I do love me a schedule…

	DEADLINES		LECTURE		LABS				
Date	**Things Due**	**Reading (15th)**	**Category**	**TOPIC**	**Lab #**	**Lab Assignment**	**Type**	**M**	**W**
Mon Aug 26		-	Orientation	Orientation	-	-	-	-	-
Wed Aug 28		chapt. 1	Orientation	Inquiry & Research	1	Empirical Claims	Topical		1
Mon Sep 2				(Labor Day)					
Wed Sep 4		chapt. 4	Orientation	Research Design	(2)	Mini-Design	Solo		(2)
Mon Sep 9		-	Orientation	SPSS Intro	3	Computing (miles)	Data	1, (2), 3	
Wed Sep 11	HW1 Article Review	chapt. 2	Orientation	Explanation & Theory	4	Writing Hypo	Topical		3, 4
Mon Sep 16		ch.17 & p.427	Orientation	Researching Research *(guest)*					
			Measurement	Recoding *(in lab?)*	5	Recoding (colleges)	Data	4, 5	
Wed Sep 18		pp. 124-137	Orientation	Conceptualization	(6)	Measuring Love	Solo		5, (6)
Mon Sep 23	HW2 Annotations	pp. 138-157	Measurement	Operationalization	7	Internal Validity	Topical	(6), 7	
Wed Sep 25		pp. 421-426	Measurement	Elementary Stats	8	Univariates (gsssub)	Data		7, 8
Mon Sep 30		chapt. 3	Orientation	Misuses of Research	EC1	Ethics (extra credit)	Solo	8, EC1	
Wed Oct 2	HW3 Dataset Basics	p.415-420, 434-438	Measurement	Quantitative Analysis	9	Coding Heroes	Topical		EC1, 9
Mon Oct 7		ch.6 & 427-428	Measurement	Indices & Scales	10	Lab (abort)	Data	9, 10	
Wed Oct 9		pp. 249-256	Measurement	Surveys: Wording	11	Draft Survey Questions	Topical		10, 11
					(12)	Apartments	Solo	11, (12)	
Wed Oct 16	MIDTERM	-	-	Project Overview (mandatory lecture)	-	-	-		(12)
					13	Crosstabs (gss84)	Data	13	
Wed Oct 23	HW5 Indices	pp. 256-263	Measurement	Surveys: Format	14	Draft Survey Format	Topical		13, 14
Mon Oct 28		-	Description	Testing (Ordinal) Association	15	Ordinal Tests (USDecline)	Data	14, 15	
Wed Oct 30		chapt. 8	Observation	Experimental Research	16	Review Survey Draft	Topical		15, 16
Mon Nov 4	HW6 Proposal	-	Description	Three Interval Tests	17	3 Tests	Data	16, 17	
Wed Nov 6		pp. 263-288	Observation	Surveys: Mode	(18)	Conduct Survey	Solo		17, (18)
Mon Nov 11				(Veteran's Day)					
Wed Nov 13		pp. 433-434	Description	Multivariate & Modeling	19	Multiple R (multireg)	Data		19
Mon Nov 18	HW7 Three Tests	ch.13 & 428-429	Orientation	Qualitative Analysis	(20)	Jaywalking Study	Solo	(18), 19	
		pp. 420-421	Observation	Data Entry	(21)	Data Entry	Solo	(20), (21)	
Wed Nov 20		chapt. 10	Observation	Fieldwork & Ethnography	(22)	Grounded Theory	Solo		(20),(21),(22)
Mon Nov 25	HW8 Experiment	chapt. 15	Description	Control Variables & Elab	EC2	Elaboration (extra credit)	Data	(22),EC2	
Wed Nov 27				(Thanksgiving)					
Mon Dec 2		chapt. 11	Observation	Unobtrusive Research	23	Content Analysis	Topical	23	
Wed Dec 4	HW9 Survey Analysis	pp. 316-318	Observation	Focus Groups	24	Evaluation Design	Topical		EC2, 23, 24
Mon Dec 9		chapt. 12	Description	Oral Presentations	(25)	Oral Presentations Eval	Solo	24,(25)	(25)
Mon Dec 16	Final Exam			(no class meeting - due 5pm, [illegible])				-	-

SOC497/L: SOCIOLOGY RESEARCH METHODS

Sampling Effects:

Errors, Bias, & External Validity

Ellis Godard

The highlight of your week so far is

A. HW4 — 30%

B. Weekend Zooms w/ Ellis — 17%

C. Studying for the midterm — 40%

D. "Catching up" on the reading — 7%

E. Watching *The Princess Bride* — 7%

0

1. Variation ratio is the % of…

A. Cases that are the mode — 53%

B. Cases that are the median — 9%

C. Cases that aren't the mode — 35%

D. Cases that aren't the mean — 3%

E. Values that are divisible by 7 — 0%

0

Outline for Today

- Overview
  - Non-probability methods (6)
  - Probability methods (4)
- Census vs. Sample
- Choosing a Sample Size (time permitting)
- Lab: Sampling methods

Why Sampling?

- Exercising a scientific vision
  - Want to order empirical facts
  - To do that, find abstractable patterns
    - Infer (generalize to) other cases from them
  - Do it iteratively, between data and theory
    - Wheel of science
- Need to sample - can't study all cases
  - Too expensive
  - Too time consuming
  - May not even know who the cases are

How Sampling Fits In

- Make inferences from sample to other cases
  - If follow rules, can use *sampling theory*
  - Allows inferences from observed to unobserved cases
- Next week starts Analysis & Presentation
  - What we DO w/ those facts
  - forms of analysis, esp. of quantitative data
  - particular focus on the elaboration method
- First: how we COLLECT those facts
  - inc. selecting cases and methods/forms of observation
  - start today w. the logic of sampling

Census vs. Sample

- Sometimes can obtain data about entire pop.
  - census = when gather data on every member
  - not always possible (see Kish 1987:140)
- Advantages of censuses
  - precise, detailed to small domains, inclusive, credible
- Advantages of samples
  - rich, complex, diverse, flexible, accurate, pertinent, inexpensive
- U.S. census uses multi-stage clusters
  - phonebooks > pages > names;
    or colleges > classes > students

Census Realities

- US census said to be "usually one of the least newsworthy of federal government activities"
  - 2000 Census, however, was front page news
  - Congress, Pres, Supreme Court, media, & others
- Several big issues, including
  - changing measurement of race (split Qs, *multiple response*, etc.)
  - whether to get rid of sampling
    - constitution requires an "actual enumeration" but undercounts and overcounts happen in practice
    - Census procedure = conduct TWO samples -> 2 datasets for all 7M census blocks
  - Long form / short form
  - weighting and adjustments
- Sampling (& apportionment) debates not new

The Problem with Samples

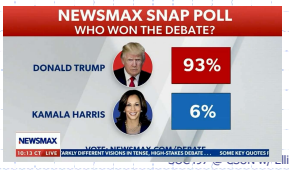
- Remember sampling error:
  - Numeric difference btwn sample statistic & pop. parameter
  - Sample 50% male, population 48% male; *sampling error* = 2%
  - Avg educ level = 18 yrs in class, 11 in state; *sampling error* = 7
  - Warning: *not the same as standard error*
- *Want* the sample to look like the population
  - Problem:
    - usually don't *know* the parameter
  - Solutions:
    - *Estimate* parameters w/ ideas about errors across multiple samples
    - Ensure methodological rigor (inc. high response rate)
    - Appropriate sample size
  - Choices – esp the *kind* of sampling employed

Non-Random Samples

Trump falsely claims every poll says he won the debate w Harris

https://www.cnn.com/2024/09/12/politics/fact-check-trump-claims-won-debate

He's citing a Newsmax poll (a terrible sample) AND he's misrepresenting *its* results



He wrote on social media on Wednesday afternoon: "Every Poll has us WINNING, in one case, 92-8, so why would I do a Rematch?"

"We won the debate according to every poll — every single poll, I think," Trump told reporters later in the afternoon.

Facts First: *Trump's claim is false. As of Thursday morning, every major scientific poll about the debate — every poll that used random sampling techniques to try to obtain a representative picture of US public opinion — had found that Harris won. The polls in which most respondents said Trump won were unscientific junk polls — open questions posted online — that allowed an unlimited number of people to click and respond no matter how old they are or what country they live in, making the results useless as a measure of US public opinion.*

Two General Types of Samples

1. **Probability**
   - probability of selection of each case is known
     - Not necessarily equal
   - More likely to be representative; no *systematic* bias
     - Probability theory estimates *degree* of representativeness
   - Few samples truly are, but we try as best as possible
2. **Nonprobability**
   - Everything else; some bias or limitation
   - probability of selection *isn't* known
     - e.g. passersby on the street

- Not necessarily survey (e.g. fingerprints)

Quota Sampling (nonprobability #6)

- Involves caps on certain proportions
- Start w/ desired target demographics
  - e.g. 50% F, 12% African-American
- The intersections are cell proportions
- The matrix of all those cells is a *quota frame*
- *Not* the same as strata or clusters (next)
  - Limits here are once the study begins
- Problems:
  - quota frame must be accurate
  - sampling errors are added for each *cell*
  - must be careful about biased selection *w/i* frame

Non-probability Samples (3 of 6*)

1. Available Subjects
   - Folks nearby – friends, colleagues, family
   - Adv/Disadv: Prone to bias, though okay for a pretest
2. Volunteer/Convenience
   - Examples:
     - People at street corner, in library, at a mall, etc.;
     - comment cards, magazine inserts, etc.
   - Disadvantages: Biased to extremes (those w/ interest) & negatives (those w/ complaints)
   - Advantages: Oft used in marketing but not much scientific use
3. Purposive/Judgmental
   - ID typical group based on knowledge of population & purpose of the study
   - Good for initial design w/ accessible subset
   - Good if focusing on deviant/trouble cases (what not fit gen. pattern)

* Note: Only 4 of these are in Babbie.

2. The opposite of probability sampling is…

A. Conventional Sampling
0%

B. Cluster Sampling
0%

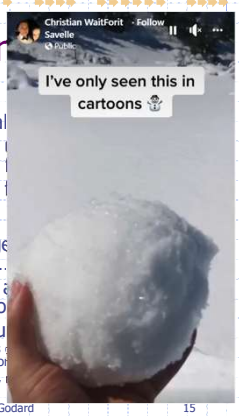
C. Modern Sampling
0%

D. Non-probability sampling
97%

E. Statistical Sampling
3%

Non-probability Sar

4. Informants
   - May need them to be typical
   - But those willing to talk are
     not know what you need to
   - may be powerful, public, or
5. Snowball
   - Sample size accumulates (ge
     through time (as it rolls on…
   - start w/ known informants; a
     help in finding 2 other respo
   - Esp. used for exploratory pu
     - Also description, esp. when members
       locate or when sampling frame not co
     - Homeless, AIDS patients in early 80s,
       immigrants, etc.



* Note: Only 4 of these are in Babbie.

Note on Representativeness

- Each of these methods introduces bias
  - Assumptions about availability of subjects
  - Assumptions about characteristics of pop.
- "Probability samples" asserts will get an *approximately* representative sample

Admin Overview Non-Prob. Probability Misc Choosing Lab

Probability Samples (4)

- Simple Random Sample
  - Random number table picks the cases
- Systematic Random Sample
  - Use a "skip number" to take every Kth case
- Stratified Random Sampling
  - Theory and data categories guide choices
- Cluster Sampling
  - Convenience and real life guide the selection

Admin Overview Non-Prob. Probability Misc Choosing Lab

Simple Random Sample

- Assumed in stat. theory (w/o replacement)
  - Each case has same chance of being selected, so
  - Each possible sample has the same chance of being selected (!!)

NOTE:

- Random is not the same as arbitrary
  - random means ignorance by choice and design
  - We don't know who we're picking

Admin Overview Non-Prob. Probability Misc Choosing Lab

Simple Random Sample, cont'd

- Assign each case in sampling frame an ID #
- Use random number table to get sampled #s
  - First decide how you'll read it – which digits (first 2?) and which direction (down each column?)
  - Doesn't matter how you do it, just be consistent
  - Pick a random starting point (*just* the starting pt.)
  - Some numbers won't be useable; just skip those
- Adv: eliminates bias, permits inference
- Disadv: time consuming, need entire sampling frame, expensive (tho less w/ PCs)

Admin Overview Non-Prob. Probability Misc Choosing Lab

Systematic Random Sample

- Every kth member is chosen
  - k is the "skip number" (Popsize/samplesize = N/n)

1. Get list of cases (randomized to elim. bias)
2. Calculate K
   - round down or will run out of cases (e.g. for N=100 and n=8, use 12 *not* 13)
3. Randomly select 1 of 1st K cases in frame
   - use the table, but only for the start number

- Adv: faster, easier, cheaper, fewer mistakes
- Disadv: need complete sampling frame with no pattern (army, city blocks); beward periodicity; may be expensive

Admin Overview Non-Prob. Probability Misc Choosing Lab

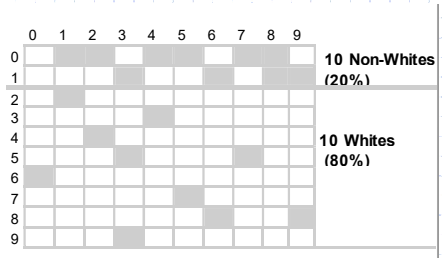
Stratified Random Sampling

- Population divided into "strata" by variable(s)
  - groups that are similar on some characteristic
  - must be mutually exclusive (i.e. at least nominal)
- SiRS or SyRS from each stratum
- Could be proportional
  - Sample proportions chosen to reflect population
    - Black respondents, southern states,
  - Like quotas, but during sampling not observation
- Could also do disproportional
  - Esp. when want to oversample a small group
- Adv: Sampling error is reduced by two factors
  - A large sample, and a homogeneous population
- Disadv: need adequate sampling frame
  - Typically, the entire population (*every* case)

Admin Overview Non-Prob. Probability Misc Choosing Lab

Disproportionate Strata

An over-sample of non-white….



Weighting & Sample Weights

- Used if sample is disproportionate
  - E.g. population is 12.8% African-American, but sample is 6.4% African-American
  - Use a data "trick": pretend there were twice as many African-Americans
  - "weight by race" in SPSS
- Something to remember & consider
- But you needn't weight data this term

Multistage Design & Errors

- A two stage cluster sample is subject to two sampling errors, one at each stage.
  - Selecting a disproportionately wealthy city block plus disproportionately wealthy residents within those blocks.
    - Elements within a natural cluster are more homogeneous than are elements composing the total population
      - Birds of a feather flock together!

Q3. For a highly disproportionate sample, it may be appropriate to…

A. Cry 3%

B. Dance 0%

C. Give up 0%

D. Start over 10%

E. Weight the data 87%

Clusters vs. Strata

Not in lecture – notes only

- Clusters:
  - chosen for convenience and should be internally *heterogeneous*
- Strata:
  - dictated by theory and should be internally *homogenous*
- Example: MTV's *Real World*:
  - The cast of each season is a cluster
    - It's a practical group of possible cases
    - Any one cast is pretty much like the rest
    - Heterogenous within each
  - Strata would be groups across casts
    - E.g. age, race, gender, sexual orientation
    - Would need list of all the members of all the casts, or at least a list of percentages of each subgroup (male and female; black white and other, etc.)

Cluster Sampling

- Steps:
  - Sampling frame is divided into clusters
    - Practical, natural groupings of cases rather than theoretical or demographic ways they differ
  - Randomly select clusters
  - Randomly select cases within those clusters
    - Sometimes use all cases within selected clusters
- Advantages
  - save time & money
  - just need list of clusters, not all cases
- Disadvantages:
  - may not have enough variation if small clusters
  - risk getting unusual clusters
    - biased if they are theoretically important (wealthy suburbs)
  - Statistical inference procedures very complicated
    - must people don't even use them

4. Sampling theory assumes…

A. Simple Random Sampling 43%

B. Systematic Random Sampling 43%

C. Cluster Sampling 10%

D. Quota Sampling 0%

E. Multi-Stage Sampling 3%

Homogeneity vs. Heterogeneity

Not in lecture – notes only

- Failure of early polls was not adequately representing the full range of voters
- Sometimes ok to assume all cases alike
  - Blood samples
  - Social psychology - Many studies based on college sophomores in a psych class
- But usually, we're concerned with heterogenous populations
  - Indeed, we *focus* on differences & variations

Considerations in Choosing N

Not in lecture – notes only

- Precision: *the width of the confidence interval*
- Confidence: *Probability that interval will contain population mean*
- Degree of Variability: *Use larger samples for more heterogeneous populations*
- Amount and Type of Analysis: *More complicated studies require larger samples*
- Resources: *How much research money do you have …*

What makes a good sample?

Not in lecture – notes only

- Ideal is to be *representative*
  - Aggregate characteristics clearly resemble same aggregate characteristics of the pop
  - Variation not *lacking* or *unbalanced*
  - In practice, may not *know* parameters
- Needn't be representative on *all* respects
  - Only on characteristics relevant to the study
  - But might not know what those are in advance
  - & Don't want criticism about hidden "control" variable
- Randomized selection increases *likelihood*
  - Does NOT guarantee it
  - Always a *chance* of sampling errors
  - Almost always *some* amount of error
  - Each *method* allows for particular errors

Sample Size Required for Estimating Population Mean

Not in lecture – notes only

If we want the "margin of error" to be smaller than some value B, with a probability of at least 1 - α, the appropriate sample size is obtained as follows

$$n = \sigma^2 \left(\frac{z_{\alpha/2}}{B} \right)^2$$

How big should my sample be?

Not in lecture – notes only

- Size matters, because it decreases error
  - But larger isn't always "better"
    - Pt. of diminishing returns
  - More important to be **unbiased**
    - Theoretical goal is representativeness
    - Best we can do is reduce barriers to that
- Best practice is to *reduce bias*
  - risks making sampling unrepresentative
  - However, any method introduces some bias
    - assumptions about availability of subjects
    - Assumes results approximately representative
- Probability sampling improves chances
  - by avoiding bias and allowing statistical inferences based on assumptions of random samples

Sample Size Required for Estimating Population Proportion

Not in lecture – notes only

If we want the "margin of error" to be smaller than some value B, with a probability of at least 1 - α , the appropriate sample size is obtained as follows

$$n = .25 \left(\frac{z_{\alpha/2}}{B} \right)^2$$

What's a "good" sample size?

Not on lecture notes only

The formula for n is the *variance*, multiplied by the square of the ratio of *the z score associated with the desired confidence level* and the *allowed margin of error*:

$$n = \hat{\sigma}^2 * \left(\frac{z_{a/2}}{B}\right)^2 = p*(1-p)*\left(\frac{z_{a/2}}{B}\right)^2$$

When the population proportions are unknown and we must use the conservative estimate of 50% (0.5), then to have a 95% level of confidence in a margin of error of not more than +/- 4% requires a sample size of 601 cases*:

$$0.5*(0.5)*\left(\frac{1.96}{.04}\right)^2 = .25(49)^2 \approx 601$$

* Per subgroup or disaggregated unit

5. What is my first name?

A. Babbie 0%

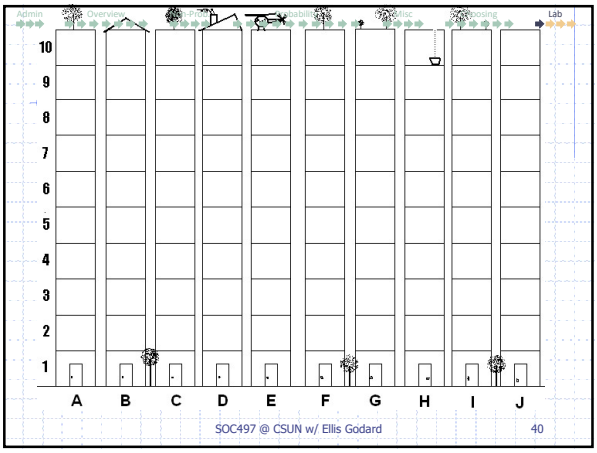
B. Donald 0%

C. Ellis 97%

D. Godard 3%

E. Zebra 0%

:00



Team Scores

Points	Team
4	Watching The Princes...
3.8	Weekend Zooms w/ Ellis
3.78	HW4
3.51	Studying for the midterm
3.5	"Catching up" on the...

Today's Lab: Sampling Effects

- See 2-page handout of apartments
  - 10 buildings w/ 10 apts each, 1st floor to penthouse
- You'll do each (every!) form probability sampling
  - Drawing same sample size from same sampling frame
  - Compare results of the four different methods
- Numbering hints (not needed in A)
  - B: Use 00-99 (not 1-100), or else need 3-digit #s
  - C: 0-9 (per row, i.e. "floor" of the buildings), 10 times
  - D: 0-9 bldgs for SiRS, then 0-9 in each of 2 for SyRS