Epi 265: Epidemiological Methods III: Sampling

Yea-Hung Chen

Monday, June 5, 2017

- about me:
  - since 2008: Epidemiologist at the San Francisco Department of Public Health, studying populations with relatively high risks for HIV infection
  - research: models for HIV incidence and behaviors associated with HIV transmission (including medication)
  - email: `yea-hung.chen@sfdph.org`
- this lecture:
  - thanks to Maria for inviting me!
  - will discuss "classic sampling methods," but I have much more experience with methods for sampling "hard-to-reach populations"
  - further reading on classic sampling methods: Lohr S. *Sampling: Design and Analysis*. Pacific Grove: Brooks/Cole Publishing Company; 1999.
  - participation and questions strongly encouraged!
  - there will be discussion reflection question and readings
  - there will be some exercises
  - there is an outline in the sidebar
  - remind me to break at $\sim$ 11:30

Introduction

Why should we care about sampling?

- epidemiological study designs rely on sampling:
  - randomized controlled trial: sample individuals, and randomly assign treatment
  - case-control study: sample cases and sample controls
  - cross-sectional study: sample individuals from the target population
  - cohort study: sample individuals and follow them over time
- other health applications:
  - disease prevalence (eg, HIV prevalence)
  - prevalence of behaviors (eg, vaccination)
  - sampling nutritional or environmental exposures (eg, testing of meat)
  - others?
- connections between missing data and sampling
- in short: studies rely on data, and data rely on sampling, and the sampling method can affect bias and variance

Bias in sampling

- suppose you take a sample from a population of interest
- use an estimator (a mathematical equation or process, such as the sample mean) to estimate an unknown value of interest, such as disease prevalence or a causal effect
- imagine repeating the process multiple times: you will have multiple estimates
- assume simplistically that you can rule out other forms of bias, such as measurement bias or confounding (the latter is not relevant in estimation of univariate quantities)
- if the average of these estimates is not equal to the true value, you have bias that is attributable to sampling (selection bias)
- many sampling methods have estimators that are, in theory at least, unbiased
- in comparison, with convenience sampling, the sample is likely not representative of the population and no special estimator is used to try to correct for that

Bias in estimation of causal effects

- sampling can cause bias in estimation of causal effects
- directed acyclic graphs provide a framework for understanding this
- this is review for you, from: Maria's lecture 2 weeks ago, Hernan and Robin's chapter on selection bias, Epi 203, . . .

Rothman, Greenland, and Nash

- Rothman, Greeland, and Nash (optional reading):

  many epidemiological studies are designed to sample subjects from a target population of particular interest, such that the study population is "representative" of the target population

- they go on to claim that there is an over-emphasis on representativeness:

  Taken to an extreme, however, the pursuit of representativeness can defeat the goal of validly identifying causal relations

Rothman, Greenland, and Nash

- but, they do close by acknowledging that representativeness can matter:

  For example, although most of the decisive data connecting smoking to lung cancer was derived from observations on men, no one doubted that the strong effects observed would carry over at least approximately to women, for the lungs of men and women appear to be similar if not identical in physiologic detail. On the other hand, given the huge sex differences in iron loss, it would seem unwise to generalize freely to men about the effects of iron supplementation observed in premenopausal women.

Bias in estimation of univariate quantities

- even if one questions how likely it is that sampling can cause selection bias, it is fairly clear that sampling can cause bias in estimation of univariate quantities
- consider some of the "other" applications of sampling from the Introduction:
  - disease prevalence (eg, HIV prevalence)
  - prevalence of behaviors (eg, vaccination)
  - sampling nutritional or environmental exposures (eg, testing of meat)
- sampling can cause bias in estimation of these quantities!

Example: prevalence of HIV and risk behaviors for HIV

- as mentioned, my work involves population with high risks for HIV infection
- these include men who have sex with men (MSM), people who inject drugs, "high-risk heterosexuals," and transexual individuals
- we use methods for sampling hard-to-reach populations (more later)
- though we do use our data to study associations, a primary purpose of the projects is to obtain prevalence estimates: the prevalence of HIV infection and the prevalence of HIV-related behaviors (eg, risk behaviors, use of antiretroviral medication, HIV testing, . . .)
- for example, we sample MSM at venues (more later)
- we estimate the HIV prevalence among San Francisco MSM to be $\sim 25\%$
- possible that HIV-infected MSM are less likely to be at the venues we sample
- we try to use venues that we think will give us a representative sample and we use math (various estimators) to try to minimize bias

Example: the Flint water crisis

- in April 2014, as a money-saving strategy, Flint switched its water source (switched to the Flint River; previously, had been buying water from Detroit Water and Sewerage Department, mostly from Lake Huron)
- Flint officials failed to properly treat the highly corrosive water, which caused lead to leach from pipes
- they then failed to adhere to federal rules for testing water in homes
- federal rules: test the "worst-case homes"
- population of interest: water from worst-case homes (those with lead pipes)
- sample actually taken: mostly from homes without lead pipes
- further reading:
  - Fonger R. Documents show Flint filed false reports about testing for lead in water. *The Flint Journal*. November 12, 2015.
  - Langkjær-Bain R. The murky tale of Flint's deceptive water data. *Significance*. 2017;14(2):16–21.

Discussion: reflection question

Explain how you might incorporate a sampling strategy into the study design (you might, for example, use respondent-driven sampling to initiate a cohort study). Briefly discuss possible logistical/practical advantages and disadvantages to this plan. Finally, discuss whether you think incorporating the sampling strategy might help (1) reduce bias in the estimation of univariate quantities (such as disease prevalence) and (2) reduce bias in the estimation of causal effects.

Bias matters!

- media/blog pundits love to exclaim "small sample size!"
- but when thinking critically about data, I would argue that it is actually more important to worry about bias
- that is because statistical methods always take into account sample size
- but bias is harder to detect and address

Variance in sampling

- that being said, variance is extremely important!
- again, suppose you take a sample from a population of interest
- use an estimator (a mathematical equation or process, such as the sample mean) to estimate an unknown value of interest, such as disease prevalence or a causal effect
- imagine repeating the process multiple times: you will have multiple estimates
- assume simplistically that you can rule out other forms of variance, such as measurement error
- the estimates may all be fairly similar (low variance) or you might have a wide range of estimates (high variance)

Variance

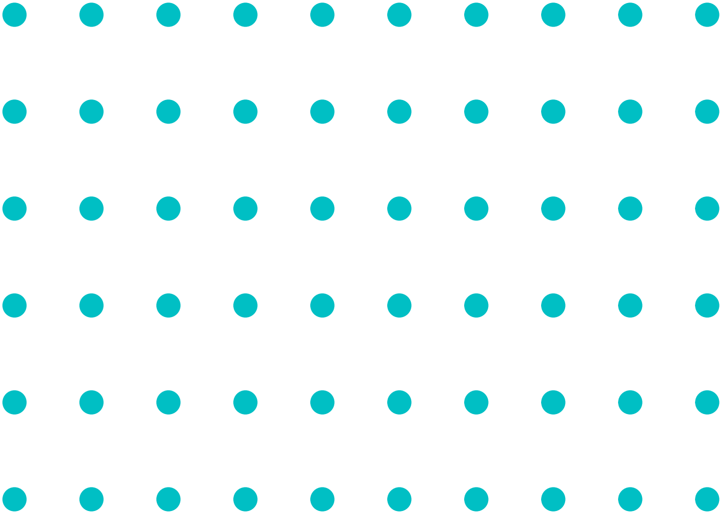
- you presumably are aware that sample size affects variance
- but, the sampling method itself affects variance!

Classic sampling methods

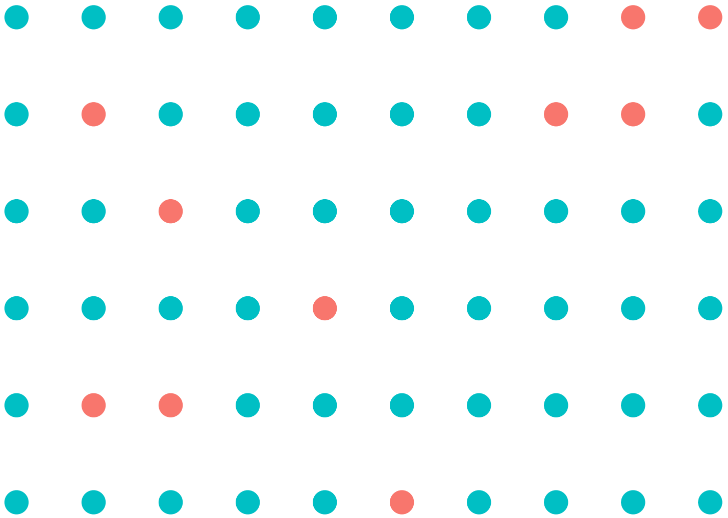
Simple random sampling

- begin with a list (sampling frame) of individuals in the population
- simple random sampling (SRS): randomly select individuals from the sampling frame, usually without replacement (remaining slides in this section assume simple random sampling without replacement)

Simple random sampling



Simple random sampling



Properties of simple random sampling

- if you think of the sampling process as a series of individual draws, the probability of being selected in a single draw depends on how many people have been sampled thus far
- probability of inclusion in the sample: n/m, where n is the sample size and m is the population size (can be shown via hypergeometric distribution)

Estimator for simple random sampling

- the estimator for the population mean is simply the sample mean:

$$\frac{1}{n}\sum_{i\in\mathcal{S}} y_i$$

- here, i is an index for individuals, n is the sample size, y_i is the value associated with person i, and summation occurs over all individuals in the sample, $\mathcal{S}$
- you have probably never questioned the logic behind this, but the intuition is that simple random sampling should, unless we are unlucky, result in a representative sample so no special math is necessary
- this principle does not hold for other forms of sampling!

Estimator for simple random sampling

- though the sample mean is not an unbiased estimator for other sampling methods, there is a connection to other sampling methods
- across sampling methods, a common estimator is to weight by the inverse by the probability of inclusion

$$\frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i}$$

- this is known as the Horvitz-Thompson estimator (H-T estimator)
- this is an important concept in sampling, with parallels in other areas as well (such as missing data)

Estimator for simple random sampling

- that idea applies to simple random sampling as well
- in simple random sampling, the probability of inclusion is the constant n/N, where N is the population size
- so, the weight is the constant $w = N/n$
- using H-T, the estimator is:

$$\frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i} = \frac{w \sum_{i \in \mathcal{S}} y_i}{wn} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$$

- in other words, the estimator for simple random sampling is a special case of the H-T estimator
- we'll revisit the H-T estimator when we examine the other methods

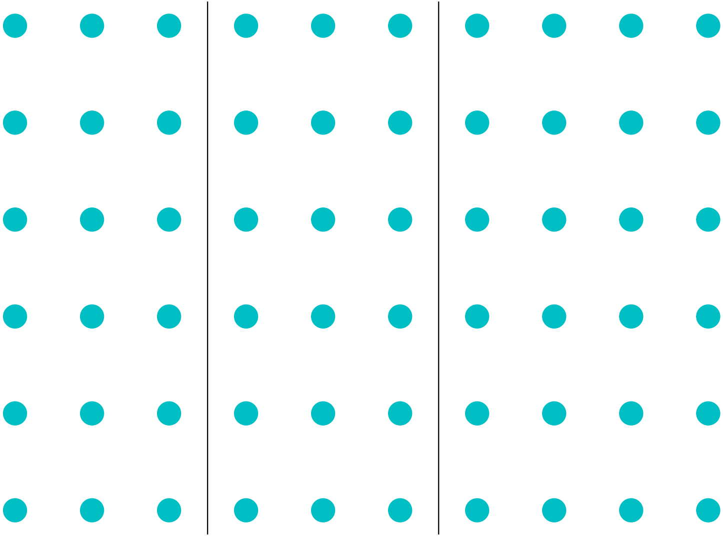
Finite population correction

- what is the variance of the estimator?
- you are probably familiar with the following expression: σ^2/n, where σ^2 is the unknown variance of the values in the population (and needs to be estimated)
- sampling folk point out that the variance of the estimator needs to be multiplied by the factor $1 - n/N$, where N is the population size
- finite population correction: the multiplication by $1 - n/N$
- the intuition is that as the sample size approaches the population size, we are more sure about our answer (if we have a census, $n = N$, there is no variance)
- the same logic applies to other sampling methods as well

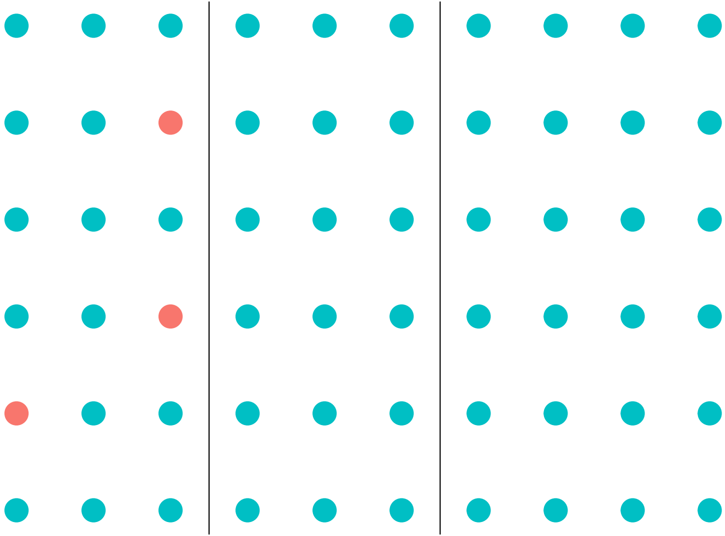
Stratified sampling

- divide the population into mutually exclusive subpopulations, called strata
- then, take a sample from each stratum (usually using simple random sampling)
- for example, suppose we want to estimate mean blood pressure in a population of adults: we could stratify by age group

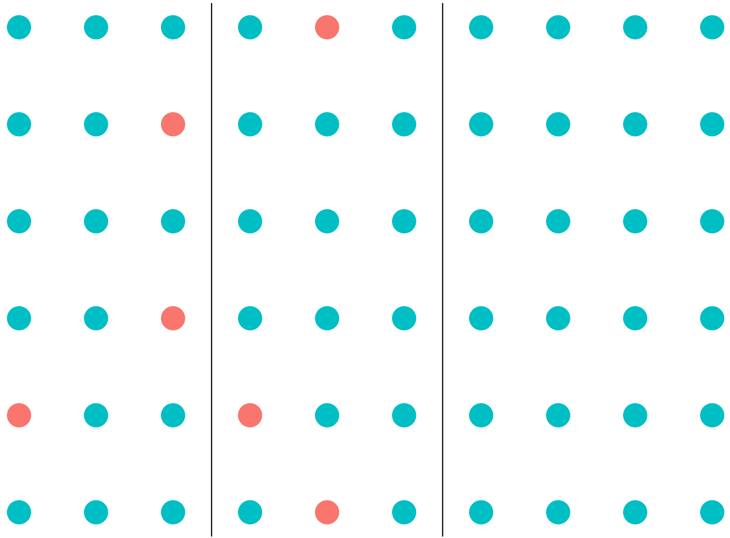
Stratified sampling



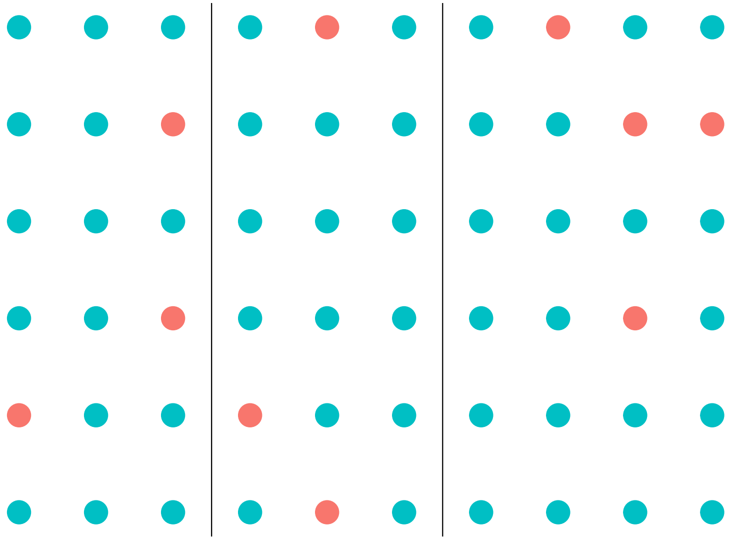
Stratified sampling



Stratified sampling



Stratified sampling



Why use stratified sampling?

- increased precision: stratified sampling will usually have greater precision than simple random sampling
- this can be shown mathematically, but the intuition is that increased precision occurs because we will usually have relatively low variability within each stratum and high variability across strata
- we won't need a large sample size from each stratum to get a good representation of the stratum, because the variability within the stratum is relatively low
- because we will sample each stratum, we will capture the potentially high variability across strata
- in comparison: an SRS of the population might over-sample the low variability within each stratum and under-sample the variability across strata

Why use stratified sampling?

- stratified sampling can help ensure representativeness of the sample
- for example, an SRS of a population might, by chance, over- or under-sample certain subpopulations
- stratified sampling can ensure precision in the analysis of subpopulations

Why not use stratified sampling?

- as mentioned, stratified sampling generally involves improved precision because we can generally expect relatively low variability within each strata and high variability across strata
- we should try to pick stratification variable such that the within-strata variability is small relative to total variability, but in practice, we may not know
- moreover, there may be many variables of interest: a stratification variable might fit the criteria well for one outcome but not for another

Why not use stratified sampling?

- the method requires:
  - a list of individuals in the population and a list of individuals in each stratum
  - the population size and the size of each stratum
- this information may not be available!

Estimator for stratified random sampling

- recall: a common technique in sampling is to use the Horvitz-Thompson estimator, which weights by the inverse of the probability of inclusion
- probability of inclusion $= n_h/N_h$, where N_h is the population size of stratum h and n_h is the size of the sample taken from stratum h
- the weight is: $w_{hi} = N_h/n_h$
- using H-T, the estimator is:

$$\frac{\sum_h \sum_{i \in \mathcal{S}_h} w_{hi} y_{hi}}{\sum_h \sum_{i \in \mathcal{S}_h} w_{hi}}$$

- you need not worry about the double summation if it confuses you
- the basic idea is the same as before: multiply each individual's value by his/her weight, sum these products, and divide by the sum of the weights

Exercise: stratified sampling

id	stratum	cvd
1	1	0
2	1	0
3	1	1
4	1	0
5	1	0
6	2	1
7	2	1
8	2	1

- size of stratum 1 is 15 and size of stratum 2 is 21
- compute the weights (what is the sum?)
- estimate the prevalence of cardiovascular disease

Variance of the estimator for stratified sampling

- the mathematical expression depends on the total population size, the total number of people in each stratum (N_h), the number of people sampled from each stratum (n_h), and the estimated variation within each stratum

Discussion: Hu et al

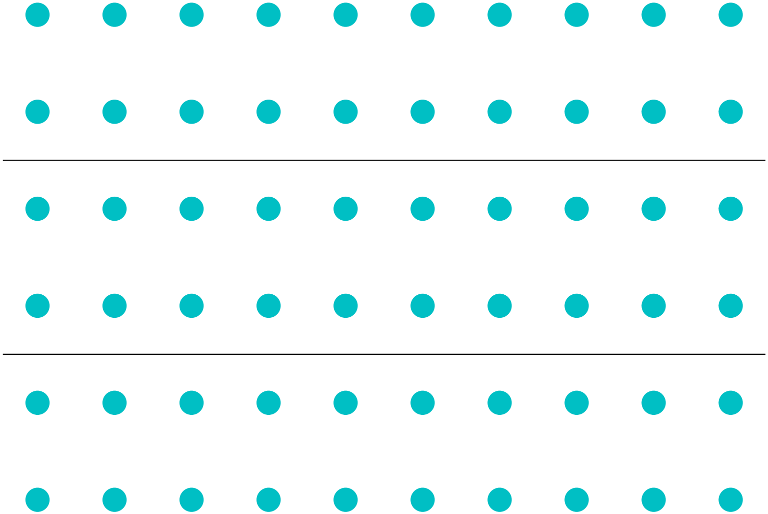
The extents of the potential biases in landline telephone surveys that exclude cell phones were estimated. Biases resulting from exclusion of adults with cell phones only from the landline-based survey were found for 9 out of the 16 health indicators.

- what were the strata?
- they were able to detect even somewhat small differences. . .
- what were some of the big differences?

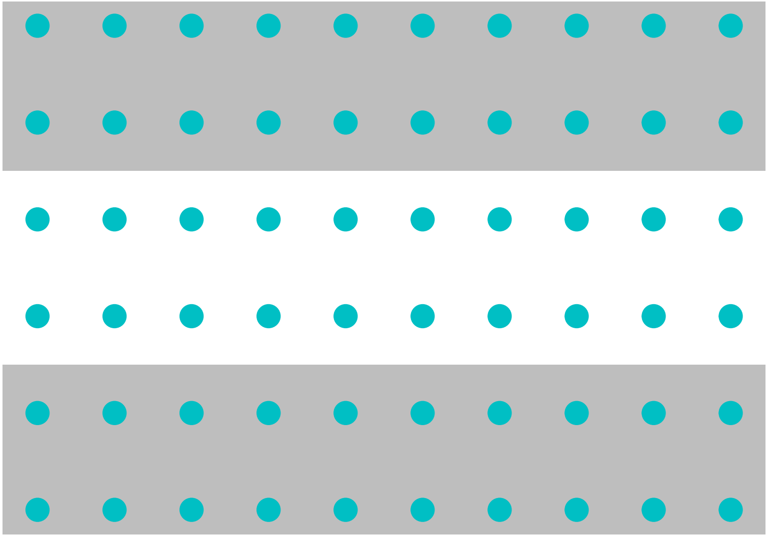
Cluster sampling

- as mentioned, stratified sampling requires a list of individuals in the population (the sampling frame)
- this requirement is dropped in cluster sampling
- cluster sampling avoids that requirement by using easily observed groups, called clusters
- often, clusters are physical/environmental: eg, schools, jails, or hospitals
- cluster sampling: sample clusters, then sample individuals from the sampled clusters

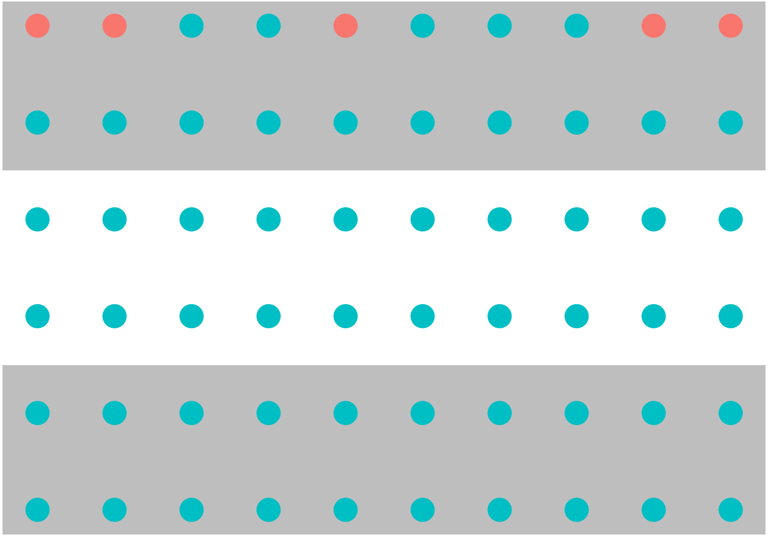
Cluster sampling



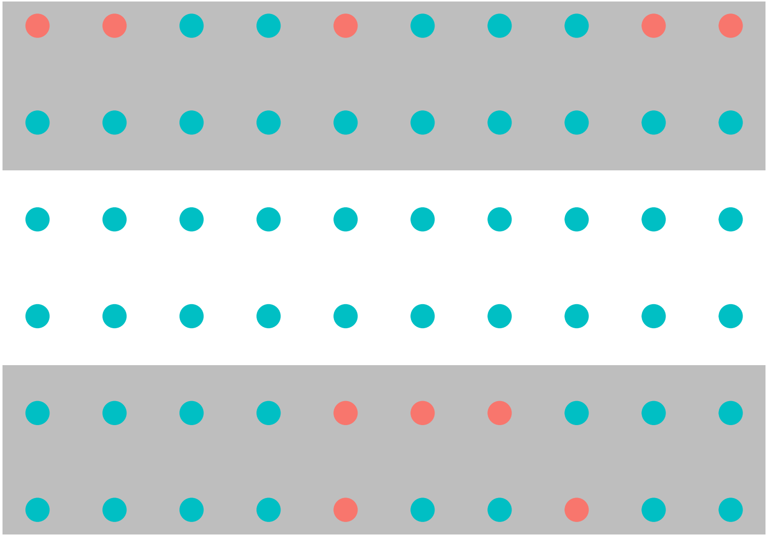
Cluster sampling



Cluster sampling



Cluster sampling



Example: patients in nursing homes

- for example, suppose the population of interest is patients in California nursing homes
- it would probably be difficult to get a list of all such patients, making simple random sampling and stratified sampling infeasible
- but, you can get a list of all nursing homes in California!
- you can then randomly sample nursing homes and then randomly sample patients

Cluster sampling vs stratified sampling: similarities

- mutually exclusive groups: in stratified sampling, a person can not be in more than one stratum; in cluster sampling, a person can not be in more than one cluster
- sample from groups: in stratified sampling, we take a sample of individuals from each stratum; in cluster sampling, we take a sample of individuals from each sampled cluster

Cluster sampling vs stratified sampling: differences

- sampling frame: in stratified sampling, you must have a list of everybody in the population; in cluster sampling, you might not have such a list
- definition of groups: in stratified sampling, the mutually exclusive groups are called strata and are ofen defined by demographic features; in cluster sampling, the mutually exclusive groups are called clusters and are usually defined by physical/environmental groups)
- sampling of groups: in stratified sampling, we sample from every single stratum; in cluster sampling, we do not take a sample from every cluster

Stages and sampling units

- the form of cluster sampling I've described is known as two-stage cluster sampling
- the first stage is sampling of the clusters (primary sampling unit), and the second stage is the sampling of individuals

Why use cluster sampling?

- the primary reason for using cluster sampling is convenience and cost
- as mentioned, cluster sampling does not require a list of individuals in the population of interest

Why not use cluster sampling?

- as mentioned, stratified sampling generally involves increased precision, relative to simple random sampling
- in comparison, cluster sampling usually involves decreased precision!
- because we do not sample every cluster, we do not capture a good representation of variability across clusters
- imagine that we take multiple cluster samples: the clusters in the first sample might be very different from the clusters in the second sample, etc... leading to wide range of estimates
- meanwhile, we use our sample size on sampling within clusters, where there is relatively little variability

Estimator for cluster sampling

- again, use the H-T estimator
- notation:
  - M: total number of clusters
  - m: number of clusters sampled
  - N_h: total number of people in cluster h
  - n_h: number of people sampled from cluster h
- probability of inclusion = probability of selecting cluster times the probability of inclusion in the cluster sample = $(m/M)(n_h/N_h)$
- the weight is: $w_{hi} = (MN_h)/(mn_h)$

Variance of the estimator for cluster sampling

- the mathematical expression depends on the total population size, the total number of clusters (M), the number of clusters sampled (m), the total number of people in each cluster (N_h), the number of people sampled from each cluster (n_h), the estimated variance within each cluster, and the estimated variance across clusters
- note that in stratified sampling, we did not need the total number of strata, the number of strata sampled, or the estimated variance across strata
- the intuition there is that in stratified sampling, we sample each stratum!

Complex sampling methods

- many large surveys actually use a combination of sampling methods
- this is primarily done for convenience and cost

Example: National Health and Nutrition Examination Survey

13 major strata were defined based on state groupings and the urban-rural population distribution of the PSUs. Fifty-two minor strata were defined based on the demographics of the PSUs.

- four stages of cluster sampling within strata:
    1. county ("in a few cases, adjacent counties were combined to keep PSUs above a certain minimum size")
    2. census blocks ("or combinations of blocks")
    3. dwelling units
    4. individuals or households
- further reading: *National Health and Nutrition Examination Survey: Sample Design, 2011–2014*

Estimators for complex sampling methods

- despite the complexity of these survey designs, point estimation is actually fairly easy
- consider each stage fo sampling and compute the probabilities of inclusion
- then use the Horvitz-Thompson estimator

Variance of estimators for complex sampling methods

- whereas with the prior methods we've discussed there are mathematical expressions for the variance, for complex sampling methods we must use approximations
- there are two primary approximation techniques:
  - Taylor-series linearization
  - replication
- both options are implemented in most software
- later, I'l show implementation of Taylor-series linearization

Sampling methods for hard-to-reach populations

Sampling methods for hard-to-reach populations

- as mentioned, my work involves sampling hard-to-reach populations
- these include men who have sex with men, people who inject drugs, and transexual individuals
- other examples of hard-to-reach populations: homeless individuals, immigrants, . . .
- what makes these populations hard to reach?
- no list exists for these populations and the populations are often marginalized

Time-location sampling

- time-location sampling (TLS): a location-based sampling method
- sample locations (venues), then sample times (events) associated with the events, and then sample individuals
- very similar to cluster sampling, with the key difference that the clusters are not necessarily mutually exclusive: an individual can not be found in more than one hospital (cluster sampling) but can be found in more than one venue (TLS)
- in other words, in cluster sampling, individuals can be mobile: they can visit more than one cluster and enter and exist the same clusters

Example of time-location sampling

- we use time-location sampling to sample men who have sex with men (MSM)
- the project is part of the CDC's National HIV Behavioral Surveillance for MSM, described in the optional reading by MacKellar et al
- our staff use qualitative methods to identify venues at which we might be able to sample MSM
- these include bars or dance clubs, parks and street locations, cafes and restaurants, and social organizations
- each month, our staff create a sampling frame of venues and associated events (times)
- they use this sampling frame to sample venues and sample events
- finally, they visit the venue at the sampled time and sample individuals

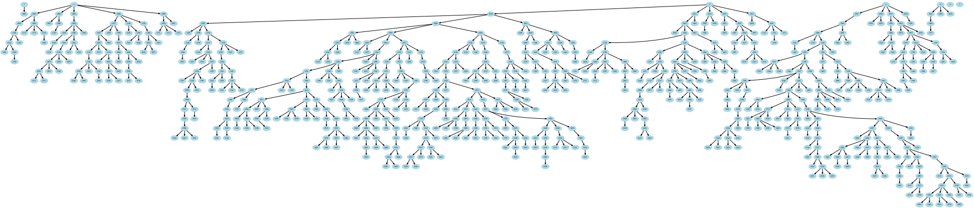
Estimators for time-location sampling

- various estimators exist, but most use the H-W approach
- estimating probabilities of inclusion can be difficult, since there are 3 stages: sampling the venue, sampling the event, and sampling the individuals
- in addition, some methods attempt to take into account frequency of venue attendance
- this addresses the mobile nature of individuals relative to venues

Respondent-driven sampling

- respondent-driven sampling (RDS) is a peer-referral sampling method
- researchers purposefully select initial participants, known as seeds
- the seeds recruit their peers, who then recruit their peers, who then recruit their peers. . .
- the number of referrals per participant is usually capped to limit the influence of any one individual on the sample
- recruitment connections are recorded, usually via the use of paper coupons and coupon numbers (however, there have been digital implementations as well)
- ask individuals what their network sizes are, since this is used in analysis
- participants are usually compensated for participation and for each recruit

Respondent-driven sampling



Respondent-driven sample of people who inject drugs in San Francisco, 2012.

Why use peer referral?

- a feature of RDS is that the burden of recruitment is placed on population members rather than on researchers
- this is generally considered desirable for hard-to-reach populations (which are often marginalized) for several reasons:
  - secrecy of venues: venues where population members may be illegal (even if not the venues are not illegal, the population members may wish to conceal the venues from outsiders)
  - identification of participants: researchers might not be able to identify participants on their own (since the population is hidden)
  - anonymity of non-participants: researchers don't know the identies of non-participants
  - safety: interviews usually occur at researcher-operated interview sites, avoiding possible dangers associated with conducting research in the field

Does selection of seeds matter?

- the theory behind RDS is that as the sample progresses and the number of recruitment waves increases, the effect of the selection of seeds on the sample is reduced
- nevertheless, researchers often try to select diverse seeds
- the rationale is that groups are not always well-connected and that individuals often recruit individuals like themselves

In what settings might we use RDS?

- individuals should be socially connected

Discussion: Firestone et al

In the absence of an accessible and accurate population-based sampling frame for urban Aboriginal people, we implemented respondent-driven sampling (RDS)

- data linkage for ER visits
- comparison with general population
- design effect

Estimators for respondent-driven sampling

- several estimators for RDS data exist, with accompanying methods for estimating variance
- most take into account network sizes (degrees)
- some also take into account recruitment patterns across and within groups
- simulation studies have compared methods

Software for respondent-driven sampling

- *RDSAT*
- *RDS* package for R: implements 3 estimation methods, including a recent method by Gile
- *RDS Analyst*: a graphical interface to the R package

Data analysis

Discussion: Korn and Graubard

Since large-scale health surveys usually have complicated sampling schemes, there is often a question as to whether the sampling design must be considered in the analysis of the data.

What was the scientific question in the Stevens et al study? What were the two primary analytic decisions considered?

Data analysis for classic sampling methods

- regardless of software, the principles for survey analysis are the same
- the data should have the following elements: (1) variable(s) for weights (more than one variable if you are using the replication method for variance estimation) and (2) variables for strata or clusters
- optionally, you might have a variable for finite population correction to adjust variance
- you then specify those variables in analysis
- I'll demonstrate analysis using data from the National Health and Nutrition Examination Survey

National Health and Nutrition Examination Survey

- data from the 2013–2014 survey
- unique files for each domain
- I downloaded and merged data on demographics, blood pressure (from examination), body measures (from examination), cardiovascular health (from questionnaire), health insurance (from questionnaire), and medical conditions (from questionnaire)

National Health and Nutrition Examination Survey

seqn	sdmvstra	sdmvpsu	wtint2yr	wtmec2yr
73557	112	1	13281.24	13481.04
73558	108	1	23682.06	24471.77
73559	109	1	57214.80	57193.29
73560	109	2	55201.18	55766.51
73561	116	2	63709.67	65541.87
73562	111	1	24978.14	25344.99

- `seqn`: ID for individuals
- `sdmvstra`: variable for stratum
- `sdmvpsu`: variable for cluster (primary sampling unit)

National Health and Nutrition Examination Survey

- why is there just one variable for cluster?

  masked variance units (MVUs) are provided for use with the public-use data files to reduce the chance of an intruder being able to match PSUs in the sample to PSUs in the population... MVUs can be used as if they were pseudo-PSUs to estimate sampling errors (similar to past NHANES).

  The MVUs or pseudo-PSUs on the data file are not the "true" design PSUs but are a collection of secondary sampling units aggregated into groups for the purpose of variance estimation.

- further reading: National Health and Nutrition Examination Survey: Estimation Procedures, 2007–2010

National Health and Nutrition Examination Survey

seqn	sdmvstra	sdmvpsu	wtint2yr	wtmec2yr
73557	112	1	13281.24	13481.04
73558	108	1	23682.06	24471.77
73559	109	1	57214.80	57193.29
73560	109	2	55201.18	55766.51
73561	116	2	63709.67	65541.87
73562	111	1	24978.14	25344.99

- `wtint2yr`: weight variable for interview data
- `wtmec2yr`: weight variable for examination data

Analysis in Stata

- specify the survey design using svyset:

```
. svyset [w=wtint2yr], psu(sdmvpsu) strata(sdmvstra)
(sampling weights assumed)

      pweight: wtint2yr
          VCE: linearized
  Single unit: missing
     Strata 1: sdmvstra
         SU 1: sdmvpsu
        FPC 1: <zero>
```

Analysis in Stata

- for percents use `svy: proportion`:

```
. svy: proportion hiq011

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
hiq011       |
           1 |   .8449897    .0099266      .8226345    .8649898
           2 |   .1539593    .0100292      .1337741    .1765695
           7 |   .0002616    .0001715      .0000647    .0010575
           9 |   .0007893    .0004709      .0002212    .0028125
--------------------------------------------------------------
```

Analysis in Stata

- means using `svy: mean`:

```
. svy: mean ridageyr

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    ridageyr |    37.5001   .4340203      36.57501    38.42519
--------------------------------------------------------------
```

Analysis in R

- use the *survey* package:

```
library(survey)
```

- specify the survey design using `svydesign()`:

```
sii<-svydesign(strata=~sdmvstra,id=~sdmvpsu,weights=~wtint2yr,
               data=ii,
               nest=TRUE)
```

- the `nest=TRUE` option isn't usually necessary but needed here, I think, due to the strange nature of the masked sampling units

Analysis in R

- for percents use `svytable()` with `Ntotal=1` or `Ntotal=100`:

```
svytable(~hiq011,sii,Ntotal=1)
```

```
## hiq011
##            1            2            7            9
## 0.8449897454 0.1539593464 0.0002616111 0.0007892971
```

- for confidence intervals use `svyciprop()`:

```
svyciprop(~hiq011==1,sii)
```

```
##                       2.5% 97.5%
## hiq011 == 1 0.845 0.823  0.86
```

Analysis in R

- for means use `svymean()`:

```
svymean(~ridageyr,sii)
```

```
##             mean    SE
## ridageyr 37.5 0.434
```

- for confidence intervals use `confint()`:

```
confint(svymean(~ridageyr,sii),df=degf(sii))
```

```
##              2.5 %   97.5 %
## ridageyr 36.57501 38.42519
```

Exercise: analyze the NHANES data

- percents and confidence intervals for:
  - gender (`riagendr`)
  - served in US Armed Forced (`dmqmiliz`)
  - pain or discomfort in chest (`cdq001`)
  - health insurance (`hiq011`)
  - asthma (`mcq010`)
  - coronary heart disease (`mcq160c`)
  - cancer (`mcq220`)
- means and confidence intervals for:
  - systolic blood pressure (`bpxsy1`) (use exam weights)
  - diastolic blood pressure (`bpxdi1`) (use exam weights)
  - body mass index (`bmxbmi`) (use exam weights)