Session I: An Introduction to Field Experiments

John A. List
U. Chicago and NBER

A. Measurement Approaches
B. Data Generation
    1. Lab Experiments
    2. Field Experiments
        a. Types of Field Experiments
        b. Uses of Field Experiments
    3. A Few Design Insights
C. Opening My Laptop
D. Concluding Remarks

A. Measurement Approaches

- **Years spent examining the evaluation problem:**

Naturally-Occurring Data
-Modeling interesting data

DID, PSM, IV, STR, etc.

- **DID:** **difference-in-differences**
- **PSM:** **propensity score matching**
- **IV:** **instrumental variables estimation**
- **STR:** **structural modeling**

Measurement

- **The goal of any evaluation method is to construct the proper counterfactual.**
- **If we could observe the counterfactual directly, then there is no evaluation problem—simply difference.**
- **The various approaches using naturally-occurring data provide alternative methods of constructing the proper counterfactual.**

B. Data Generation

Controlled Data
-Generate interesting data

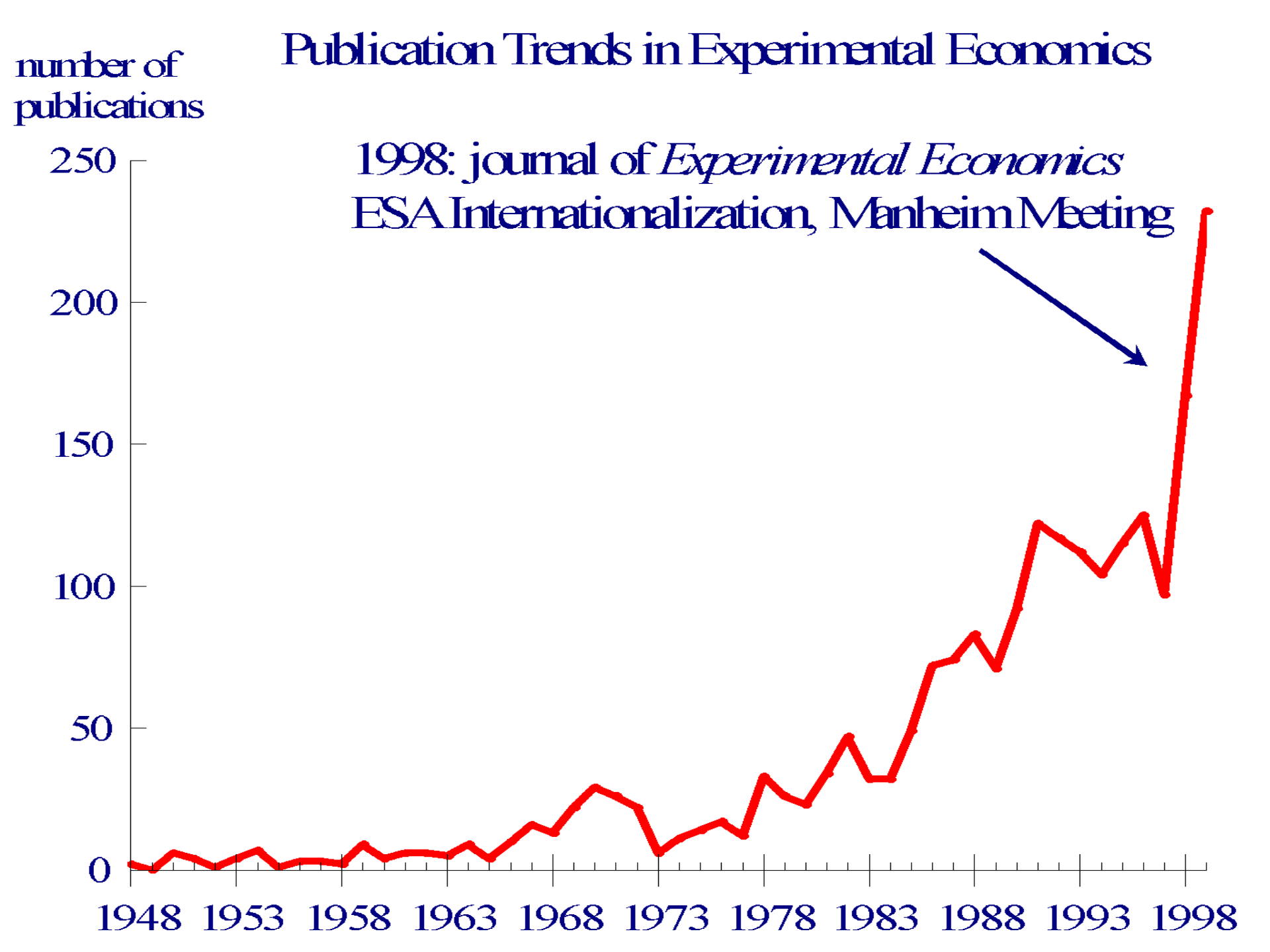
Naturally-Occurring Data
-Model interesting data

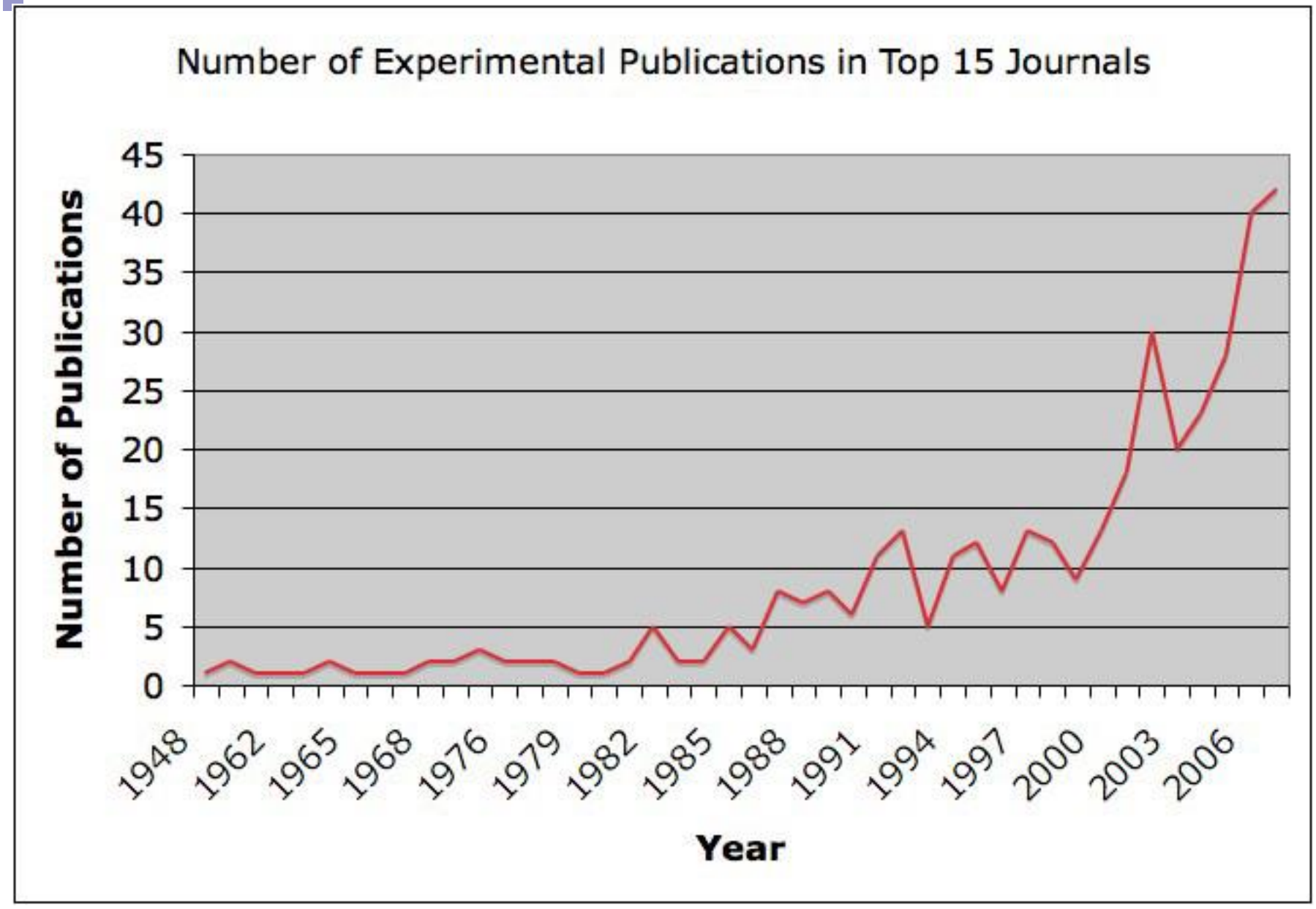
Lab **NE, PSM, IV, STR, etc.**

- **Lab:** controlled lab experiments
- **NE:** natural experiments
- **PSM:** propensity score estimation
- **IV:** instrumental variables estimation
- **STR:** structural modeling

Brief History in Economics

- 1. Bernoulli (Daniel) seems to be the first to run an economic experiment (1738)—St. Petersburg Paradox. Interestingly, his resolution was in the spirit of expected utility theory.
- 2. Individual Choice: Thurston (1931) tested ordinal utility theory with thought experiments.
- 3. Markets: Chamberlin (1948) tested competitive equilibrium using induced values.
- 4. Early experimental tests of game theory by Nash and others.


Publication Trends in Experimental Economics
number of publications
1998: journal of Experimental Economics
ESA Internationalization, Manheim Meeting
250
200
150
100
50
0
1948
1953
1958
1963
1968
1973
1978
1983
1988
1993
1998


Number of Experimental Publications in Top 15 Journals
Number of Publications
0
5
10
15
20
25
30
35
40
45
1948
1962
1965
1968
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
Year

Oft-Heard Criticism

- While working at the CEA I argued that certain lab results should be considered when revising the benefit/cost guidelines. An official from the White House commented:

  "even though these results appear prevalent, they are suspiciously drawn…….by methods similar to scientific numerology…..because of students…….who are not 'real' people"

Next Line of Skepticism

- Cross (1980) notes: "it seems to be extraordinarily optimistic to assume that behavior in an artificially constructed "market" game would provide direct insight into actual market behavior."

Commodity, stakes, environment--"hot" versus "cold" (or short-run versus long-run), task, etc., may vary from the lab to the field

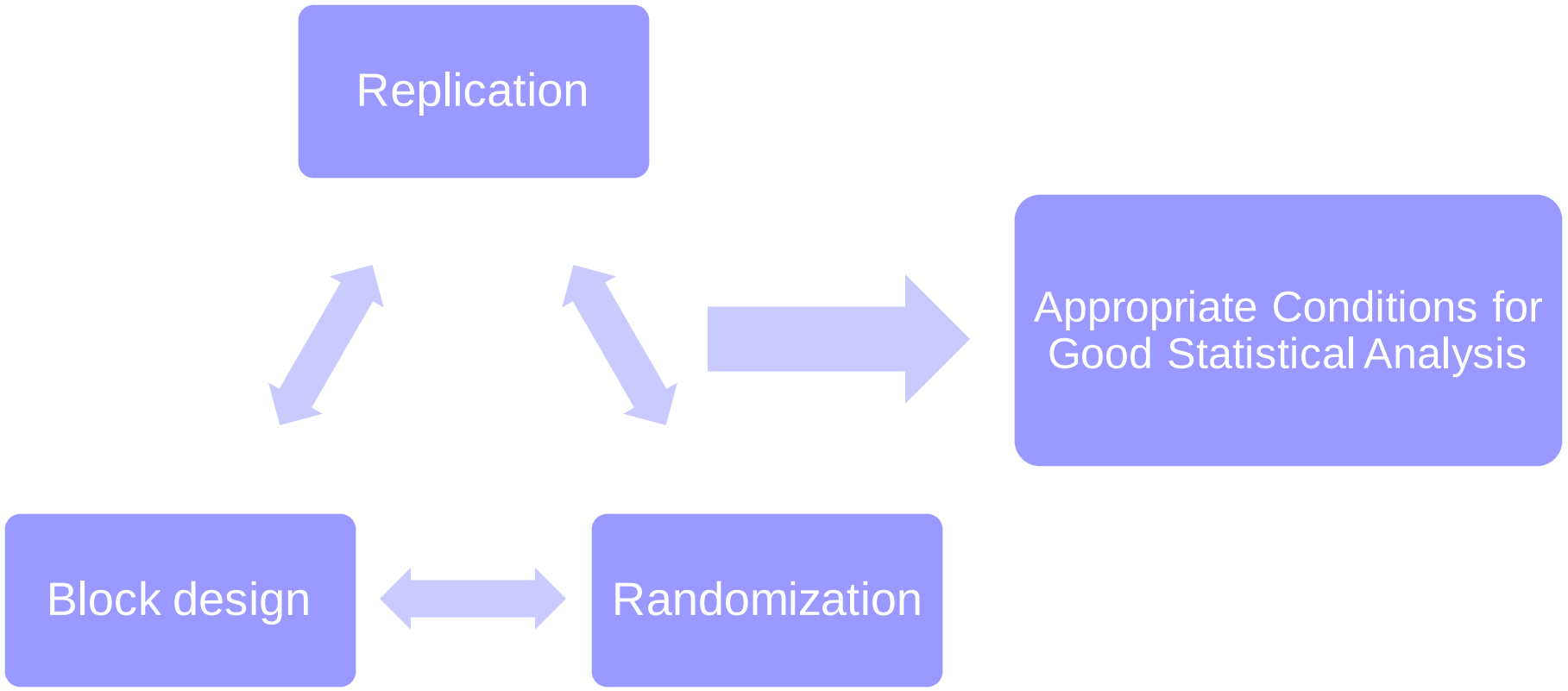
- Smith responds to Cross (1980 AER): "Experiments are sometimes criticized for not being 'realistic'….are there field data to support the criticism, i.e., data suggesting that there may be differences between laboratory and field behavior. If not, then the criticism is pure speculation."

- One avenue to help inform this debate is to run field experiments

A. The Dawn of “Field” Experiments

(Levitt and List, 2009 EER)

Neyman (1923) and Fisher (1926) laid the groundwork to use the physical act of randomization.



Wave II: Social Experiments

US: Heather Ross's $5 million MIT Ph.D. thesis is an early example. Beginning in late 60s, she tested the behavioral effects of Friedman's proposed negative income tax

Europe: Late 1960s, electricity pricing schemes in Great Britain.

Beyond showing the power of randomization, the profusion of social experiments helped kick off a series of methodological debates

*Plots of land cannot excuse themselves from treatment or select whether to be available for treatment

*More broadly, touched off the debate between structural econometricians and experimental advocates, which thrives today

Wave III: Recent Field Experiments

Harrison and List (2004 JEL)

_____AFE_____FFE_____NFE_____

Lab [field experiments] NE, PSM, IV, STR, etc.

- conventional lab experiment (Lab)
  - employs a standard subject pool of students, an abstract framing, and an imposed set of rules
- artefactual field experiment (AFE)
  - same as a conventional lab experiment but with a non-standard subject pool
- framed field experiment (FFE)
  - same as an artefactual field experiment but with field context in the commodity, task, information, stakes, time frame, etc.
- natural field experiment (NFE)
  - same as a framed field experiment but where the environment is the one that the subjects naturally undertake these tasks, such that the subjects do not know that they are in an experiment

Underlying Idea

Lab Field Experiments Models Using Nat. Data

A deeper economic understanding is possible by taking advantage of the myriad of settings in which economic phenomena present themselves.

In many cases experimentation in small-scale field settings is quite useful in developing a first understanding when observational data is limited or experimentation in more "important" markets is not possible.

After which, one explores how the key features of the studied domain compare to more distant domains.

Some Uses of Field Experiments

1. **Testing theory, measuring key parameters, collecting information to construct a theory, speaking to policymakers, making money**

2. **Methodological**

   a. Whether lab behavior is a good indicator of behavior in the field (e.g., List, 2006 JPE).

   b. The role of the market (market experience) in mitigating various forms of behavior (e.g., List, 2002 AER; List and Lucking-Reiley, 2000 AER).

   Are we making proper inference from lab and naturally-occurring data? (Levitt and List, 2007, JEP)

A. Inference

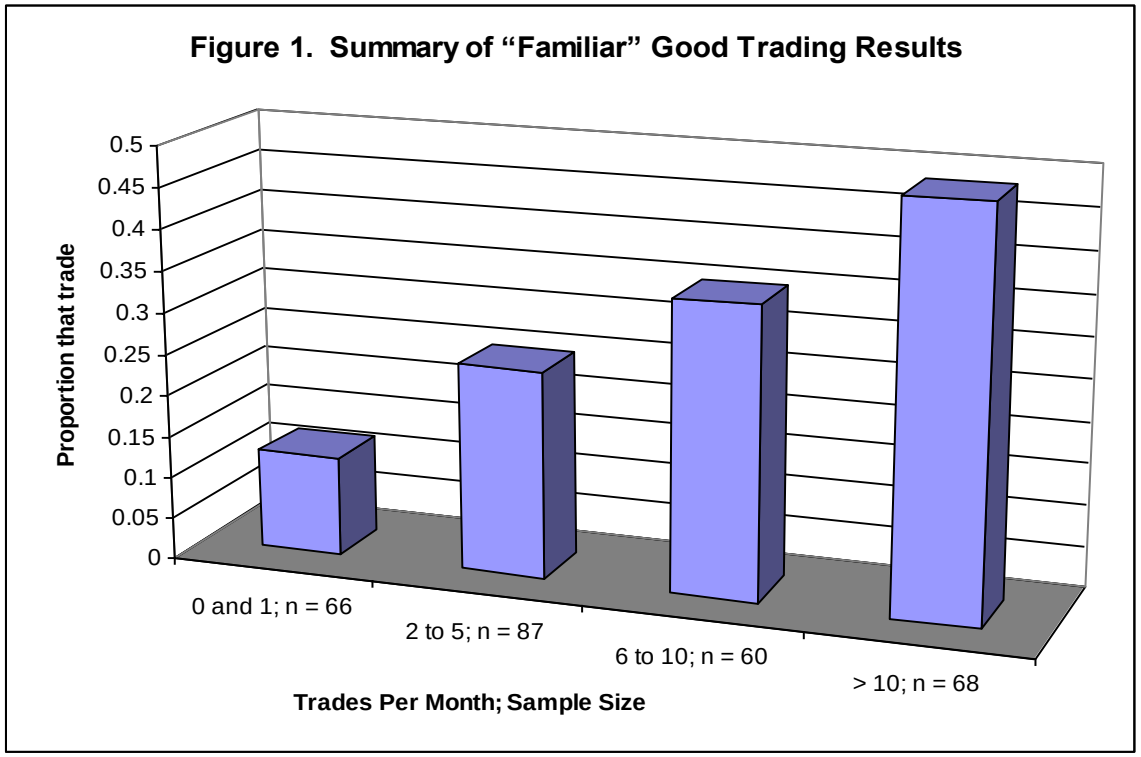
An Example:

Field experiments might highlight that certain results from lab experiments or naturally-occurring data should be defined more narrowly than first believed

or even might cause the initial insights from the lab or the field to be reinterpreted.

Consider the endowment effect experiments: mugs and candy bars

From List (2003; QJE; 2004 ECTA)



- Combine these results with those from the lab to use two distinct observations to create a deeper theory of human behavior than either could alone (e.g., Boyarchenko; Koszegi and Rabin; Chambers and Melkonyan; Neilson; Wilson).
- Replicate and then detail the types of variables that cause, attenuate, or exacerbate the endowment effect (e.g., Sugden, Starmer, Loomes, Lusk, Murphy, etc.). Go back to the field for more testing.

B. Causes

Field experiments can shed light on the causes and underlying conditions necessary to produce data patterns observed in the lab or in uncontrolled field data.

Parsing type of discrimination observed (List, 2004 QJE)

Discrimination Experiment

Study seller side discrimination:

12 disabled and 12 non-disabled testers approached various body shops in Chicago with different cars (identical cars across disabled and abled) that were in need of repair

Offer differences: disabled receive roughly 30% higher offers in repair market

Complementary Evidence

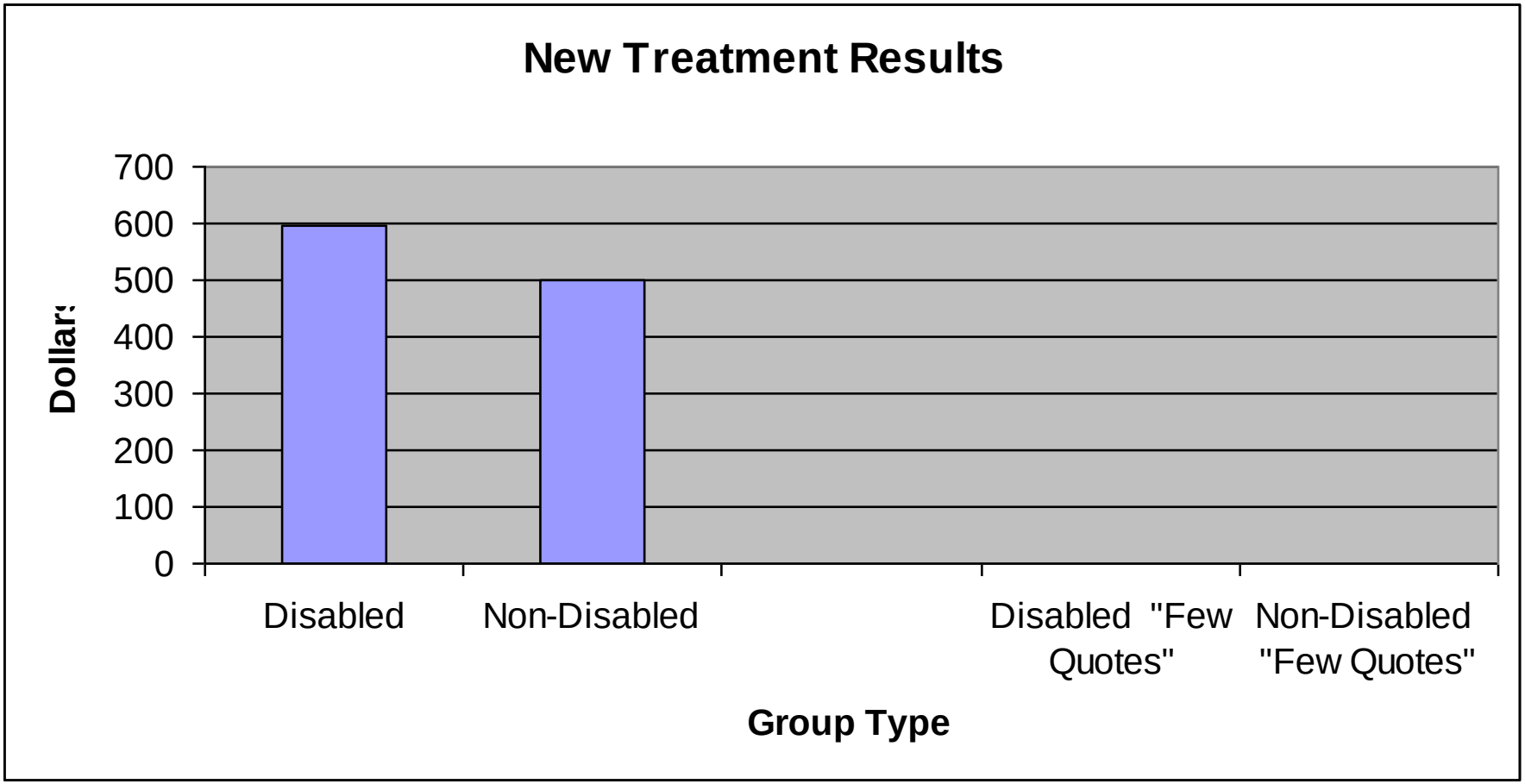
Survey

- Disabled search less
- Body shop mechanics believe there are search differences.

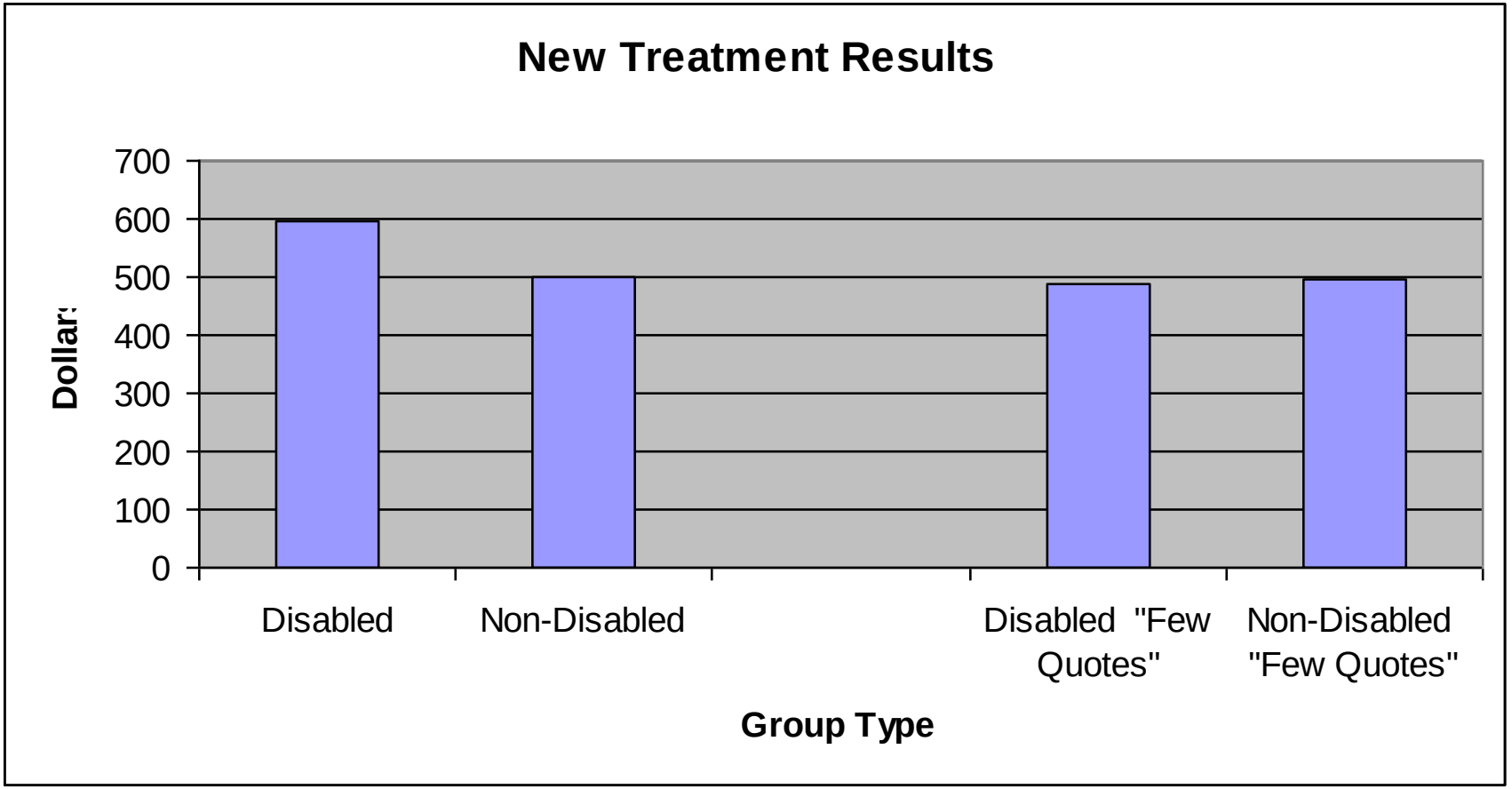
New Treatments

- Re-send different pairs to receive price quotes
- One treatment replicates above treatment
- Another treatment is identical except that it has both agent types explicitly noting that “I’m getting a few price quotes”

Replication Treatment



“Few Quote” Treatment



Back to the Basics: Generating Data

- If you were to go out and run an experiment tomorrow, what would I want you to know about design?

3. Some Design Insights

A. 0/1 treatment, equal outcome variances

B. 0/1 treatment, unequal outcome variances

C. Treatment Intensity—no longer binary

D. Clusters

Some Design Rules of Thumb for Differences in between-subject experiments

Assume that X_0 is $N(\mu_0,\sigma_0^2)$ and X_1 is $N(\mu_1, \sigma_1^2)$; and the minimum detectable effect $\mu_1 - \mu_0 = \delta$. H_0: $\mu_0 = \mu_1$ and H_1: $\mu_1 - \mu_0 = \delta$. We need the difference in sample means $X_1 - X_0$ to satisfy:

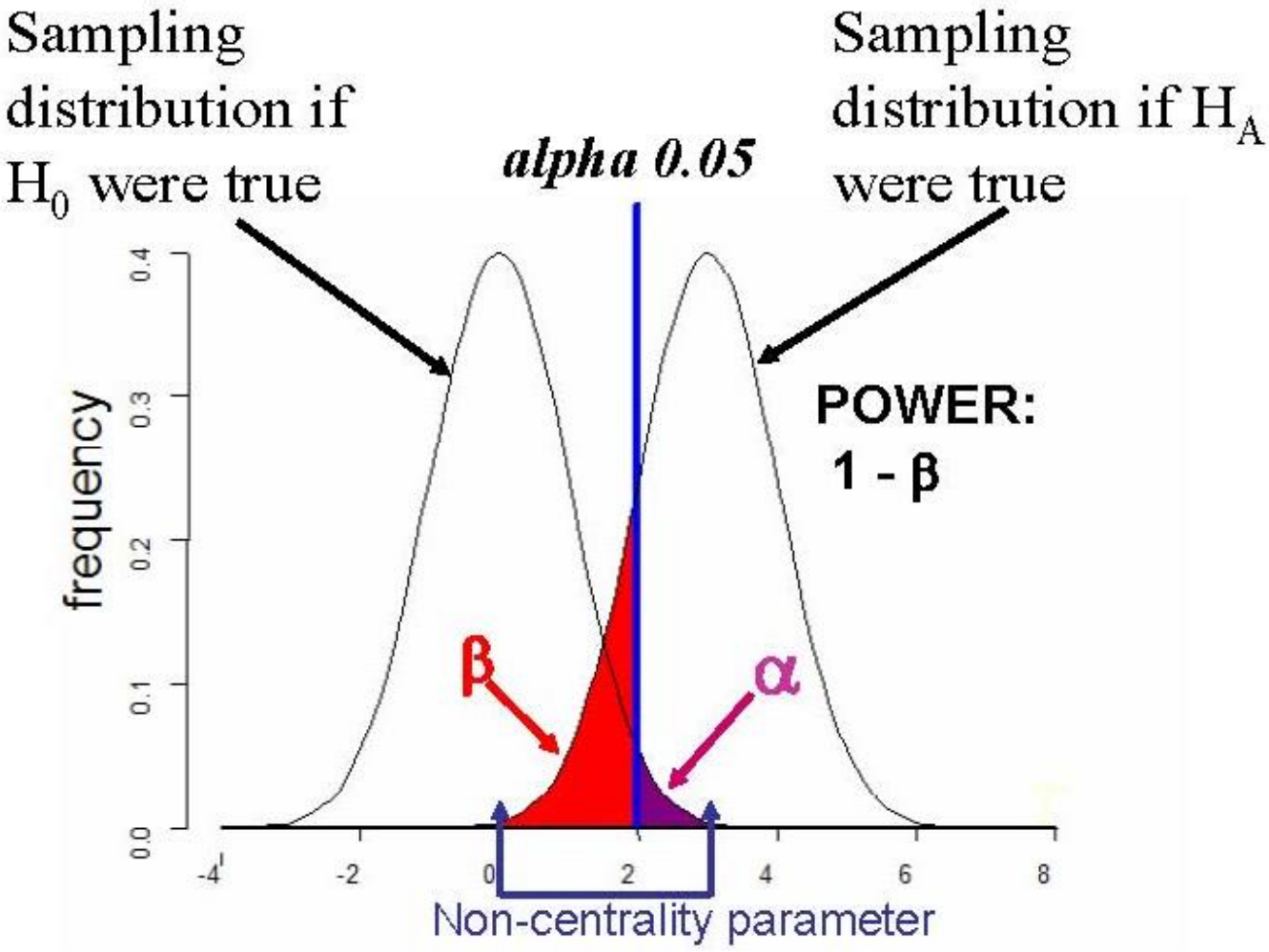
1. Significance level (probability of Type I error) = α:

$$\frac{X_1 - X_0}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = t_{\alpha/2} \Rightarrow X_1 - X_0 = t_{\alpha/2}\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$

2. Power (1 – probability of Type II error) = 1-β:

$$\frac{(X_1 - X_0) - \delta}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = -t_\beta \Rightarrow X_1 - X_0 = \delta - t_\beta\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$

Standard Case



Power

A. Our usual approach stems from the standard regression model: under a true null what is the probability of observing the coefficient that we observed?

B. Power calculations are quite different, exploring if the alternative hypothesis is true, then what is the probability that the estimated coefficient lies outside the 95% CI defined under the null.

Sample Sizes for Differences in Means (Equal Variances)

- Solving equations 1 and 2 assuming equal variances $\sigma_1^2 = \sigma_2^2$:

$$n_0^* = n_1^* = n^* = 2(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- Note that the necessary sample size
  - Increases rapidly with the desired significance level and power.
  - Increases proportionally with the variance of the outcomes.
  - Decreases inversely proportionally with the square of the minimum detectable effect size.

- Sample size depends on the ratio of effect size to standard deviation. Hence, effect sizes can just as easily be expressed in standard deviations.

- Standard is to use α=0.05 and have power of 0.80 (β=0.20).
- So if we want to detect a one-standard deviation change using the standard approach, we would need:
- $n = 2(1.96 + 0.84)^2*(1)^2 = 15.68$ observations in each cell
- ½ std. dev. change is detectable with 4*15.68 ~ 64 observations per cell
- n=30 seems to be the magic number in many experimental studies: ~ 0.70 std. dev. change.

Sample Size “Rules of Thumb”:

- Assuming α =0.05 and β = 0.20 requires n subjects:
  - α = 0.05 and β = 0.05 -> 1.65 × n
  - α = 0.01 and β = 0.20 -> 1.49 × n
  - α = 0.01 and β = 0.05 -> 2.27 × n

Sample Sizes for Differences in Means (unequal variances)

- Another Rule of Thumb—if the outcome variances are not equal then:

The ratio of the optimal proportions of the total sample in control and treatment groups is equal to the ratio of the standard deviations.

What about Treatment Levels?

- Assume that you are interested in understanding the intensity of treatment :
  - Number of ads., various incentive levels, etc.
  - Assume that the outcome variance is equal across various cells.
- How should you allocate the sample if incentives can be 1, 3, 5, 7, and 9?
- Assume that you have 1000 subjects available.

Reconsider what we are doing:

$$Y = XB + e$$

- One goal in this case is to derive the most precise estimate of *B* by using exogenous variation in X.

- Recall that the standard error of *B* is =

  var(e)/n*var(X)

Rules of Thumb

- Linear: ½ sample at X=1 and ½ at X=9.

- Quadratic: ¼, ½, ¼ at X=1, X=5, and X=9
- Intuitively: the test for a quadratic effect compares the mean of the outcomes at the extremes to the mean of the outcome at the midpoint

Intra-cluster Correlation

- What happens when the level of randomization differs from the unit of observation? Think of randomization at the village level, or at the store level, and outcomes are observed at the individual level.

An Example:

We convinced UAL to run an incentive experiment but they insisted the randomization occur at the city level. They allowed us four cities in which to run the experiment, with each city either receiving the high or low incentive. Thus, the randomization occurs at the city level, but our analysis is at the individual level—how many tickets the individual purchases.

- In each city we ran the incentive experiment on an equal number of consumers. But, responses of consumers might be correlated within the city.

Intracluster Correlation

- Real Sample Size (RSS) = mk/CE

m = number of subjects in a cluster

k = number of clusters

CE = $1 + \rho(m-1)$

ρ = intracluster correlation coefficient

= $s^2_B/(s^2_B + s^2_w)$

s^2_B = variance between clusters

s^2_w = variance within clusters

Intracluster Correlation

- What does $\rho \rightarrow 0$ mean?

No correlation of responses within a cluster

►No need to adjust optimal sample sizes

What does $\rho \rightarrow 1$ mean?

All responses within a cluster are identical

►Large adjustment needed: RSS is reduced to the number of clusters

Example

- Pilot testing confirms our suspicion, yielding $\rho = 0.04$.
- They wish to detect a 1/10 std. dev. change.
- Using the standard approach, what should the sample size equal?

$\rho \rightarrow 0$:

What is *n*?

Sample Size Formula:

$n = 2*(t_a + t_B)^2 * [\sigma/\delta]^2$

n = 1568 at each level; 3136 total.

Example

- RSS = mk/CE
  =784*4/(1+.04(784-1))
  ~97!

What is the required sample size?
$= 2*(t_a + t_B)^2 * 100(1+783(0.04))$

$= 15.68*3232$ (note that $\rho \rightarrow 0$: 15.68*100)

=50,678 at each incentive level!

Fractional Factorial Design
A Graphical Illustration