An Introduction to Sampling and Weighting

Marc N. Elliott

January 27, 2011

Course Origins

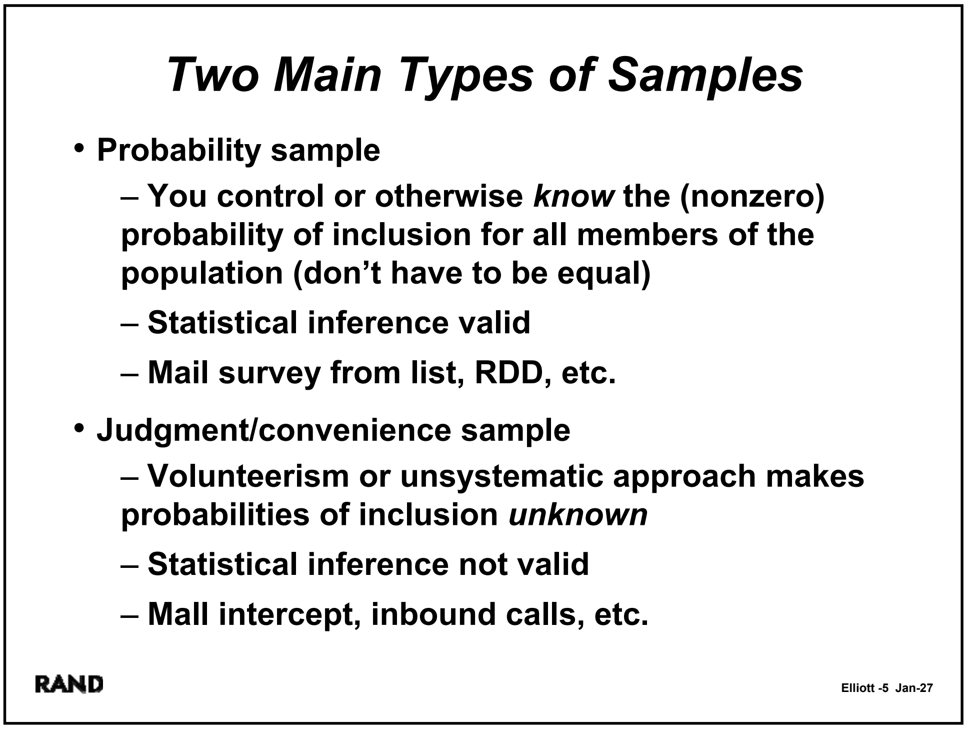
- This review course was developed and first presented in 1998.
- Presented a second time in 2005.
- In response to input from newer staff, we are re-offering the course

Outline

- SAMPLING
  - Probability versus non-probability samples
  - Types of probability samples (e.g., simple random samples, stratified random samples, cluster samples)
  - Systematic and unsystematic survey error
  - Types of survey bias (e.g., frame bias, nonresponse bias, self-selection bias)
- Development and use of simple survey weights.
- Concepts in Sample Size and Power Calculations

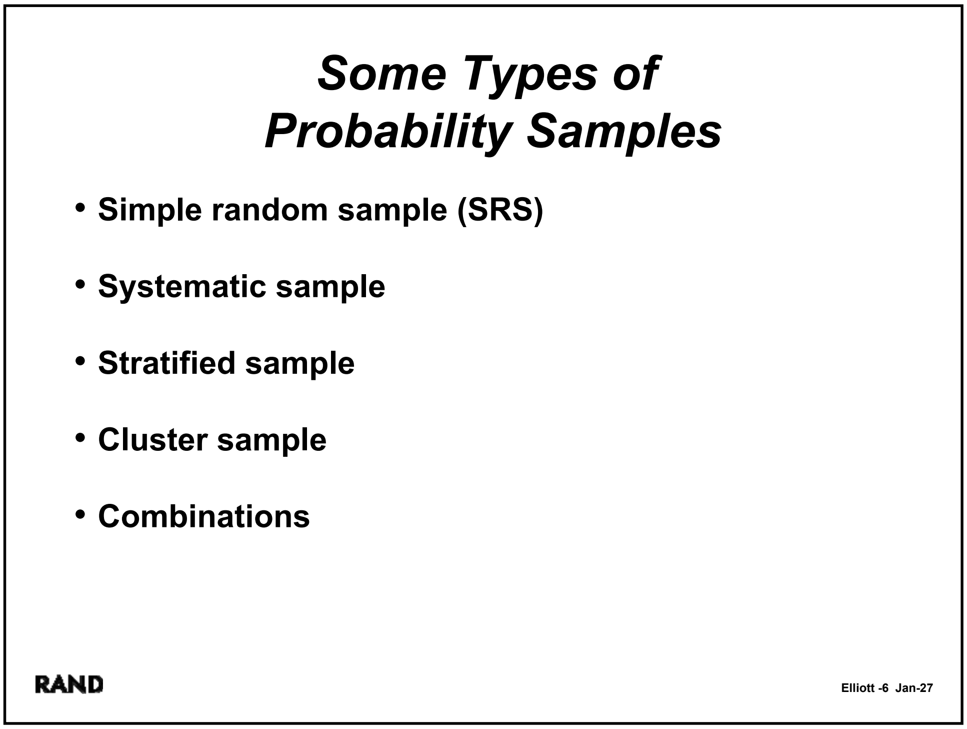
Purpose of Sampling

- Valid and reliable inferences about characteristics (*parameters*) of a large *population* of interest from a smaller *sample*
- One survey may be used to make inferences about multiple parameters for multiple subpopulations
- Parameters can be means, proportions, regression coefficients, . . .
- *Statistics* are calculated on the sample
- Sampling links the sample to the population



Two Main Types of Samples

- Probability sample
  - – You control or otherwise *know* the (nonzero) probability of inclusion for all members of the population (don’t have to be equal)
  - – Statistical inference valid
  - – Mail survey from list, RDD, etc.
- Judgment/convenience sample
  - – Volunteerism or unsystematic approach makes probabilities of inclusion *unknown*
  - – Statistical inference not valid
  - – Mall intercept, inbound calls, etc.



Some Types of Probability Samples

- Simple random sample (SRS)
- Systematic sample
- Stratified sample
- Cluster sample
- Combinations

Sample Statistics are Random Variables Estimating Population Parameters

- Take a SRS and calculate the sample mean for that particular SRS
- Doing this many times produces many sample means
- Draw a histogram of those sample means
- This histogram is an approximation of the density of the Sample Mean, a random variable (R.V.)

$\text{R.V. Sample Mean } \overline{\mathrm{X}}$

$\mathrm{E}(\overline{\mathrm{X}}) = \mu_{\overline{X}} = \mu_X$

$\mathrm{Var}(\overline{\mathrm{X}}) = \sigma^2_{\overline{X}} = \dfrac{\sigma^2_X}{\mathrm{n}} \qquad \mathrm{SD}(\overline{\mathrm{X}}) = \dfrac{\sigma_{\mathrm{X}}}{\sqrt{\mathrm{n}}}$ **also called the standard error**

Central Limit Theorem (C.L.T.)

- As the sample size n becomes large,
- $\overline{\mathbf{X}}$ is approximately Normally distributed with mean $\mu_{\mathbf{X}}$
- and variance $\dfrac{\sigma^2_{\mathbf{X}}}{\mathbf{n}}$
- *regardless* of the underlying distribution of X.
- Good Rule of Thumb:
- Sample size n > 30 for continuous, roughly symmetric
- Might need 100 or even 1000 if really skewed, discrete, etc.

Estimation Error

- *Bias*: Systematic error
  - – Expected difference between sample and population parameter
  - – From sample design or estimator properties
- *Variance*: Unsystematic (sampling) error
- *Mean squared error* (MSE): Total estimation error
  - – Variance + squared bias

Bias and Sample Size

- Bias is typically “invisible” to software packages
- Larger sample sizes reduce variance component, but not bias
- As sample sizes get bigger, bias dominates variance and becomes most of the MSE

Types of Biases

- Whom you select: Selection bias
- Who responds: Non-response bias
- How they respond: Response bias

Whom you select: Selection bias

–Are some people systematically omitted, over-represented, or under-represented?

–Example: Landon vs. Roosevelt

–Solutions

- Alter sampling approach
- Design weights
- Redefine population

Who responds: Nonresponse bias

–Are the people who respond like those who do not in important ways?

–Solutions

- Drive response rate up/alter survey approach
- Nonresponse weights

How they respond: Response bias

– When people answer, do the answers themselves systematically differ from the truth?

– Solutions

- Redesign questions to deal with social desirability, etc.
- Regression adjustment/ calibration

A little bias goes a long way

- **If we use ε to represent standardized bias ($\varepsilon = b / \sigma$)**
- **It can be shown that the effective sample size (ESS) for an estimate is**
  - **$n / (1 + n\varepsilon^2)$**

ε = 0.02 (1% bias on 50% outcome, 0.6% bias on 10% outcome)

-kills 4% of the “efficiency” (MSE-based ESS)

at n=100

-kills 67% of the efficiency at n=5000

ε = 0.10 (5% bias on 50% outcome, 3% bias on 10% outcome)

-kills 67% of the efficiency at n=100

-kills 98% of the efficiency at n=5000

Stratified Sampling

-Proportionate

-Disproportionate

Stratified Sampling Using Proportionate Stratification

- Proportions selected within each stratum are proportionate to the proportion of the population they comprise

- Produces a *self-weighting sample*

- Simple

- Usually improves precision as compared to an SRS (can't worsen)
  - –Depends on relationship of stratification variable to what you are estimating

Chlamydial Infection Example

- 7% of the population have infection, so SRS has

$$\sigma^2 = (.07)(0.93) = 0.0651$$

- 25% of a 10% subpopulation have infection; while remaining 90% of population have a 5% infection rate

- To reduce the variance from that of an SRS, we can take a proportionate stratified sample so that our sample has *exactly* 10% from the high risk subpopulation

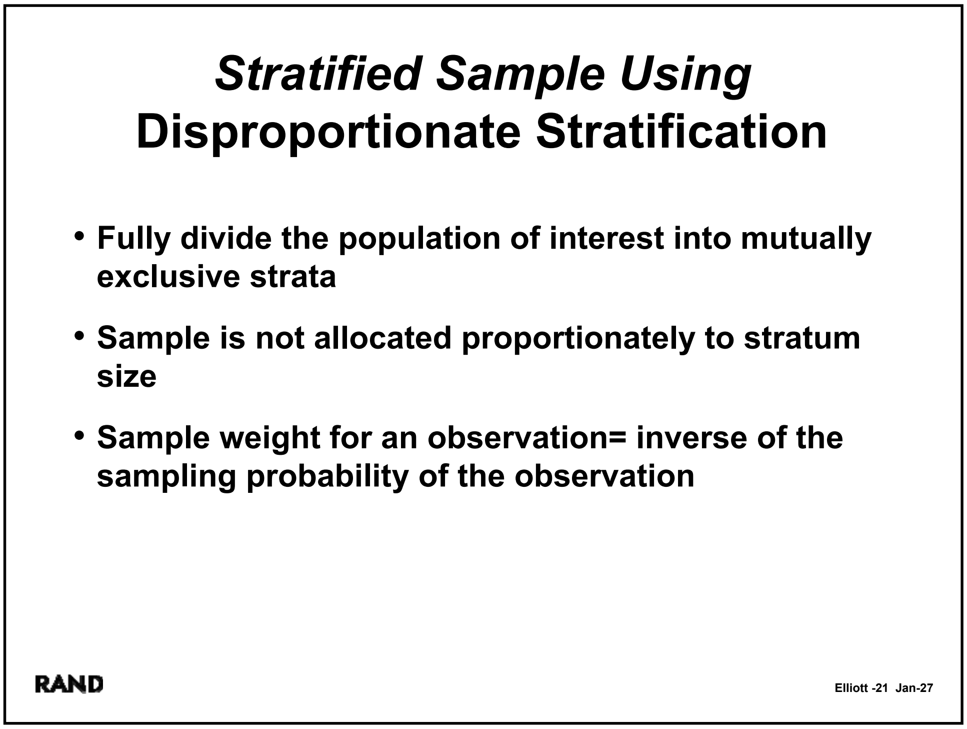
Relationship between SRS and PSS Variances

- For proportionate stratification sampling (PSS), the variance is σ_w^2 , the weighted average of the within strata variances
- The relationship between the SRS variance and the PSS variance σ_w^2 for strata i=1,…,k is

$$\sigma^2 = \sigma_w^2 + \sum_{i=1}^{k} w_i (\mu_i - \mu)^2$$

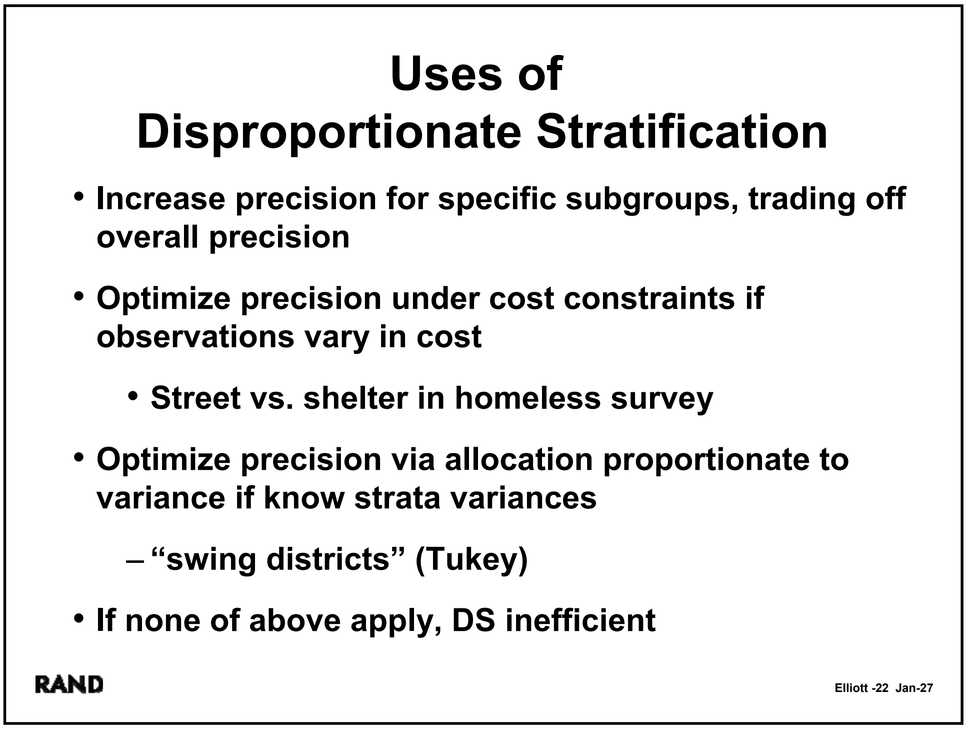
Proportionate Stratification Results in a Smaller Variance

$$w_1 = .1 \quad w_2 = .9$$
$$\mu = .1(.25) + .9(.05) = .07$$
$$\mu_1 = .25 \quad \mu_2 = .05$$
$$\text{Recall SRS } \sigma^2 = .0651$$
$$\text{For PSS } \sigma_w^2 = (.1)(.25)(.75) + (.9)(.05)(.95) = .0615$$
$$\sum_{i=1}^{k} w_i (\mu_i - \mu)^2 = .1(.25 - .07)^2 + .9(.05 - .07)^2 = .0036$$
$$\text{and} \quad .0651 = .0615 + .0036 \text{ (relationship seen on previous slide)}$$
$$\sigma_w^2 = .0615 = 0.945(.0651) = 0.945\sigma^2$$



Stratified Sample Using Disproportionate Stratification

- Fully divide the population of interest into mutually exclusive strata
- Sample is not allocated proportionately to stratum size
- Sample weight for an observation= inverse of the sampling probability of the observation



Uses of Disproportionate Stratification

- Increase precision for specific subgroups, trading off overall precision
- Optimize precision under cost constraints if observations vary in cost
  - Street vs. shelter in homeless survey
- Optimize precision via allocation proportionate to variance if know strata variances
  - “swing districts” (Tukey)
- If none of above apply, DS inefficient

Cambodian stratification example
(Elliott McCaffrey et al 2009 POQ)

- Cambodians in Long Beach
- -People who lived in Cambodia at a certain time are target population; screening of general area for 12% group
- -Allocate to census tracts according to 2000 census
- -Define strata for households based on expert's judgment of whether HH is Cambodian
- -86% sensitivity, 91% specificity
- -Undersample low-prob HHs by a factor of 4
- -Tremendous improvements in ESS/$

Combining Complete and Incomplete Frames-1
(Elliott Finch et al. 2008 Stat in Med)

- Want national probability sample of Chinese Americans in telephone survey
- RDD can result in a complete frame, but inefficient
- Run phone directories through Chinese surname list to generate a listed sample that is an incomplete frame, but efficient
- If just use listed sample do not have a probability sample
- But can define strata to make this a disproportionately stratified probability sample
  - We can account for differences among listed!

Combining Complete and Incomplete Frames-2

- Define two strata (mutually exclusive and exhaustive):
  - (1) Chinese on lists- N1 in population
  - (2) Chinese not on lists- N2 in population
- RDD of sample size n1 should reach n1*N1/(N1+N2) in stratum 1 and n1*N2/(N1+N2) in stratum 2
- Surname list supplement of size n2 is entirely in stratum 2
- Probabilities of being selected
  - Stratum 1: n1/(N1+N2)
  - Stratum 2: n1/(N1+N2)+n2/N2

Post-stratification

- Stratification variables may only be available *after* data collection
- Strata may be defined after sampling
  - May reduce non-response or selection bias
  - Increases variance
- *Iterative Proportionate Fitting/Raking*

Cluster Sampling

Pros & Cons of Cluster Sampling

- **Advantages**
  - **–Feasibility (two-stage access to units)**
  - **–Cost-effective (travel)**
  - **–Want clusters for analysis (market, neighborhood, etc.)**
- **Disadvantage:**
  - **–Loss of information b/c of homogeneity within groups (variance inflation)**

Design Effects of Complex Surveys

- Design effect (DEFF) of a complex design =

(variance of the estimate obtained via the complex design)

divided by

(variance of the estimate obtained via a SRS with the same sample size)

- May come from stratification/weighting
  - Applies to all outcomes equally
- May come from clustering
  - Affects all outcomes, but unequally

Meaning of Design Effects

- DEFF>1 ---> loss of precision relative to an SRS
  - Most common for complex survey design
- DEFF=1 ---> precision is equivalent to that of an SRS
- DEFF<1 ---> gain in precision relative to an SRS
  - may happen with a proportionately stratified sample

Effective Sample Size

- ***Effective Sample Size (ESS)*** is the sample size of simple random sample to which the current sample size is equivalent given the DEFF from the complex survey design

ESS=N/DEFF

- **Effective Sample Size Translates Design effects (DEFF) into Sample Size terms.**
  - **The ESS of a SRS is the nominal Sample Size**
  - **Allows comparison alternative design in common terms**
    - **Minimize cost per ESS**
    - **Maximize ESS under cost constraints**

- **We sometimes really want ESS (and DEFF) based on MSE, rather than variance, but there is no standard term for this (MSE-based ESS?)**

Design Effects in Cluster Sampling

- **Size of the clusters (B) and the degree of similarity of items within a cluster, as measured by the *intra-class correlation coefficient =ICC=r,* increase the loss of precision in a cluster sample**

- **ICC is the proportion of variance of individual scores attributable to clusters**

r = 0 ---> items are as heterogeneous within clusters as between (random assignment to clusters)

r > 0 ---> items are more homogeneous within clusters than between (most common)

Typical values of r are between 0.01 and 0.15

Formula for DEFF for Cluster Samples

$$r \approx \frac{B(Var(cluster\ means)) - Var(X)}{(B-1)Var(X)}$$

$$DEFF \approx 1 + (B-1)r$$

- If Var(cluster means)=Var(X), then r=0
- Related to F-Stat in 1-way ANOVA (with clusters as groups)
- Can derive from PROC VARCOMP (SAS) or simple hierarchical models

Cluster Sampling Example

- Suppose r=0.05 and a cluster sample of 15 students from each of 20 schools is chosen
- total sample size=300
- DEFF=1+(15-1)(0.05)=1.7
- The variance of a the sample mean estimated from the cluster design is 1.7 times as large as it would be based on an SRS of 300 students
- ESS=300/1.7=176.5
- If 25 students are chosen from each of 12 schools, which also produces a total sample size of 300, then DEFF=2.2 and ESS=136.4

ICC limits usefulness of more observations within a cluster

- Let r be the ICC, B be the cluster size
- 1/r is the maximum ESS per cluster
- Br is a measure of how “saturated” the clusters are relative to the ICC
- As Br increases, the marginal value of observations added to a cluster drops rapidly
  - >50% of maximum ESS is achieved at Br=1; >75% at Br=3; >90% at Br=9
- For example, at r=0.05, 20 is the maximum ESS per cluster
  - B=19 -> ESS of 10 /cluster; B=57->ESS of 15/cluster; B=171->ESS of 18/cluster
  - If extra observations within a cluster have marginal cost, they are wasteful beyond a certain point

Combining Design Effects: Example

- Rate of Chlamydia in high school students
- Survey 1 high risk and 9 low risk students from each of 30 schools; r=0.1
- DEFF(PSS)=0.945
- DEFF(Clustering)=1+(10-1)(0.1)=1.9
- DEFF(Overall)=DEFF(PSS)*DEFF(Clustering)=1.80
- ESS=30(10)/1.80=167

Cluster Sampling and Group Randomized Trials

- As an aside, the issues of ICCs in cluster sampling have strong parallels in group randomized trials
  - Same losses in power with high ICCs, large sample size per randomized group
  - “Saturation” heuristics apply in a similar manner
  - See Torgerson (2001 BMJ) for trade-offs vs. contamination

Weighting

- Calculating and using sample weights
- Design weights and non-response weights
- “Design effects” of weighting (variance inflation)
- Trade-offs in weighting
  - Fixing problems
  - When not to use weights

Purpose of Weights-1

- Imagine that we want to know the proportion of recent inpatients who would recommend their hospital to friends and family.
- From a list, we send out a survey to a subset, which is completed by a subset of that group
- How do we estimate the above parameter from our survey responses?

Purpose of Weights-2

- If all on the list were *equally likely to be surveyed* and *equally likely to respond*, we could simply average the outcome among respondents.
- If, however, the above conditions are not met, *and* there is some association between these probabilities and the characteristics we are measuring, simple averages will be *biased*.
- Weights can reduce or eliminate these biases.

Weights Have Limitations

- Weights cannot turn a convenience sample into a probability sample
- In a probability sample, weights are not always needed when sampling probabilities are unequal
- Poorly designed weights can make inference *less* accurate

Types of Weights-1

- Design weights
  - Correct for *known* differing probabilities of selection of population members into the sample we attempt to contact
  - Used with disproportionate stratified random sampling (e.g., attempt to contact 10% of one subgroup but 20% of another subgroup)
- Non-response weights
  - Correct for *estimated* differing probabilities of participation among those we attempt to contact, using information available for both non-respondents and respondents
  - Example: perhaps 74% of females and 52% of males respond

Types of Weights-2

- Post-stratification weights
  - Correct for *estimated* differing probabilities of population members into the sample of respondents using characteristics that are known for the general population but which are not known about the individual members of the population until they respond
  - Example: suppose we did not know the race/ethnicity of non-respondents, but we did know the true distribution of race/ethnicity for the population from a separate data source, and we want our sample to be representative of race/ethnicity

RAND

Creating Weights

Creating Weights-1

- At their simplest, weights are the inverse of the probability of a population member being included in the sample of respondents
  - If we have responses from 10% of those in subpopulation A, 20% of those in subpopulation B, and 25% of those in subpopulation C, . . .
  - We can assign each respondent from subpopulation A a weight of 1/.10=10.
  - Likewise, members of subpopulations B and C would receive weights of 5 and 4, respectively.

RAND

Creating Weights-2

- In a typical survey with non-response, both design and non-response weights are involved.
  - Design weights reflect the probability of selection from the population into the subset we attempt to contact (p_{sel})
    - DW=1/ p_{sel}
  - Non-response weights reflect the estimated probability of people like the respondent responding, given that we attempted to contact them (p_{res})
    - NRW=1/ p_{res}

RAND

Creating Weights-3

- Overall weights reflect the probability of selection from the population into sample of respondents $(p_{sel})*(p_{res})$
- $OW=1/((p_{sel})*(p_{res}))= DW*NRW$
- Overall weights can be estimated directly as the inverse of the fraction of population members who respond within each stratum
- They can also be created in stages: design weights, then non-response weights, then multiply to create overall weight; This approach often has advantages

Example 1:Design weights for HHs for the Cambodian Study

- Within sampled blocks, we selected all HHs “likely” to contain eligible, ¼ of “unlikely” HHs.
- At this stage, an eligible found in a “likely” HH had P=1.00 of being selected ; P=0.25 for an eligible found in an “unlikely” HH
- Design weights at this stage are therefore 1 and 4, respectively.
- 18% of HHs were “likely;” they contained 96% of the eligibles in the sample, the mean weight was 0.96*1+0.04*4=1.11 (w/o rounding)
- Standardized (mean one) weights were 1/1.11 and 4/1.11 = 0.90 and 3.59, respectively

Example 2: Design weights for Chinese Listed/RDD Sample

- Let’s say 30% of population is picked up on lists, total population is 3,000,000, we get 1000 completes by RDD and 2700 from listed sample
- Expect 700 completes for the 2,100,000 unlisted, all from RDD
- Expect 3000 completes for the 900,000 listed (300 from RDD)
- Weights are 2,100,000/700=3000 and 900,000/3000=300, respectively
- Mean weight is 3,000,009/3700=810.8
- Standardized (mean one) weights were 3000/810.8 and 300/810.8 = 3.70 and 0.37, respectively

Creation of Nonresponse Weights

- Design weights are based on known probabilities (you chose the number per strata), so these are just ratios in a simple stratified sample. Often simple.
- Non-response weights involve estimating the probability that people like the respondent respond when contacted.
  - These estimates are based on an implicit or explicit *model* of non-response
  - Incorporating additional variables that are predictive of non-response can further decrease bias
  - Incorporating un-predictive variables (including strata that are too fine for non-response) can add noise to the weights and the estimates

Response Rates and Non-response Bias

- The main reason there is a push for high response rates (beyond the desire for larger sample sizes) is that higher response rates leave less room for differential non-response by personal characteristics

- While it is possible that non-response is entirely random with a 10% response rate and highly selective with a 70% response rate, the reverse is more often true

- But the relationship is weak (Groves and Petcheya 2008); good probability sampling process is most important

Statistical Significance of Non-response Models

- A statistically significant non-response model can correct for differential non-response on the basis of the predictor variables in the model

- A statistically non-significant non-response model can provide some evidence that non-response does not differ on the basis of the predictor variables in the model if the sample size is large enough for adequate power

Design Effects from Weights

Design Effects of Weighting-1

- Weighting may correct bias, but generally at a price-increased variance
- Statistics calculated on unequally weighted observations are less stable (have larger standard errors) than statistics calculated on equally weighted observations
- The more variable the weights, the more variance is added to estimates and inference
- The amount of additional variance added is measured by the *design effect*

Design Effect of Weighting-2

- Design effect (DEFF) of a weighted design =

(variance of the estimate obtained via the weighted design)

divided by

(variance of the estimate obtained via a SRS with the same sample size)

- DEFF>1 ---> loss of precision relative to an SRS
- DEFF=1 ---> precision is equivalent to that of an SRS

Design Effect of Weighting: Kish Approximation

- For disproportionately stratified or post-stratified designs $DEFF=1+(CV_w)^2$ where CV_w is the coefficient of variation (SD/mean) of the weights.
- If the weights are standardized to have mean 1 then DEFF=1+Var(weights).
- Kish approximation overstates DEFF if weights are highly predictive of the outcome (Little & Vartivanian 2005)
  - Weights can be a win-win, reducing both bias and variance

How do weights increase variance?

Variance of a convex combination of two i.i.d. random variables X_1 and X_2

$$Var(tX_1 + (1-t)X_2) = t^2 Var(X) + (1-t)^2 Var(X)$$

$$= 2Var(X)\left(\left(t - \frac{1}{2}\right)^2 + \frac{1}{4}\right)$$

Design Effect Example: Cambodian Study

- 96% of the sample has standardized weights of 0.89
- 4% has standardized weights of 3.59
- DEFF= $1+0.96[(1-0.89)^2]+0.04[(1-3.59)^2]$ =1.26

Design Effect Example: Chinese Telephone Survey

- 81% of the sample has standardized weights of 0.37
- 19% has standardized weights of 3.70
- DEFF=1+0.81[(1-0.37)^2]+0.19[(1-3.70)^2] =1.70
- ESS=3700/1.70=2175
  - Adding 2700 listed sample to 1000 RDD was equivalent to adding 1175 RDD cases- good news if these cases were less than 43% as expensive as RDD cases
  - Marginal value of oversampled listed cases would decrease as the size of the supplement increased

Design Effects From Non-Sampling Weights

- Same formulas for DEFF apply to non-sampling weights (e.g. propensity score weights)

Trade-offs

Trade-offs: Intuition

- **Weights only reduce bias if the basis of over- or under-representation is associated with what you are trying to measure.**
- **If sampling or response probabilities differ by something irrelevant, *no bias is saved* by weighting unequally, but variance is still inflated**
- **Example: If somehow we sampled people whose SSNs were odd at 10 times the rate as people whose SSNs were even, we would NOT want to weight to correct for this inequality, as it almost certainly would be irrelevant to any outcome we might want to measure**
- **Sometimes we can evaluate these trade-offs empirically**

Trade-offs: The MSE Metric

- Total mean squared error of estimation (MSE) =

(variance of observations / n) + (bias of estimates) ^{2}

- Because weights reduce bias but increase variance, there is a trade-off that should be evaluated before weights are used
- Effective weights improve precision (reduce mean square error) by eliminating more bias than they add variance
- We can approximate bias by the difference in weighted and un-weighted estimates (which somewhat overestimates bias), so that change in MSE from weights =
- (DEFF – 1) *(variance of observations / n) - (weighted estimate – un-weighted estimate)^2

Trade-offs and Sample Size

- Design effects (variance inflation) apply to all outcomes equally; bias reduction varies by outcome, so one has to evaluate this trade-off across several important outcomes (See Ghosh-Dastidar Elliott et al. 2009 POQ)

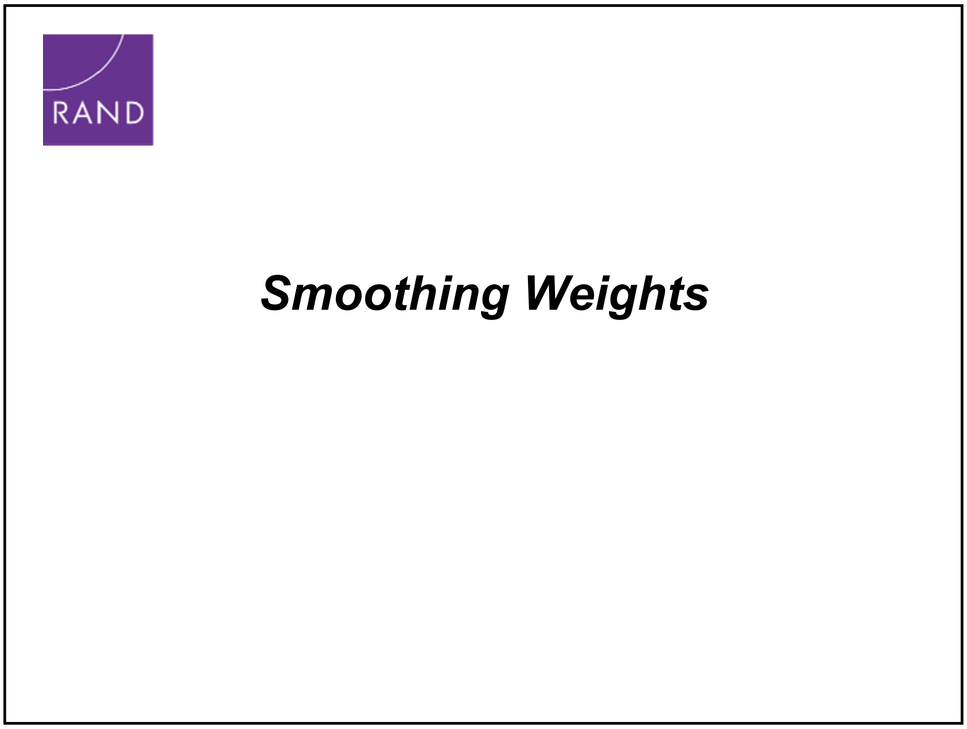
- Because large sample sizes reduce variance but not bias, even small bias reductions are a good trade-off in large sample sizes (See Elliott & Haviland 2007 Survey Methodology)

Bias-Variance Tradeoffs: Cambodia Example

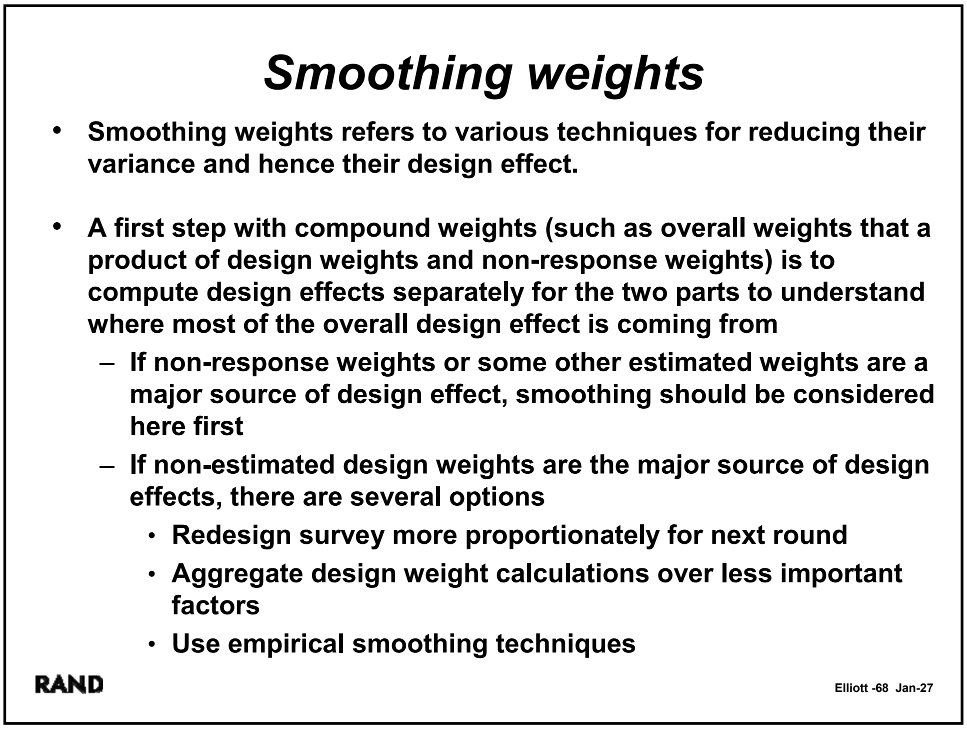
- Mean estimated bias reduction was 0.005, 0.000, 0.020, 0.012 standard deviations for demographics, trauma incidence, psychological disorder incidence, and regression coefficients predicting Dx, respectively
- At our n=500, breakeven DEFFs would have 1.02, 1.00, 1.21, and 1.07 in terms of MSE
- So weights with DEFF of 1.26 would worsen MSE across the board- do not use?
- Appears that errors were mainly mobility between rating and interview (and mobility unimportant for parameter estimates).

Hypothesis Tests of Whether Weights Reduce Bias

- Test significance (and magnitude) of correlation of weights with outcomes
- Test interaction terms between independent variables and weights in regression
- Null Hypothesis is that weights do not reduce bias, p<0.05 suggests that they may



Smoothing Weights



Smoothing weights

- Smoothing weights refers to various techniques for reducing their variance and hence their design effect.

- A first step with compound weights (such as overall weights that a product of design weights and non-response weights) is to compute design effects separately for the two parts to understand where most of the overall design effect is coming from
  - If non-response weights or some other estimated weights are a major source of design effect, smoothing should be considered here first
  - If non-estimated design weights are the major source of design effects, there are several options
    - Redesign survey more proportionately for next round
    - Aggregate design weight calculations over less important factors
    - Use empirical smoothing techniques

Smoothing non-response weights: Fixing The Model

- Aggregate to higher level strata (e.g., practices rather than physicians, annual rather than quarterly)

- Drop non-significant terms (or even significant terms with small standardized odds ratios) from logistic regression models of nonresponse; ignore higher-order interaction terms

- Shrink out sampling error in non-response weights if sample sizes/bins too small

Empirical smoothing techniques

- Capping: After standardizing weights to mean 1, limit maximum value to a = 5 or 10, then re-standardize weights

- Shrinkage: After standardizing weights to mean 1, let new weight =

- (1-a) * old weight + a * 1, 0 < a < 1
  - larger values of a shrink more, multiplying DEFF-1 by $(1\text{-a})^2$
  - generally better than capping

- Example of shrinkage: 500 observations with weight = 0.2, 100 observations with weight=0.44, 24 observations with weight = 20.0; DEFF = 15.45, ESS=40

- With a=0.8, weights are 0.84, 0.888, 4.80; DEFF= 1.58; ESS=395

Shrinking Sampling Error

(Haviland & Elliott, in prep)

- **Design effect from sampling error alone is**

  $DEFF_0$=1+(1-p)/(p(n-1))

  if strata have size *n* per stratum and true response rate *p*

- **Suppose our original strata weights W_i have mean 1 and that we create new shrunken weights Z_i by shrinking the weights toward the mean of 1 in a linear combination where**

Z_i=a(1)+(1-a)W_i, 0<=a<=1.

- **Reduce design effect by amount corresponding to sampling error with**

a=1-sqrt((DEFF-$DEFF_o$)/(DEFF-1))

Smoothing design weights

- **Redesign survey more proportionately for next round**
  - **Best approach because maximizes ESS and statistical power**
  - **Limited by subgroup analysis needs**
- **Aggregate design weight calculations over less important factors**
- **Use empirical smoothing techniques**
  - **Last choice, because weights don’t reproduce the population and you are knowingly leaving some bias present to control extremely variable weights.**

One approach for Evaluating Weights

- **If DEFF is small and sample size is large, weights are probably worth it and no further examination is needed**
- **Otherwise**
  - **Test Ho: that weights remove no bias**
  - **If reject Ho, compare magnitude of estimated bias reduction and variance inflation.**
  - **If reject Ho and there is more bias reduction than variance inflation, weights are probably worth it**
  - **Otherwise**
    - **Try to reduce DEFF of weights by aggregating or shrinking, examining components of compound weights individually. Could solve to minimize MSE.**

Other considerations when evaluating weights

- **Evaluate weights for several key outcomes; look for a pattern**
- **Easiest to make one decision about weights for all outcomes**
- **Remember that smoothing weights or not weights can reduce the robustness to misspecification that comes with the design-based approach**
  - **But sometimes trivially, and at too high a price**

Sample Size, Power, Precision, and Confidence Intervals

- **What sample size do I need (for a certain amount of power/precision)?**

- **What can I say with a given sample size?**

- **What matters and how much?**
  - **Proportions vs. Means**
  - **Effects of N, confidence level, significance level, power level, allocation/balance, pairing**

RAND

Power Versus Precision

- **Power has to do with the ability to detect differences of a given magnitude in hypothesis testing**

- **Precision refers to the amount of variability present in point estimates in estimation**

- **Precision is a more basic concept, if we understand it, we understand power**
  - **The power you have is largely determined by the precision you have**

RAND

Sample Size and Precision

Width of a confidence interval is inversely related to the square root of the sample size

For a CI that is 1/3 as wide, multiply n by 9

1/2	4
2/3	2.25
3/4	1.78

Cutting sample size by 10% multiplies CI width by 1.05

20%	1.12
30%	1.20
40%	1.29
50%	1.41

Precision/Power "Worse" for Proportions

- Standard deviations are "large" for proportions
- 50% at 50%
- 40% at 20% / 80%
- 30% at 10% / 90%
- "Small" / "Medium" / "Large" effect sizes are 0.2 / 0.5 / 0.8 standard deviations (Cohen)
- These are 6-10% / 15-25% / 24-40% for proportions in the 10-90% range

Precision and p

- Precision is greatest when the proportion p is near 0 or 1 and least when p is near 0.5 in terms of percentages points, but NOT relative to p
- Let n be the sample size required for a CI of width E when p=.25 and of width ap when p=0.25

p	Obs for CI width E	Obs for CI Width ap
.05	0.25n	6.33n
.1	0.48n	3.00n
.2	0.85n	1.33n
.25	1.00n	1.00n
.3	1.12n	0.78n
.4	1.28n	0.50n
.5	1.33n	0.33n

Estimating Differences in Means Requires Larger Ns

- Estimates of differences are much less precise than estimates of a single proportion or mean
- If we are estimating a single mean and have a CI of width E with sample size n -- to obtain a CI of width E for the difference of that mean and the mean of another population with the same SD, we need a total of 4n observations (2n for each of the two groups)
- Differences-of-differences (and 2-way interactions) are even worse: 16n= 4n in each of 4 cells needed

Equal Sample Allocation Helps

- Assuming equal SDs, precision is least when the sample is allocated unevenly among the two populations
- Let a sample size n*=n1+n2 result in a CI of width E when n1=n2

r=n2/n1	Total sample size needed for a difference in mean CI of width E	Sample in rare group
1	1.00n*	0.50n*
1.5	1.04n*	0.42n*
2	1.12n*	0.38n*
3	1.33n*	0.33n*
4	1.58n*	0.32n*
5	1.80n*	0.30n*
10	3.02n*	0.27n*

Small Departures from Balance Don't Cost Much

- Let d=|p-0.5|, where p is the proportion allocated to group 1 of 2.
- Proportionate loss in effective sample size compared to perfect balance is 4d^2
  - 4% for a .6/4 split
  - 16% for a .7/.3 split
- "Flat" function near .5/.5

50/50 "Coin Flips" Provide Good Power for n>100

(Elliott, McCaffrey, & Lockwood, 2007)

- CLT and flatness work together in simple randomization
  - Imbalance is both rare and rarely consequential
- Expected power loss is equal to the loss of ONE OBSERVATION
- 5% chance of loss of power equivalent to 4 or more observations
- Stratified randomization or blocking not needed except for small samples

Greater Confidence Levels are Costly

- A 99% CI is 32% wider than a 95% CI

  A 90% CI is 84% as wide as a 95% CI

  An 80% CI is 65% as wide as a 95% CI
- To have the same width as a given 95% CI with sample size n you need

  1.73n for a 99% CI

  0.71n for a 90% CI

  0.43n for an 80% CI

Power

Least Practically Significant Difference

- We can’t have good power against all alternatives
- Some alternatives might be *statistically* significant with large n, but not *practically* significant
- We typically want 80% (or 90%) power versus the *least practically significant difference*
  - This is the minimum power over the range of alternatives we care about.
  - Higher power often requires enormous *n:*
    - Compared to n for 80% power
      - 1.33n for 90%power; 1.66n for 95%, 2.46 n for 99%

Effect Sizes are like CI widths

- Effect size detectable is inversely related to the square root of the sample size
- To be able to detect an effect size 1/3 as large, multiply n by 9

1/2	4
2/3	2.25
3/4	1.78

- Cutting sample size by 10% multiplies CI width by 1.05

20%	1.12
30%	1.20
40%	1.29
50%	1.41

Power for percentage points, RR, and OR

- Like precision, power to detect a given percentage point change is greatest when the proportion p is near 0 or 1 and least when p is near 0.5
- Power to detect a given RR or OR *increases* as p rises from 0 to 0.5. Can oversample rare outcomes to ameliorate.
- Let n be the sample size required for 80% power to detect a change of *a* points, a RR of *b*, and an OR of *c* near p=.25

p	Obs for a point change	Obs for RR of b	Obs for OR of c
.05	0.25n	6.33n	10.16n
.1	0.48n	3.00n	4.32n
.2	0.85n	1.33n	1.52n
.25	1.00n	1.00n	1.00n
.3	1.12n	0.78n	0.68n
.4	1.28n	0.50n	0.32n
.5	1.33n	0.33n	0.15n

Pairing Helps A Lot With Continuous Outcomes

- Repeated measures, pre-/post-, propensity matching etc.
- Let r be the correlation of paired observations
- Paired data can achieve same power as unpaired data using only 1-r as many case.
  - E.g. if r=0.7, two-sample t-test with n=100 in each group equivalent to 30 pairs.

Pairing Helps Less With Dichotomous Outcomes

- R is Tetrachoric Correlation
- Table displays sample size needed for equivalent power in paired data
- Least gains for rare dichotomous data
- Alternatively, dichotomization hurts the most when data are paired

	Cont	Dich P=0.5	Dich P=0.1
R=0.0	A	B	C
R=0.2	0.80A	0.87B	0.92C
R=0.4	0.60A	0.74B	0.81C
R=0.6	0.40A	0.59B	0.68C
R=0.8	0.20A	0.41B	0.49C

Covariates can Increase (or Decrease) Power

- **If X2,…,Xk are correlated with the residual of Y|X1 , but are not correlated with X1, power improves ("cleaning the error term")**
- **If a=R^2 of X2,…, Xk with residuals of Y|X1 and X2,…,Xk are not correlated with X1, need only n(1-a) observations for the same power as n without covariates**
- **If X2,…,Xk are correlated with X1, the effect size can increase or decrease ("reducing confounding")**

Costs of lost N on power in p-value terms

Say that I had done power calculations for a given sample size, then lost q% of the sample.

What p-value under the original N becomes .05 now?

q	**Original p-value**
9%	**0.04**
18%	**0.03**
29%	**0.02**
42%	**0.01**
51%	**0.005**
65%	**0.001**
75%	**0.0001**

Increase in N needed for power to detect a true effect ns in a smaller pilot

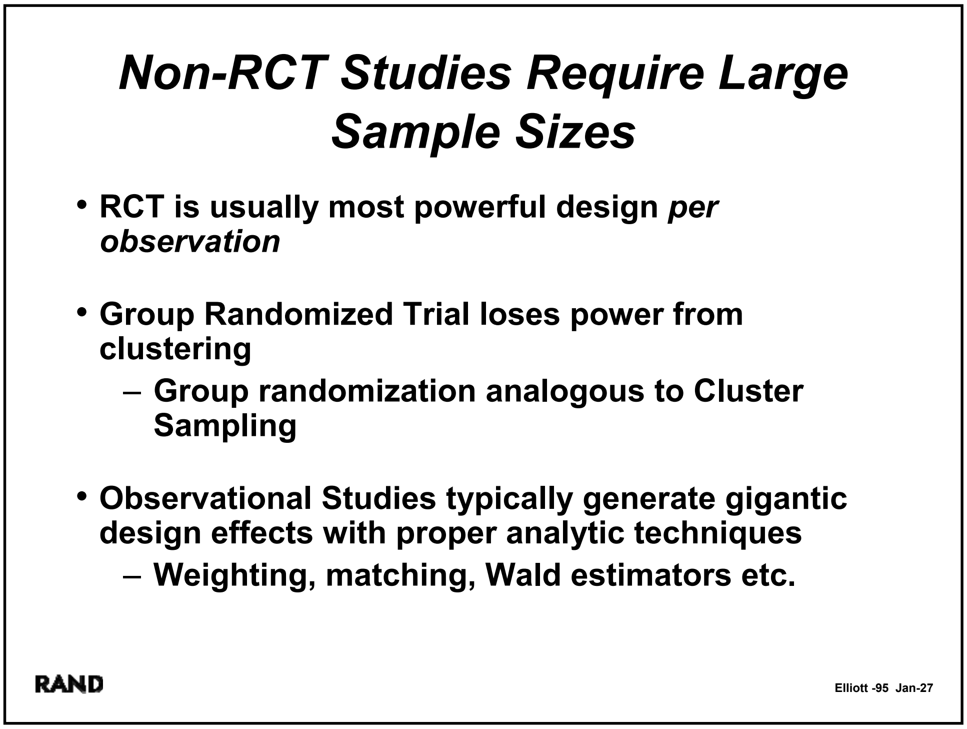
Suppose an effect in a small pilot is not significant at that N but is real.

By what % must we increase N to achieve 50% or 80% power at 0.05, 2-sided?

Pilot p-value	% add for 50% power	% add for 80% power
0.06	9%	122%
0.07	17%	139%
0.08	25%	156%
0.09	34%	173%
0.10	42%	190%
0.125	63%	233%
0.15	85%	278%
0.20	134%	377%

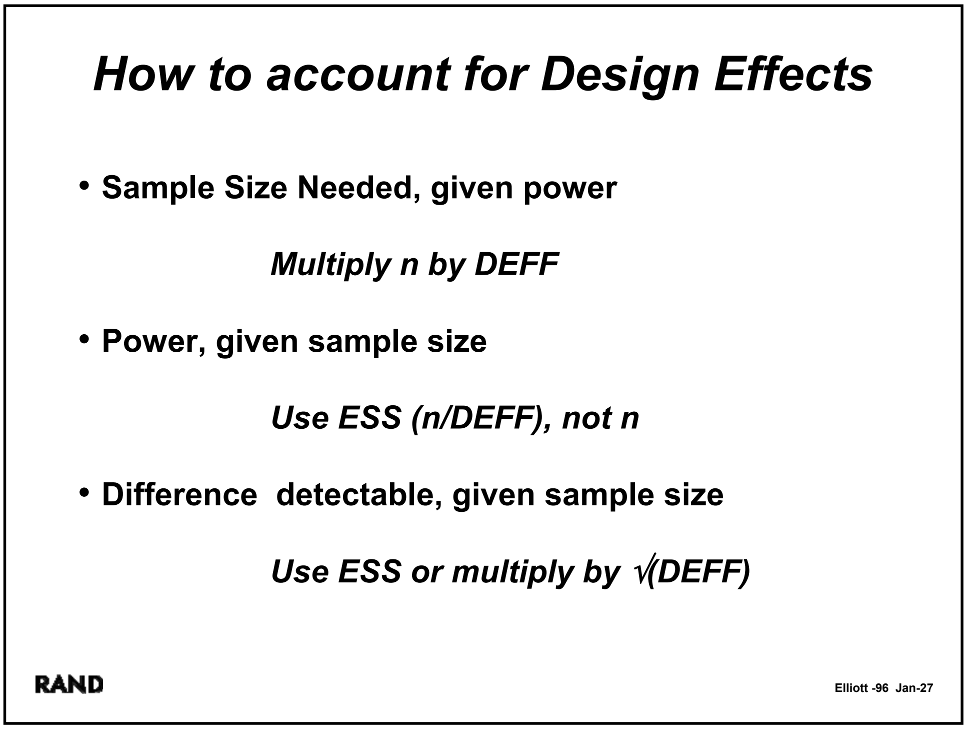
Other factors influencing power

- **Power decreases as Type I error rate decreases**
  - **In order to maintain 80% power while decreasing α from   .05 to .01 on a 2-sided test, sample size must be increased by 49%!**
  - **Slightly smaller effects at higher power levels; bigger effects for 1-sided tests**
- **Power is greater for 1-sided* than 2-sided tests**
  - **Switching from a 2-sided to a 1-sided test (in the correct direction) increases 70% power->80% and 80%->87.5%**
  - **A 1-sided* test requires 21% less sample size for 80% power at α =.05 than a 2-sided test**
  - **A 1-sided test in the wrong direction has virtually no power**



Non-RCT Studies Require Large Sample Sizes

- RCT is usually most powerful design *per observation*
- Group Randomized Trial loses power from clustering
  - Group randomization analogous to Cluster Sampling
- Observational Studies typically generate gigantic design effects with proper analytic techniques
  - Weighting, matching, Wald estimators etc.



How to account for Design Effects

- Sample Size Needed, given power

Multiply n by DEFF

- Power, given sample size

Use ESS (n/DEFF), not n

- Difference detectable, given sample size

Use ESS or multiply by √(DEFF)

References-Methods

- *Introduction to Survey Sampling*, Graham Kalton, 1983 (Sage)
- *Survey Sampling*, Leslie Kish, 1995 (Wiley)
- *Survey Methodology*, Robert Groves et. al. 2004 (Wiley)
- *Statistical Power Analysis for the Behavioral Sciences*, Jacob Cohen, 1988 (Lawrence Erlbaum)
- Rod Little's website http://www.sph.umich.edu/~rlittle/

References-Applications-1

- Elliott MN, Finch BK, Klein DJ, Ma S, Do P, Beckett MK, Orr N, Lurie N. (2008). "Sample Designs for Measuring the Health of Small Racial Ethnic Subgroups." *Statistics in Medicine*, 27 (20): 4016-4029.
- Elliott MN, Haviland A. (2007). "Use of a Web-based Convenience Sample to Supplement and Improve the Accuracy of a Probability Sample." *Survey Methodology*, 33(2): 211-215.
- Elliott MN, McCaffrey D, Lockwood JR. (2007). "How Important is Exact Balance in Treatment and Control Sample Sizes to Evaluations?" *Journal of Substance Abuse Treatment*, 33(1):107–110.
- Elliott MN, McCaffrey D, Perlman J, Marshall GN, Hambarsoomian K. (2009). "Use of Expert Ratings as Sampling Strata for a More Cost-Effective Probability Sample of a Rare Population." *Public Opinion Quarterly*, 73(1): 56-73.

References-Applications-2

- **Ghosh-Dastidar B, Elliott MN, Haviland A, Karoly L. (2009). “Composite Estimates from Incomplete and Complete Frames for Minimum-MSE Estimation in a Rare Population: An Application for Families with Young Children.” *Public Opinion Quarterly*, 10.1093/poq/nfp064*.***

- **Groves, R. M., & Peytcheya, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly, 72*(2), 167-189.**

- **Torgerson DJ. Contamination in trials: Is cluster randomisation the answer? *BMJ.* 2001;322:355-357.**