Kristen Himelein

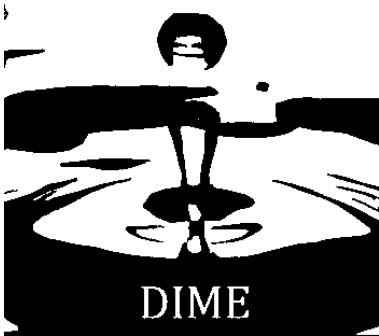
Practical Sampling for Impact Evaluations

Development Impact Evaluation

Field Coordinator Training

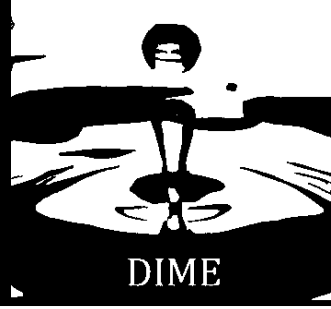
Washington, DC

April 22-25, 2013



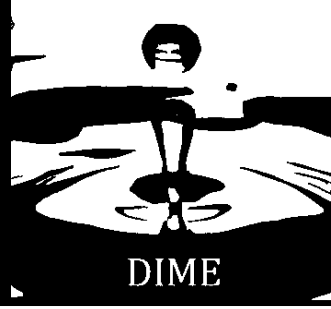
introduction

Introduction



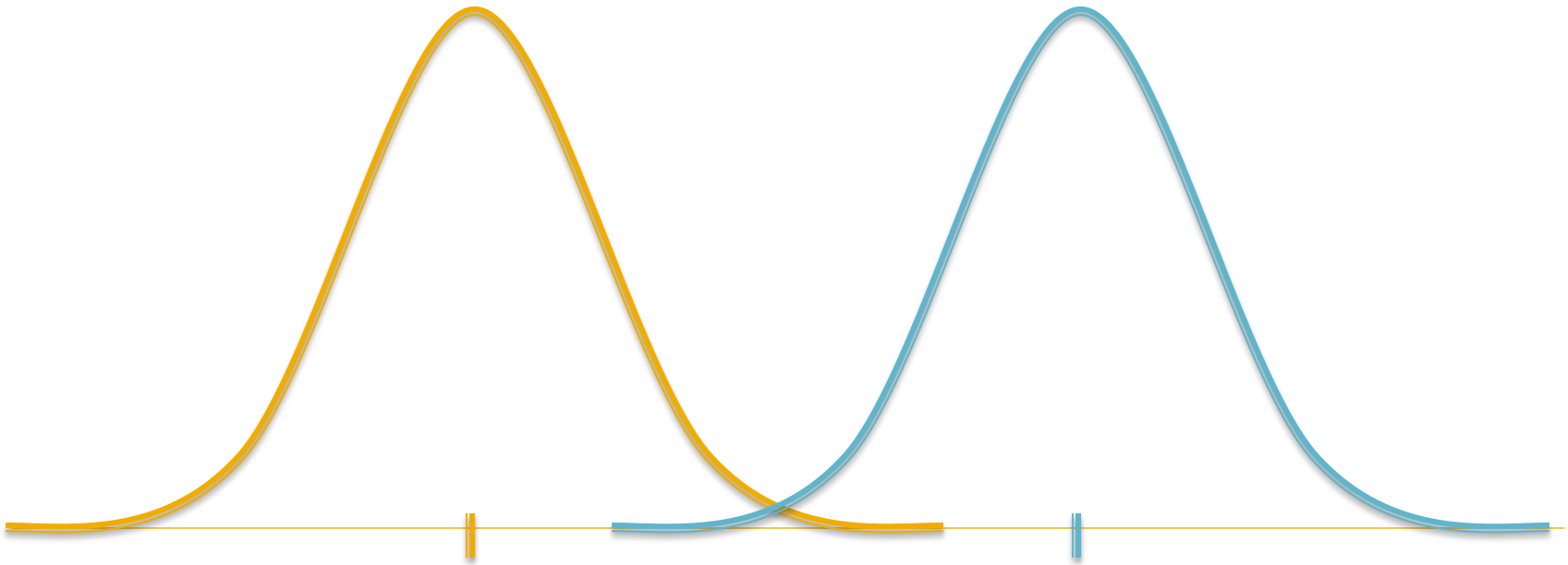
- This is a basic introduction to sampling for impact evaluation.
- Focuses mainly on sample size calculations for randomized cluster samples but basic ideas are transferrable to more complex randomized designs and non-random sample designs.
- **Main Question: how do we construct a sample to credibly detect a given effect size within our evaluation budget constraints?**

Purpose of Sampling

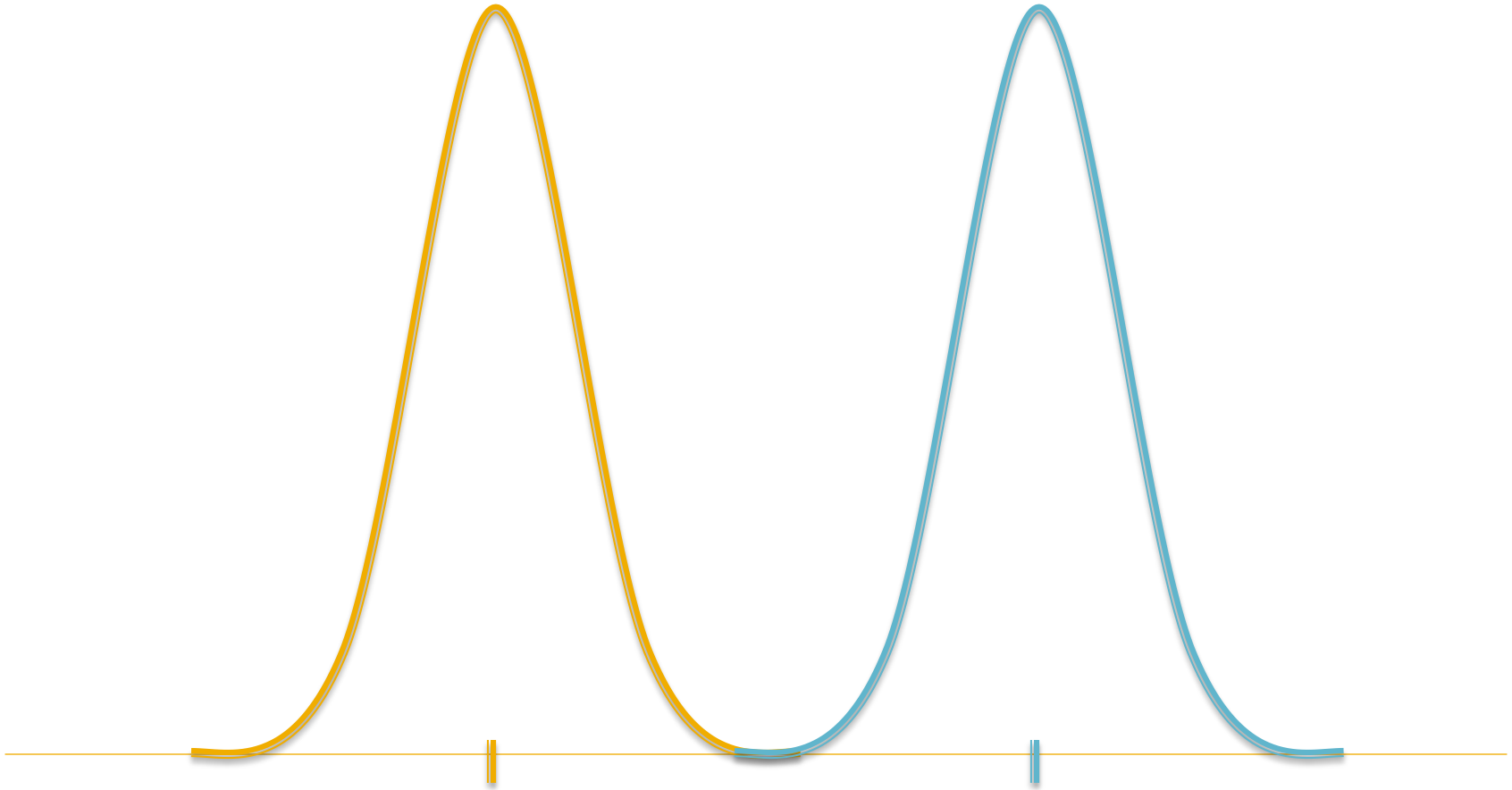


- Why sampling?
  - Saves **cost** compared to full enumeration
  - Easier to control **quality** of sample
  - More **timely** results from sample data
  - Measurement can be destructive

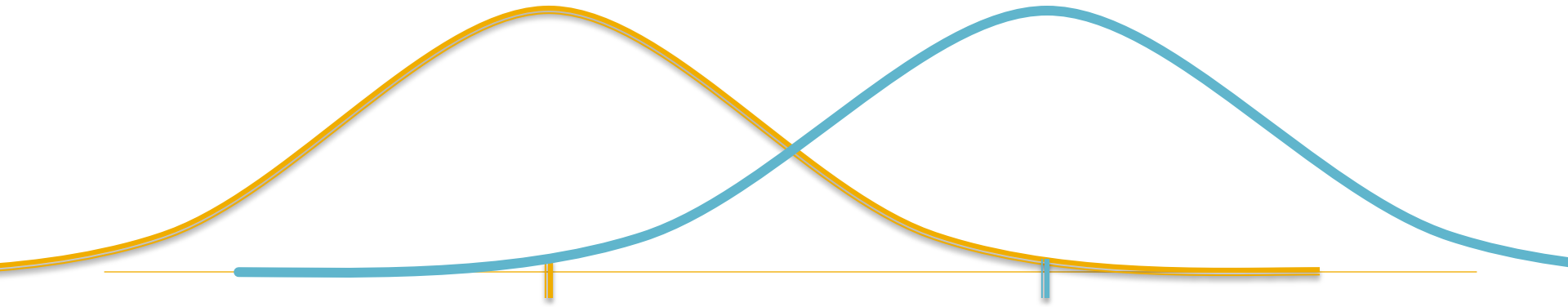
Sample Size



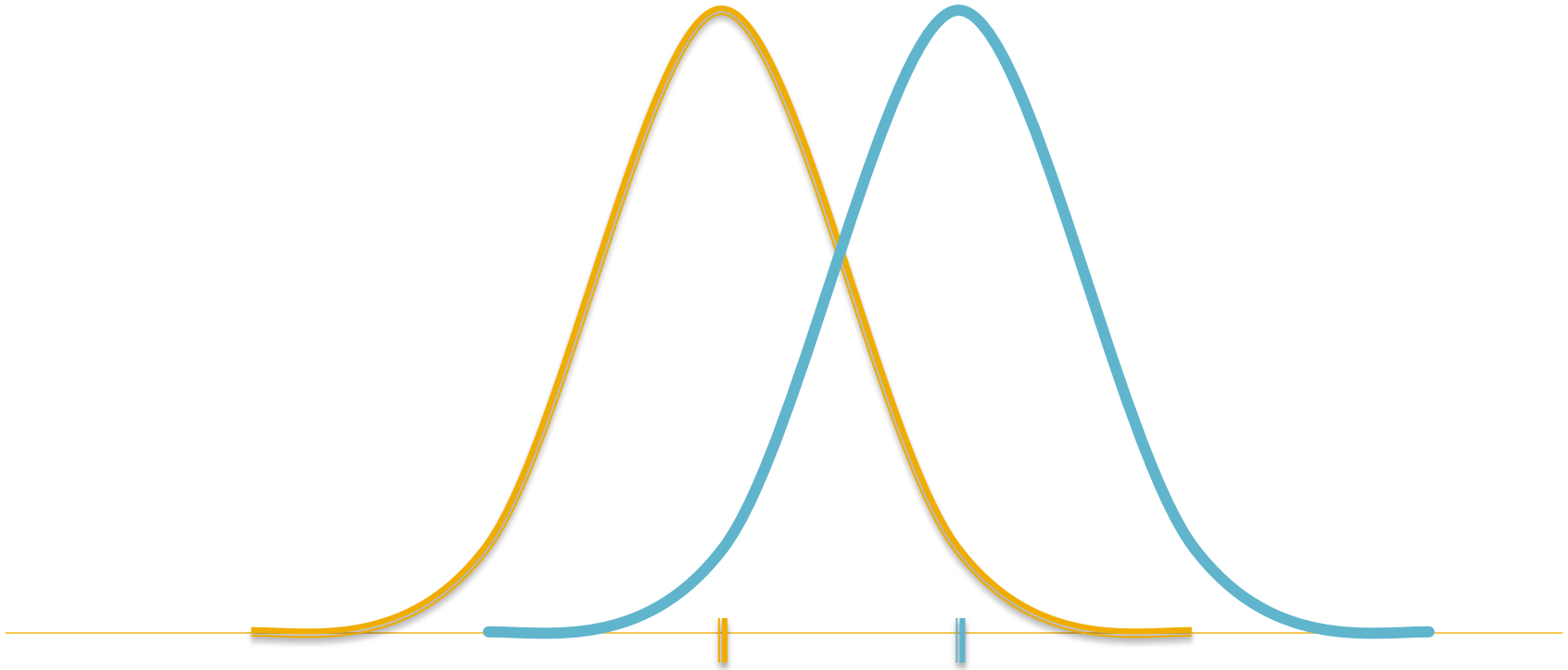
Sample Size



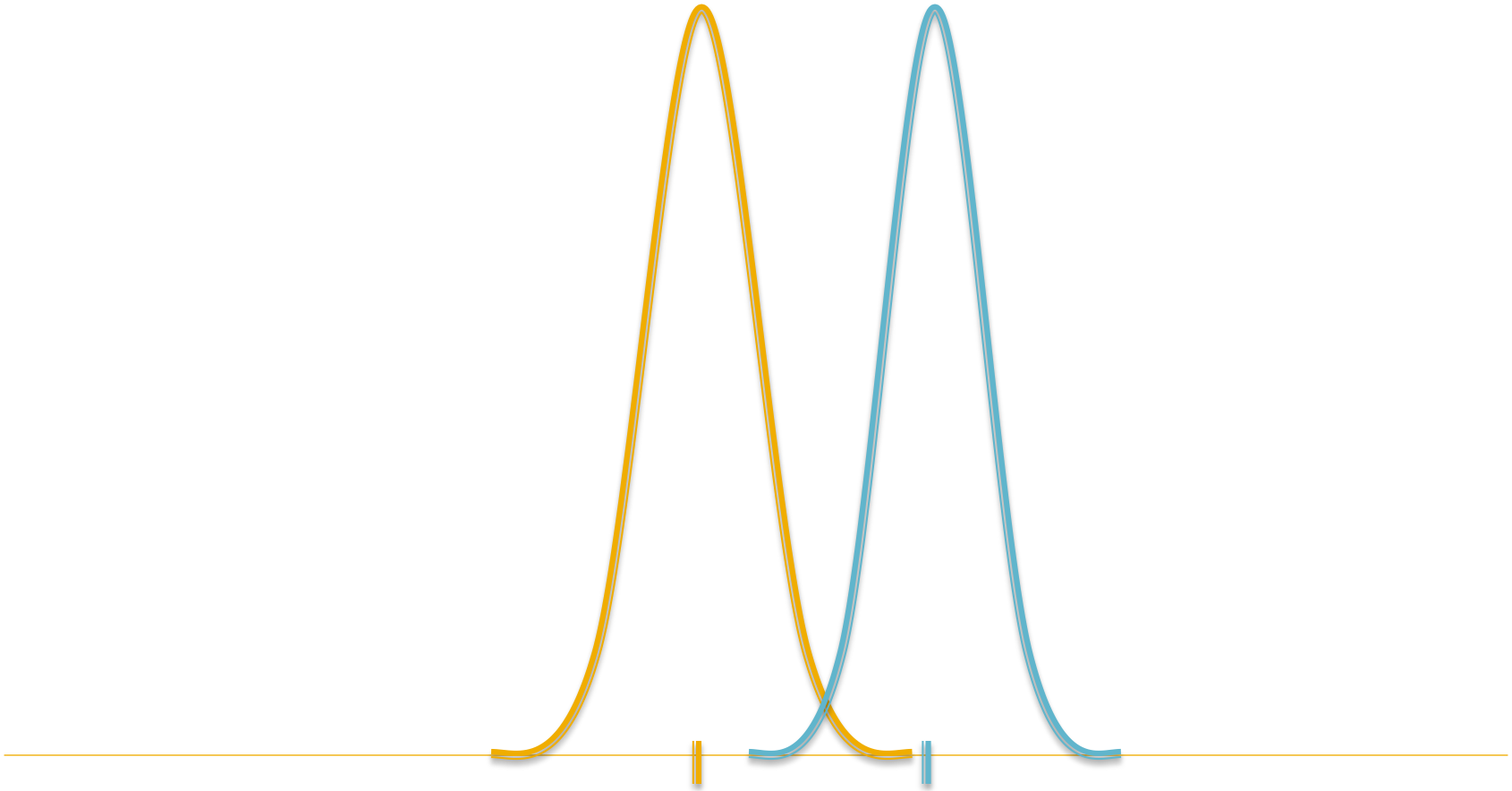
Sample Size



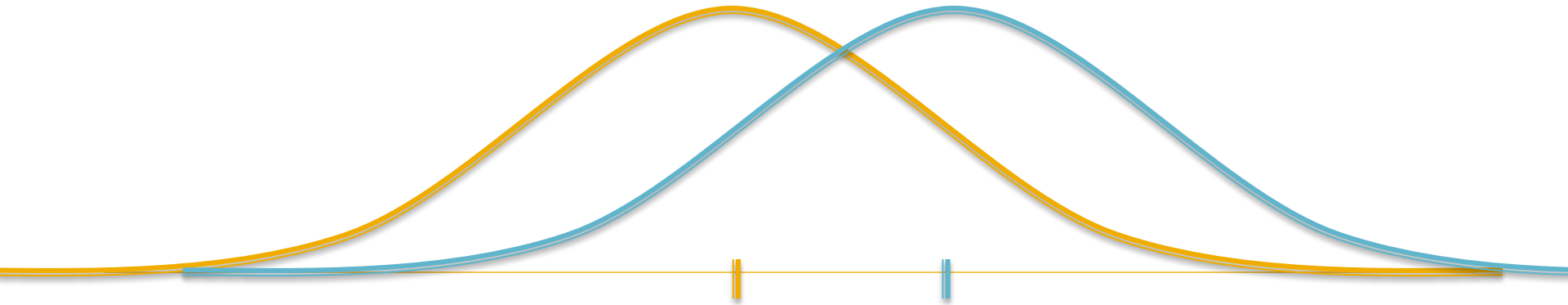
Sample Size



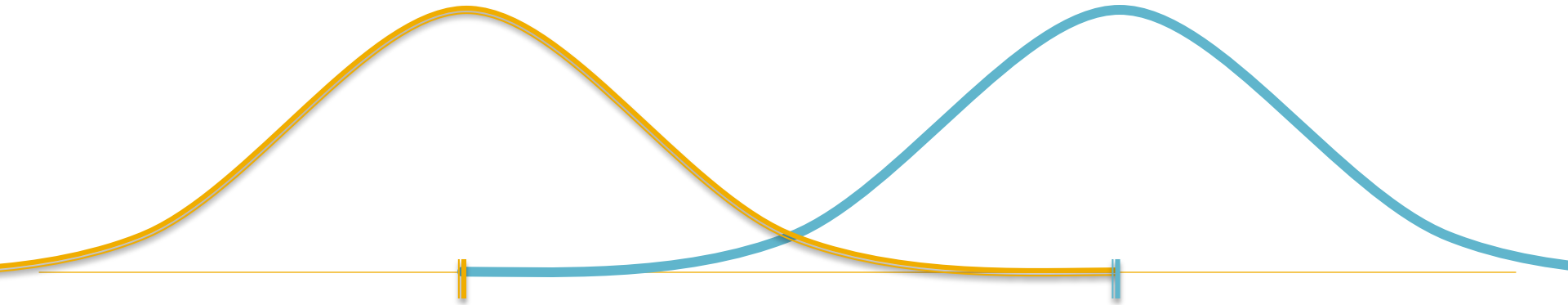
Sample Size



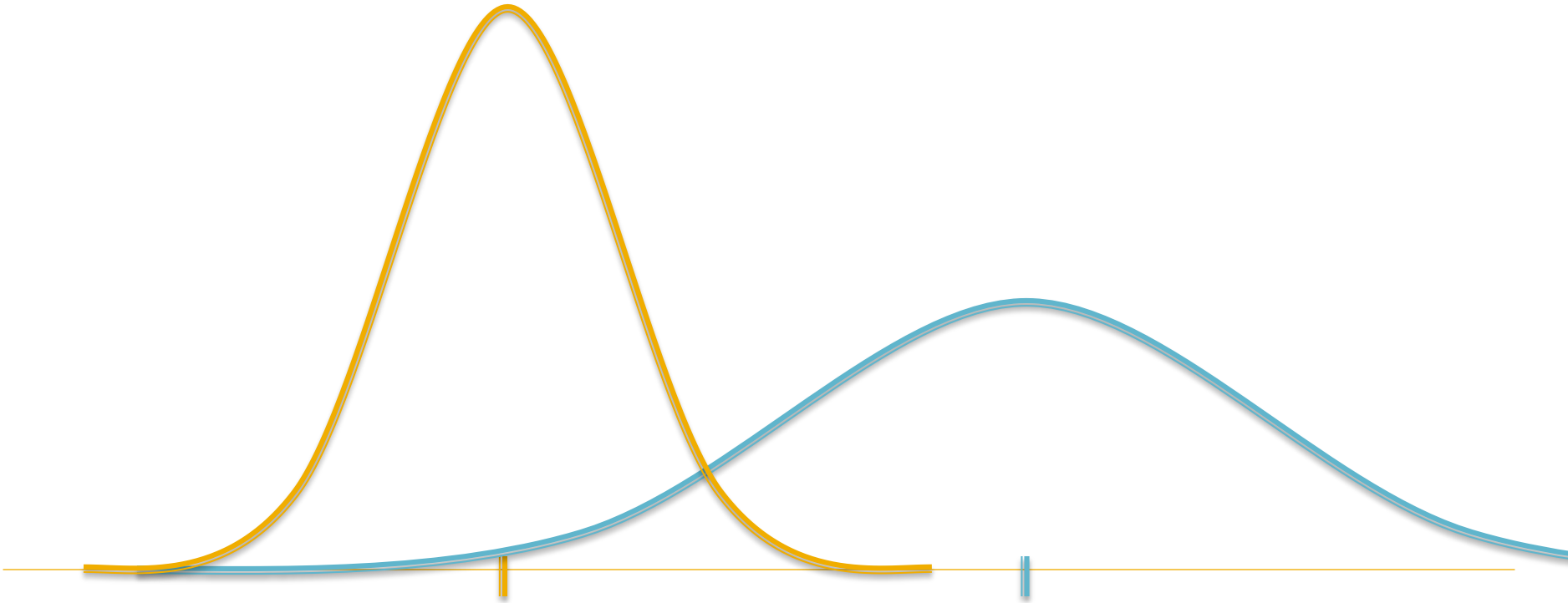
Sample Size



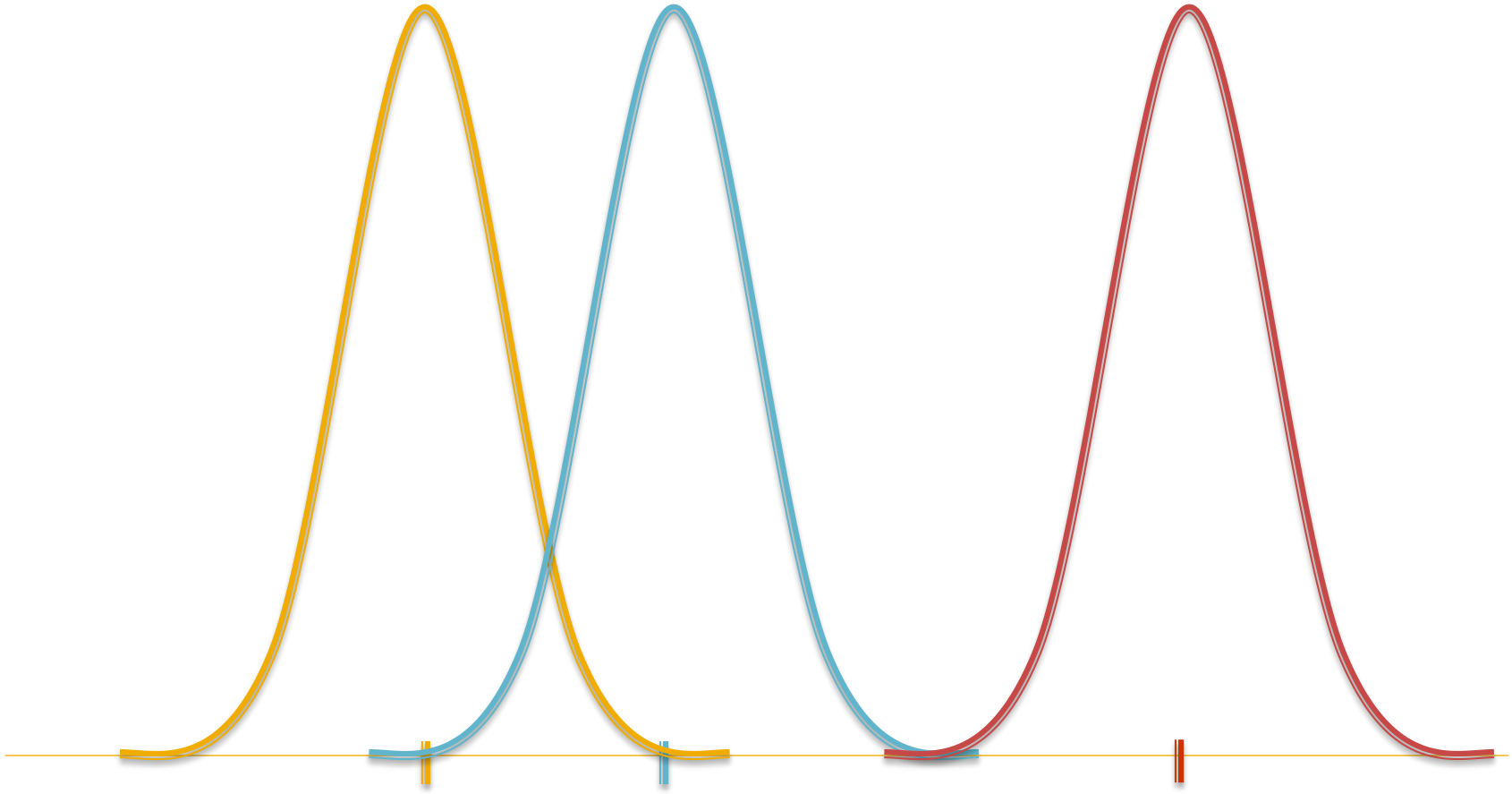
Sample Size



Sample Size

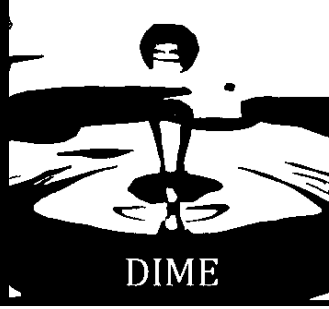


Sample Size



definitions

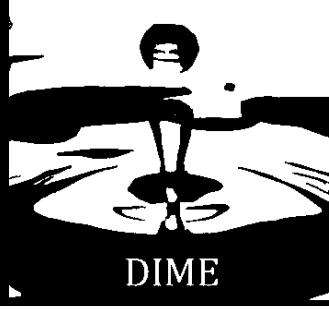
Sampling Concepts and Definitions



Unit of analysis

- The level at which a measurement is taken
- Most common units of analysis are persons, households, farms, and economic establishments

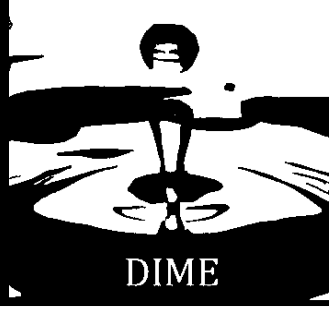
Sampling Concepts and Definitions



Target population or universe

- The complete collection of *all the units of analysis* to study.
- Examples: population living in households in a country; students in primary schools

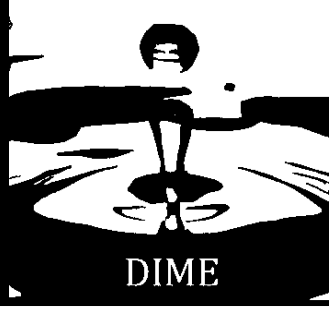
Sampling Concepts and Definitions



Sampling frame

- List of all the units of analysis whose characteristics are to be measured
- Comprehensive, non-overlapping and must not contain irrelevant elements
- Units must be identifiable (often linked to cartography)
- Should be updated to ensure complete coverage
- Examples: list of establishments; census; civil registration

Sampling Concepts and Definitions



Parameter / Estimate

- Objective of sampling is to estimate parameters of a population
- Quantity computed from all N values in a population set
- Typically, a descriptive measure of a population, such as mean, variance
  - Poverty rate, average income, etc.

Sampling Concepts and Definitions

Unbiased Estimator

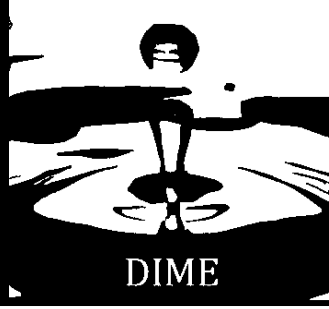
- **Estimator** - mathematical formula or function using sample results to produce an estimate for the entire population

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

- When the mean of individual sample estimates equals the population parameter, then the estimator is unbiased
- Formally, an estimator is unbiased if the expected value of the (sample) estimates is equal to the (population) parameter being estimated (where k is the number of experiments).

$$\frac{\hat{\theta}_1 + \hat{\theta}_2 + \ldots + \hat{\theta}_k}{k} \xrightarrow[k \to \infty]{} \theta$$

Sampling Concepts and Definitions



Standard Deviation (Population)

σ^2 = variance of the population

$\frac{\sigma}{\sqrt{N}}$ = standard deviation around the mean

Standard Error (Sample)

s^2 = variance of the sample

$\frac{s}{\sqrt{n}}$ = standard error

Difference: The standard deviation is a descriptive statistic. It is degree to which individuals in the population differ from the mean of the population. The standard error is an estimate of how close to the population mean your sample mean is likely to be.

Standard errors decrease with sample size. Standard deviations are left unchanged.

sample size calculations

Sample Size

- Starting point…

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right]$$

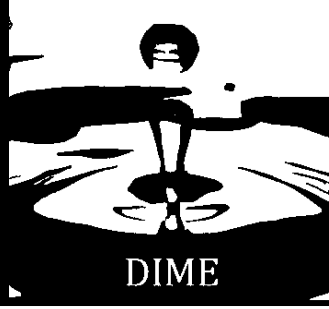
LSMS
Living Standards Measurement Study

Sample Size

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right]$$

- Sigma (σ) is the standard deviation in population outcome metric – σ^2 is the variance.
- Basically means how wide of a range of differences you expect in the outcome that you will measure.
- This can be difficult to calculate – the best way is if you have data collected previously (national household survey, project assessment, piloting data, etc).
- If not, estimations can be made using "(high-low)/4" as a rule of thumb.

Sample Size



$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right]$$

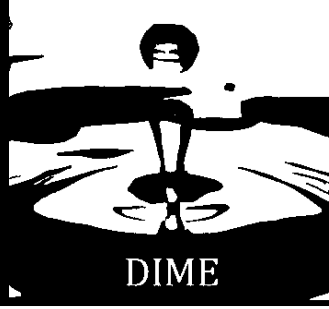
- D is the effect size or how much of an impact your project will have.
- Trade off between sample size and effect – the smaller an effect is the bigger a sample size that you will need.
- **Be careful about picking too big of an effect size as you are setting yourself up for failure.**

Sample Size

$$n = \left[\frac{4\sigma^2 \left(z_{1-\alpha/2} + z_{1-\beta} \right)^2}{D^2} \right]$$

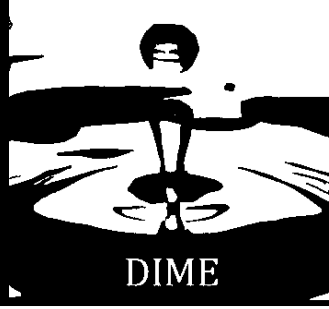
- Z's are from standard normal cumulative distribution function and they relate to the certainty of your conclusions.
- The values of z are taken from a table depending on the values of α and β.
- α relates to "type I error" and β relates to "type II error"

Type I Error (α)



- Significance level: Probability that you will falsely conclude that the program has an effect when in fact it does not.
- Type I error: Conclude that there is an effect, when in fact there are no effect.
- You select level of 5%, you can be 95% confident in the validity of your conclusion that the program had an effect
- For policy purpose, you want to be very confident of the answer you give: the level will be set fairly low.
- The more confident you want to be in your answer, the lower level you will need to select and the bigger your sample will need to be.

Type II Error (β)



- Power: Probability to find a significant effect if there truly is an effect
- Type II error: Fail to reject that the program had no effect when it fact it does have an effect
- Common values used are 80% or 90%.
- One minus the power is the probability to be disappointed. (So if you pick a power of 80%, there is a 20% chance that even though your project does have an impact, the evaluation will fail to detect it.)
- The more power you want your test to have, the larger a sample size you will need.

Sample Size

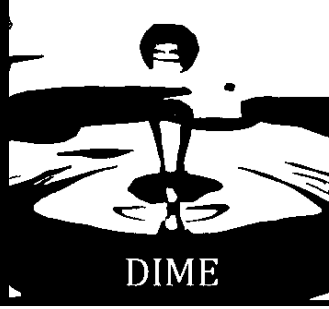
$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right]$$

- Final note: Beware the square!
- There is not a 1 to 1 relationship between sample size and most of the terms that are used to calculate it.
- So halving the size of the effect that you are looking for will raise required sample size by 4 times.

LSMS
Living Standards Measurement Study

clustering

Clustering



- In the real world, it is almost impossible to do a SRS due to cost and logistics limitations.
- Therefore almost all surveys are clustered.
- Clustering requires that intermediary stages are picked in advance of the final units
  - Households within a village
  - Schools within a district
  - Members within a household
- You can have one or more layers to your clustering design, but there are implications for your statistical power.

Clustering

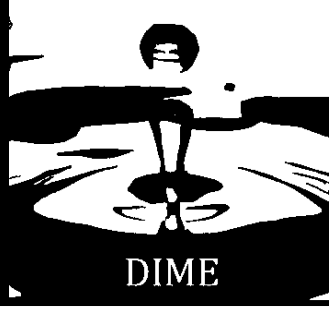
$$e^2_{TSS} = e^2_{SRS} \times ceff = e^2_{SRS} \times \left[1 + \rho(m-1)\right]$$

where:

ρ = intraclass correlation coefficient – measure of homogeneity within a cluster

m = number of units per cluster

Clustering



- The cluster effect increases with the intraclass correlation coefficient (ρ) and the number of sampling units per cluster
- The intraclass correlation coefficient is
  - Very high (> 0.2) for variables of infrastructure
  - High (~ 0.05) for socioeconomic variables
  - Low (< 0.02) for demographic variables
- Typical number of households per cluster:
  - 8 to 10 sample households for socioeconomic and LSMS surveys
  - 20 to 25 households for Demographic Surveys

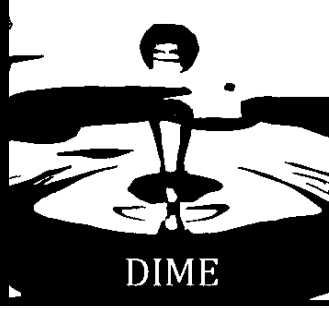
Clustering

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] [1 + \rho(m-1)]$$

- This part of the equation relates to how many clusters and households you select into your sample.
- Rho (ρ) is the intracluster correlation effect. This is a measure of how similar your observations within each PSU tend to be.
- m is the number of observations in each cluster.
- The more similar households are to each other and the more households you have in each cluster, the higher overall sample size you will need.

stratification

Stratification



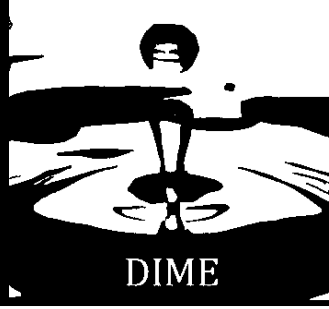
- The population is divided up into subgroups or "**strata**".
- A separate sample of units is then selected from each stratum.
- There are two primary reasons for using a stratified sampling design:
  - To potentially reduce sampling error by gaining greater control over the composition of the sample.
  - To ensure that particular groups within a population are adequately represented in the sample.

These objectives are often contradictory in practice.

- The sampling fraction generally varies across strata.

Sampling weights need to be used to analyze the data.

Sample allocation under stratified sampling



- Three major types of sample allocation of sample units among the strata:
  - Proportional allocation
  - Equal allocation
  - "Optimum" allocation

Proportional allocation

- The sample allocated to each stratum is proportionally to the number of units in the frame for the stratum:

$$n_h = n \times \frac{N_h}{N}$$

- Simplest form of sample allocation
- Provides self-weighting sample
- Efficient sample design for national-level results when variability is similar for the different strata

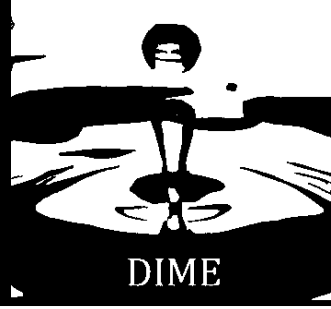
Equal allocation

- Each stratum is allocated an equal number of sample units:

$$n_h = \frac{n}{L}$$

- Used when same level of precision is required for each stratum
- Example: estimates of similar quality required for each region

Neyman allocation

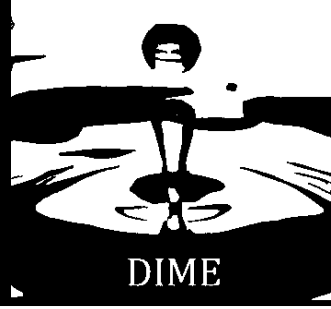


- Provides minimum total error and minimum cost for a fixed sample size

$$n_h = n \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}}$$

- S_h = standard deviation in stratum h
- c_h = cost per unit in stratum h

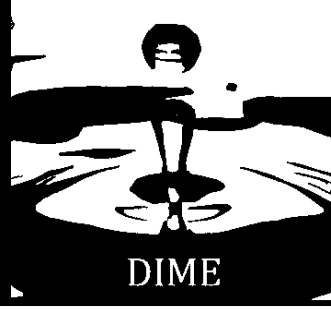
Practical allocation criteria



- For national household surveys, sometimes allocation is a compromise between proportional, equal and Neyman allocation; e.g. we start with a proportional allocation and then we increase the sample size in the smaller regions

- In countries with high proportion of rural population, sometimes a higher sampling rate is used for the urban stratum, to increase the urban sample size and because of the lower cost of data collection in urban areas

Second Stage Stratification



- Sometimes it is desirable to stratify the sample in the last stage (household or individual level)
- Examples: male/female headed households, program beneficiaries, households with OVCs.
- Beware of the DANGERS! Second stage stratification increases the need for close supervision of field teams

miscellaneous

Sample Size (for proportions)

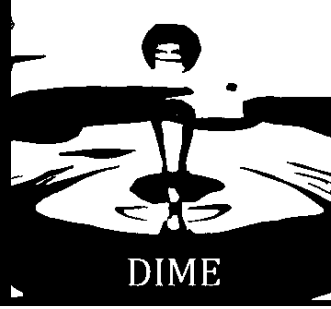
The previous formula was for continuous variables – for proportions use:

$$n = \frac{\left[\left(z_{1-\alpha/2}\sqrt{\bar{p}\bar{q}}\right) + \left(z_{1-\beta}\sqrt{p_1 q_1 + p_2 q_2}\right)\right]^2}{(p_1 - p_2)^2}\left[1+\rho(m-1)\right]$$

where

$$p = 1 - q \qquad \bar{p} = \frac{p_1 + p_2}{2} \qquad \bar{q} = 1 - \bar{p}$$

Multiple Treatment Arms



- Having multiple treatment arms in a program increases the required sample size quickly.
- The sample size calculations give you the total sample size for a two-arm evaluation. If you decide you want to add a third arm – you will need another 50% jump in the sample.

Finite Population Correction

- No where in the discussion of sample size calculation did the size of the population enter into the equation. This is because it largely does not matter.

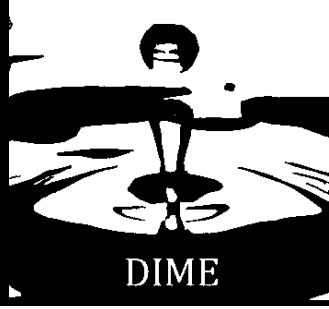
$$n_N = \frac{n_\infty}{1 + n_\infty / N}$$

Panel Surveys

- Panel surveys have higher power than repeated cross-sections for a given sample size. The factor in the equation to take that into account is the inter-wave correlation (r).

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] \left[1 + \rho(m-1) \right] \left[1 - r \right]$$

Unit Non-Response



- Nearly every survey has some degree of non-response for which we can make adjusts in the weights.
- It is important to note that this is a non-response adjustment, not a "correction." Without perfect information on all variables, it is not possible to completely correct for non-response.
- The best we can do is try to estimate the bias and make the best adjustment possible based on the information available.

Unit Non-Response

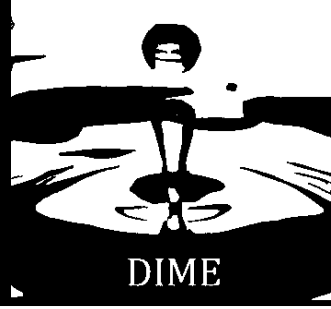
- Simplest form: If your cluster has 12 households, but one refuses, a simple calculation for the *nr* component of the weights would be $\left(\frac{11}{12}\right)^{-1}$ or 1.09.
  - Each remaining household counts a bit more than 1 to make up for its missing neighbor.
- Other common methods of adjustment for non-response:
  - Weighting class: divides the data into cells (such as age x gender x geography) and assigns a correction factor based on the cell response.
  - Propensity adjustment: uses a basic regression to model non-response (based on propensity score)

Post Stratification

- Post stratification is generally the last step in the process and adjusts weighted totals to known population totals and has been shown in the literature to reduce overall variance.
- Example:

State	Weighted Total Population from Survey	Known Population	Adjustment Factor
Maryland	5,245,757	5,699,478	1.0865
Virginia	8,475,901	7,882,590	0.9300
DC	662,842	599,657	0.9047

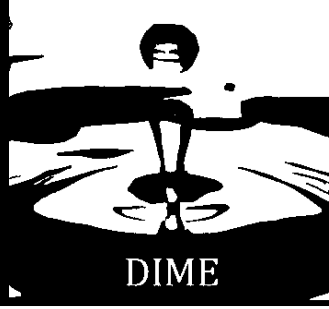
Raking



- Similar to post-stratification, most commonly used for non-response correction
- Example:

  As a break from the problems of the developing world, you decide to tackle the problems right here in the District of Columbia. As a baseline, you conduct a household survey to get a poverty measure. You, however, know from the literature that men and minority populations have low response rates – what to do?

Raking



- Weighted totals from your survey:

	White	African American	Latino/ Hispanic	Other	Total
Male	79,586	125,489	22,566	4,581	232,222
Female	97,089	185,057	22,689	5,422	310,757
Total	176,675	310,546	45,255	10,003	542,479

- Totals from census bureau

	White	African American	Latino/ Hispanic	Other	Total
Male					281,839
Female					317,818
Total	179,897	359,794	47,973	11,993	599,657

Raking

- Rake across:

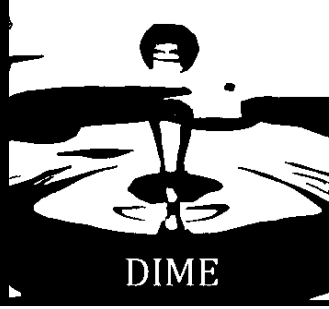
	White	African American	Latino/ Hispanic	Other	Total	
Male	96,591	152,301	27,388	5,560	281,839	1.214
Female	132,876	253,268	31,052	7,421	317,818	1.369
Total	229,466	405,570	58,440	12,980	599,657	

- Rake down:

	White	African American	Latino/ Hispanic	Other	Total
Male	0.784	0.887	0.821	0.924	
Female					
Total	179,897	359,794	47,973	11,993	599,657

- Repeat until factors converge to 1.

Raking



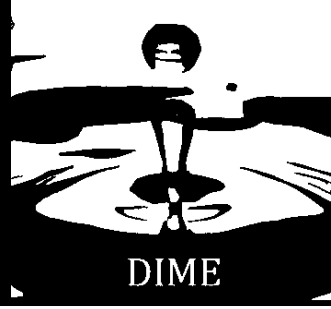
- After convergence:

	White	African American	Latino/ Hispanic	Other	Total
Male	88,927	160,865	26,028	6,019	281,839
Female	90,970	198,929	21,945	5,974	317,818
Total	179,897	359,794	47,973	11,993	599,657

- Divide sample totals by raking totals to find adjustment factors

	White	African American	Latino/ Hispanic	Other
Male	88,927/79,586	160,865/125,489	26,028/22,566	6,019/4,581
Female	90,970/97,089	198,929/185,057	21,945/22,689	5,974/5,422

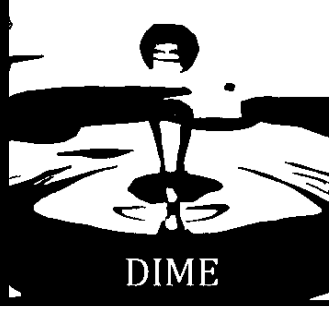
Raking



- These adjustment factors would then be used as a factor in the weight calculations.

	White	African American	Latino/ Hispanic	Other
Male	1.117	1.282	1.153	1.314
Female	0.937	1.075	0.967	1.102

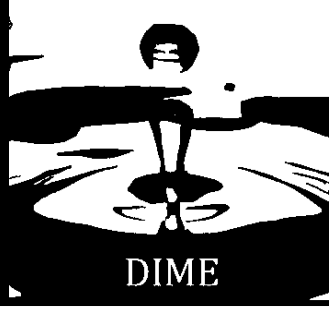
Non Random Sample Design



- Randomization in an impact evaluation is not always possible – may want to consider other designs such as Propensity Score Matching or Regression Discontinuity.
- In PSM, basic rule of thumb is to collect as many observations as possible to get best match for treatment.
  - See David McKenzies' blog from November 2011 for more details http://blogs.worldbank.org/impactevaluations/node/693
- In RDD, design effects dramatically increase sample size. Individual calculations necessary but can be estimated at roughly 3-4 times random sample.

conclusion

"Cheat Sheet"



Decrease Necessary Sample Size
Lower Variance
Bigger Effect Size
More Clusters
Panel Data
Increase Necessary Sample Size
Higher Confidence (α)
More Power (β)
Clusters More Similar
More Observations Per Cluster