A three-day short course sponsored by the Social & Economic Research Institute, Qatar University

Introduction to Survey Sampling

James M. Lepkowski
& Michael Traugott
Institute for Social Research
University of Michigan

April 29 – May 1, 2013

Day 1 April 29th 2013 Qatar University New Library Computer Lab room # 204	
8:30 am	Registration and Coffee
9:00 am – 11:30 am	Background: Course introduction, Simple random sampling methods Exercise 1. The sampling distribution Element sampling: A brief history of survey sampling Estimation of population means and proportions
11:30 am – 12:30 pm	Light Lunch/ Prayer break
12:30 pm – 2:00 pm	Sampling variance Sample size determination Exercise 2 Element sampling: Systematic sampling Exercise 3

Day 2: April 30th, 2013	
8:30 am	Registration and Coffee
9:00 am – 11:30am	Cluster sampling: Equal sized clusters Subsampling Design effects and intracluster homogeneity Exercise 4
11:30 am – 12:30 pm	Light Lunch/Prayer break
12:30 pm – 2:00 pm	Sampling unequal sized clusters Probability proportionate to size selection Exercise 5 Stratification: Purpose of stratification Stratified sampling estimates

Day 3: May 1st, 2013	
8:30 am	Registration and Coffee
9:00 am – 11:30 am	Determining sample allocation Exercise 6 Sampling problems: Frame problems Objective respondent selection
11:30 am – 12:30 pm	Light lunch/ Prayer break
12:30 pm – 2:00 pm	Weighting Exercise 7 General issues in variance estimation SESRI Sampling Methods

1. Overview of Surveys & Survey Sampling

- Where does sampling fit in?
- Sampling topics to be covered
  - Probability v. non-probability sampling
  - Population of inference
  - Sampling frames
  - Sample designs for list frames or widespread populations
  - Sample deficiencies
  - Weighting
  - Variance estimation for complex sample surveys

WHERE DOES SAMPLING FIT IN?

- **During conceptualization, a researcher considers the RELEVANT POPULATION for evaluating the theory/hypothesis**

- **In designing the data collection, the researcher has two concerns in mind:**
  - **External validity**
  - **Cost/benefit calculations for the overall cost of the study**

DIFFERENCES BETWEEN CENSUSES AND SAMPLES

A census involves an enumeration of a population. When the population is large:

1. **It is costly**
2. **It is time consuming**
3. **May not be feasible with precision**

(US Census as an example)

A sample involves a selection of a representative subset of a population in order to draw inferences to the population

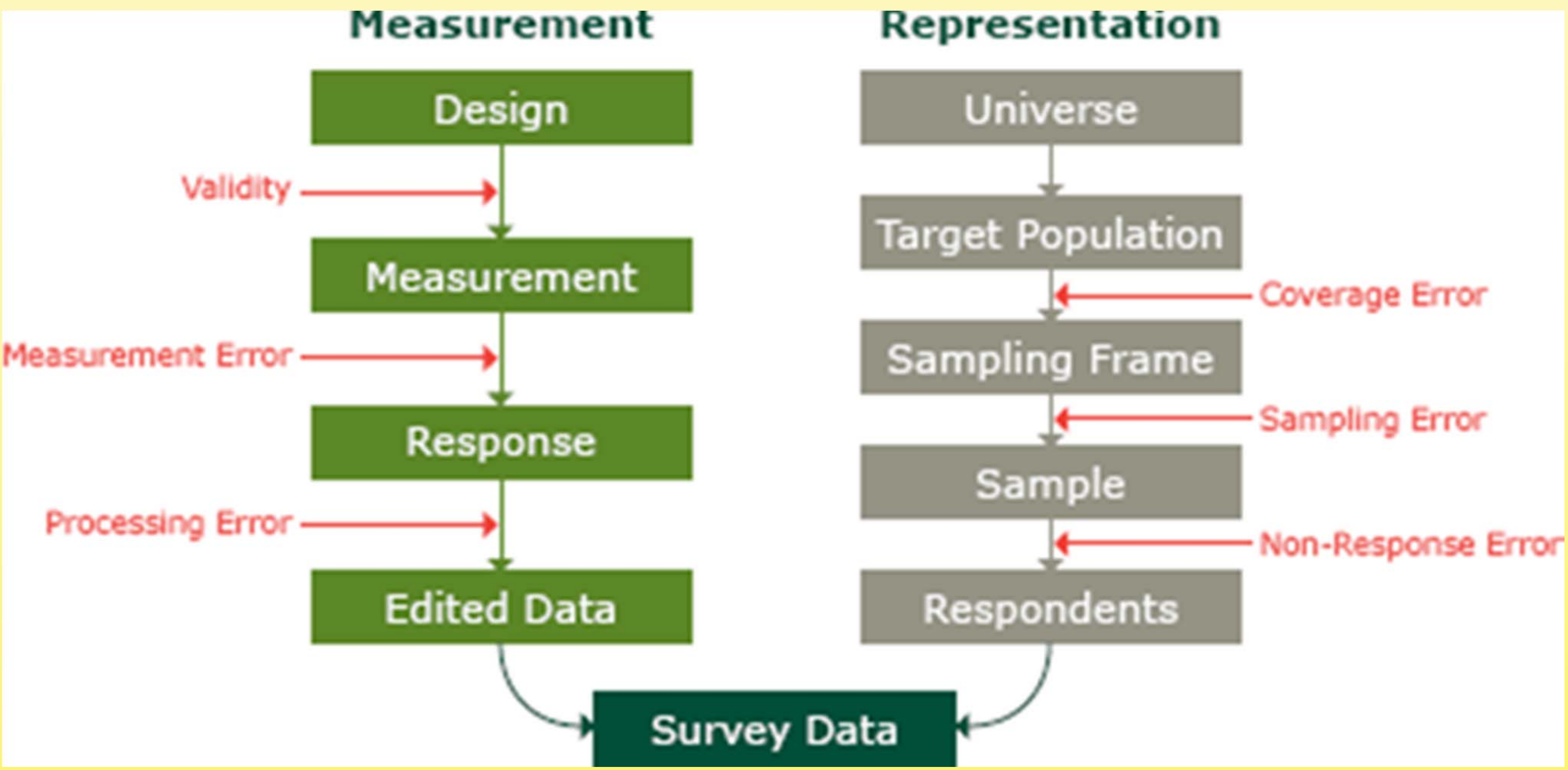
Collecting data from a sample of a large population is FAR LESS costly and FAR LESS time consuming

Greater Accuracy

- Because of the cost savings, sampling allows a researcher to devote
  - More resources to the collection of more data (variables)
  - The reduction of error in measurement (reliability and validity)
  - Better coverage of the units of analysis

- This fits in with the Total Survey Error perspective

Total Survey Error

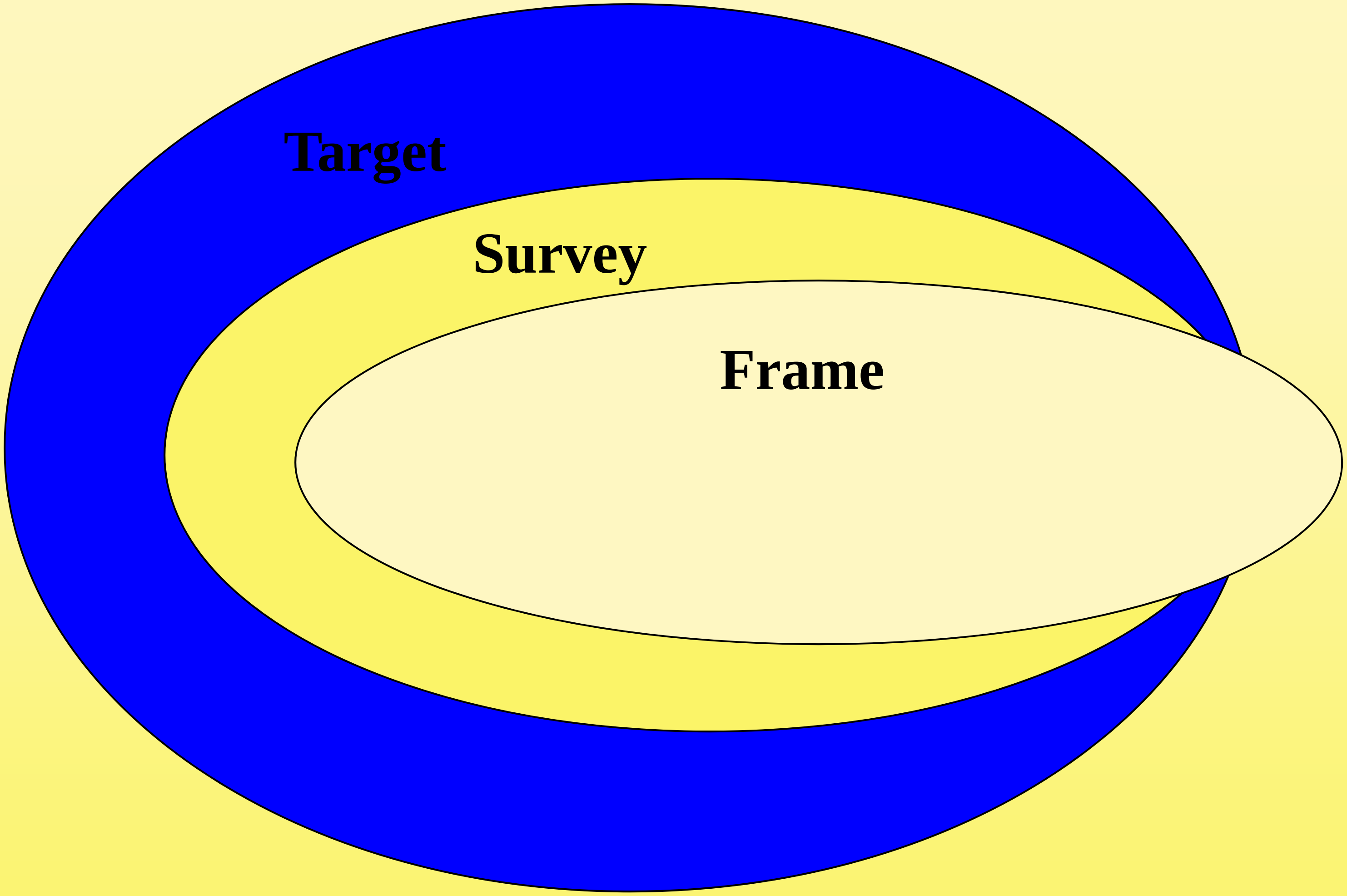


Probability v. non-probability sampling

- Non-probability sampling
  - Haphazard, convenience, or accidental sampling
  - Purposive sampling or expert choice
  - Quota sampling
- Probability sampling

Population of inference

- Target population
  - Geographical boundaries
  - Age limits
  - Date
- Survey population
  - Possible exclusions from target population
    - Institutionalized
    - Homeless
    - Nomads
    - Remote sparsely settled areas


Target
Survey
Frame

Sampling frame

- List frame
- Area frame
- Problems
  - Missing elements
  - Duplicate listings
  - Clusters
  - Blanks or ineligibles

Sample designs for compact populations

- Simple random sampling
- Systematic sampling
- Stratified sampling
  - Proportionate allocation
  - Disproportionate allocation

Sample designs for widespread populations

- Cluster sampling
  - One-stage (take all)
  - Two-stage (subsampling)
  - Multi-stage
- Probability proportionate to size sampling
- Stratified cluster sampling
- Systematic sampling of clusters

Sample deficiencies

- Nonresponse
  - Total/unit
  - Item
- Noncoverage
- Compensation: weighting
  - Unequal probabilities
  - Nonresponse
  - Noncoverage (poststratification)
    - Make the sample distribution conform to known population distribution

Variance estimation

- Standard software cannot handle complex sample designs correctly
- Methods of variance estimation
  – Taylor series approximation
  – Balanced or Jackknife repeated replication
- Computer software available for these methods
  – Requires stratum, cluster, and weight on each sample record

2. Simple random sampling

- Simple random sampling
- Exercise 1
- Table of random digits
- Faculty salaries

Simple random sampling (SRS)

- Rarely used in practice for large scale surveys
- Theoretical basis for other sample designs
- Sample size *n* from population size *N*
- Every element of the population has the same probability of selection (*epsem*) and every combination of size *n* has the same probability of selection

Selection and estimation

- Use a table of random numbers to select SRS samples
- Sample mean estimates population mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \longrightarrow \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Exercise 1

- The following table (3 pages) lists the salaries of n = 370 faculty members at a major midwestern university in 2013 in the U.S.
- For each faculty member there is a sequence number, an ID, division, rank, and 2010-2011 salary
- The list is ordered alphabetically by surname and given name (which are not shown)

Exercise 1

This is a group exercise.

Each group should select a simple random sample of n = 20 from the list.

Use the accompanying table of random numbers to select the sample.

Then compute the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

One member of the group should report the sample mean on behalf of the group.

Exercise 1:
Starting columns for groups

Group	Columns
1	1-3
2	41-43
3	81-83
4	11-13
5	31-33
6	51-53
7	21-23
8	36-38
9	56-68
10	61-63
11	66-68
12	71-73

Exercise 1: Table of Random Digits

Row	Column																	
	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80	81-85	86-90
1	49018	34042	72000	49522	85941	84723	51072	56454	67420	05025	25234	10671	05579	90906	54706	79486	57057	40468
2	97294	25351	12331	82557	13834	91334	32510	47165	08535	27491	87064	23579	72223	45164	98781	20189	17391	75145
3	97638	18356	31198	39366	37340	76043	77528	21714	44751	81797	28670	50973	07915	45259	45334	88904	47365	37249
4	34525	30477	75462	34635	51422	60669	62413	52524	79883	26235	46933	23381	72335	74702	77289	83419	28761	68996
5	79619	43993	89902	64817	88397	35390	44558	91500	87656	83603	00491	37693	75524	04058	77373	61598	60059	32241
6	54778	70353	54134	19513	89074	07807	74520	59684	47494	58194	29810	91489	45410	28737	55504	50467	94953	25565
7	12256	17900	33754	11853	65033	24106	41833	68345	62300	33076	70119	60498	70180	06929	34567	37075	57735	44602
8	33297	14796	91080	67108	85984	81892	37533	24643	37522	71461	96220	16177	04449	38396	09675	64290	96410	49117
9	75083	44991	46851	46383	00695	54453	34156	49854	68163	83123	89928	39667	15632	43854	04707	41766	01876	20016
10	66288	63908	74090	52902	69701	72959	64480	78123	81841	92675	08731	20577	94939	43211	63438	93640	75825	57922
11	84578	05698	92016	94285	26563	36372	55989	94790	36338	30640	81337	56599	05695	42896	57115	73143	49959	84903
12	55699	23402	30639	39508	41495	44462	11924	70471	97867	82637	18031	38020	70819	64948	17274	67345	31672	66155
13	51917	88538	58239	58633	80392	89447	81230	97654	52579	34888	06454	94398	16452	76723	00902	81924	73166	85669
14	36779	68538	88591	96616	84918	29413	99116	66987	41334	43877	00185	90070	43292	01754	01505	25362	39548	60933
15	49852	36333	84789	65346	46181	61218	54131	57370	64814	44430	43774	72286	11644	33071	74301	02154	37021	04828
16	66752	08578	57498	17884	83667	59532	73254	83347	85751	18536	55969	73265	06726	80734	29351	36800	77081	10687
17	61689	45570	53663	66779	85627	27662	34436	58824	18902	49414	05020	98033	85987	53127	72623	00983	92504	54686
18	19111	76703	32467	51391	85381	48433	68754	89843	02166	59177	80856	71628	27731	90073	04233	34913	46188	28778
19	46913	70576	16918	46675	02304	83330	55894	39684	20753	48885	72907	37048	80065	58931	78214	36397	97252	69593
20	22224	48264	96826	15434	52010	22811	07914	89541	61620	83346	96204	52742	27485	37716	71756	79244	04517	20831
21	84119	49920	29328	03239	15832	72406	94946	45797	70566	19586	26419	40852	70097	02276	93410	87952	71018	96533
22	75594	56191	18861	44995	44764	76960	12585	01842	19324	46085	33903	77234	07418	42805	21925	86305	12510	87281
23	34821	90491	28843	85959	72301	14576	94229	43353	55740	86145	73278	89446	36093	39173	07384	32388	17494	52734
24	23378	01578	09081	20536	31412	00632	16380	14876	26249	00449	26441	14765	05223	08297	54280	35937	02965	79389
25	09985	71346	32130	58906	97244	07003	91231	23396	47378	19064	01118	04376	83218	01890	94316	40309	41332	30966
26	43814	09227	11841	44516	62348	31284	58895	88559	19567	82425	00614	68626	10523	96822	79297	16858	52693	63887
27	26724	80216	75905	54725	46995	75504	79112	50571	57115	02600	35097	04329	78514	02663	48700	57166	30316	97649
28	37876	85859	19333	87221	44809	50700	57889	43075	99310	32235	62624	88356	51865	21946	52479	69599	29065	26434
29	23634	07454	63628	30531	52979	28534	03208	75663	33587	27738	04018	32256	32259	14042	27624	94889	91414	72658
30	10906	61337	16571	98829	96434	25748	01518	97758	93725	64532	79331	25961	82782	23354	47052	36078	12780	78331
31	09372	97239	72017	99537	99977	96404	04824	64248	68816	02734	38384	87274	18213	67600	18730	17870	02026	34180
32	86659	47171	96123	33853	64659	76657	53911	09900	70918	07733	89084	42345	22250	13583	52020	96144	25382	10875
33	78209	23140	94532	89438	43271	89616	63137	85026	15799	62580	70837	50071	74496	94191	45858	13545	66999	77390
34	15430	43742	77673	21745	34854	31505	05275	16758	58996	70211	97794	60918	98986	14446	72130	43056	13412	86691
35	64947	43432	14105	78393	03682	47498	75738	76250	69143	19799	31068	31261	31912	47359	26853	62917	40581	40772
36	71143	09505	65318	29034	89055	17744	48752	69171	08426	08827	14816	61969	68694	19168	67081	26010	68211	80384
37	03104	54280	49703	72368	99964	68555	57769	27567	55962	31100	26364	61603	48176	04177	00935	05130	83625	66323
38	56085	69548	50876	92855	52293	11580	22797	94044	67994	50651	26397	01782	73341	80486	72738	66943	75883	10106
39	41842	68437	92724	67791	21113	47124	28279	50647	09809	26717	48925	14686	24824	38530	62429	57330	33340	07994
40	28521	08035	30260	91407	04111	18581	84777	87116	96280	09202	31360	02923	83625	19821	35903	86927	36021	90593
41	85133	15310	42745	84831	82992	73756	67473	62066	83254	02735	55402	39765	92121	07338	39944	36882	74892	00148
42	28122	35506	71104	96492	90721	22225	23256	30415	63671	27160	19768	08441	38172	15357	73851	53381	20093	42073
43	56665	12467	44282	00817	58668	70312	66617	75720	93458	74491	72624	45673	68051	53523	58745	13730	93676	87636
44	19871	89889	70142	63766	71799	97398	23855	08350	11993	16729	23096	75940	45632	05786	46643	52563	30407	28338
45	48253	37932	79566	98774	02523	54942	15195	01354	03979	36909	21991	08828	45452	75565	90933	08713	36319	70259
46	80828	98357	85671	69918	30878	48784	81471	43729	60566	81014	68445	82593	59634	16601	05712	80642	26928	11496
47	09863	88615	26990	94808	32784	51992	60048	09830	75745	30593	64917	90209	55266	57533	68877	37486	91998	30055
48	05754	47499	53052	86074	01045	90121	12938	84746	55683	64345	22413	08513	04316	38192	73202	99160	56397	77063
49	32883	01773	11423	07799	12268	59983	60446	16744	12452	81457	56278	49040	31680	66267	05187	69329	28067	78017
50	82869	70040	36427	18798	57316	09565	11637	30597	11151	46114	30048	60952	48736	39133	79698	90272	80447	88785

Faculty Member Salaries (in $1,000)

Seq. No.	ID	Division	Sex	Ran	Salary	Seq. No.	ID	Division	Sex	Ran	Salary	Seq. No.	ID	Division	Sex	Ran	Salary
1	1	Eng&Prof	m	3	$88	51	155	Eng&Prof	m	3	$55	101	217	Lit&SocSci	m	2	$55
2	2	Medicine	f	3	$45	52	156	Biol&Sci	m	1	$49	102	218	Medicine	m	3	$80
3	9	Medicine	m	3	$57	53	157	Eng&Prof	m	3	$57	103	219	Eng&Prof	m	1	$114
4	11	Medicine	m	1	$133	54	158	Medicine	m	1	$118	104	220	Lit&SocSci	m	1	$63
5	12	Eng&Prof	f	2	$71	55	159	Medicine	m	3	$84	105	221	Medicine	m	1	$112
6	13	Lit&SocSci	m	1	$113	56	160	Eng&Prof	m	3	$52	106	222	Medicine	m	1	$93
7	14	Medicine	f	3	$65	57	161	Medicine	m	3	$64	107	223	Lit&SocSci	m	2	$47
8	15	Biol&Sci	m	3	$47	58	162	Eng&Prof	m	1	$75	108	224	Biol&Sci	m	1	$127
9	16	Lit&SocSci	f	3	$39	59	163	Medicine	f	1	$87	109	225	Eng&Prof	m	2	$121
10	17	Biol&Sci	m	1	$74	60	164	Eng&Prof	m	3	$58	110	226	Medicine	m	3	$58
11	18	Medicine	m	1	$88	61	165	Medicine	f	3	$39	111	227	Biol&Sci	f	3	$97
12	19	Lit&SocSci	m	1	$62	62	166	Medicine	m	3	$69	112	228	Lit&SocSci	m	1	$71
13	37	Lit&SocSci	m	1	$49	63	167	Medicine	f	2	$46	113	229	Eng&Prof	m	1	$72
14	38	Medicine	m	3	$88	64	179	Eng&Prof	f	1	$86	114	230	Lit&SocSci	m	3	$29
15	39	Medicine	m	1	$181	65	180	Medicine	m	3	$87	115	231	Medicine	m	2	$167
16	40	Eng&Prof	m	3	$63	66	181	Medicine	m	3	$59	116	232	Lit&SocSci	m	3	$36
17	41	Medicine	m	2	$94	67	182	Eng&Prof	f	3	$44	117	233	Medicine	m	1	$57
18	42	Eng&Prof	m	1	$91	68	183	Medicine	m	2	$123	118	234	Biol&Sci	m	1	$107
19	43	Medicine	m	1	$60	69	184	Lit&SocSci	f	3	$37	119	235	Medicine	m	2	$88
20	44	Eng&Prof	m	3	$55	70	185	Lit&SocSci	m	1	$106	120	236	Medicine	m	2	$87
21	45	Biol&Sci	m	2	$55	71	186	Lit&SocSci	m	1	$91	121	237	Lit&SocSci	f	2	$43
22	46	Medicine	f	1	$106	72	187	Lit&SocSci	m	1	$78	122	238	Lit&SocSci	m	1	$79
23	47	Medicine	m	1	$116	73	188	Biol&Sci	m	1	$77	123	239	Medicine	m	2	$113
24	48	Medicine	m	3	$79	74	189	Medicine	m	1	$90	124	240	Medicine	m	3	$55
25	49	Lit&SocSci	m	1	$61	75	190	Eng&Prof	m	2	$71	125	280	Medicine	m	3	$57
26	50	Lit&SocSci	f	3	$37	76	191	Medicine	f	3	$42	126	281	Eng&Prof	m	3	$56
27	51	Medicine	m	2	$72	77	192	Medicine	f	2	$59	127	282	Eng&Prof	m	2	$65
28	52	Eng&Prof	m	1	$105	78	193	Eng&Prof	m	2	$49	128	283	Medicine	m	2	$42
29	59	Medicine	m	2	$79	79	194	Biol&Sci	m	1	$83	129	284	Medicine	m	1	$102
30	133	Medicine	m	1	$61	80	195	Lit&SocSci	m	1	$34	130	285	Medicine	f	3	$40
31	134	Medicine	m	1	$86	81	196	Medicine	f	3	$42	131	286	Eng&Prof	m	3	$53
32	135	Biol&Sci	m	1	$103	82	197	Medicine	m	2	$97	132	287	Medicine	m	3	$82
33	136	Lit&SocSci	m	1	$48	83	198	Medicine	m	1	$109	133	288	Medicine	m	2	$64
34	137	Eng&Prof	m	2	$64	84	199	Lit&SocSci	f	2	$48	134	289	Eng&Prof	m	1	$72
35	138	Eng&Prof	m	1	$78	85	200	Medicine	m	1	$47	135	290	Biol&Sci	f	3	$36
36	139	Medicine	f	2	$53	86	201	Eng&Prof	m	2	$45	136	291	Lit&SocSci	f	1	$66
37	140	Biol&Sci	m	1	$85	87	202	Medicine	m	3	$83	137	292	Medicine	f	3	$66
38	141	Eng&Prof	m	1	$61	88	203	Medicine	m	2	$51	138	293	Medicine	m	2	$102
39	142	Medicine	m	1	$106	89	204	Biol&Sci	m	1	$78	139	294	Biol&Sci	m	1	$103
40	143	Lit&SocSci	m	2	$60	90	205	Lit&SocSci	m	1	$70	140	295	Medicine	m	1	$148
41	144	Biol&Sci	f	1	$73	91	206	Eng&Prof	f	2	$46	141	296	Lit&SocSci	f	1	$60
42	145	Medicine	m	1	$70	92	207	Eng&Prof	m	1	$85	142	297	Lit&SocSci	f	3	$46
43	147	Medicine	f	3	$32	93	208	Lit&SocSci	m	1	$53	143	298	Lit&SocSci	f	1	$57
44	148	Lit&SocSci	m	2	$49	94	209	Medicine	f	3	$40	144	299	Medicine	f	2	$50
45	149	Eng&Prof	m	3	$43	95	210	Eng&Prof	m	1	$87	145	300	Lit&SocSci	m	1	$90
46	150	Medicine	m	1	$75	96	211	Lit&SocSci	m	1	$71	146	301	Eng&Prof	m	3	$63
47	151	Lit&SocSci	m	1	$92	97	212	Medicine	m	1	$75	147	303	Eng&Prof	m	1	$80
48	152	Medicine	m	2	$107	98	214	Biol&Sci	m	1	$85	148	304	Medicine	m	3	$56
49	153	Biol&Sci	m	2	$57	99	215	Lit&SocSci	m	2	$50	149	305	Medicine	m	1	$72
50	154	Medicine	m	2	$114	100	216	Medicine	m	3	$118	150	306	Eng&Prof	m	1	$96

Faculty Member Salaries (Continued)

Seq. No.	ID	Division	Sex	Ran	Salary	Seq. No.	ID	Division	Sex	Ran	Salary	Seq. No.	ID	Division	Sex	Ran	Salary
151	307	Medicine	m	3	$65	201	440	Medicine	m	1	$108	251	496	Medicine	m	3	$60
152	308	Lit&SocSci	m	3	$37	202	441	Lit&SocSci	m	1	$48	252	497	Eng&Prof	m	1	$86
153	309	Eng&Prof	m	1	$127	203	442	Medicine	m	3	$85	253	498	Medicine	m	1	$134
154	310	Lit&SocSci	m	1	$90	204	443	Lit&SocSci	m	1	$59	254	499	Medicine	f	3	$63
155	311	Lit&SocSci	m	3	$45	205	444	Lit&SocSci	f	1	$63	255	500	Medicine	m	1	$123
156	312	Eng&Prof	f	1	$75	206	445	Lit&SocSci	f	2	$46	256	501	Medicine	m	3	$85
157	313	Medicine	m	2	$60	207	446	Medicine	f	3	$41	257	502	Medicine	f	3	$42
158	314	Lit&SocSci	m	2	$57	208	447	Medicine	m	3	$71	258	503	Medicine	f	2	$83
159	315	Medicine	m	1	$129	209	448	Eng&Prof	f	3	$44	259	504	Lit&SocSci	m	1	$54
160	316	Eng&Prof	m	1	$102	210	449	Lit&SocSci	m	2	$46	260	505	Lit&SocSci	f	1	$66
161	317	Eng&Prof	m	3	$57	211	450	Medicine	m	3	$85	261	506	Medicine	m	1	$84
162	318	Eng&Prof	m	3	$61	212	452	Medicine	m	1	$119	262	507	Eng&Prof	m	3	$46
163	319	Eng&Prof	m	1	$93	213	453	Medicine	m	2	$69	263	508	Eng&Prof	m	1	$90
164	320	Medicine	f	3	$41	214	454	Eng&Prof	m	3	$74	264	509	Medicine	m	2	$76
165	321	Medicine	m	1	$181	215	455	Biol&Sci	m	1	$59	265	510	Eng&Prof	m	1	$88
166	322	Medicine	f	2	$69	216	456	Biol&Sci	m	1	$53	266	515	Medicine	f	1	$87
167	323	Lit&SocSci	m	1	$81	217	457	Medicine	f	3	$49	267	516	Eng&Prof	m	3	$75
168	324	Biol&Sci	m	1	$94	218	459	Eng&Prof	m	1	$78	268	517	Eng&Prof	m	3	$64
169	325	Lit&SocSci	m	2	$53	219	460	Biol&Sci	m	1	$68	269	518	Biol&Sci	f	3	$52
170	326	Medicine	m	3	$48	220	461	Eng&Prof	m	1	$83	270	519	Medicine	m	2	$109
171	327	Lit&SocSci	m	1	$83	221	462	Eng&Prof	m	1	$105	271	520	Lit&SocSci	m	1	$144
172	328	Lit&SocSci	m	1	$47	222	463	Lit&SocSci	m	3	$37	272	521	Eng&Prof	m	2	$79
173	329	Lit&SocSci	m	3	$45	223	464	Medicine	m	1	$111	273	522	Biol&Sci	m	1	$56
174	330	Medicine	f	1	$75	224	465	Medicine	f	2	$70	274	530	Biol&Sci	m	1	$60
175	331	Medicine	m	3	$49	225	466	Eng&Prof	m	1	$57	275	531	Biol&Sci	m	3	$52
176	333	Medicine	m	3	$53	226	467	Eng&Prof	m	1	$71	276	532	Lit&SocSci	f	2	$45
177	334	Eng&Prof	m	1	$84	227	468	Biol&Sci	m	3	$36	277	533	Lit&SocSci	m	1	$59
178	335	Eng&Prof	m	1	$78	228	469	Eng&Prof	f	3	$43	278	534	Eng&Prof	m	3	$56
179	336	Lit&SocSci	m	1	$102	229	470	Eng&Prof	m	1	$120	279	535	Medicine	m	1	$123
180	337	Lit&SocSci	f	2	$50	230	471	Lit&SocSci	m	1	$66	280	536	Medicine	m	2	$75
181	338	Medicine	f	2	$49	231	472	Eng&Prof	m	1	$84	281	537	Eng&Prof	m	1	$84
182	339	Medicine	m	1	$54	232	473	Medicine	m	2	$99	282	538	Medicine	m	2	$70
183	340	Medicine	m	3	$35	233	474	Biol&Sci	f	1	$91	283	539	Medicine	m	3	$84
184	341	Medicine	m	2	$87	234	475	Eng&Prof	m	2	$105	284	540	Eng&Prof	m	1	$63
185	342	Lit&SocSci	m	1	$52	235	476	Medicine	f	2	$60	285	541	Eng&Prof	m	1	$121
186	343	Lit&SocSci	m	1	$75	236	477	Medicine	f	3	$34	286	542	Medicine	m	1	$52
187	344	Medicine	f	3	$41	237	478	Medicine	f	3	$42	287	543	Biol&Sci	m	1	$73
188	345	Eng&Prof	m	2	$62	238	479	Medicine	m	2	$80	288	544	Eng&Prof	f	3	$32
189	346	Medicine	m	1	$79	239	480	Medicine	m	1	$94	289	545	Eng&Prof	f	3	$40
190	347	Biol&Sci	m	3	$37	240	481	Biol&Sci	m	1	$57	290	546	Biol&Sci	m	3	$47
191	348	Lit&SocSci	m	3	$44	241	482	Medicine	m	1	$82	291	547	Medicine	m	1	$112
192	349	Lit&SocSci	m	3	$47	242	483	Lit&SocSci	m	1	$70	292	548	Biol&Sci	m	1	$68
193	353	Medicine	m	1	$70	243	484	Lit&SocSci	m	1	$75	293	550	Medicine	m	2	$93
194	433	Lit&SocSci	m	1	$113	244	485	Medicine	m	1	$139	294	551	Medicine	m	1	$124
195	434	Medicine	m	3	$55	245	486	Lit&SocSci	m	1	$40	295	552	Lit&SocSci	f	2	$49
196	435	Lit&SocSci	m	1	$50	246	488	Lit&SocSci	m	2	$60	296	556	Medicine	f	3	$65
197	436	Lit&SocSci	f	2	$54	247	489	Eng&Prof	f	1	$128	297	557	Eng&Prof	m	1	$84
198	437	Eng&Prof	m	3	$53	248	490	Medicine	m	3	$47	298	558	Medicine	f	2	$71
199	438	Biol&Sci	m	1	$79	249	491	Eng&Prof	m	3	$67	299	559	Medicine	f	3	$40
200	439	Biol&Sci	m	2	$53	250	495	Eng&Prof	m	1	$90	300	560	Medicine	m	2	$70

Faculty Member Salaries (Continued)

Seq. No.	ID	Division	Sex	Ran	Salary	Seq. No.	ID	Division	Sex	Ran	Salary	Seq. No.	ID	Division	Sex	Ran	Salary
301	561	Eng&Prof	m	1	$98	351	636	Lit&SocSci	f	1	$72						
302	562	Lit&SocSci	m	1	$89	352	637	Eng&Prof	m	1	$94						
303	563	Medicine	f	3	$36	353	638	Eng&Prof	m	3	$52						
304	564	Medicine	m	1	$63	354	639	Biol&Sci	m	1	$66						
305	565	Eng&Prof	m	2	$74	355	640	Eng&Prof	m	3	$68						
306	566	Medicine	f	3	$38	356	641	Lit&SocSci	m	1	$89						
307	567	Eng&Prof	m	3	$76	357	642	Medicine	m	2	$148						
308	568	Medicine	m	3	$97	358	643	Medicine	m	1	$159						
309	569	Medicine	m	1	$76	359	644	Biol&Sci	m	1	$62						
310	570	Eng&Prof	m	1	$86	360	645	Lit&SocSci	m	1	$70						
311	571	Medicine	m	3	$59	361	646	Medicine	f	3	$109						
312	572	Medicine	f	2	$60	362	647	Eng&Prof	m	1	$120						
313	573	Lit&SocSci	m	2	$45	363	648	Eng&Prof	m	1	$112						
314	595	Biol&Sci	m	2	$56	364	649	Medicine	m	2	$90						
315	596	Lit&SocSci	m	1	$63	365	650	Medicine	m	1	$108						
316	597	Lit&SocSci	m	1	$69	366	651	Eng&Prof	m	1	$152						
317	598	Eng&Prof	m	1	$138	367	652	Medicine	f	2	$47						
318	599	Lit&SocSci	f	3	$31	368	653	Medicine	m	1	$116						
319	600	Medicine	f	2	$50	369	654	Biol&Sci	m	1	$77						
320	601	Eng&Prof	m	1	$89	370	655	Biol&Sci	M	1	$57						
321	602	Eng&Prof	m	1	$148												
322	603	Lit&SocSci	m	3	$55												
323	604	Lit&SocSci	m	1	$81												
324	605	Lit&SocSci	m	1	$52												
325	606	Medicine	m	3	$85												
326	607	Medicine	m	1	$132												
327	608	Lit&SocSci	m	1	$85												
328	609	Eng&Prof	m	1	$66												
329	610	Eng&Prof	f	1	$94												
330	611	Eng&Prof	m	2	$77												
331	612	Medicine	f	2	$76												
332	613	Medicine	m	1	$109												
333	614	Lit&SocSci	m	1	$99												
334	616	Eng&Prof	f	2	$78												
335	617	Eng&Prof	m	1	$98												
336	618	Medicine	f	3	$41												
337	619	Medicine	f	3	$37												
338	620	Eng&Prof	m	3	$89												
339	622	Biol&Sci	m	2	$55												
340	623	Lit&SocSc	m	1	$52												
341	624	Eng&Prof	m	3	$42												
342	625	Biol&Sci	m	2	$52												
343	626	Lit&SocSc	m	1	$63												
344	627	Lit&SocSc	m	1	$95												
345	628	Medicine	f	3	$75												
346	629	Medicine	f	3	$106												
347	630	Lit&SocSc	f	3	$44												
348	631	Lit&SocSc	m	1	$58												
349	632	Lit&SocSc	m	1	$79												
350	633	Lit&SocSc	m	1	$135												

3. Historical perspective

- Historical development
- The beginnings
- Development
- Divergence
- Framework for comparison
- Selection bias
- Development, part II
- What should *we* do?

Historical development

- Sampling practice:
  - Result of attempts to solve practical problems
- Function of theory
  - Formalize implicit assumptions, and confirm, correct, or extend practice
- Origins
  - Data gathering
    - health and social problems
    - social physics
  - Census
  - Monography

The beginnings

- Berne, 1895
  - Kaier at ISI: Representative method
    - Miniature of country
    - Large number of units
    - Use prior information in selection
  - Von Mayr and others
    - No calculation where observation is possible
    - Cf. Godambe, Basu after 1950
  - Cheysson and others
    - Monography: detailed examination of typical cases

Development

- 1903 ISI Resolution
  - Four implicit principles
    - Representative
    - Objective
    - Measurability
    - Specification
  - Actuality
    - Multistage proportionate stratified samples (no theory)

Divergence

- Representative
  - Purposive sampling
  - Expert choice
  - Balanced sampling
- Objective
  - Randomized selection
  - Bowley, 1906 (colleague of R.A. Fisher)

Separation

- ISI Commission 1926 report
  - Sampling established as basis for information collection
  - Equal status given to random and purposive sampling
  - No theory for unequal sized clusters
- No basis for comparing the two methodologies

Framework for comparison

- Neyman, 1934
- The sampling distribution
  - Properties of sample under repeated sampling
    - All possible samples and their associated probabilities of occurrence
  - The sampling distribution of an estimator

Conditions for inference

- Conditions under which different procedures will produce valid estimates
  - Probability sampling
    - “Unbiased” irrespective of population structure
  - Purposive/balanced/quota sampling
    - Tough assumptions about population structure, unlikely to be achieved in practice

Selection bias

- Italian census storage problem
- Sample of completed forms to be retained
- Gini and Galvani, 1929
  - Matched sample communes on 7 variables
  - Other variables, even aspects other than means of 7 variables, showed wide deviations from population values

What should *we* do?

- Probability sampling for objectivity
- Stratification for precision (representativeness)
- Variance estimation from the sample
- Complete and comprehensible description of the sampling procedure

4. Element samples

- Element samples
- The sampling distribution
- Properties of the sampling distribution
- Central limit theorem
- Properties of the sample mean for SRS
- Estimation of variance
- Determination of sample size
- Formulas
- Exercise 2

Element samples

- A sample design for which the unit of selection is the population element
- Basic framework: Neyman, 1934
  - Must be applicable to all populations
  - Must not depend on assumptions about the population structure
  - Appropriate for large populations of elements

Element samples

- Repeated sampling
  - Objective (mechanical) selection of elements
  - Consider possible outcomes of the sampling process
  - Evaluation of the whole set of possible outcomes

The sampling distribution

- The set of all possible values of the estimator that can be obtained with a given sample design
  - For a given sample we obtain a particular value, the estimate (such as $\bar{y}$)
- We want to know …
  - … how likely is the estimate to be close to the population value

Sample realization

- In fact, we select just one sample
- The estimate may be correct, or incorrect
- Want to maximize the probability of a satisfactory estimate

Properties of the sampling distribution

- Unbiasedness
  - Expected value (average value): $E\left(\overline{y}\right)$
- Variability from one sample to another
  - Variance of the estimator $Var(\overline{y})$
  - The square root of the variance is called the standard error of the estimator: $\sqrt{Var(\overline{y})}$
- Measurable design
  - A design for which the variance can be estimated from the sample itself

Central limit theorem

- For large samples, the sampling distribution of $\overline{y}$ is Normal
- Confidence intervals $\overline{y} \pm z_{(1-\alpha/2)} \sqrt{Var(\overline{y})}$

Properties of the sample mean for SRS

- Unbiased $E\left(\bar{y}\right)$
- Variance
  - Consider $Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$
    - $1 - f = 1 - \frac{n}{N}$
    - S^2
    - n
  - Where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$ and $S^2 = P\left(1 - P\right)$

Estimation of variance

- Can use s^2 (sample) to estimate S^2 (population)
- Estimate of $Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$ (population)
  - $var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$ (sample)
- From a single sample we can not only estimate $\bar{Y}$ using $\bar{y}$ but also estimate the precision of $\bar{y}$ using $var(\bar{y})$
- Note that $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ and $s^2 = p\left(1-p\right)$ for a proportion

Determination of sample size

- What sample size do we need to obtain a given standard error of the estimator?
- S^2 population variance known (or guessed)
  - Census
  - Other surveys
  - Administrative records
- Desired standard error
  - Policy requirements in terms of $\sqrt{Var(\bar{y})}$
  - Decision making requirements

Sample size formulas

- In general, $Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$

- For an infinitely large population (or for sampling with replacement), this is
$$Var(\bar{y}) = \frac{S^2}{n}$$
- We can calculate the necessary sample size to achieve variance $Var(\bar{y})$ as $n = S^2 / Var(\bar{y})$

Sample size formulas (continued)

- In general (that is, not assuming *N* is large), the variance may be expressed as

$$Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \frac{S^2}{n'}$$

  – Where $n' = n \Big/ \left(1 - \frac{n}{N}\right)$

Sample size formulas (continued)

- We can compute the necessary n' as

$$n' = \frac{S^2}{Var(\bar{y})}$$

- To calculate the *n* necessary for a population of a particular size, we use the formula

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

Exercise 2

- The variability in income levels is comparable across many countries
- For a country with a value of $S = 2,000$ (which would give $S^2 = 4,000,000$), we want an estimate of the mean income which has a standard error ($\sqrt{Var(\bar{y})}$) of 50.
- Answer the following questions in groups:

Exercise 2 (continued)

Calculate the sample size needed in China with N = 1,400,000,000?

What about in the US where N = 320,000,000?

What about in Qatar where N = 1,700,000?

What about in a small city where N = 100,000?

What about in a small town where N = 10,000?

5. Systematic sampling

- Systematic sampling
- Problems with intervals in systematic sampling
- Solutions
- Exercise 3

Systematic sampling

- A simple method of selecting a sample from a list
- Once the first element is chosen, every *k* th element is selected by counting through the list sequentially
- In probability sampling, the first element is chosen at random

Sampling intervals

- Determine the sampling interval $k = N/n$
- Select a random number (RN) from 1 to *k*
- Add *k* repeatedly
- Example:
  - *N* = 12,000 dwellings in a city
  - Sample of *n* = 500 required
  - *k =* 12,000/500 = 24
  - Take a RN from 01 to 24, say 03
  - Take the third dwelling, and every 24th thereafter: 3, 27, 51, *etc*.

Problems with intervals

- Take 1 in *k* where $k = N/n$
- *k* may not be an integer
- Examples
  - N = 9, n = 2, and k = 4.5
  - N = 952, n = 200, and k = 4.76
  - N = 170,345, n = 1,250, and k = 136.272

Solutions: round sampling interval

- Round the fractional interval
  - Let the sample size vary, depending on the choice of the "integer interval" k
  - Example: N = 9, n =2, take k = 4 or 5
    - If k = 4 and RN = 1, the sample is elements 1, 5, 9.
    - If RN = 2, 3, or 4, the sample has only two elements
    - If k = 5 and RN = 1, 2, 3, or 4, the sample has two elements
    - If RN = 5, the sample has only one element
  - Under this method, what happened when N = 952 and n = 200?
  - What about for N = 170,345 and n = 1,250?

Solutions: elimination or duplication

- Eliminate, or duplicate, population elements by *epsem* to get exact multiple
  - Example: N = 9 and n = 2. Eliminate one of 9 at random, and take 1 in 4 of remaining 8.
  - If N = 952 and n = 200, duplicate 48 at random, and take 1 in 5 from the 1,000 listed elements
  - If N = 170,345 and n = 1,250, eliminate 345 at random, and take 1 in 136 of the remainder

Solutions: circular list

- Treat the list as circular
- Select one element at random from anywhere on the list
- Take every [*k*]th thereafeter, where [*k*] is an integer near *N*/*n*, until *n* selections are made

Exercise 3

- Consider again the list of 370 faculty member salaries given in Exercise 1 (slides 23-25)
- Suppose again we seek a sample of n = 20 from this list

Each group should select two systematic samples of *n* = 20 from the list using as random starts the next appropriate numbers from the random number table (slide 22) -- that is, the next random number after the last one used in Exercise 1

Exercise 3 (continued)

Each group should select two systematic samples of n = 20 from the list using as random starts the next appropriate numbers from the random number table (slide 22) -- that is, the next random number after the last one used in Exercise 1

Since N/n is not an integer, use for one sample the rounding method (letting the sample size vary depending on the choice of k) for the first sample

And the circular list method for the second sample

For each sample, compute the mean salary $\bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i$

6. Cluster sampling

- Cluster sampling
- Equal-sized cluster sampling
- Effective sample size
- Design effect
- Intra-class correlation
- Exercise 4

Cluster sampling

- Populations widely distributed geographically
- Cannot afford to visit *n* sites drawn randomly from the entire area
- Cluster sampling reduces the cost of data collection
  - Sample schools and children within them
  - Sample blocks and households within them

Cluster sampling

- Cluster sampling is also useful when the sampling frame lists clusters and not elements
  - Select clusters and list elements in selected clusters
  - Frame of blocks: list households within selected blocks
- Clusters are often naturally occurring units
  - Facilitates sample selection

Cluster sampling

- Suppose we select an SRS of *a* = 10 classrooms from *A* = 1,000, and examine the immunization history of all *b* = 24 children in selected classrooms
- Here $n = a \cdot b = 240$
- We refer to the *A* classrooms as primary sampling units or PSU's

Cluster sampling

- For each of the *a* = 10 selected PSU's, we record the number of children immunized:

$$\frac{9}{24},\frac{11}{24},\frac{13}{24},\frac{15}{24},\frac{16}{24},\frac{17}{24},\frac{18}{24},\frac{20}{24},\frac{20}{24},\frac{21}{24}$$

- Adding the numerators, there are 160 immunized children
- The overall proportion immunized is

$$p = 160/240 = 0.67$$

Cluster sampling

- Recall for SRS (without replacement selection of *n* elements), the sample mean was $\bar{y} = \sum_{i=1}^{n} y_i \Big/ n$
- The estimated sampling variance is

$$\text{var}\left(\bar{y}\right) = \left(1 - f\right) s^2 \big/ n$$

- But for an SRS of *a* equal-sized clusters from *A*, we have a p_α for each selected PSU

Cluster sampling: variance estimation

- In cluster sampling, treat the sample as an SRS of *a* units from *A*:

$$\mathrm{var}(p) = \frac{(1-f)}{a} s_a^2$$

  – Where $s_a^2 = \sum_{\alpha=1}^{a} (p_\alpha - p)^2 \Big/ (a-1)$

$$f = a / A$$

- That is, $\mathrm{var}(p) = \frac{(1-f)}{a} \frac{\sum_{\alpha=1}^{a} (p_\alpha - p)^2}{a-1}$

Cluster sampling: estimated variance

- For the illustration,

$$s_a^2 = \frac{1}{10-1}\left[\left(\frac{9}{24}-\frac{160}{240}\right)^2+\left(\frac{11}{24}-\frac{160}{240}\right)^2+\ldots\right]$$

$$= 0.02816$$

$$\mathrm{var}(p) = (1-f)\,s_a^2/a = 0.002760$$

$$se(p) = \sqrt{\mathrm{var}(p)} = 0.0525$$

Design effect

- If the sample had instead been an SRS of *n* = 240 children from all schools, then

$$p = 160/240$$

$$\text{var}_{SRS}(p) = (1-f)\frac{p(1-p)}{n-1}$$

$$= 0.0009112$$

Design effect

- Compared to cluster sampling, the estimated variance of *p* is considerably smaller for SRS
- A ratio quantifies the comparison:

$$deff\left(p\right)=\frac{\text{var}\left(p\right)}{\text{var}_{SRS}\left(p\right)}=\frac{0.002760}{0.0009112}=3.029$$

roh

- The design effect is a function of …
  - the size of the clusters *b*
  - the degree of homogeneity of elements within clusters
- The homogeneity is measured by the intra-cluster correlation *roh*
- The design effect is given by

$$deff\left(p\right)=1+\left(b-1\right)roh$$

Estimating roh

- The intra-cluster correlation can be estimated from the design effect:

$$
\begin{aligned}
roh &= \frac{deff\left(p\right)-1}{b-1} \\
&= \frac{3.029-1}{24-1} \\
&= 0.088
\end{aligned}
$$

Features of roh

- *roh* is a property of the clusters and the variable under study
- *roh* is substantive, not statistical
- *roh* is nearly always positive
  – Elements in a cluster tend to resemble one another
- Source of *roh*
  – Environment
  – Self-selection
  – Interaction

Magnitude of r*oh*

- Magnitude depends on
  - The characteristic (variable) under study (*e.g.*, disease status, age)
  - The nature of the clusters (*e.g.*, households, establishments)
  - The size of the cluster (*e.g.*, household, blocks of household, census tracts)

Effective sample size

- Alternatively, the actual sample size is n = 240 in the cluster sample, but an SRS that is equally precise would only have to have

$$n_{eff} = \frac{240}{3.209} = 79$$

Examples

- Consider alternative outcomes for our sample of *a* = 10 classrooms
  - Homogeneity with, heterogeneity between

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

$$s_a^2 = 0.2222 \quad \text{var}\left(p\right) = 0.02178$$

$$deff = 23.90 \qquad roh = \frac{23.90-1}{24-1} = 0.996$$

$$n_{eff} = 240/23.9 = 10$$

Examples

- Heterogeneity within, homogeneity among:

$$\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24}$$

$$s_a^2 = 0.0 \quad \mathrm{var}\left(p\right) = 0.0$$

$$deff = 0$$

$$n_{eff} = 240/0$$

Exercise 4

- An equal probability (*epsem*) sample of n = 2,400 was obtained from a one-stage sample of 60 equal-sized clusters selected by SRS
- In a journal article describing survey results, we found the following information
  - For a key proportion, $p = 0.40$
  - And $\operatorname{var}(p) = 0.00021795$

Estimate *deff* and *roh*

7. Two-stage sampling

- Two-stage sampling
- Portability of *roh*
- Exercise 5

Two-stage sampling

- Selecting many elements per cluster increases variances
- Even small values of *roh* can be magnified by large *b* since

$$deff\left(p\right)=1+\left(b-1\right)roh$$

- Consider the following for

$$n = a \cdot b = 240$$

$$f = \frac{a}{1000} \cdot \frac{b}{24} = \frac{a \cdot b}{24000} = \frac{240}{24000} = \frac{1}{100}$$

Subsamples of size *b*

- Sample *a* = 20 classrooms and *b* = 12:

$$deff\left(p\right)=1+\left(12-1\right)\times 0.088=1.97 \quad n_{eff}=122$$

- Sample *a* = 30 classrooms and *b* = 8:

$$deff\left(p\right)=1+\left(8-1\right)\times 0.088=1.62 \quad n_{eff}=148$$

- Sample *a* = 80 classrooms and *b* = 3:

$$deff\left(p\right)=1+\left(3-1\right)\times 0.088=1.18 \quad n_{eff}=204$$

Portability of *roh*

- Estimation

$$\text{var}_{(1)}(p) = \frac{(1-f)}{a} s_a^2$$

$$\text{var}_{(1),SRS}(\bar{y}) = \frac{p(1-p)}{n_{(1)}}$$

$$deff_{(1)} = \frac{\text{var}_{(1)}(p)}{\text{var}_{(1),SRS}(p)}$$

$$roh = \frac{deff_{(1)} - 1}{b_{(1)} - 1}$$

- Design

$$\text{var}_{(2)}(p) = deff_{(2)} \times \text{var}_{(2),SRS}(p)$$

$$\text{var}_{(2),SRS}(\bar{y}) = \frac{p(1-p)}{n_{(2)}}$$

$$deff_{(2)} = 1 + (b_{(2)} - 1) roh$$

roh

Exercise 5

- Suppose the sample described in Exercise 4 (with n = 2,400 and a = 60) is to be repeated with a smaller sample of n =1,200 and in only a = 30 equal-sized clusters

Project how large the sampling variance of p will be under this new design.

Exercise 5 (continued)

- Now suppose the reduced size of n = 1,200 is retained, but we want to consider a = 60 equal-sized clusters.

Project how large the sampling variance of p will be under this new design.

8. Probability proportionate to size sampling

- Unequal-sized cluster sampling
- Sampling with fixed rates
- Control of subsample size
- Selection of fixed size subsamples
- PPS sampling
- Systematic PPS sampling
- Exercise 6

Unequal-sized cluster sampling

- Naturally occurring clusters tend to be unequal in size
- Fixed sampling rates and unequal sized clusters result in variation in sample size

Consider the following sample of 12 schools:

School	B_a	**School**	B_a
1	308	7	393
2	823	8	148
3	146	9	321
4	809	10	393
5	827	11	207
6	775	12	850

Fixed rate sample

- An *epsem* sample of n = 100 students is to be selected from the N = 6,000 students in the 12 schools: $f = 100/6000 = 1/60$
- Two stages: Select a = 2 schools, say an SRS of a = 2 schools (a rate of 2/12 = 1/6)
- And then choose students at the rate 1/10 within the selected schools

$$f = (1/6) \cdot (1/10) = 1/60$$

Unequal subsample sizes

- Suppose schools 3 and 8 are chosen
  - Subsampling at the rate of 1/10 yields sample size

$$n = (146 + 148)/10 = 14.6 + 14.8 = 29.4$$

- On the other hand, if schools 5 and 12 were chosen instead,

$$n = (727 + 750)/10 = 72.7 + 75 = 142.7$$

- Subsample size varies from 29 to 143 …
  - Sample administration becomes difficult

Sample size variation

- Variation in the overall sample size is undesirable
- Since *n* is a random variable, $\bar{y} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} y_i$ no longer applies
- We need to use a ratio estimator

$$r = \frac{\sum_{\alpha=1}^{a} y_\alpha}{\sum_{\alpha=1}^{a} x_\alpha} = \frac{y}{x}$$

Control of subsample size

- In the survey literature, we need to find a way to control the sample size – keep it from varying
- A controlled sample size provides administrative convenience in fieldwork
- It also has greater statistical efficiency
- Several methods – we discuss two
  - Select exactly *b* elements per cluster
  - Probability proportionate to size (PPS)

Selection of fixed subsample sizes

- Suppose a = 2 schools are chosen at random
- And b = 50 students are chosen at random per selected school
- Sample size is n = 2 x 50 =100
  – Sample size does not vary across samples!
- But this design, on average across, all possible samples, over-represent students in small schools
  – Why?

Selection of fixed subsample sizes

- For example, for school 3,

$$f = (1/6)(50/146) = 1/17.52$$

- While for school 12,

$$f = (1/6)(50/750) = 1/90$$

- If students in large schools are different than those in small, we have bias
- The bias can be taken care of through weighting (later discussion)

PPS

- Require a method that is equal chance for students (*epsem*)
- And still achieves equal sized subsamples
  - And thus achieves fixed sample sizes
- Again, consider a = 2 and b = 50
- "Selection equation:"

$$f = \frac{1}{60} = P\{\alpha\} \cdot \frac{50}{B_\alpha}$$

PPS: Achieving *epsem*

- For example, if school 1 is chosen, then

$$f = \frac{1}{60} = P\{\alpha\} \cdot \frac{50}{308} = P\{\alpha\} \cdot \frac{1}{6.16}$$

- In order to make this *epsem* for students, we need for each school to be selected with probability …

$$\frac{1}{60} = P\{\alpha\} \cdot \frac{50}{B_\alpha} \quad OR \quad P\{\alpha\} = \frac{1}{60} \cdot \frac{B_\alpha}{50} = \frac{B_\alpha}{3000}$$

PPS: Selection by size

- Re-expressing this in terms of selecting both schools,

$$P\{\alpha\} = \frac{2 \cdot B_\alpha}{6000} = \frac{2 \cdot B_\alpha}{\sum_\alpha B_\alpha}$$

- In general, this becomes, across two stages,

$$f = P\{\alpha \ and \ \beta\} = \frac{a \cdot B_\alpha}{\sum_a B_\alpha} \cdot \frac{b}{B_\alpha} = \frac{a \cdot b}{\sum_a B_\alpha} = \frac{n}{N}$$

PPS selection of schools

School	B_α	Cum. B_α		
1	308	308		
2	823	1131	√	702
3	146	1277		
4	809	2086	√	1744
5	827	2913		
6	775	3688		
7	393	4081		
8	148	4229		
9	321	4550		
10	393	4943		
11	207	5150		
12	850	6000		

PPS:Choosing schools

- Select Random Numbers (RN's) from 1 to 6000:
  - RN = 702
  - RN = 1744
- Find the first school with cumulative sum greater than or equal to the first RN
- Find the next school with sum greater than the second RN
- These choose hospitals 2 and 4:

Systematic PPS

- How can we avoid selecting the same school twice?
- Systematically: select one RN from 1 to the interval 6000/2 = 3000
  - Say RN = 702
- Find the selected school, as above (school 2)
- Add the interval to the RN to obtain 702 + 3000 = 3702
- Find the second school with this selection number, as above, school 7
- RN 702 leads to the selection of schools 2 & 7

Exercise 6

- A two-stage *epsem* sample of 200 students is to be selected from the following 10 schools with 4,588 total students

Select two schools from this list with PPS using two Random Numbers (taken from the Table of Random Digits for Exercise 1).

What is the within school sampling rate for the first selected school?

Select two schools using systematic PPS.

School	B_α
Um Hakeem	261
Ahmad Bin Hanbal Independent	677
AlShamal	965
Khaleefa	406
Lusail	427
AlTijara	661
Qatar Independent Campus	169
Bilal Bin Rabah	285
Al Shahhaniya Independent	662
AlFatat AlMuslima	75
Total	**4,588**

9. Stratified random sampling

- Stratification
- Advantages
- Stratification – an example
- Stratified sample
- SRS
- Design effect
- Effective sample size
- Problems
- Multipurpose surveys
- Domains of study
- Proportionate stratified sampling
- Disproprotionate stratification
- Exercise 7

Stratification

- Procedure
  - Form strata
  - Independent selection within each
  - Estimate for stratum *h,*
  - Overall estimate $\bar{y}_h$

$$\bar{y} = \sum_{h=1}^{H} W_h \bar{y}_h$$

- Where

$$W_h = N_h / N$$

Variance

- For the overall sample estimate

$$Var\left(\bar{y}\right) = \sum_{h=1}^{H} W_h^2 Var\left(\bar{y}_h\right)$$

- With estimated variance

$$var\left(\bar{y}\right) = \sum_{h=1}^{H} W_h^2 var\left(\bar{y}_h\right)$$

Formation of strata

- Strata should be internally homogeneous
- Strata should differ as much as possible from each other
- Advantages
  - Gains in precision
  - Administrative convenience
  - Guaranteed representation of important domains
  - Acceptability/credibility
  - Flexibility

Stratification – an example

Population	Stratum 1 Qatari	Stratum 2 White & Blue Collar Expatriate (Other)
Size $\boldsymbol{N}$ 1,000,000	N_1 200,000	N_2 800,000
Variance S^2 1,800,000	S_1^2 4,000,000	S_2^2 1,000,000
Mean $\bar{Y}$ 1,400	$\bar{Y}_1$ 3,000	$\bar{Y}_2$ 1,000

Stratified sample

- What will be $Var(\bar{y})$?

$$n_1 = 240, n_2 = 960$$

$$Var(\bar{y}) = \sum_{h=1}^{2} (1 - f_h) W_h^2 S_h^2 / n_h$$

$$\approx W_1^2 S_1^2 / n_1 + W_2^2 S_2^2 / n_2$$

$$= (0.2)^2 (4000000) / 240 + (0.8)^2 (1000000) / 960$$

$$= 666.7 + 666.7$$

$$= 1333$$

SRS

- For $n = 1200$ What will be $Var_{SRS}(\bar{y})$?

$$Var_{SRS}(\bar{y}) = (1 - f)S^2/n$$

$$= (1 - 1200/1000000)\frac{1800000}{1200}$$

$$\approx \frac{1800000}{1200} = 1500$$

Design effect

- As for cluster sampling,

$$deff\left(\bar{y}\right)=\frac{Var\left(\bar{y}\right)\ for\ a\ given\ design}{Var_{SRS}\left(\bar{y}\right)of\ \ same\ size}$$

- For this example,

$$deff\left(\bar{y}\right)=\frac{Var\left(\bar{y}\right)}{Var_{SRS}\left(\bar{y}\right)}$$

$$=\frac{1333}{1500}$$

$$=0.89$$

Effective sample size

- What sample size with SRS would be necessary to achieve the same precision (variance) as the given design?
- Effective sample size:
- For our example,

$$n_{eff} = n / deff\left(\bar{y}\right)$$

$$n_{eff} = \frac{1200}{0.89}$$
$$= 1348$$

Problems

- Availability of data
  - Census
  - Administrative reports
  - Other surveys
- Multipurpose surveys
  - Survey of households in Qatar
  - Fixed assets, buildings, use of expatriate labor, expenditures, income, health, health care use, psychological well-being, social integration

Problems

- Domains of study
  - Subpopulations for which separate estimates are required
  - Geographic subdivisions such as provinces, districts, subdistricts
  - Socio-demographic characteristics, such as age groups, occupation, income, education

Proportionate stratified sampling

- Same sampling fraction in all strata

$$f = n/N = n_h/N_h = f_h$$

- Variance

$$Var(\bar{y}) = (1-f)\sum_{h=1}^{H} W_h^2 S_h^2 / n_h = \frac{(1-f)}{n}\sum_{h=1}^{H} W_h S_h^2$$

- Compare $Var_{SRS}(\bar{y}) = \frac{(1-f)}{n} S^2$

$$deff(\bar{y}) = \frac{\sum_{h=1}^{H} W_h S_h^2}{S^2}$$

Disproportionate stratification

- Purposes
  - Gains in precision for overall estimator
  - Precision for comparisons
  - Precision for domains
- Factors to consider
  - Size of strata W_h
  - Variability within strata S_h^2
  - Cost within strata c_h

Exercise 7

Calculate $Var\left(\overline{y}\right)$ for each of the following combinations of sample sizes across the two strata:

$$n_1 = 100 \quad n_2 = 1100$$
$$n_1 = 240 \quad n_2 = 960$$
$$n_1 = 400 \quad n_2 = 800$$
$$n_1 = 600 \quad n_2 = 600$$
$$n_1 = 960 \quad n_2 = 240$$

10. Frame problems

- Frame problems
- Objective respondent selection

Frame problems

- Frame: set of materials used to designate a sample of units
- Simple list, or set of materials such as maps, lists, rules for linking frame elements to population elements, *etc.*
- Accurate, up-to-date frames in single location, arranged suitably for selection
  - Numbered or computerized lists useful

Four types of frame problems

- Consider the following list of housing units in Doha
- Interested in sampling persons within these housing units
- The question is whether there are any of the following types of problems on the frame:
  - Non-coverage
  - Blanks
  - Duplicates
  - Clusters

ResidenceID	City	Street	ResidenceType	Nationality	Persons
1	Doha	Wahb	Villa	Non-Qataris	3
2	Doha	Wahb	Villa	Non-Qataris	6
3	Doha	Wahb	Villa	Non-Qataris	3
4	Doha	Wahb	Villa	Qataris	5
5	Doha	Wahb	Villa	Non-Qataris	5
6	Doha	Wahb	Villa	Non-Qataris	5
7	Doha	Wahb	Villa	Non-Qataris	3
8	Doha	Wahb	Villa	Non-Qataris	5
9	Doha	Wahb	Villa	Qataris	13
10	Doha	Wahb	Villa	Non-Qataris	6
11	Doha	Wahb	Villa	Non-Qataris	3
12	Doha	Wahb	Villa	Non-Qataris	5
13	Doha	Wahb	Villa	Non-Qataris	4
14	Doha	Wahb	Villa	Non-Qataris	3
15	Doha	Al Quds	Villa	Non-Qataris	4
16	Doha	Al Quds	Villa	Non-Qataris	5
17	Doha	Al Quds	Villa	Qataris	8
18	Doha	Al Quds	Villa	Non-Qataris	2
19	Doha	Al Quds	Villa	Non-Qataris	3
20	Doha	Al Quds	Villa	Non-Qataris	5

ResidenceID	City	Street	ResidenceType	Nationality	Persons
21	Doha	Al Quds	Villa	Non-Qataris	4
22	Doha	Al Quds	Villa	Non-Qataris	4
23	Doha	Al Quds	Villa	Non-Qataris	4
24	Doha	Al Quds	Villa	Qataris	3
25	Doha	Al Quds	Villa	Non-Qataris	1
26	Doha	Al Quds	Villa	Non-Qataris	4
27	Doha	Al Quds	Villa	Qataris	5
28	Doha	Al Quds	Villa	Non-Qataris	3
29	Doha	Al Quds	Villa	Non-Qataris	3
30	Doha	Al Quds	Villa	Non-Qataris	5
31	Doha	Murwab	Villa	Qataris	4
32	Doha	Murwab	Villa	Non-Qataris	2
33	Doha	Murwab	Villa	Non-Qataris	5
34	Doha	Murwab	Villa	Non-Qataris	2
35	Doha	Murwab	Villa	Non-Qataris	5
36	Doha	Murwab	Villa	Non-Qataris	2
37	Doha	Murwab	Villa	Non-Qataris	3
38	Doha	Murwab	Villa	Non-Qataris	5
39	Doha	Murwab	Villa	Non-Qataris	4
40	Doha	Murwab	Villa	Non-Qataris	4

Non-coverage

- Some elements of the population are not contained on the frame
  - Housing units not appearing on the list
  - Remedies
    - Use a frame that provides complete coverage
    - Supplement the existing frame with other frames
    - Use “population control adjustment weights” to compensate in analysis

Blanks

- List elements for which there are no eligible members of the population
  - Voter has moved
  - Remedies
    - Reject blank listings
    - Variation in sample size (smaller than desired): select additional listings
    - Avoid selecting next element on list

Duplicates

- Population element appears more than once on the list
  - Introduces unequal probabilities of selection
  - Housing unit appears more than once
  - Person living in two different addresses
  - Remedies
    - Determine number of times element is on list, and weight
    - Modify address list to eliminate duplicates

Clustering

- More than one population element is associated with a single list element
  - Variation in sample size
  - Remedies
    - Subsample clusters, and weight results by the inverse of the probability of selection
    - Accept variation in sample size

Within Household Selection: Objective Respondent Selection

- Remedy for selecting elements from small clusters, objectively in field settings
- Not *epsem*
- Suppose there are a maximum of four age-eligible persons per household
- Consider the following listing and selection table:

	Relationship to informant	Age	Gender
1			
2			
3			
4			

Respondent selection table

If number of eligible subjects is …	… then select subject number …
1	1
2	2
3	3
4	3

Interviewer instructions

- Interviewer:
  - List eligible household members by gender and age
  - Follow the instructions on the selection table to determine whom to interview
- This scheme is based on a set of 6 tables which are rotated among households to achieve the desired probabilities of selection for each subject:

Respondent selection tables

Table A (1/4)	
If number of eligible subjects is	**Select subject number**
1	1
2	1
3	1
4	1

Table B (1/12)	
If number of eligible subjects is	**Select subject number**
1	1
2	1
3	1
4	2

Table C (1/6)	
If number of eligible subjects is	**Select subject number**
1	1
2	1
3	2
4	2

Table D (1/6)	
If number of eligible subjects is	**Select subject number**
1	1
2	2
3	2
4	3

Table E (1/12)	
If number of eligible subjects is	**Select subject number**
1	1
2	2
3	3
4	3

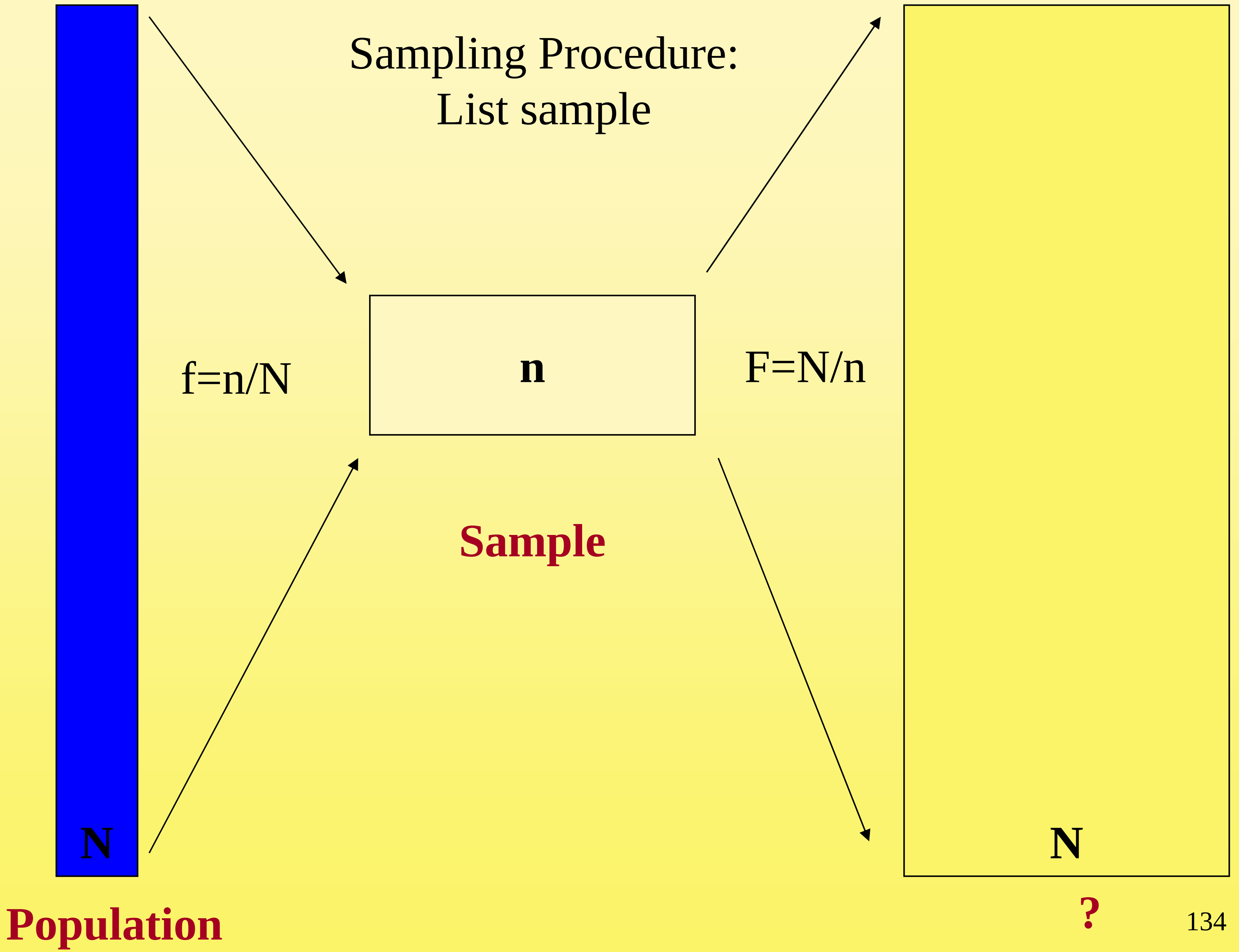
Table F (1/4)	
If number of eligible subjects is	**Select subject number**
1	1
2	2
3	3
4	4

11. Weighting

- Weighting to compensate for within household selection
- Exercise 8
- Weighting to compensate for unequal selection probabilities: over- and under-sampling
- Weighting to compensate for nonresponse
- Poststratification

Weighting

- Among four problems, two remedies involve weighting to compensate for unequal selection probabilities
- Weights common in survey practice
  - Within household selection
  - *Duplication of elements on the frame*
  - Over or under sampling
  - Nonresponse
  - Poststratification


Sampling Procedure:
List sample
f=n/N
n
F=N/n
Sample
N
N
Population
?

Weighting for within household selection

- As long as the sampling is *epsem* …
  –
- Then $\pi_i = \pi = f = n/N$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{1 + 1 + \cdots 1}$$

- For example, from *N* = 2000 adults, select *n* = 20 with *epsem*

$$\pi_i = \frac{20}{2000} = \frac{1}{100} \quad and \quad w_i = 100$$

- Each adult represents themselves and 99 others

Non-*epsem* estimation

- But the mapping may not be equal for every element
- A weighted estimator is required:

$$\bar{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{100 \cdot y_1 + 100 \cdot y_2 + \cdots + 100 \cdot y_{20}}{100 \cdot 1 + 100 \cdot 1 + \cdots + 100 \cdot 1}$$

- When the weights are constant, they cancel

Within household sampling

- Suppose a sample of 20 households are selected
- For 8 households, 1 adult: 3 reported being outside the country in the past year
- For 6 households, 2 adults: 3 outside
- For 4 households, 3 adults: 3 outside
- For 2 households, 4 adults: 2 outside

Probability of selecting adults

- When 1 adult in the household, two stages of selection and

$$\pi_i = (20/2000)(1/1) = 1/100 \quad w_i = 100$$

- When 2 adults in the household,

$$\pi_i = (20/2000)(1/2) = 1/200 \quad w_i = 200$$

- When 3 adults in the household,

$$\pi_i = (20/2000)(1/3) = 1/300 \quad w_i = 300$$

- When 4 adults in the household,

$$\pi_i = (20/2000)(1/4) = 1/400 \quad w_i = 400$$

ID	Response (Y)	Housing unit prob.	No. persons 18+	Weight
1	1	0.01	1	100
2	1	0.01	1	100
3	0	0.01	1	100
4	0	0.01	1	100
5	0	0.01	1	100
6	0	0.01	1	100
7	1	0.01	1	100
8	1	0.01	1	100
9	0	0.01	2	200
10	0	0.01	2	200
11	1	0.01	2	200
12	0	0.01	2	200
13	0	0.01	2	200
14	1	0.01	2	200
15	0	0.01	3	300
16	1	0.01	3	300
17	1	0.01	3	300
18	1	0.01	3	300
19	1	0.01	4	400
20	1	0.01	4	400

Weighted or unweighted estimate

- This can be represented in the weighted mean (proportion of adults who recycle) as

$$\bar{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{100\times1+100\times1+\cdots+400\times1}{100\times1+100\times1+\cdots+100\times4} = 0.65$$

- The corresponding unweighted mean is

$$\bar{y} = \frac{\sum_i y_i}{n} = \frac{1+1+1+0+0+\cdots+1}{20} = 0.55$$

Exercise 8

- Selected a sample of 20 households
- Selected one person 15 years or older (15+) in each
- Asked them whether they had been outside Qatar in the past year:

ID	Response (Y)	Housing unit prob.	No. persons 18+	Weight
1	1	Unknown, but equal	5	
2	0	Unknown, but equal	4	
3	1	Unknown, but equal	4	
4	0	Unknown, but equal	4	
5	0	Unknown, but equal	3	
6	0	Unknown, but equal	4	
7	1	Unknown, but equal	11	
8	1	Unknown, but equal	5	
9	0	Unknown, but equal	2	
10	0	Unknown, but equal	2	
11	1	Unknown, but equal	4	
12	0	Unknown, but equal	3	
13	0	Unknown, but equal	2	
14	1	Unknown, but equal	6	
15	0	Unknown, but equal	2	
16	1	Unknown, but equal	5	
17	1	Unknown, but equal	3	
18	1	Unknown, but equal	3	
19	1	Unknown, but equal	4	
20	1	Unknown, but equal	2	

Exercise 8 (continued)

Compute the weights for each sample person.

Compute an unweighted estimate of the proportion who have been outside in the past year

Compute a weighted estimate of the proportion who have been outside in the past year

Over- and under- sampling

- The basic approach above has been to weight by $1/\pi_i$
  - Count an element $1/\pi_i$ times
- Consider the following population and sample distribution for persons 15 years and older (15+) in Qatar comparing Qatari and White and Blue Collar Expatriates (Other):

Group	N	n	Sampling rate	Weight A	Weight B
Qatari	150,000	125	1/1,500	1,500	1
Other	1,350,000	875	1/1,500	1,500	1
Total	**1,500,000**	**1,000**	1/1,500	1,500	1

Sample selection

- This is a proportionate allocation, with equal probabilities in each group
- Some investigators might prefer that the distribution in the sample be equal across the two groups:

Group	N	n	Sampling rate	Weight A	Weight B
Qatari	150,000	500	1/300	300	1
Other	1,350,000	500	1/2,700	2,700	9
Total	**1,500,000**	**1,000**	1/1,500	1,500	**--**

Proportionate v. equal allocation

- The equal allocation would be used for comparing the two groups
- The proportionate allocation would be used to represent the population
- Consider the consequences of the equal allocation when estimating "proportion never married" among, again, 15+, across the two groups:

Proportionate allocation

Group	Never married	Proportionate allocation		Weights	
		n	Never married	A	B
Qatari	0.400	170	0.400	1,500	1
Other	0.305	830	0.305	1,500	1
Total	**0.315**	**1,000**	**0.315**		

Equal allocation

Group	**Never married**	**Dispro-portionate allocation**		**Weights**		**Weighted estimate**
		n	Never married	A	B	
Qatari	0.400	500	0.400	300	1	(500)(1)(0.400)
Other	0.305	500	0.305	2,700	9	(500)(9)(0.305)
Total	**0.315**	**1,000**	**0.353**	**--**	**--**	**0.315**

Restoring the balance

- Weights will restore the balance to the population distribution:

$$\overline{y} = \frac{\sum y_i}{n} = \frac{500\ x\ 0.400 + 500\ x\ 0.305}{500 + 500} = 0.353$$

$$y_{w(B)} = \frac{\sum w_{i(B)} y_i}{\sum w_{i(B)}}$$

$$= \frac{500\ x\ 1\ x\ (0.400) + 500\ x\ 9\ x\ (0.305)}{500\ x\ 1 + 500\ x\ 9} = 0.315$$

$$y_{w(A)} = \frac{\sum w_{i(A)} y_i}{\sum w_{i(A)}} = \frac{500\ x\ 300\ x\ (0.400) + 500\ x\ 2700\ x\ (0.305)}{500\ x\ 300 + 500\ x\ 2700} = 0.315$$

Weights in practice

- Is it necessary to weight, even when unequal probabilities are involved?
- Descriptive statistics require weights
  – Otherwise, estimates will be biased
- Analytic statistics are more controversial
  – Comparing income between Latino and non-Latino groups – no need to weight
  – Comparing income between male and female respondents in the same sample requires weighting

Effect of weights

- Often the effect of weights is not large for descriptive statistics
- If not large, analysts may decide not to use weights
  - Use of weights more difficult historically because of lack of software to handle weights
  - Duplication factors used

Weighting for nonresponse

- Suppose that not everyone in the sample of 1,000 drawn from our two groups responded
- Ignoring nonresponse produces slightly biased estimates when averaging across the now disproportionately distributed groups:

Group	n	r	Weight A	Never married	Weighted estimate
Qatari	500	450	1	0.400	(450)(1)(0.400)
Other	500	350	9	0.305	(350)(9)(0.305)
Total	**1,000**	**800**	**--**	**0.315**	**0.317**

Nonresponse weights

- Compute weighted response rates in each group
- Adjust the base weights (those computed to compensate for unequal probabilities of selection) for nonresponse
- Assumption: data is missing at random (MAR) within subgroups
- Response rate in each group is a “sampling rate” under the MAR assumption

Group	w_{1i}	n_h	r_h	$(r_h)^{-1}$	$w_i = w_{1i} / r_h$
Qatar	1	450	0.90	1.11	1.11
Other	9	350	0.70	1.43	12.86
Total		**800**	**0.80**		

Nonresponse weights

- These nonresponse adjusted weights 'restore the balance':

$$\bar{y} = \frac{\sum y_i}{n} = \frac{450\ x\ 0.400 + 350\ x\ 0.305}{450 + 350} = 0.358$$

$$y_{w(B)} = \frac{\sum w_{i(B)} y_i}{\sum w_{i(B)}}$$

$$= \frac{450\ x\ 1.11\ x\ (0.400) + 350\ x\ 12.86\ x\ (0.305)}{450\ x\ 1.11 + 350\ x\ 12.86} = 0.315$$

Poststratification

- Poststratification is used to make the weighted sample distribution conform to a known population distribution
- Adjust the nonresponse adjusted weights
- Suppose that gender in the sample does not agree with known gender distributions in the population:

Gender	n_g	p_g	N_g	P_g	$w_g = P_g / p_g$
Male	500	0.615	1,222,000	0.815	1.320
Female	300	0.375	278,000	0.185	0.490
Total	**800**	**1.000**	**1,500,000**	**1.000**	**--**

A final weight

- In poststratification, the weights for the individuals in groups are adjusted up or down to obtain the distribution of the sum of weights that corresponds to the population distribution
- The final weight is an adjustment of the baseline weight for nonresponse and poststratification:

Group/Gender	n_{hg}	w_{hg}
Qatari		
Male	215	1.11 x 1.320 = 1.465
Female	235	1.11 x 0.490 = 0.549
Other		
Male	285	12.86 x 1.320 = 16.975
Female	65	12.86 x 0.490= 6.301
Total	**800**	

12. Variance estimation

- Sampling error
- General sample design
- Variance estimation
- Simple replicated sampling
- Problems with simple replicated estimates
- Three methods of variance estimation
- Comparison of methods
- Computer software

Sampling error

- Problem
  - Many variables in a single survey
  - Many subclasses (domains) of interest
  - Fairly complex designs
  - Enormous computing task
- Requirement
  - Practical and efficient methods of variance estimation
  - Computer programs to implement them

General sample design

- Stratified
- Clustered
  - Primary stage units
  - *b* elements within each PSU
- Weights
- Sampling methods
  - Over representation of domains
  - Optimum allocation (rarely)
- Nonresponse
- Poststratification

Variance estimation

- Durbin, 1952
  - If clusters (PSU's) selected independently, variance can be estimated using only PSU totals
  - Variance estimate contains the contribution of later stages of subsampling
  - For rapid methods of variance estimation, no components of variance are needed

Simple replicated subsampling

- Alternative approaches based on ‘repetition’
- *c* independent subsamples (replicates) selected under same design from population
- Estimate some statistic *Z*
- Each replicate provides z_i
- Compute

$$\bar{z} = (1/c) \sum_i z_i$$

$$\mathrm{var}(\bar{z}) = (1/c(c-1)) \sum_i (z_i - \bar{z})^2$$

Three general estimators

- Taylor series expansion
  - Approximate analytic solution
- Balanced repeated replication (BRR)
  - Based on replicated sampling, but actually replicated subsampling
- Jackknife repeated replication (JRR)
  - Simplified form of replicate formation: drop out one
  - General methodology developed for another purpose – has broad application

Comparison of methods

- Empirical studies conducted for variety of statistics and methods of variance estimation
  - Mean square errors (MSE) of variance estimates favor Taylor series
  - Coverage properties of confidence intervals favor BRR
- All three methods reasonably good for
  - Correlation coefficients
  - Ratio means
  - Regression coefficients
- Taylor series most versatile, with respect to sample designs
  - Jackknife is the most general approach

Computer programs

- Standard statistical packages such as SPSS, SAS, Stata, assume SRS by default
- Necessary input to compute sampling errors
  - PSU for every element
  - Stratum for every element
  - Weight for every element
  - At least two PSU's per stratum
- See American Statistical Association web site for comprehensive review: http://www.hcp.med.harvard.edu/statistics/survey-soft/

13. Survey sampling textbooks

- Barnett, V. (1974). *Elements of Sampling Theory*. London: English Universities Press. A short introduction to topics in sampling theory.
- Cassell, C-M., Sarndal, C-E., and Wretman, J.J. (1977). *Foundations of Inference in Survey Sampling*. New York: J.W. Wiley and Sons, Inc. Theoretical treatment of survey sampling inference, including issues such as admissibility of estimators.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition. New York: J.W. Wiley and Sons, Inc. Excellent and widely used text on the basic theory for sampling techniques.
- Deming, W.E. (1950). *Some Theory of Sampling*. New York: Dover. Text on sampling theory and practice.
- Deming, W.E. (1960). *Sample Design in Business Research*. New York: J.W. Wiley and Sons, Inc. Text on sampling theory and practice, with emphasis on replicated sampling methods. Recently released by Wiley as a paperback Classics edition.

- Hajek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker. A monograph on sampling theory from an advanced perspective.
- Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*. *Volume I: Methods and Applications. Volume II: Theory.* New York: J.W. Wiley and Sons, Inc. Classic two volume text on sampling practice and theory that is considered still to be the standard.
- Jessen, R.J. (1978). *Statistical Survey Techniques.* New York: J.W. Wiley and Sons, Inc. An intermediate text on sampling with a presentation of lattice sampling methods.
- Kalton, G. (1983). *Introduction to Survey Sampling.* Beverly Hills, CA: Sage Publications. Short non-mathematical treatment of sampling. A Sage mongraph.
- Kish, L. (1965). *Survey Sampling*. New York: J.W. Wiley and Sons, Inc. Comprehensive text on sampling practice, about to be issued as a paperback Classic edition.

- Konijn, H.S. (1973). *Statistical Theory of Sample Survey Design and Analysis*. New York: American Elsevier. Advanced text on sampling theory.
- Levy, P.S. and Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications.* New York: J.W. Wiley and Sons, Inc. Intermediate level text on sampling methods.
- Lohr, Sharon L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press. Intermediate level text blending theory and practice, including exercises and sample data sets for analysis of survey data.
- Moser, C.A. and Kalton, G. (1971). *Survey Methods in Social Investigation,* 2nd edition. London: Heinemann. Text on survey methods with a non-mathematical introduction to sampling methods.
- Murthy, M.N. (1967). *Sampling Theory and Methods.* Calcultta: Statistical Publishing Society. Advanced text on sampling theory and practice.
- Raj, D. (1968). *Sampling Theory*. New York: McGraw Hill. Advanced text on sampling theory.

- Raj, D. (1972). *The Design of Sample Surveys.* New York: McGraw-Hill, Inc. Two part text: the first is an intermediate-level text on sampling practice, and the second presents surveys applications.
- Särndal, C-E. SwenÑson, B. and Wretman, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag. Advanced text on sampling methods.
- Scheaffer, R.L., Mendenhall, W., and Ott, L. (1990). *Elementary Survey Sampling,* 4th edition. Boston: PWS Kent. Elementary text requiring minimal mathematical background.
- Stuart, A. (1984). *The Ideas of Survey Sampling,* revised edition. London: Griffin. Short text that illustrates the basic concepts of sampling with a small numerical example.
- Sudman, S. (1976). *Applied Sampling.* New York: Academic Press. Intermediate-level text on sampling practice.
- Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C. (1984). *Sampling Theory of Surveys with Applications*, 3rd edition. Ames, Iowa: Iowa State University Press. Advanced text on sampling theory with important treatments on ratio estimation.

- Thompson, S.K. (1992). *Sampling*. New York: J.W. Wiley and Sons, Inc. Intermediate-level text on sampling methods, including a number used widely in the natural sciences, and a discussion of adpative sampling techniques.
- Williams, W.H. (1978). *A Sampler on Sampling*. New York: J.W. Wiley and Sons, Inc. Intermediate-level treatment of sampling methods.
- Yamane, T. (1967). *Elementary Sampling Theory*. Englewood Cliffs, NJ: Prentice Hall. An introductory text that provides a mix theory and simple illustrations; useful for students with limited mathematical backgrounds.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag. Comprehensive treatment of variance estimation for survey sampling.
- Yates, F. (1981). *Sampling Methods for Censuses and Surveys,* 4th edition. London: Griffin. Advanced text on sampling practice.