“How to reject your most hated hypothesis”
or
Assuring enough Statistical Power

Mead Over, Center for Global Development
and
Sergio Bautista, INSP

Impact evaluation workshop
July 16, 2011
Outline

• Why are sampling and statistical power important to policymakers?
• Sampling and power for efficacy evaluation
• Sampling and power for effectiveness evaluation
• The impact of clustering on power and costs
• Conclusions
To researchers, statistical power is important because you need it to find a statistically significant impact, to publish your findings, to be cited by other researchers.

Finding an effect in your study group is enough.
Why are sampling and statistical power important to policymakers?

Because they are the tools you can use to reject the claims of skeptics.
What claims will skeptics make about MC rollout?

They might say:

• ‘Circumcision has no impact’
• ‘Circumcision has too little impact’
• ‘Intensive Circumcision Program has no more impact than Routine Circumcision Program’
• ‘Circumcision has no benefit for women’

Which of these do you hate the most?
So make sure the researchers design MC rollout so that you will have the evidence to reject your most hated hypothesis when it is false.

If it turns out to be true, you will get the news before the skeptics and can alter the program accordingly.
Hypotheses to reject

• Circumcision has no impact
• Circumcision has too little impact
• Intensive Circumcision Program has no more impact than Routine Circumcision Program
• Circumcision has no benefit for women
Efficacy Evaluation
Hypothesis to reject

• Circumcision has no impact
• Circumcision has too little impact
• Intensive Circumcision Program has no more impact than Routine Circumcision Program
• Circumcision has no benefit for women
Statistical power in the context of efficacy evaluation

• Objective: To reject the hypothesis of “no impact” in a relatively pure setting, where the intervention has the best chance of succeeding – to show “proof of concept”.

• In this context, statistical power can be loosely defined as:
  – Power: the probability that you find a benefit of male circumcision when there really is a benefit.
Confidence, power, and two types of mistakes

• Confidence describes the test’s ability to minimize type-I errors (false positives)
• Power describes the test’s ability to minimize type-II errors (false negatives)
• Convention is to be more concerned with type-I than type-II errors
  – (ie, more willing to mistakenly say that something didn’t work when it actually did, than to say that something worked when it actually didn’t)
• We usually want confidence to be 90 – 95%, but will settle for power of 80 – 90%
Power

- As power increases, the chances of saying “no impact” when in reality there is a positive impact, decline.
- Power analysis can be used to calculate the minimum sample size required to accept the outcome of a statistical test with a particular level of confidence.
The problem

Time of Experiment

Impact?

1 person, 1 year

Sample Size

All men in the country, 20 years

Impact!
The problem

• In principle, we would like:
  – The minimum sample size
  – The minimum observational time
  – The maximum power

• So we are “sufficiently” confident about the impact we find, at minimum cost
Things that increase power

• More person-years
  – More persons
  – More years

• Greater difference between control and treatment
  – Control group has large HIV incidence
  – Intervention greatly reduces HIV incidence

• Good cluster design: Get the most information out of each observed person-year
  – Increase number of clusters
  – Minimize intra-cluster correlation
Power is higher with larger incidence in the control group or greater effect.
Gaining Precision

Effectiveness (% reduction in HIV incidence)

Precision we got

Estimated Average

66

60

38

N in efficacy trial

Person-Years
With more person-years, we can narrow in to find the “real” effect.

- Effectiveness (% reduction in HIV incidence)

```
<table>
<thead>
<tr>
<th>Person-Years</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAL</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>60%</td>
</tr>
</tbody>
</table>
```

N in efficacy trial vs. Person-Years
The “real” effect might be higher than in efficacy trials ...

![Graph showing the comparison between real and real N in efficacy trials.](image)
... or the “real” effect might be lower

- Person-Years
- Effectiveness (% reduction in HIV incidence)

REAL

N in efficacy trial

Person-Years
Effectiveness Evaluation
Hypothesis to reject

• Circumcision has no impact
• Circumcision has too little impact
• Intensive Circumcision Program has no more impact than Routine Circumcision Program
• Circumcision has no benefit for women
What level of impact do you want to reject in an effectiveness study?

• For a national rollout, you want the impact to be a lot better than zero!
• What’s the minimum impact that your constituency will accept?
• What’s the minimum impact that will make the intervention cost-effective?
The “Male Circumcision Decisionmaker’s Tool” is available online at:

http://www.healthpolicyinitiative.com/index.cfm?id=software&get=MaleCircumcision
Cost and Impact of Male Circumcision

About the Male Circumcision Model
This model is intended to support policy development and planning for scaling up the provision of male circumcision services. It enables analysts and decisionmakers to understand the costs and impacts of policy options. It is a part of a larger tool kit developed by the United Nations Program on HIV/AIDS (UNAIDS) and World Health Organization (WHO) that provides guidelines on comprehensive approaches to male circumcision, including types of surgical

The key policy areas addressed by the model are the following:
1. Target populations
   - Adults, young adults, adolescents, newborns, most-at-risk groups (all male)
2. Target coverage levels and rates of scale-up
3. Service delivery modes
   - Service delivery modes (hospital, clinic, mobile van; public, private, nongovernmental organization, and "other")
4. Task shifting

The costing data should already be prepared in the companion costing spreadsheets. When opening this model, be sure to choose "Update" to update the costing data.

For more information, please contact:
Health Policy Initiative, Task Order 1
Futures Group
One Thomas Circle, NW, Suite 200
Washington, DC 20005 USA
Tel: (202) 775-9680
Fax: (202) 775-9694
http://www.healthpolicyinitiative.com
policyinfo@futuresgroup.com

USAID Office of HIV/AIDS
Technical Leadership and Research Division
Emmanuel F. Njeuhmeli, MD, MPH, MRA
Using the “MC Decisionmaker’s Tool,” let’s compare 60% effect ...

<table>
<thead>
<tr>
<th>Reduction in annual probability of infection when circumcised</th>
<th>Female to male</th>
</tr>
</thead>
<tbody>
<tr>
<td>- General population</td>
<td>60%</td>
</tr>
<tr>
<td>- Most-at-risk population</td>
<td>69%</td>
</tr>
<tr>
<td>- Other population 1</td>
<td>69%</td>
</tr>
<tr>
<td>- Other population 2</td>
<td>69%</td>
</tr>
<tr>
<td>- Other population 3</td>
<td>69%</td>
</tr>
</tbody>
</table>

Male to female: 0%

Sources:

Calculated force of infection for those not circumcised (For information only, no inputs required)

<table>
<thead>
<tr>
<th></th>
<th>Force of Infection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male 15-29</td>
<td>0.13</td>
</tr>
<tr>
<td>Female 15-29</td>
<td>2.05</td>
</tr>
<tr>
<td>Male 30-49</td>
<td>0.92</td>
</tr>
<tr>
<td>Female 30-49</td>
<td>2.27</td>
</tr>
</tbody>
</table>
Using the “MC Decisionmaker’s Tool,” let’s compare 60% effect ...

### Effectiveness of male circumcision

<table>
<thead>
<tr>
<th>Reduction in annual probability of infection when circumcised</th>
<th>60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female to male</td>
<td></td>
</tr>
<tr>
<td>General population</td>
<td>60%</td>
</tr>
<tr>
<td>Most-at-risk population</td>
<td>69%</td>
</tr>
<tr>
<td>Other population 1</td>
<td>69%</td>
</tr>
<tr>
<td>Other population 2</td>
<td>69%</td>
</tr>
<tr>
<td>Other population 3</td>
<td>69%</td>
</tr>
<tr>
<td>Male to female</td>
<td>0%</td>
</tr>
</tbody>
</table>

Suppose the effect is 60% ???

**Calculated force of infection for those not circumcised** (For information only, no inputs required)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Force of Infection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male 15-29</td>
<td>0.13</td>
</tr>
<tr>
<td>Female 15-29</td>
<td>2.05</td>
</tr>
<tr>
<td>Male 30-49</td>
<td>0.92</td>
</tr>
<tr>
<td>Female 30-49</td>
<td>2.27</td>
</tr>
</tbody>
</table>

   Protective effect = 60% (32%-77%)

   Protective effect = 55% (22%-75%)

   Protective effect = 55% (22%-75%)

   Protective effect among most-at-risk men = 71% (59%-80%)
To a 20% effect.

Effectiveness of male circumcision

<table>
<thead>
<tr>
<th>Reduction in annual probability of infection when circumcised</th>
<th>Male to female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female to male</td>
<td></td>
</tr>
<tr>
<td>General population</td>
<td>20%</td>
</tr>
<tr>
<td>Most-at-risk population</td>
<td>20%</td>
</tr>
<tr>
<td>Other population 1</td>
<td>20%</td>
</tr>
<tr>
<td>Other population 2</td>
<td>20%</td>
</tr>
<tr>
<td>Other population 3</td>
<td>20%</td>
</tr>
<tr>
<td>Male to female</td>
<td>0%</td>
</tr>
</tbody>
</table>

Sources:

Calculated force of infection for those not circumcised
(For information only, no inputs required)

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-29</td>
<td>0.13</td>
</tr>
<tr>
<td>15-29</td>
<td>2.05</td>
</tr>
<tr>
<td>30-49</td>
<td>0.92</td>
</tr>
<tr>
<td>30-49</td>
<td>2.27</td>
</tr>
</tbody>
</table>
To a 20 % effect.

Suppose the effect is only 20 % ???

**Effectiveness of male circumcision**

<table>
<thead>
<tr>
<th>Reduction in annual probability of infection when circumcised</th>
<th>Return to Menu</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female to male</strong></td>
<td></td>
</tr>
<tr>
<td>General population</td>
<td>20%</td>
</tr>
<tr>
<td>Most-at-risk population</td>
<td>20%</td>
</tr>
<tr>
<td>Other population 1</td>
<td>20%</td>
</tr>
<tr>
<td>Other population 2</td>
<td>20%</td>
</tr>
<tr>
<td>Other population 3</td>
<td>20%</td>
</tr>
<tr>
<td><strong>Male to female</strong></td>
<td>0%</td>
</tr>
</tbody>
</table>

| Calculated force of infection for those not circumcised      |                |
| (For information only, no inputs required)                   |                |
| Male 15-29                                                   | 0.13           |
| Female 15-29                                                 | 2.05           |
| Male 30-49                                                   | 0.92           |
| Female 30-49                                                 | 2.27           |

Less effective intervention means less reduction in incidence

60% Effectiveness

20% Effectiveness
Less effective intervention means less cost-effective.
Less effective intervention means less cost-effective

At 20% effectiveness, MC costs about $5,000 per HIV infection averted in example country.

60% Effectiveness

20% Effectiveness
Hypothesis to reject in effectiveness evaluation of MC

• Circumcision has no impact
• Circumcision has too little impact
• Intensive Circumcision Program has no more impact than Routine Circumcision Program
• Circumcision has no benefit for women
Additional challenges of effectiveness evaluation

• Investigate **determinants** of effectiveness
  – Supply side
  – Demand side

• Investigate impact on secondary outcomes and their determinants
  – Such as benefits to women, children

• Seek “external validity” on effectiveness issues
Sample size must be larger to show that the effect is at least 20 Person-Years.
Sample size has to be larger to distinguish impacts in two different settings.
Sampling methods for effectiveness evaluation

- **Probability sampling**
  - Simple random: each unit in the sampling frame has the same probability of being selected into the sample
  - Stratified: first divide the sampling frame into strata (groups, blocks), then do a simple random sample within each strata
  - Clustered: sample clusters of units. Eg. villages with all the persons that live there
    - One stage: Random sample of villages, then survey all men in selected villages
    - Two stage: Random sample of villages, then random sample of men in selected villages
Sampling (→ Representative data)

• Representative surveys
  – Goal: learning about an entire population
    • Ex. LSMS/ national household survey
      – Sample: representative of the national population

• Impact evaluation
  – Goal: measuring changes in key indicators for the target population that are caused by an intervention
    – In practice: measuring the difference in indicators between treatment and control groups
    – We sample strategically in order to have a representative sample in the treatment and control groups
    – Which is not necessarily the same as a representative sample of the national population
Cluster Sampling Design
Cluster Sampling

• In some situations, individual random samples are not feasible
  – When interventions are delivered at the facility/community level
  – When constructing a frame of the observation units may be difficult, expensive, or even impossible
    • Customers of a store
    • Birds in a region
  – When is of interest to identify community level impact
  – When budget constraints don’t allow it
Clustering and sample size

• Clustering reduces efficiency of the design
  – Standard sample size calculation for individual-based studies only accommodate for variation between individuals
  – In cluster studies, there are two components of variation
    • Variation among individuals within clusters
    • Variation in outcome between clusters
Clustering and sample size

• Individual-based studies assume independence of outcomes among individuals

• In cluster randomization:
  – Individuals within a cluster are more likely to be similar

• Measure of this intracluster dependence among individuals is intracluster correlation (ICC)
  – Based in within-cluster variance
    • High when individuals in cluster are more “similar”

• Not taking ICC into account may lead to underpowered study (too small sample)
Taking ICC into account

• In a cluster randomized design, in order to achieve the equivalent power of an individual random study, the sample size must be inflated by a factor called a “design effect”: 
  \[ Deff = 1 + (\tilde{n} - 1) \rho \]

• $\tilde{n}$ = average cluster size
• $\rho$ = ICC
• Assuming clusters of similar size
How big is the impact of cluster design on sample size

Effectiveness (% reduction in HIV incidence)

With 6 clusters, 20 cannot be rejected

With 12 clusters, 20 is rejected

At a given number of person-years

Person-Years
Increasing number of clusters vs increasing number of individuals per cluster

- Increasing the number of clusters has a much stronger effect on power and confidence
  - Intuitively, the sample is the number of units (clusters) at the level where the random assignment takes place. It is not the same as the number of people surveyed

- Challenge is to engineer the logistics to maximize the number of clusters, given budget
How big is the impact of cluster design on sample size
Conclusions

• Philosophy of sample design is different for efficacy and effectiveness studies
  – Efficacy: narrow & deep
  – Effectiveness: broad & shallow
• Many of the special requirements of effectiveness sampling will increase sample size
• Clustering reduces data collection costs, but at a sacrifice of power
• Survey costs also affected by:
  – Number of indicators collected
  – Number of non-index cases interviewed
• Most cost-effective way to reject your “hated hypothesis” is through randomized, efficiently powered, sampling