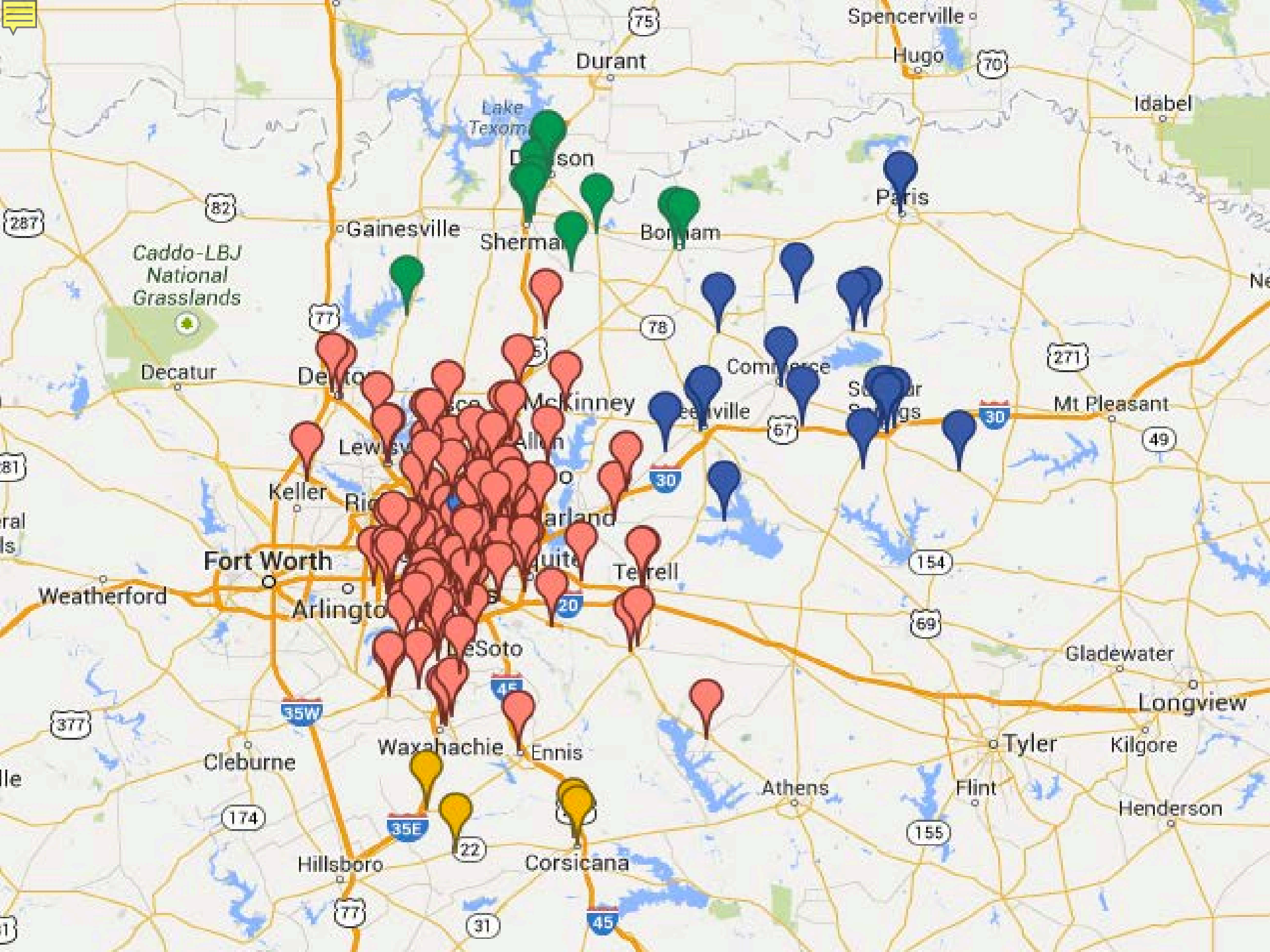
Development of Cost Function

- Assume that interviewers could conduct a constant number of interviews per day, and that the financial cost was constant up to travel reimbursement.
- Travel reimbursement is proportional to the miles the interviewers drive.

Cost = Miles x (Reimbursement + Hourly Wage/Average Speed)

Development of Sample Design

- Stratification
  - Ensures representation of both urban and rural areas
  - No objective definition of urban and rural
  - Instead, we ensured that food banks from the northern, southern, and eastern regions would be represented.
  - Four Geographical Strata
    - Metropolitan Region
    - North Region
    - South Region
    - East Region


Durant
Spencerville
Hugo
Idabel
Lake Texoma
Paris
Gainesville
Sherman
Bonham
Caddo-LBJ National Grasslands
Decatur
McKinney
Commerce
Mt Pleasant
Keller
Fort Worth
Weatherford
Arlington
Terrell
DeSoto
Gladewater
Longview
Tyler
Kilgore
Cleburne
Waxahachie
Ennis
Athens
Flint
Henderson
Hillsboro
Corsicana
75
70
82
287
77
78
271
30
67
49
20
154
69
35W
35E
45
377
174
155
22
31

Development of Sample Design

- PPS Sampling
  - Assigned probabilities to each bank proportional to the number of clients served
  - Larger banks have a higher probability of selection
  - Sampling the same number of clients from each bank ensures approximately equal probability of selection for clients (makes analysis easier)

Development of Objective Function

- Goal: Predict food insecurity using decision trees
  - Closest Statistic: Proportion
- Objective: Minimize variance with respect to cost
- Variance:

$$Var_{total} = \sum [Var_{within} + Var_{between}]$$

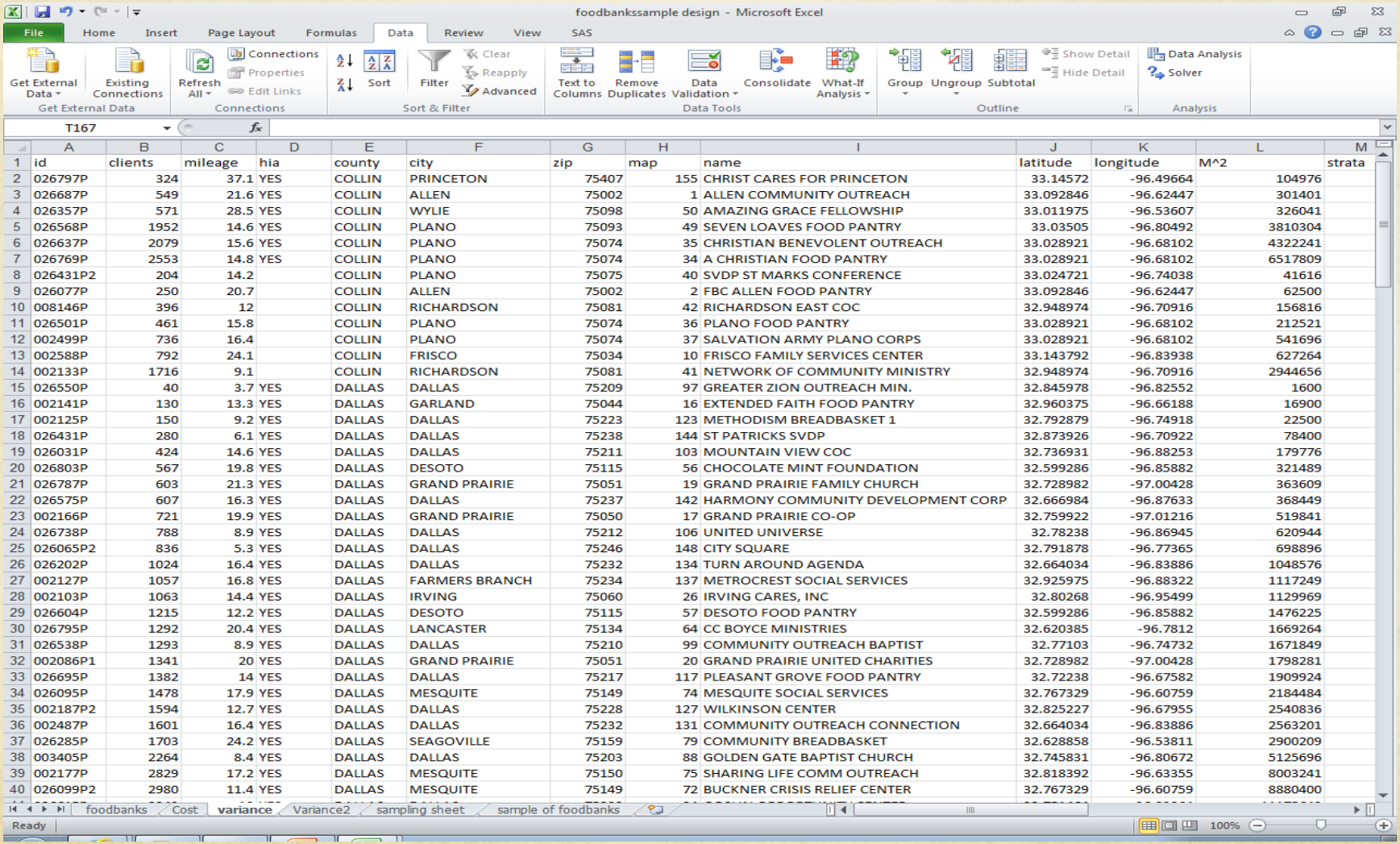
Objective Function

$$Var_{between} = N(1-\frac{n}{N})(\frac{\sum m^2(\pi(2\alpha-1)-E(\alpha))^2}{n-1})$$

$$Var_{within} = \frac{N}{n}\sum m^2(1-\frac{\min(m, N/n)}{m})V(\alpha)$$

where α represents the proxy variable, unemployment rate,
π represents the sampling weight based on the number of clients per bank,
and m represents the number of clients served at the bank.

Excel Solver



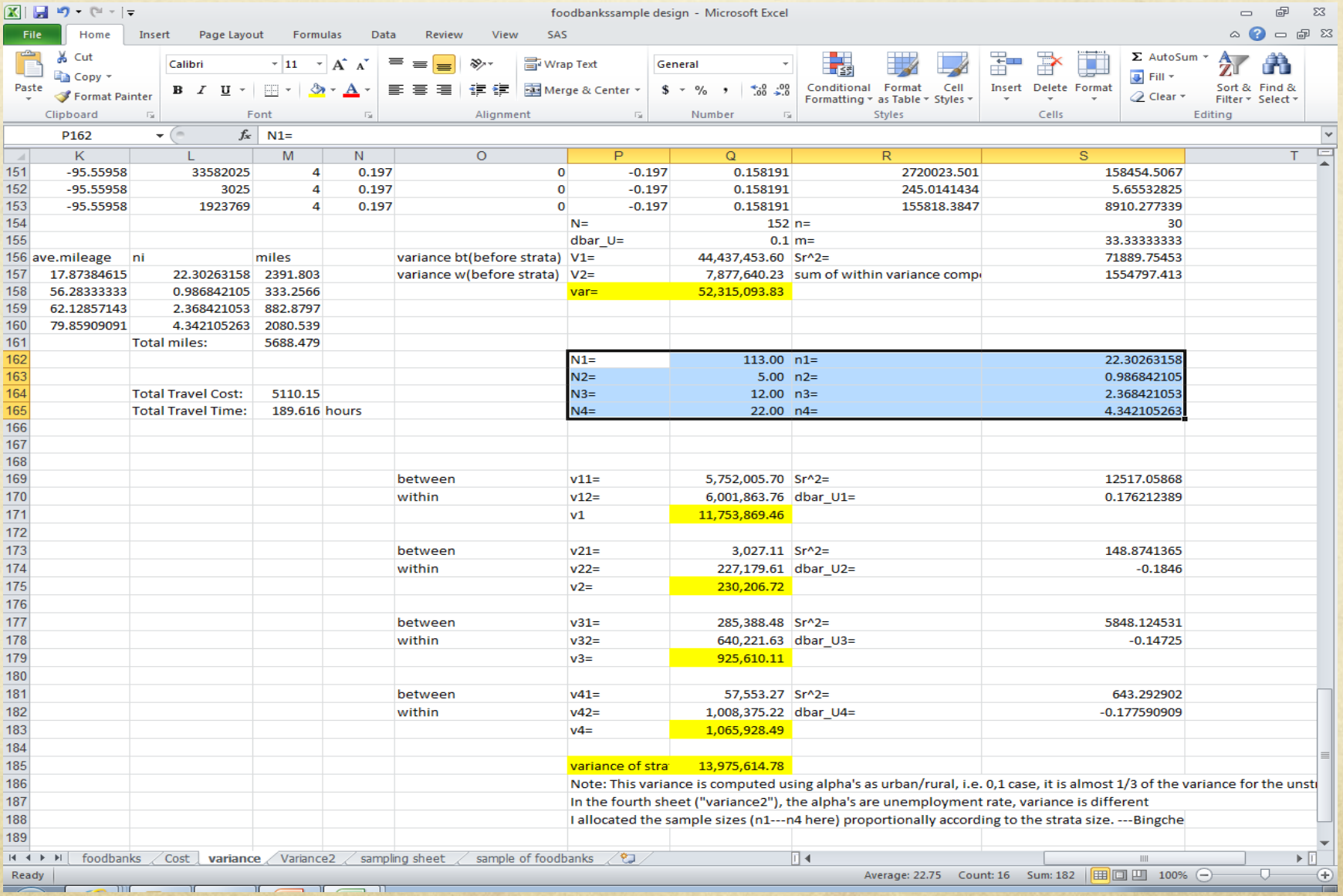
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	clients	mileage	hia	county	city	zip	map	name	latitude	longitude	M^2	strata
2	026797P	324	37.1	YES	COLLIN	PRINCETON	75407	155	CHRIST CARES FOR PRINCETON	33.14572	-96.49664	104976	
3	026687P	549	21.6	YES	COLLIN	ALLEN	75002	1	ALLEN COMMUNITY OUTREACH	33.092846	-96.62447	301401	
4	026357P	571	28.5	YES	COLLIN	WYLIE	75098	50	AMAZING GRACE FELLOWSHIP	33.011975	-96.53607	326041	
5	026568P	1952	14.6	YES	COLLIN	PLANO	75093	49	SEVEN LOAVES FOOD PANTRY	33.03505	-96.80492	3810304	
6	026637P	2079	15.6	YES	COLLIN	PLANO	75074	35	CHRISTIAN BENEVOLENT OUTREACH	33.028921	-96.68102	4322241	
7	026769P	2553	14.8	YES	COLLIN	PLANO	75074	34	A CHRISTIAN FOOD PANTRY	33.028921	-96.68102	6517809	
8	026431P2	204	14.2		COLLIN	PLANO	75075	40	SVDP ST MARKS CONFERENCE	33.024721	-96.74038	41616	
9	026077P	250	20.7		COLLIN	ALLEN	75002	2	FBC ALLEN FOOD PANTRY	33.092846	-96.62447	62500	
10	008146P	396	12		COLLIN	RICHARDSON	75081	42	RICHARDSON EAST COC	32.948974	-96.70916	156816	
11	026501P	461	15.8		COLLIN	PLANO	75074	36	PLANO FOOD PANTRY	33.028921	-96.68102	212521	
12	002499P	736	16.4		COLLIN	PLANO	75074	37	SALVATION ARMY PLANO CORPS	33.028921	-96.68102	541696	
13	002588P	792	24.1		COLLIN	FRISCO	75034	10	FRISCO FAMILY SERVICES CENTER	33.143792	-96.83938	627264	
14	002133P	1716	9.1		COLLIN	RICHARDSON	75081	41	NETWORK OF COMMUNITY MINISTRY	32.948974	-96.70916	2944656	
15	026550P	40	3.7	YES	DALLAS	DALLAS	75209	97	GREATER ZION OUTREACH MIN.	32.845978	-96.82552	1600	
16	002141P	130	13.3	YES	DALLAS	GARLAND	75044	16	EXTENDED FAITH FOOD PANTRY	32.960375	-96.66188	16900	
17	002125P	150	9.2	YES	DALLAS	DALLAS	75223	123	METHODISM BREADBASKET 1	32.792879	-96.74918	22500	
18	026431P	280	6.1	YES	DALLAS	DALLAS	75238	144	ST PATRICKS SVDP	32.873926	-96.70922	78400	
19	026031P	424	14.6	YES	DALLAS	DALLAS	75211	103	MOUNTAIN VIEW COC	32.736931	-96.88253	179776	
20	026803P	567	19.8	YES	DALLAS	DESOTO	75115	56	CHOCOLATE MINT FOUNDATION	32.599286	-96.85882	321489	
21	026787P	603	21.3	YES	DALLAS	GRAND PRAIRIE	75051	19	GRAND PRAIRIE FAMILY CHURCH	32.728982	-97.00428	363609	
22	026575P	607	16.3	YES	DALLAS	DALLAS	75237	142	HARMONY COMMUNITY DEVELOPMENT CORP	32.666984	-96.87633	368449	
23	002166P	721	19.9	YES	DALLAS	GRAND PRAIRIE	75050	17	GRAND PRAIRIE CO-OP	32.759922	-97.01216	519841	
24	026738P	788	8.9	YES	DALLAS	DALLAS	75212	106	UNITED UNIVERSE	32.78238	-96.86945	620944	
25	026065P2	836	5.3	YES	DALLAS	DALLAS	75246	148	CITY SQUARE	32.791878	-96.77365	698896	
26	026202P	1024	16.4	YES	DALLAS	DALLAS	75232	134	TURN AROUND AGENDA	32.664034	-96.83886	1048576	
27	002127P	1057	16.8	YES	DALLAS	FARMERS BRANCH	75234	137	METROCREST SOCIAL SERVICES	32.925975	-96.88322	1117249	
28	002103P	1063	14.4	YES	DALLAS	IRVING	75060	26	IRVING CARES, INC	32.80268	-96.95499	1129969	
29	026604P	1215	12.2	YES	DALLAS	DESOTO	75115	57	DESOTO FOOD PANTRY	32.599286	-96.85882	1476225	
30	026795P	1292	20.4	YES	DALLAS	LANCASTER	75134	64	CC BOYCE MINISTRIES	32.620385	-96.7812	1669264	
31	026538P	1293	8.9	YES	DALLAS	DALLAS	75210	99	COMMUNITY OUTREACH BAPTIST	32.77103	-96.74732	1671849	
32	002086P1	1341	20	YES	DALLAS	GRAND PRAIRIE	75051	20	GRAND PRAIRIE UNITED CHARITIES	32.728982	-97.00428	1798281	
33	026695P	1382	14	YES	DALLAS	DALLAS	75217	117	PLEASANT GROVE FOOD PANTRY	32.72238	-96.67582	1909924	
34	026095P	1478	17.9	YES	DALLAS	MESQUITE	75149	74	MESQUITE SOCIAL SERVICES	32.767329	-96.60759	2184484	
35	002187P2	1594	12.7	YES	DALLAS	DALLAS	75228	127	WILKINSON CENTER	32.825227	-96.67955	2540836	
36	002487P	1601	16.4	YES	DALLAS	DALLAS	75232	131	COMMUNITY OUTREACH CONNECTION	32.664034	-96.83886	2563201	
37	026285P	1703	24.2	YES	DALLAS	SEAGOVILLE	75159	79	COMMUNITY BREADBASKET	32.628858	-96.53811	2900209	
38	003405P	2264	8.4	YES	DALLAS	DALLAS	75203	88	GOLDEN GATE BAPTIST CHURCH	32.745831	-96.80672	5125696	
39	002177P	2829	17.2	YES	DALLAS	MESQUITE	75150	75	SHARING LIFE COMM OUTREACH	32.818392	-96.63355	8003241	
40	026099P2	2980	11.4	YES	DALLAS	MESQUITE	75149	72	BUCKNER CRISIS RELIEF CENTER	32.767329	-96.60759	8880400	

Excel Solver

- Solver is an available add-in to MS Excel that handles sampling design
  - Objective Variable
  - Changing Variables
  - Constraints

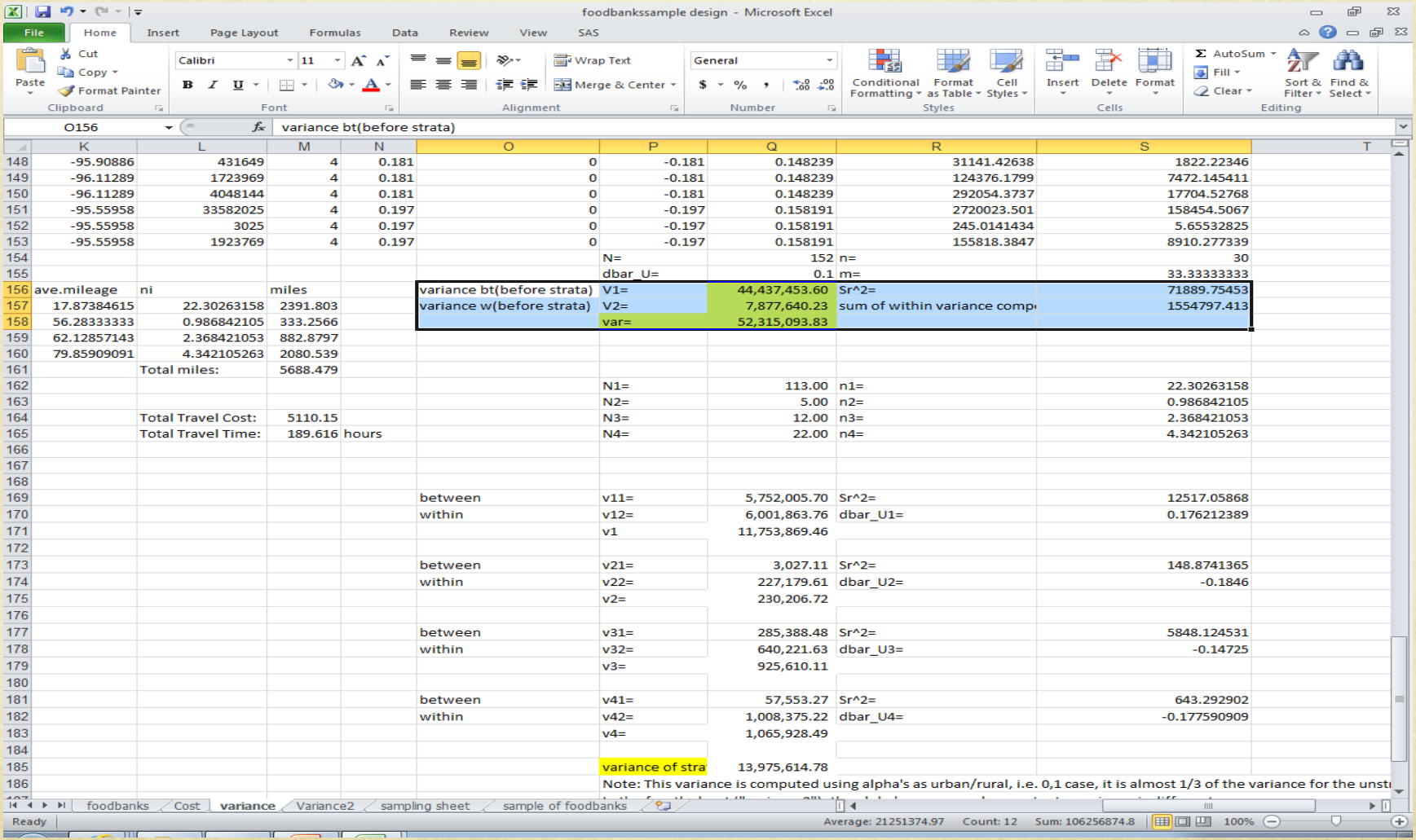
- Obtains an optimal solution when it exists

Excel Solver

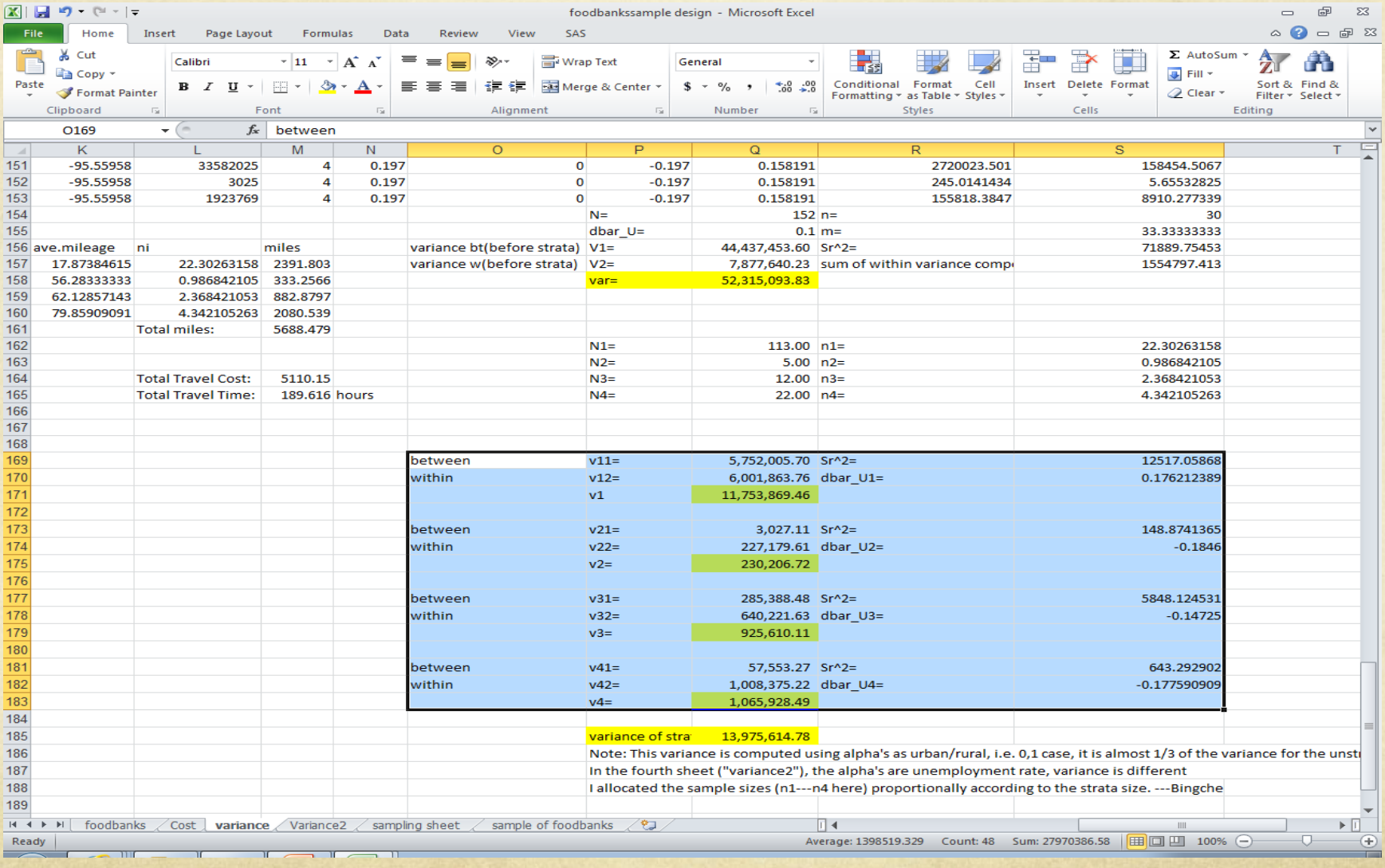


	K	L	M	N	O	P	Q	R	S	T
151	-95.55958	33582025	4	0.197	0	-0.197	0.158191	2720023.501	158454.5067	
152	-95.55958	3025	4	0.197	0	-0.197	0.158191	245.0141434	5.65532825	
153	-95.55958	1923769	4	0.197	0	-0.197	0.158191	155818.3847	8910.277339	
154						N=	152	n=	30	
155						dbar_U=	0.1	m=	33.33333333	
156	ave.mileage	ni	miles		variance bt(before strata)	V1=	44,437,453.60	Sr^2=	71889.75453	
157	17.87384615	22.30263158	2391.803		variance w(before strata)	V2=	7,877,640.23	sum of within variance comp	1554797.413	
158	56.28333333	0.986842105	333.2566			var=	52,315,093.83			
159	62.12857143	2.368421053	882.8797							
160	79.85909091	4.342105263	2080.539							
161		Total miles:	5688.479							
162						N1=	113.00	n1=	22.30263158	
163						N2=	5.00	n2=	0.986842105	
164		Total Travel Cost:	5110.15			N3=	12.00	n3=	2.368421053	
165		Total Travel Time:	189.616	hours		N4=	22.00	n4=	4.342105263	
166										
167										
168										
169					between	v11=	5,752,005.70	Sr^2=	12517.05868	
170					within	v12=	6,001,863.76	dbar_U1=	0.176212389	
171						v1	11,753,869.46			
172										
173					between	v21=	3,027.11	Sr^2=	148.8741365	
174					within	v22=	227,179.61	dbar_U2=	-0.1846	
175						v2=	230,206.72			
176										
177					between	v31=	285,388.48	Sr^2=	5848.124531	
178					within	v32=	640,221.63	dbar_U3=	-0.14725	
179						v3=	925,610.11			
180										
181					between	v41=	57,553.27	Sr^2=	643.292902	
182					within	v42=	1,008,375.22	dbar_U4=	-0.177590909	
183						v4=	1,065,928.49			
184										
185						variance of stra	13,975,614.78			
186						Note: This variance is computed using alpha's as urban/rural, i.e. 0,1 case, it is almost 1/3 of the variance for the unst				
187						In the fourth sheet ("variance2"), the alpha's are unemployment rate, variance is different				
188						I allocated the sample sizes (n1---n4 here) proportionally according to the strata size. ---Bingche				
189										

Excel Solver - Variance



Excel Solver - Variance



foodbankssample design - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View SAS

Cut Copy Format Painter Paste — Clipboard
Calibri 11 — Font
Wrap Text Merge & Center — Alignment
General — Number
Conditional Formatting Format as Table Cell Styles — Styles
Insert Delete Format — Cells
AutoSum Fill Clear Sort & Filter Find & Select — Editing

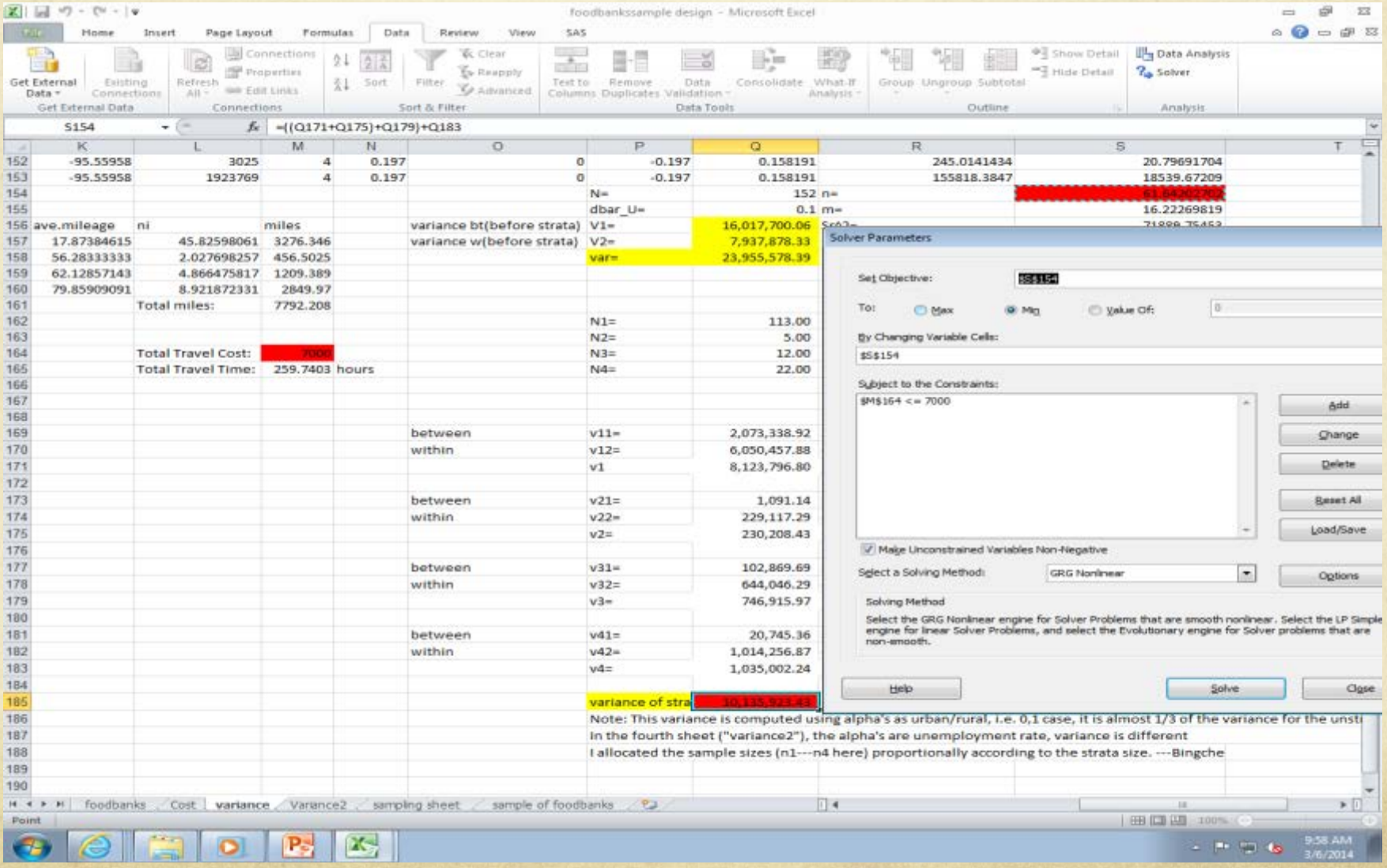
O169 between

	K	L	M	N	O	P	Q	R	S	T
151	-95.55958	33582025	4	0.197	0	-0.197	0.158191	2720023.501	158454.5067	
152	-95.55958	3025	4	0.197	0	-0.197	0.158191	245.0141434	5.65532825	
153	-95.55958	1923769	4	0.197	0	-0.197	0.158191	155818.3847	8910.277339	
154						N=	152	n=	30	
155						dbar_U=	0.1	m=	33.33333333	
156	ave.mileage	ni	miles		variance bt(before strata)	V1=	44,437,453.60	Sr^2=	71889.75453	
157	17.87384615	22.30263158	2391.803		variance w(before strata)	V2=	7,877,640.23	sum of within variance compo	1554797.413	
158	56.28333333	0.986842105	333.2566			var=	52,315,093.83			
159	62.12857143	2.368421053	882.8797							
160	79.85909091	4.342105263	2080.539							
161		Total miles:	5688.479							
162						N1=	113.00	n1=	22.30263158	
163						N2=	5.00	n2=	0.986842105	
164		Total Travel Cost:	5110.15			N3=	12.00	n3=	2.368421053	
165		Total Travel Time:	189.616	hours		N4=	22.00	n4=	4.342105263	
166										
167										
168										
169					between	v11=	5,752,005.70	Sr^2=	12517.05868	
170					within	v12=	6,001,863.76	dbar_U1=	0.176212389	
171						v1	11,753,869.46			
172										
173					between	v21=	3,027.11	Sr^2=	148.8741365	
174					within	v22=	227,179.61	dbar_U2=	-0.1846	
175						v2=	230,206.72			
176										
177					between	v31=	285,388.48	Sr^2=	5848.124531	
178					within	v32=	640,221.63	dbar_U3=	-0.14725	
179						v3=	925,610.11			
180										
181					between	v41=	57,553.27	Sr^2=	643.292902	
182					within	v42=	1,008,375.22	dbar_U4=	-0.177590909	
183						v4=	1,065,928.49			
184										
185						variance of stra	13,975,614.78			
186						Note: This variance is computed using alpha's as urban/rural, i.e. 0,1 case, it is almost 1/3 of the variance for the unst				
187						In the fourth sheet ("variance2"), the alpha's are unemployment rate, variance is different				
188						I allocated the sample sizes (n1---n4 here) proportionally according to the strata size. ---Bingche				
189										

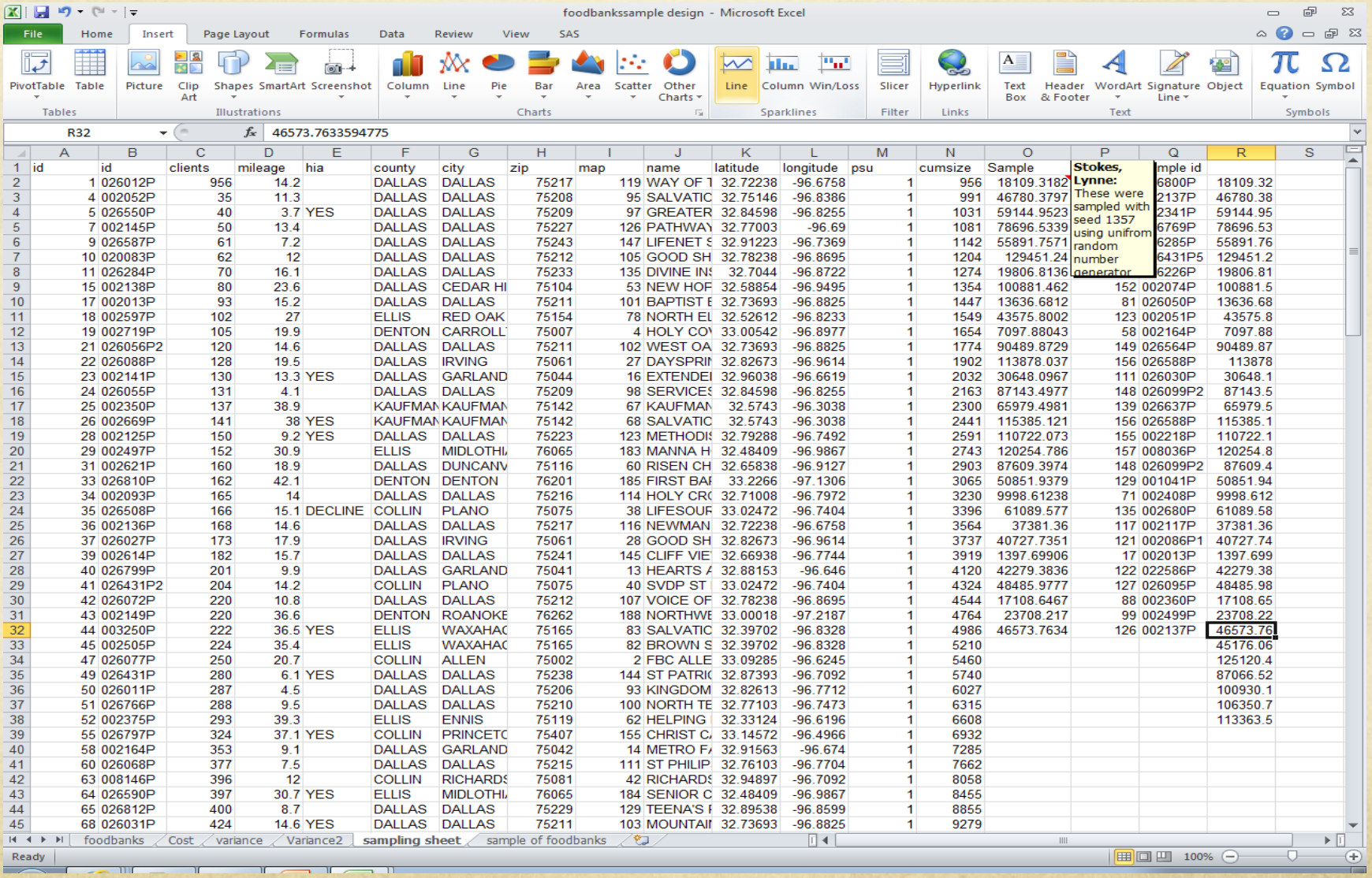
foodbanks Cost variance Variance2 sampling sheet sample of foodbanks

Ready — Average: 1398519.329 Count: 48 Sum: 27970386.58 — 100%

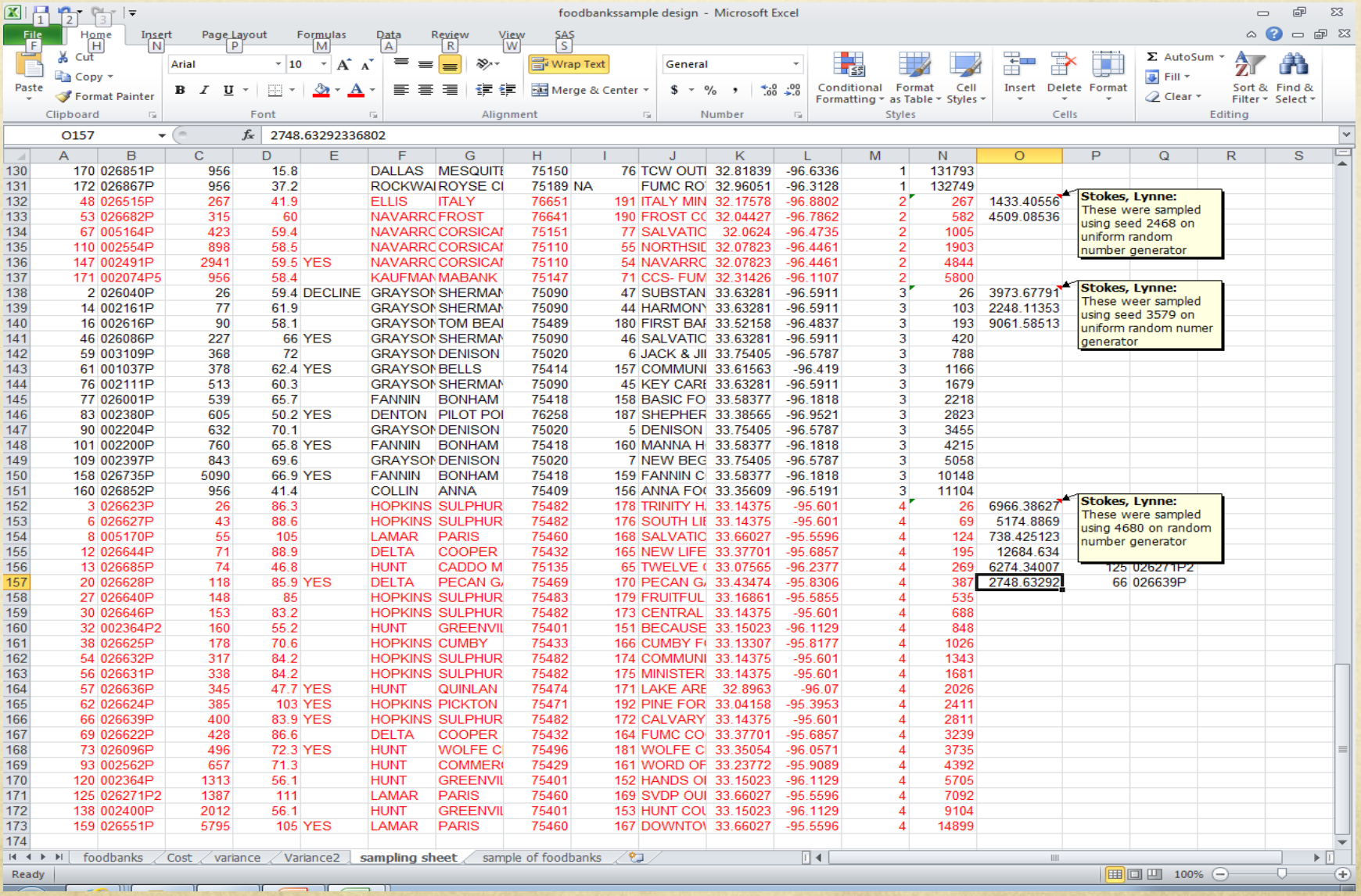
Excel Solver – Minimize Sample Size Based on Variance



Excel Solver – Final Sample



Excel Solver – Final Sample



Final Sample

- Eight Interviews per Day
  - 42 Food Banks
  - 24 Clients per Food Bank
- Twelve Interviews per Day
  - 30 Food Banks
  - 33 Clients per Food Bank

Other Practical Issues

- How do we select clients once we're actually inside the food banks?
  - As randomly as possible
  - No possible frame
  - Choose them systematically, i.e., every tenth person
- Non-response: Clients may choose not to participate

Concluding Remarks

- When in doubt, it's helpful to think of what we would do if we had more data.
- It's also helpful to start from a simple scenario where you know all the formulas and analysis and work your way up to a more complicated scenario.

Concluding Remarks

- This is not a math problem! We rarely have enough information to find the optimal solution, and if we do, it may be too complicated. Our goal is to do something reasonable.
- Part of doing something reasonable is to take a look at different scenarios and see how the results may be affected.
- If you spend time making a complicated design, you need to make sure the analysis is also done the right way.

References

- Stokes, L., & Plummer, J. (2004). Using Spreadsheet Solvers in Sample Design. *Computational Statistics & Data Analysis* , *44* (3), 527-546.

Special thanks to the STAT 6380 students, especially Chelsea Allen and Charles South, for helping finish this project and providing materials for this presentation.