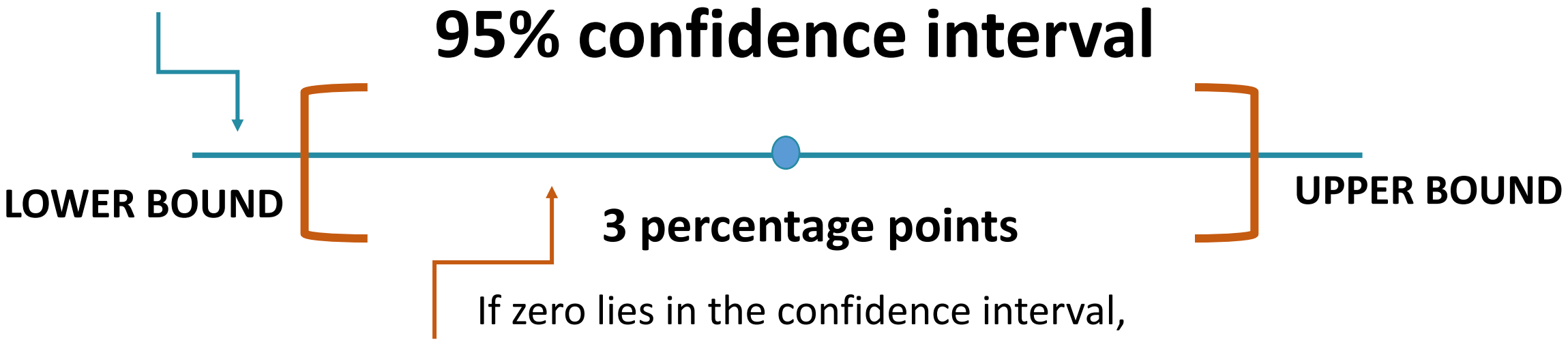
Sample size and statistical power

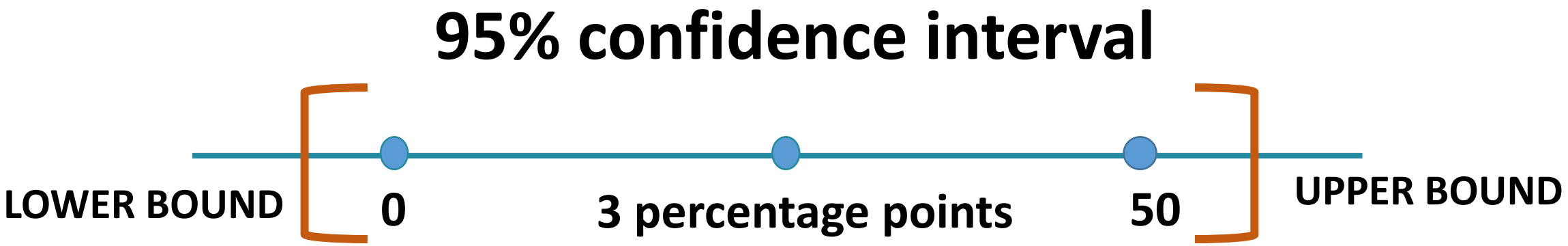
We accept error from randomness….

We know that we have no detectable impact if zero is in the confidence interval…

If zero lies outside the confidence interval, 0 and 3 can be statistically distinguished.



But how large of a confidence interval is ok?



We cannot distinguish a very large impact from zero impact.

We should minimize errors of inference

Type 1 error

You say there is a program impact when there really isn't one.

Can minimize this with choice of confidence interval (95%-99%) at time of data analysis

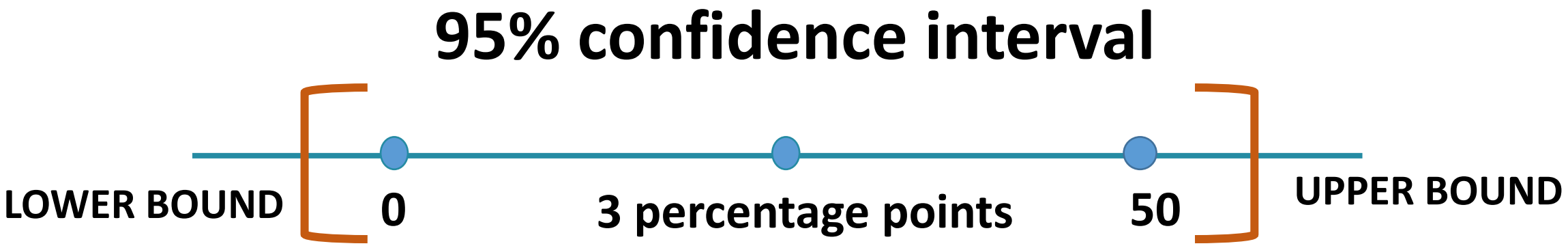
We should minimize errors of inference

Type 2 error

There is a program impact but you cannot detect it.

Can minimize this by thinking about statistical power while designing the impact evaluation.

Why are Type 2 errors so bad?



Why do an impact evaluation if you're going to learn nothing?

We cannot distinguish a very large impact from zero impact.

Agenda

1. What is statistical power?

2. How large of a sample size do I need?

3. How can I increase statistical power?

Defining statistical power

Type 2 error

There is a program impact but you cannot detect it.

Statistical power = 1 – Prob (Making a Type 2 error)

We don’t want to make Type 2 errors so we need to maximize statistical power.

What determines statistical power?

Sample size plays a large role

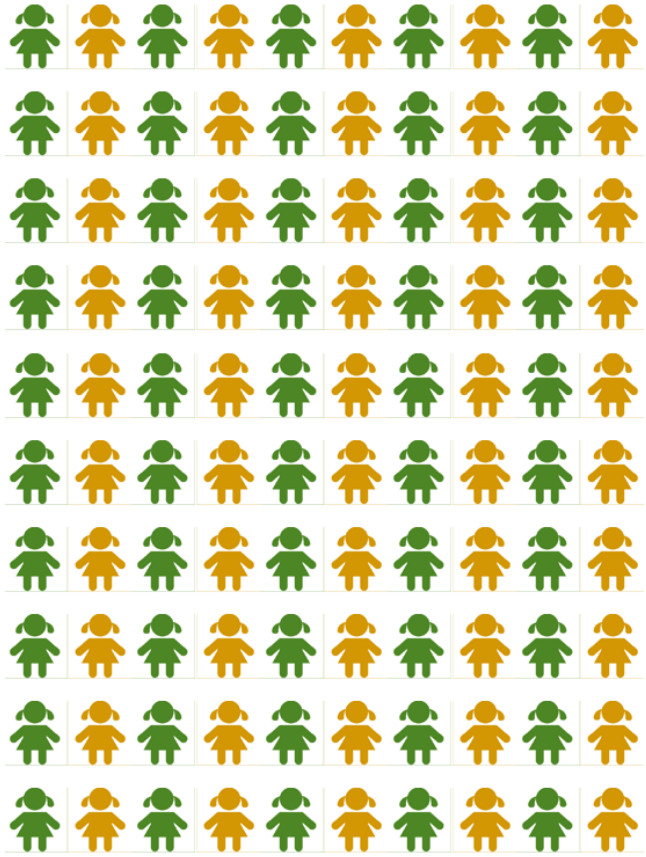
Why do we sample?

Usually we cannot gather data from entire target population.

Random sampling allows you to infer characteristics of the target group from the smaller sample.

Random sampling

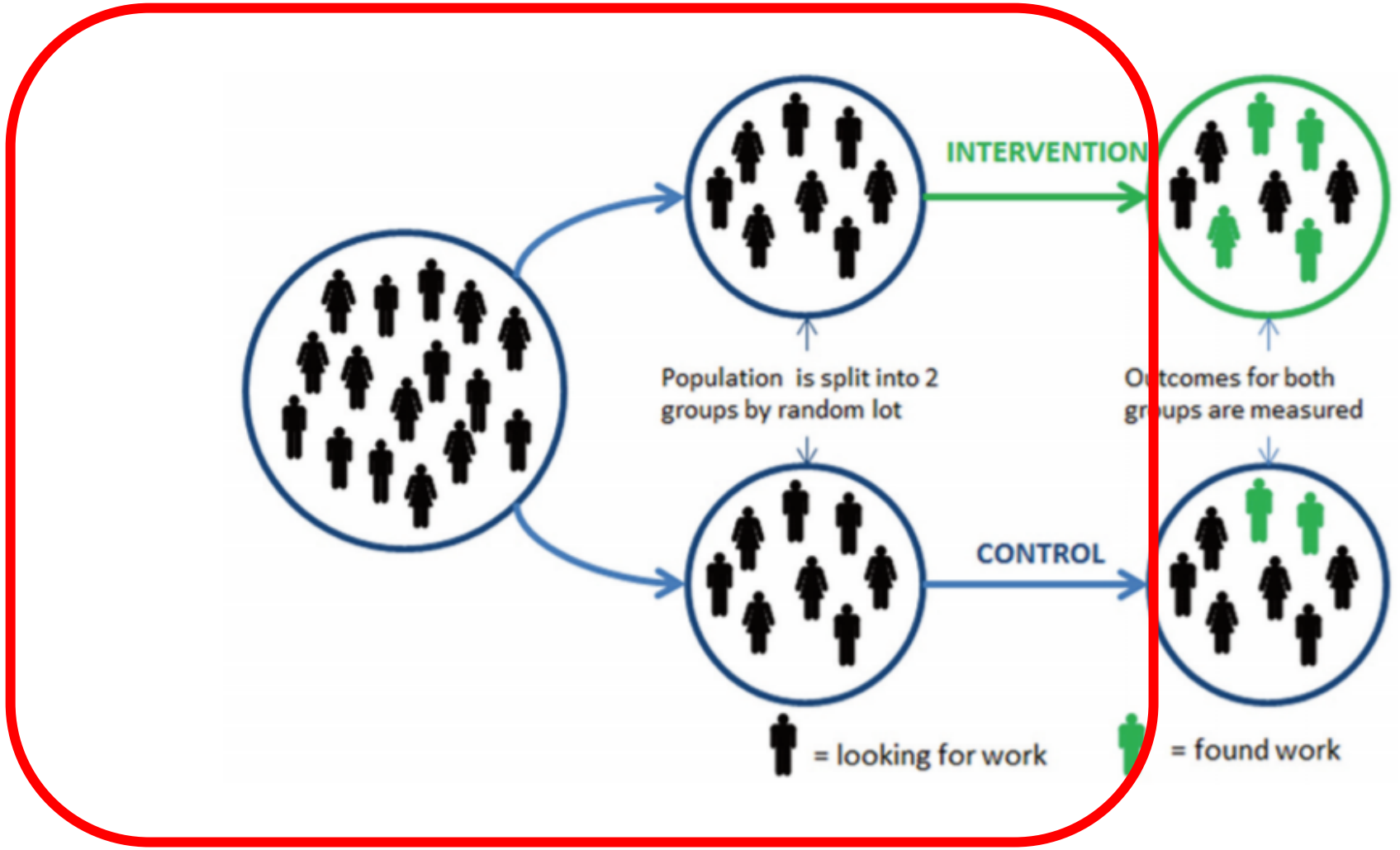
Target group



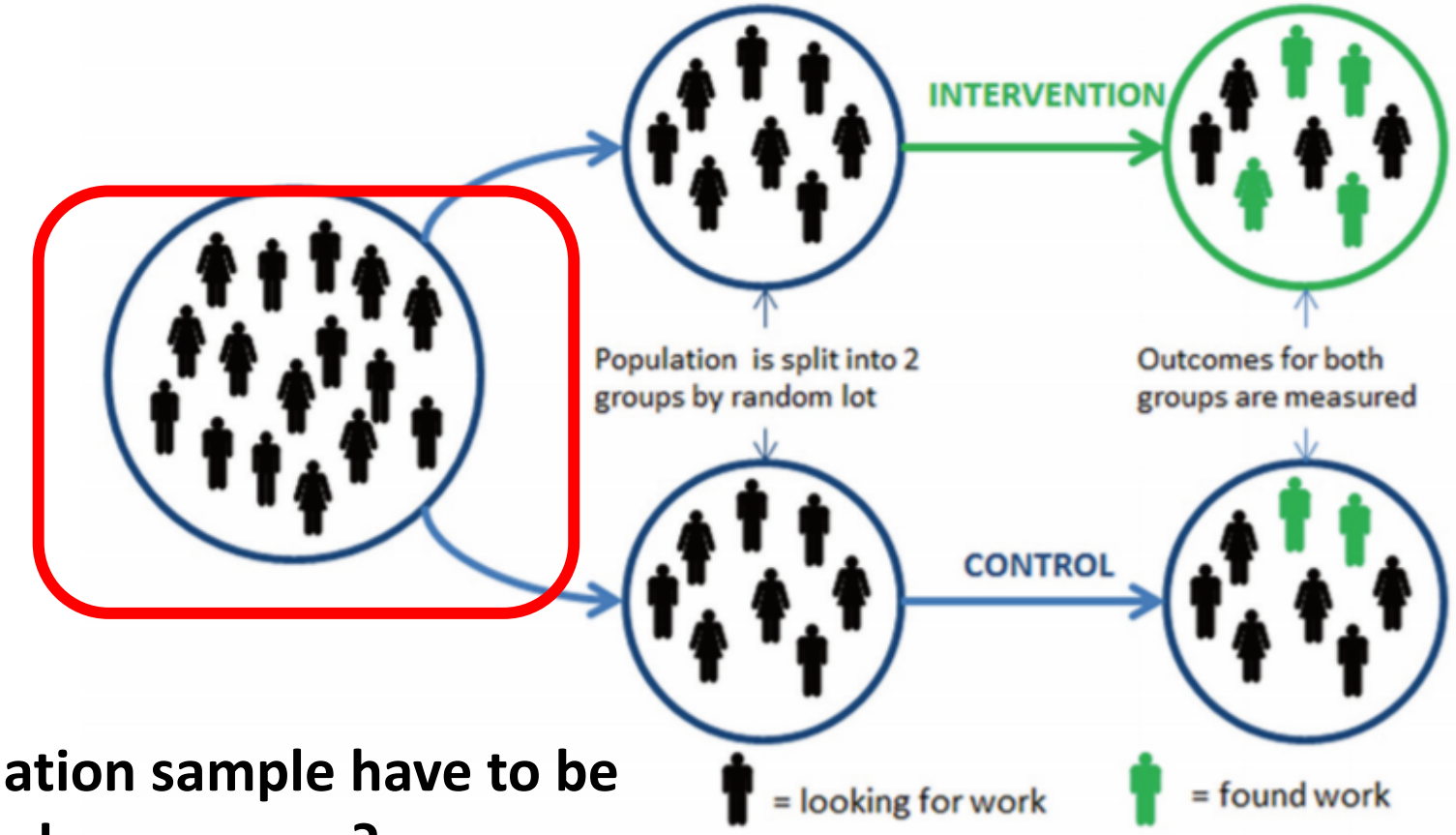
Sample

The sample has the same average characteristics as the target group.

Random assignment



Do we have to do any sampling for an impact evaluation?



Does the evaluation sample have to be sampled from a larger group?

Do we have to sample for an impact evaluation?

No. We sample when it's not feasible to collect data on a large group.

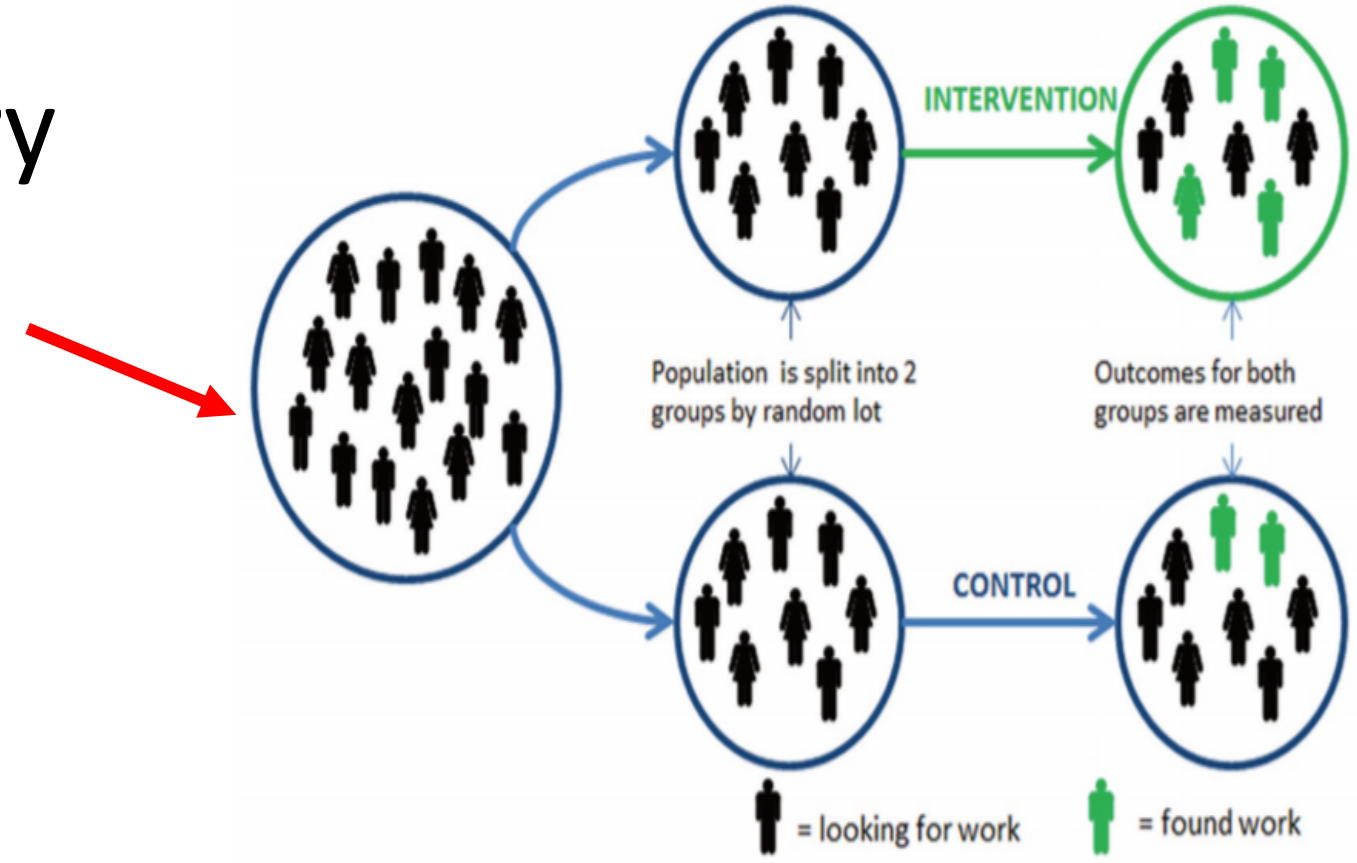
No need to sample if:

Evaluation sample is already small

Obtaining and analyzing data for all is not expensive

Not always necessary to sample, but.....

We need to worry about the size of this group



Why should we care about the size of the evaluation sample?

Let's think about something simpler than program impact.

What is the average height of women in Tanzania?

- Survey 1: Let's ask 5 people
- Survey 2: Let's ask 1,000 people

Which survey will give us an answer closer to the true average?

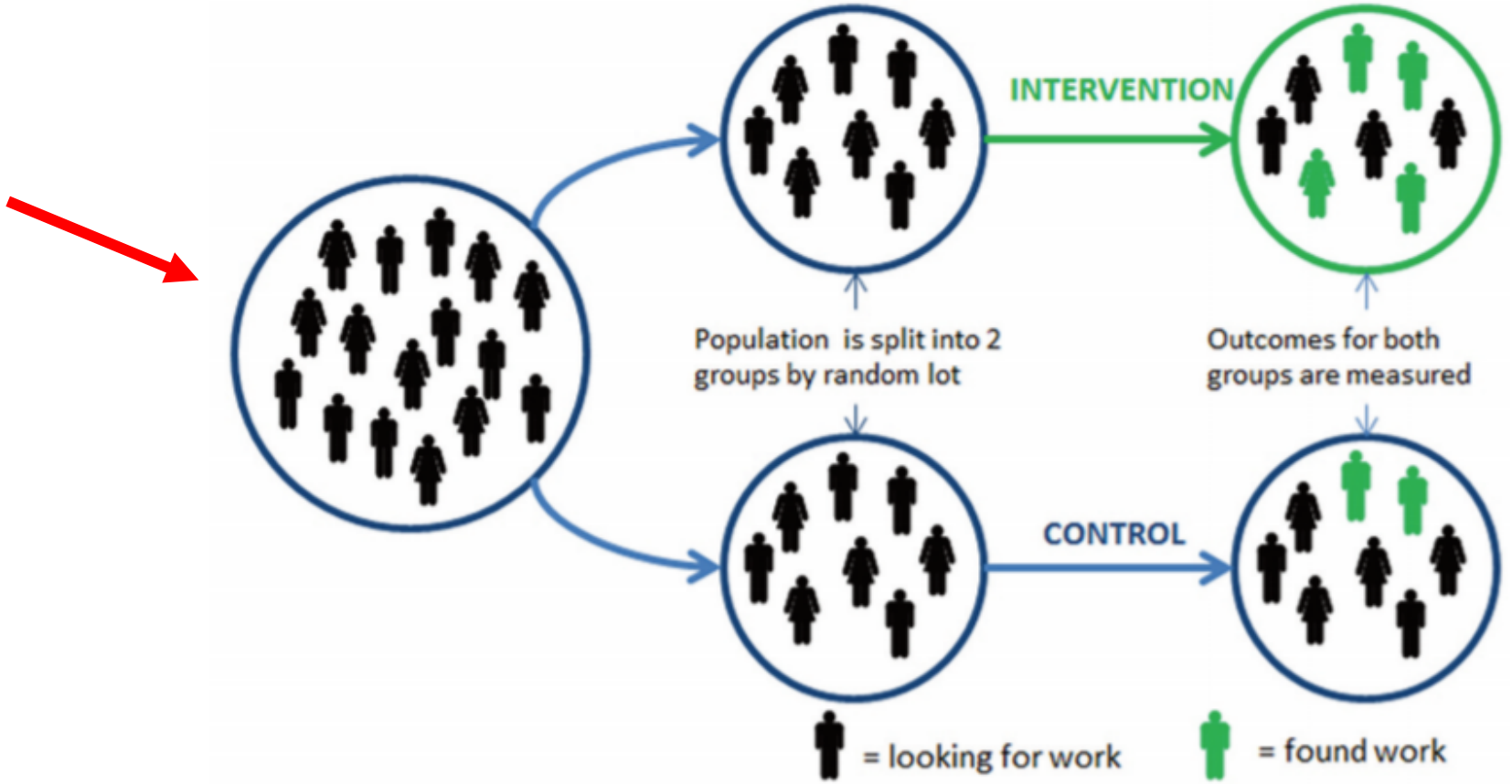
Asking 5 people

Height	Number of responses
140 -144cm	1
145-149cm	2
150-154cm	1
155-159cm	0
160-164cm	1

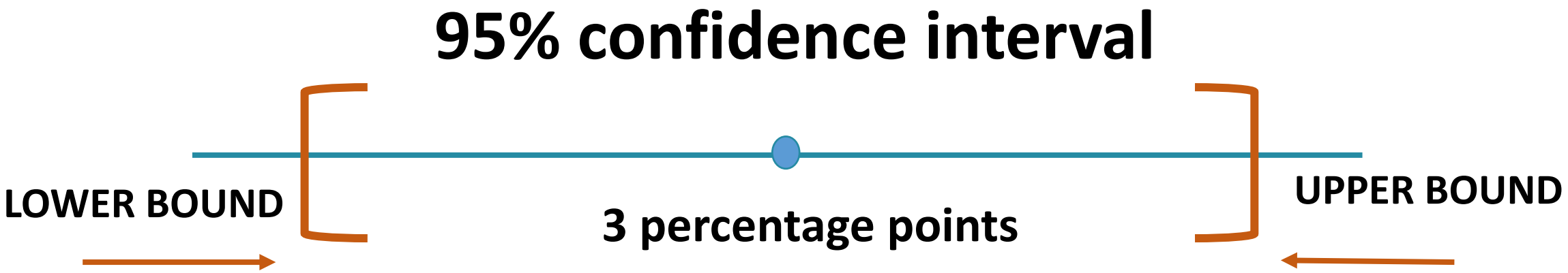
Asking 1000 people

Height	Number of responses
140 -144cm	70
145-149cm	150
150-154cm	125
155-159cm	650
160-164cm	5

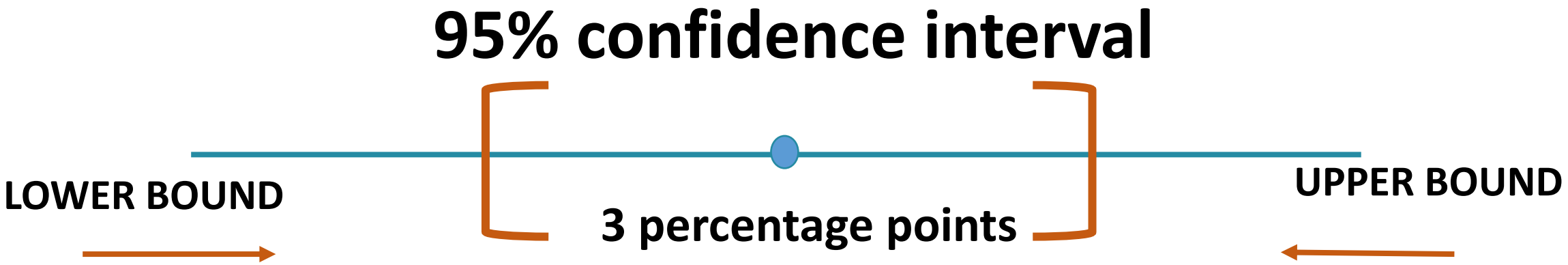
Larger sample sizes will yield estimates of impact we're more confident of.



Larger sample sizes will tighten our confidence interval



Larger sample sizes will tighten our confidence interval



How large is large?

1. Probability of Type 1 and Type 2 errors.
2. How similar or different is the population?
3. How large of an impact is expected?
4. What is the unit of randomization?

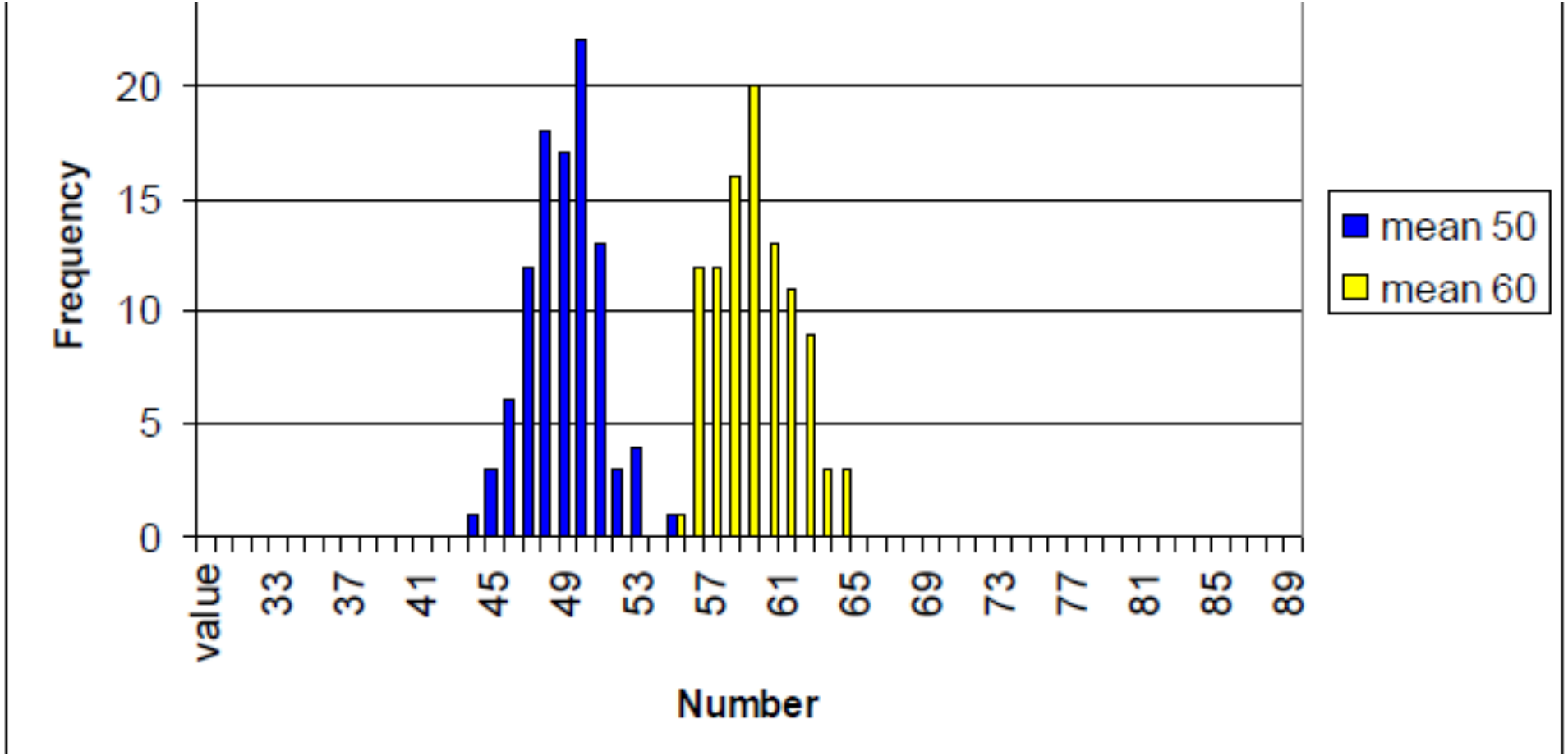
$$N = \left[\frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[1 + \rho(H - 1) \right]$$

How large is large?

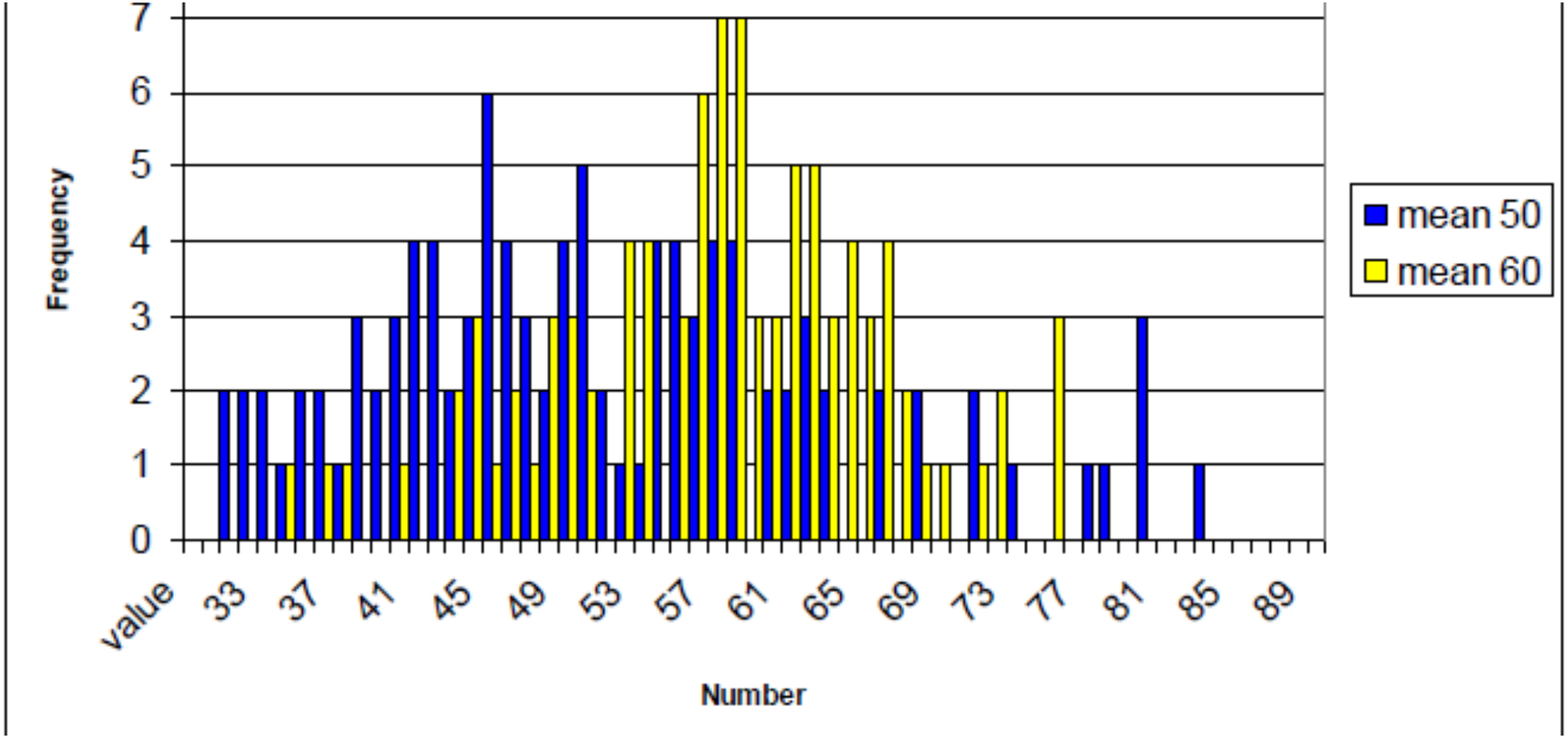
1. Probability of Type 1 and Type 2 errors.
2. **How similar or different is the population?**
3. How large of an impact is expected?
4. What is the unit of randomization?

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

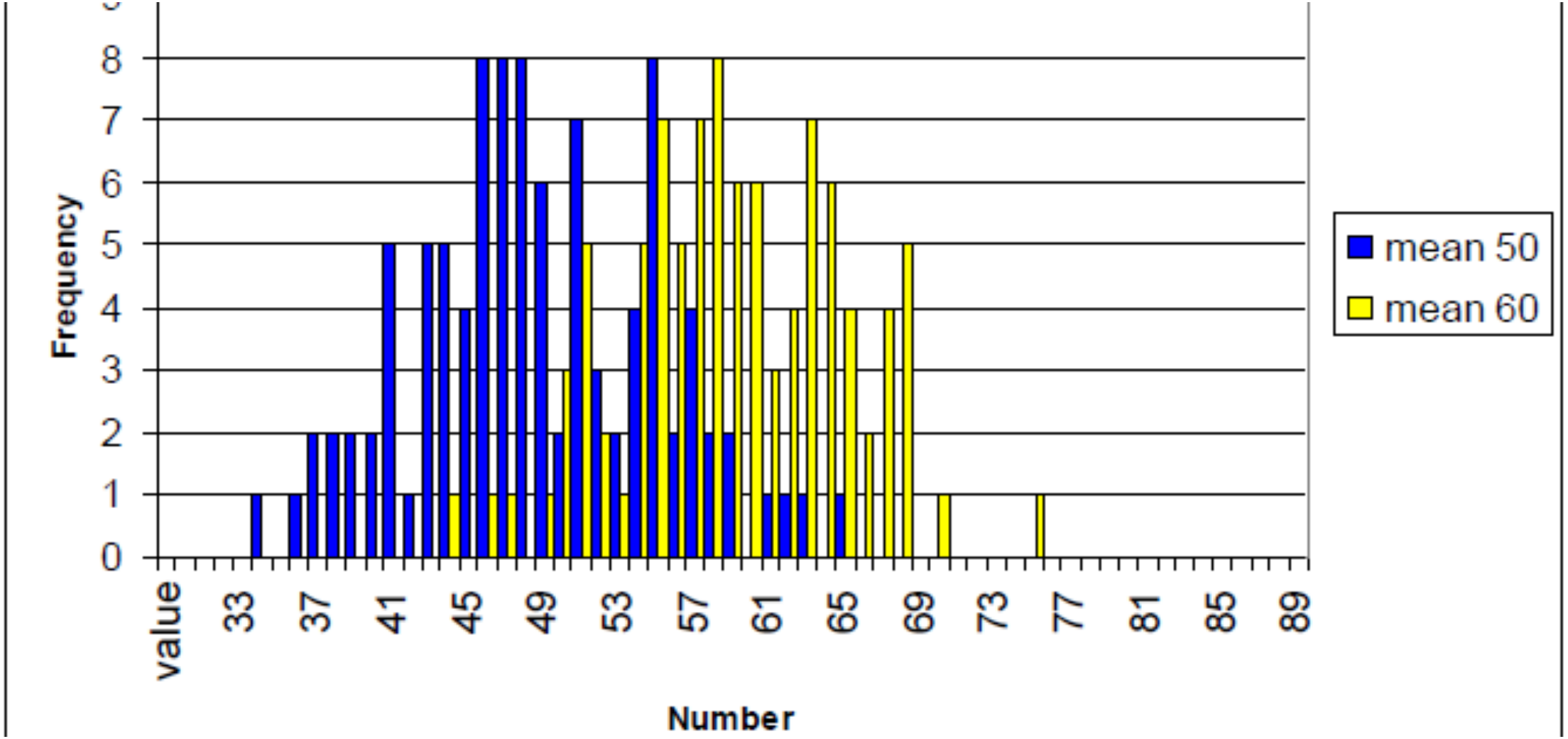
Less variance makes it easier to detect difference



Less variance makes it easier to detect difference



Less variance makes it easier to detect difference



How large is large?

1. Probability of Type 1 and Type 2 errors.
2. How similar or different is the population?
3. **How large of an impact is expected?**
4. What is the unit of randomization?

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

How many times do you need to look to see who is taller?

Distinguishing a 50 percentage point increase from 0

Distinguishing a 3 percentage point increase from zero

So, should we just pick really large effect sizes?

No.

Choose the lowest value you would consider a success.

This is your new zero effect.

Would a 3 percentage point increase be considered a success?

How large is large?

1. Probability of Type 1 and Type 2 errors.
2. How similar or different is the population?
3. How large of an impact is expected?
4. **What is the unit of randomization?**

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

What is a cluster?

Examples

School, clinic, village

Clustering creates similarity

Pupils in same school will be similar because school-level factors are similar

Sampling more students does not give more information

Pupils in same school will be similar because school-level factors are similar

What if I asked the same person 1000 times?

Asking 5 people

Height	Number of responses
140 -144cm	1
145-149cm	2
150-154cm	1
155-159cm	0
160-164cm	1

Asking 1000 people

Height	Number of responses
140 -144cm	70
145-149cm	150
150-154cm	125
155-159cm	650
160-164cm	5

What does this mean for us?

Goal is to maximize number of clusters

Not number of individuals per cluster

Larger samples of individuals in total needed if unit of randomization is a cluster

Measuring the degree of similarity from clustering

Intra-cluster correlation

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

If ρ=1

All individuals within the same cluster are exactly the same.

Increasing number of individuals no impact on power

If ρ=0

Same as performing individual level randomization.

Endline variable	8 children per cluster	
	Number of clusters	ICC
Treatment effectiveness		
Perceived there was a free access offer	223	.71
Perceived there was any financial help	223	.77
Perceived there was a large financial help	223	.85
Assisted to some information sessions	233	.59
Assisted to 3 or more information sessions	233	.5
Parental perceptions on the benefits of KG		
Primary school will be completed	235	.26
Secondary school will be completed	229	.25
Child will not be bullied in primary school	230	.29
Child will be treated with respect by teachers in primary school	230	.27
KG participation		
Ever attended preschool or kindergarten	236	.25
Registered according to kindergarten registers	219	.24
Registered in a kindergarten (self-reported)	236	.21
Self-reported attendance rate over the past 5 days	206	.2
% of unannounced visits where child was present	236	.25

More clusters the better

Same power

- **80 clusters, 20 individuals per cluster**
- **40 clusters, 1067 individuals per cluster**

[intra-cluster correlation=0.05]

That's **1,600** individuals compared to **42,680**!

Randomizing individuals vs. clusters

Same power

- **Individual level: 393 in treatment and 393 in control**
- **Cluster level: 80 clusters each for treatment and control, 20 individuals per cluster** (1,600 individuals in treatment and 1,600 in control]

How can we increase statistical power?

What do we have control over?

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

1. Sample size
2. Expected effect size
3. Number of clusters
4. Number of units per cluster
5. Variance?
   - Stratification
   - Baseline data
   - Rounds of follow-up data

Strategic Impact
Evaluation Fund

THANK YOU