HANDOUT # 1

INTRODUCTION TO DESIGN OF EXPERIMENTS

TOPICS

1. Experimental Design Principles
2. Experimental Design Terminology
3. Example of Designed Experiment
4. Variations in Randomization for Similar Experiments
5. Common Problems in Experimental Design
6. Selecting the Appropriate Design
7. Randomization
8. Permutation Tests

EXPERIMENTAL DESIGN

Types of Studies

A scientific study may be conducted in various ways.

In the social sciences, environmental sciences, epidemiology, quality control, etc., many of the studies involve collecting information (data) on natural processes which are undisturbed by the observer. For example, political polls, factors affecting the rate of criminal behavior in young offenders, the impact on the environment of various types of coal mining, the factors affecting global warming, consumer purchasing behaviors, a study of the factors affecting the occurrence of certain diseases, a quality control study of why the proportion of product not meeting specifications has suddenly increased, etc.

In other areas of research, the study may be conducted in an artificial environment with highly controlled conditions. For example, greenhouse studies of the appropriate level of growth retardant to apply to rose plants, laboratory studies of the optimal temperature to heat treat an alloy to obtain a specified degree of hardness, a quality control study of a medical lab to determine which factors have the greatest impact on the variability in the measurement of a specified blood component, etc.

Finally, there are many fields of science in which there is a degree of control but the populations/processes under study are influenced by many uncontrolled factors. For example clinical trials in medicine, agricultural field studies, studies of the manufacturing process while commercial production is ongoing, etc.

Pure Observational Studies, for example, many epidemiology studies, are defined as observing and measuring units within populations without any interference with the natural process. The data collected from such studies is often used to just provide a description of the population but can also be used to establish a baseline for comparison with envisioned changes that are anticipated. An environmental impact report on a wetland near a proposed new residential development is such a study. In many of these studies, pure observation is all the researcher can obtain from the data. It is often nearly impossible to ascribe differences between observed populations, because the differences in the populations may be confounded with historical events which may not be completely known or whose impact is uncertain. Weather conditions, land usage, subsurface soil differentials may have a large impact on the wetlands which would have occurred even without the new residential development.

Sample Surveys are studies in which a very small proportion of a population are selected for interviews on various questions of interest. The method of analyzing the data and making inferences from surveys depends on the manner in which the subjects are selected from the population. Many potential sources of variation and hidden biases in the surveys must be addressed to obtain valid conclusions from the survey. Scientific surveys of non-human populations are widely used: behavior of dolphins when stressed by recreational boaters, a determination of diversity of plant communities in rain forests after large scale deforestation, soil and leaf microbe micro-cultures, observation of black holes in the universe. Surveys encounter the same difficulties that occur with pure observational studies.

Deming, data and observational studies

A process out of control and needing fixing

"Any claim coming from an observational study is most likely to be wrong." Startling, but true. Coffee causes pancreatic cancer. Type A personality causes heart attacks. Trans-fat is a killer. Women who eat breakfast cereal give birth to more boys. All these claims come from observational studies; yet when the studies are carefully examined, the claimed links appear to be incorrect. What is going wrong? Some have suggested that the scientific method is failing, that nature itself is playing tricks on us. But it is our way of studying nature that is broken and that urgently needs mending, say **S. Stanley Young** and **Alan Karr**; and they propose a strategy to fix it.

Science works by experiments that can be repeated; when they are repeated, they must give the same answer. If an experiment does not replicate, something has gone wrong. In a large branch of science the experiments are observational studies: we look at people who eat certain foods, or take certain drugs, or live certain lifestyles, and we seem to find that they suffer more from certain diseases or are cured of those diseases, or – as with women who eat more breakfast cereal – that more of their children are boys. The more startling the claim, the better. These results are published in peer-reviewed journals, and frequently make news headlines as well. They seem solid. They are based on observation, on scientific method, and on statistics. But something is going wrong. There is now enough evidence to say what many have long thought: that any claim coming from an observational study is most likely to be wrong – wrong in the sense that it will not replicate if tested rigorously.

As long ago as 1988[1,2] it was noted that there were contradicted results for case–control studies in 56 different topic areas, of which

Table 1. We have found 12 papers in which claims coming from observational studies were tested in randomised clinical trials. Many of the trials are quite large. In most of the observational studies multiple claims were tested, often in factorial designs, e.g. vitamin D and calcium individually and together along with a placebo group. Note that none of the claims replicated in the direction claimed in the observational studies and that there was statistical significance in the opposite direction five times

ID no.	*Pos.*	*Neg.*	*No. of claims*	*Treatment(s)*	*Reference*
1	0	1	3	Vit E, beta-carotene	*NEJM* 1994; **330**: 1029–1035
2	0	3	4	Hormone Replacement Ther.	*JAMA* 2003; **289**: 2651–2662, 2663–2672, 2673–2684
3	0	1	2	Vit E, beta-carotene	*JNCI* 2005; **97**: 481–488
4	0	0	3	Vit E	*JAMA* 2005; **293**: 1338–1347
5	0	0	3	Low Fat	*JAMA*. 2006; **295**: 655–666
6	0	0	3	Vit D, Calcium	*NEJM* 2006; **354**: 669–683
7	0	0	2	Folic acid, Vit B6, B12	*NEJM* 2006; **354**: 2764–2772
8	0	0	2	Low Fat	*JAMA* 2007; **298**: 289–298
9	0	0	12	Vit C, Vit E, beta-carotene	*Arch Intern Med* 2007; **167**: 1610–1618
10	0	0	12	Vit C, Vit E	*JAMA* 2008; **300**: 2123–2133
11	0	0	3	Vit E, Selenium	*JAMA* 2009; **301**: 39–51
12	0	0	3	HRT + Vitamins	*JAMA* 2002; **288**: 2431–2440
Totals	0	5	52		

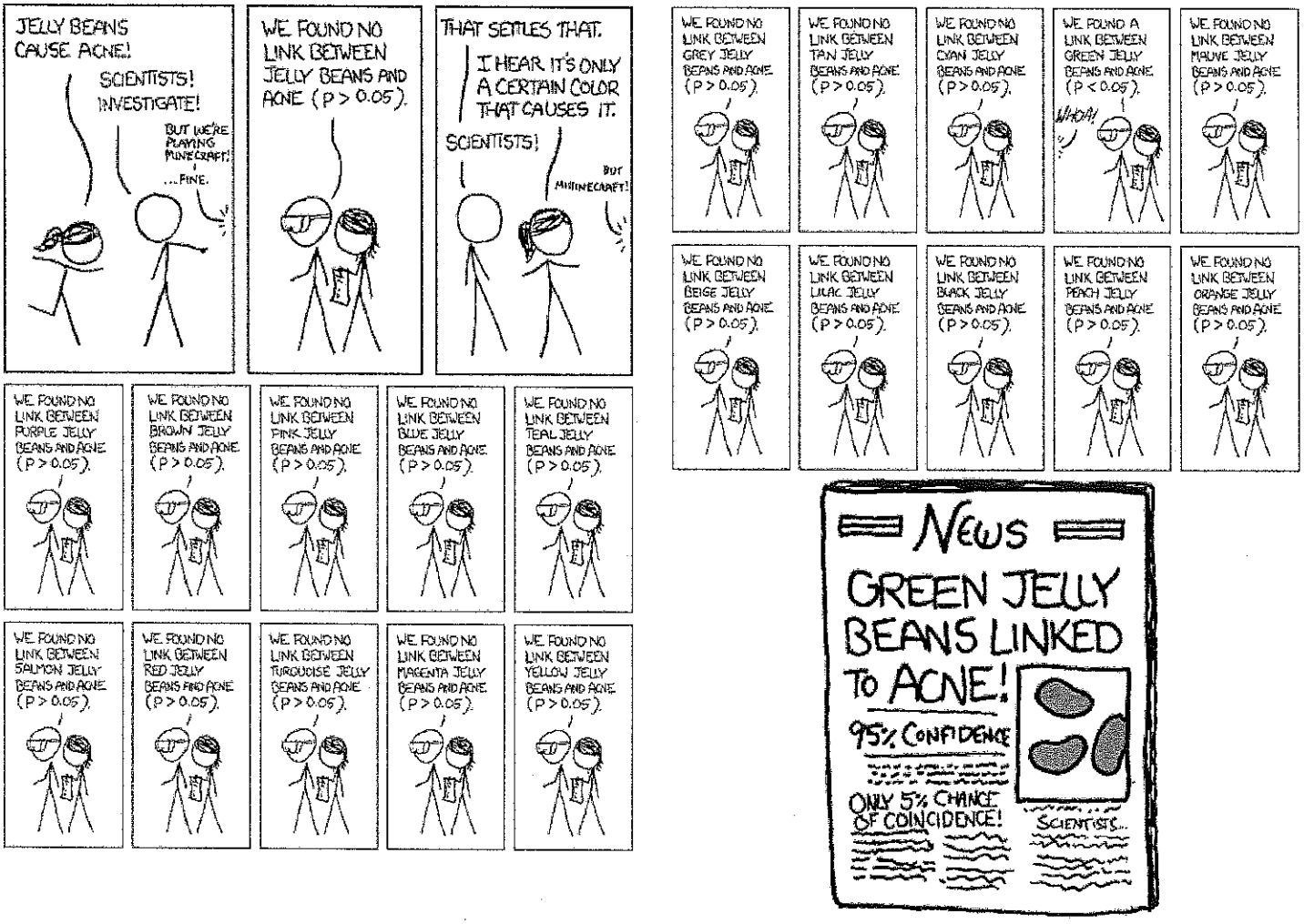


Figure 1. There is no overall effect of jelly beans on acne. Bummer. How about subgroups? Often subgroups are explored without alerting the reader to the number of questions at issue. Courtesy xkcd, http://xkcd.com/882/

cancer and things that cause it or cure it were by far the most frequent. An average of 2.4 studies supported each association – and an average of 2.3 studies did not support it. For example, three studies supported an association between the anti-depressant drug reserpine and breast cancer, and eight did not. It was asserted[2] that "much of the disagreement may occur because a set of rigorous scientific principles has not yet been accepted to guide the design or interpretation of case-control research". Problems extend to essentially all observational studies. Little progress has been made to adopt rigorous scientific principles. Some journal article titles give a flavour of the sentiments: "Epidemiology faces its limits", "Is it time to call it a day?", "Have we learned from our mistakes, or are we doomed to compound them?". In the popular press, an article by Jonah Lehrer in the *New Yorker*[3] bore the subheading "Is there something wrong with the scientific method?" and seemed to imply that replicability was no longer occurring; it concluded with the phrase: "When the experiments are done, we still have to choose what to believe." No. In the Lehrer example the motivating finding was wrong and therefore should not be expected to replicate.

It may not be appreciated how often observational claims fail to replicate. In a small sample in 2005[4], of 49 claims coming from highly cited studies, 14 either failed to replicate entirely or the magnitude of the claimed effect was greatly reduced (a regression to the mean). Six of these 49 studies were observational studies, and in these six, in effect, randomly chosen observational studies, five failed to replicate. This last is an 83% failure rate. In an ideal world in which well-studied questions are addressed and statistical issues are accounted for properly, few statistically significant claims are false positives. Reality for observational studies is quite different.

We ourselves carried out an informal but comprehensive accounting of 12 randomised clinical trials that tested observational claims – see Table 1. The 12 clinical trials tested 52 observational claims. They all confirmed no claims in the direction of the observational claims. We repeat that figure: 0 out of 52. To put it another way, 100% of the observational claims failed to replicate. In fact, five claims (9.6%) are statistically significant in the clinical trials *in the opposite direction* to the observational claim. To us, a false discovery rate of over 80% is potent evidence that the observational study process is not in control. The problem, which has been recognised at least since 1988, is systemic.

The cause of it all

The cause is elusive and can be considered both technically and operationally. Individual researchers, the workers, respond rationally to incentives by publishing papers in peer-reviewed journals and securing funding for their research. The quality of their papers is judged by funding agencies and journal editors, the important managers of the observational study production system. We can turn here to statistician W.

Box 1. Amplification of W. Edwards Deming's thinking

It is worth contrasting control of an observational study with that of a production process. When Deming first looked at manufacturing, it was common to inspect only the final product, be it a screw or a car, to maintain product quality. There was little or no systematic feedback from problems with the final product to places in the process where these defects occurred. This inspection of the final product works, but it is frightfully expensive. Deming's insight was to control each step of the process where errors occur so that the final frequency of bad product is greatly reduced. Now, world-wide, industrial production is *process control*. Control the steps of the process and the final product will largely take care of itself. Consider the production of an observational study: Workers – that is, researchers – do data collection, data cleaning, statistical analysis, interpretation, writing a report/paper. It is a craft with essentially no managerial control at each step of the process. In contrast, management dictates control at multiple steps in the manufacture of computer chips, to name only one process control example. But journal editors and referees inspect only the final product of the observational study production process and they release a lot of bad product. The consumer is left to sort it all out. No amount of educating the consumer will fix the process. No amount of teaching – or of blaming – the worker will materially change the group behaviour. Deming's insight was to admonish management to redesign an out-of-control process.

Edwards Deming[5], the most visionary innovator ever on quality control and the man who transformed first Japanese car manufacturing then manufacturing quality control worldwide (see Box 1). Deming said: "The worker is not the problem. The problem is at the top! Management!" To Deming, blaming the workers – individual researchers – is as incorrect as it is useless. Bringing the system under control is the responsibility of those managing it.

What is needed to fix the system? Among Deming's famous "Fourteen Points for Management", the third is most directly relevant: *cease dependence on inspection to achieve quality*. Every successful company today relies on control of the process; they do not wait until the end of the process and then throw away bad product. That would be product control, not process control. It is wasteful to make something, then inspect and throw away the bad product. Instead, every step of the process is monitored and controlled, so that bad product is not made. The "observational studies industry" must build a good product; journal editors cannot inspect bad product out at the publication stage, let alone the replication stage. If the processes are controlled by management, the products can be sound studies. Control of the processes is feasible, and requires attention to the incentives, publications and grants. First we examine three of the main technical difficulties with observational studies: Multiple testing, bias, and multiple modelling.

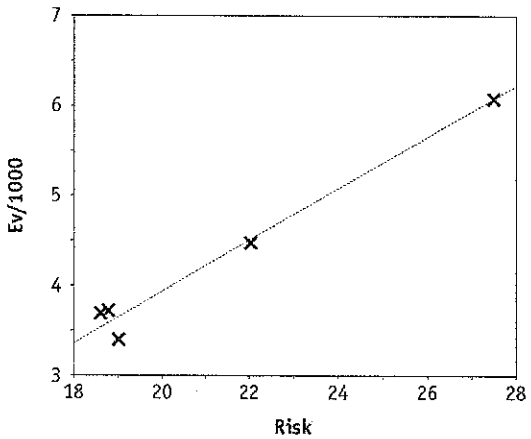
Multiple testing

False positives do occur, even in an ideal world. When many questions are asked of the same data, some of those questions will by chance come up positive. Producing at least one false positive becomes a near certainty unless the data analysis accounts for the multiple questions. Figure 1, from the excellent website `xkcd.com`, brilliantly explains the basic problem. The "females eating cereal leads to more boy babies" claim translated the cartoon example into real life. The claim appeared in the *Proceedings of the Royal Society, Series B*. It makes essentially no biological sense, as for humans the Y chromosome controls gender and comes from the male parent. The data set consisted of the gender of children of 740 mothers along with the results of a food questionnaire, not of breakfast cereal alone but of 133 different food items – compared to only 20 colours of jelly beans. Breakfast cereal during the second time period at issue was one of the few foods of the 133 to give a positive. We reanalysed the data[6], with 262 *t*-tests, and concluded that the result was easily explained as pure chance.

For those who want more than cartoons, a simple web simulation[7] is convincing that multiple testing needs to be controlled. Although many workers who are thought leaders of researchers doing observational studies argue against any correction of the analysis for multiple testing[8], managers can require that authors deal with multiple testing.

Bias

Whereas multiple testing is random error, bias is systematic error. To illustrate it, consider channelling, where doctors steer certain patients to particular treatments. For example, doctors directed HIV patients at high cardiovascular risk to a particular HIV treatment, abacavir, and lower-risk patients to other drugs, preventing a simple assessment of abacavir compared to other treatments. An analysis that did not correct for this bias unfairly penalised the abacavir, since its patients were more high-risk so more of them had heart attacks (Figure 2). Another problem is that covariate adjustment is widely used, but is vulnerable to manipulation and is well known to give unreliable results when the treatment groups are not comparable; see "Multiple modelling" below. Missing factors, unmeasured confounders, and loss to follow-up can also lead to bias. For example, in a study published in *Pediatrics*[9], offspring IQ was the issue, yet IQ of the fathers was not measured and of the 505 children starting the study, 256 (50.7%) were lost to follow-up. By selecting papers with a significant *p*-value, negative studies are selected against – which is publication bias (see Box 2).



Drug	% high-risk patients	MI events per 1000 patient years
1 zidovudine	19.0	3.43
2 stavudine	18.8	3.72
3 lamivudine	18.6	3.68
4 didanosine	22.1	4.47
5 abacavir	27.5	6.13

Figure 2. Events per thousand patient-years are plotted against estimated risk of a heart attack. Risky patients were channelled to the HIV drug ABC, abacavir, and those patients had more heart attacks, as shown by the uppermost point on the graph. Risk-adjusted, all the drugs appear to be of equal risk. Source: *Lancet* 371, 1417 ff.

Box 2. Publication bias

There is general recognition that a paper has a much better chance of acceptance if something new is found. This means that, for publication, the claim in the paper has to be based on a p-value less than 0.05. From Deming's point of view[5], this is quality by inspection. The journals are placing heavy reliance on a statistical test rather than examination of the methods and steps that lead to a conclusion. As to having a p-value less than 0.05, some might be tempted to game the system[10] through multiple testing, multiple modelling or unfair treatment of bias, or some combination of the three that leads to a small p-value. Researchers can be quite creative in devising a plausible story to fit the statistical finding.

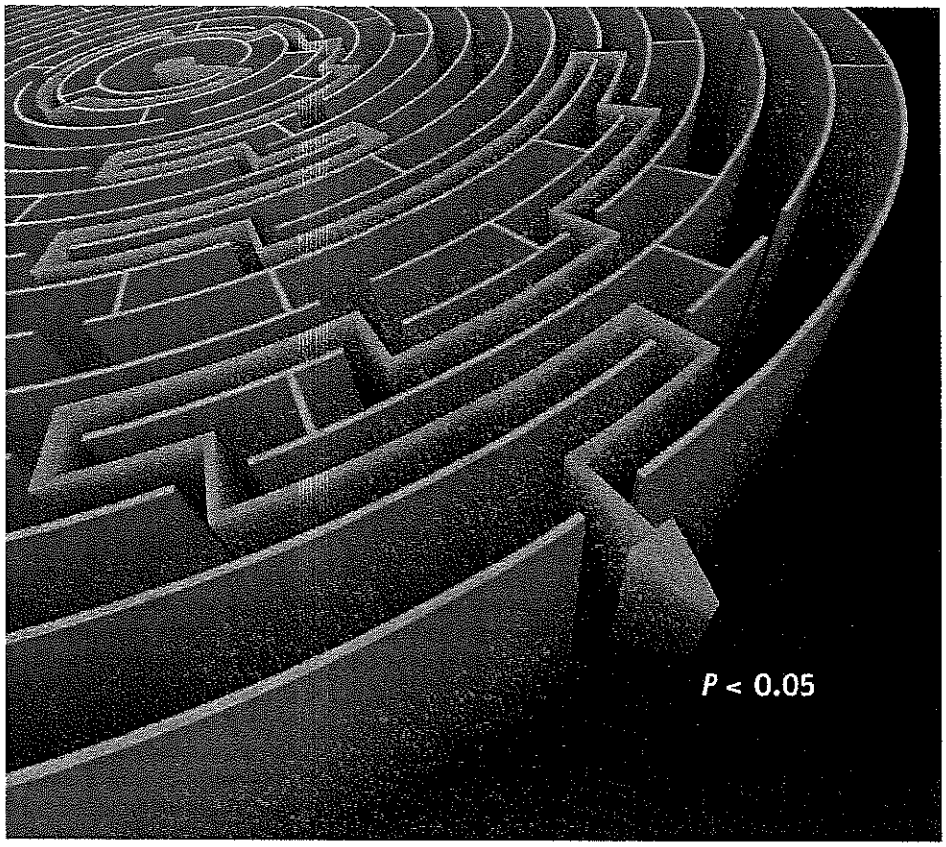


Figure 3. The path through a complex process can appear quite simple once the path is defined. Which terms are included in a multiple linear regression model? Each turn in a maze is analogous to including or not a specific term in the evolving linear model. By keeping an eye on the p-value on the term selected to be at issue, one can work towards a suitably small p-value. © ktsdesign – Fotolia

Multiple modelling

This problem is akin to – but less well recognised and more poorly understood than – multiple testing. For example, consider the use of linear regression to adjust the risk levels of two treatments to the same background level of risk. There can be many covariates, and each set of covariates can be in or out of the model. With ten covariates, there are over 1000 possible models. Consider a maze as a metaphor for modelling (Figure 3). The red line traces the correct path out of the maze. The path through the maze looks simple, once it is known. Returning to a linear regression model, terms can be put into and taken out of a regression model. Once you get a p-value smaller than 0.05, the model can be frozen and the model selection justified after the fact. It is easy to justify each turn.

The combination of multiple testing and multiple modelling can lead to a very large search space, as the example of bisphenol A in Box 3 shows. Such large search spaces can give small, false positive p-values somewhere within them. Unfortunately, authors and consumers are often like a deer caught in the headlights and take a small p-value as indicating a real effect.

How can it be fixed? A new, combined strategy

It should be clear by now that more than small-scale remedies are needed. The entire system of observational studies and the claims that are made from them is no longer functional, nor is it fit for purpose. What can be done to fix this broken system? There are no principled ways in the literature for dealing with model selection, so we propose a new, composite strategy. Following Deming, it is based not upon the workers – the researchers – but on the production system managers – the funding agencies and the editors of the journals where the claims are reported.

We propose a multi-step strategy to help bring observational studies under control (see Table 2). The main technical idea is to split the data into two data sets, a modelling data set and a holdout data set. The main operational idea is to require the journal to accept or reject the paper based on an analysis of the modelling data set without knowing the results of applying the methods used for the modelling set on the holdout set *and* to publish an addendum to the paper giving the results of the analysis of the holdout set. We now cover the steps, one by one.

1 The data collection and clean-up should be done by a group separate from the analysis group. There can be a temptation on the part of the analyst to do some exploratory data analysis during the data clean up. Exploratory analysis could lead to model selection bias.

2 The data cleaning team creates a modelling data set and a holdout set and gives the modelling data set, less the item to be predicted, to the analyst for examination.

Table 2. Steps 0–7 can be used to help bring the observational study process into control. Currently researchers analysing observational data sets are under no effective oversight

Step	*Process / Action*
0	Data are made publicly available
1	Data cleaning and analysis separate
2	Split sample: A, modelling; and B, holdout (testing)
3	Analysis plan is written, based on modelling data only
4	Written protocol, based on viewing predictor variables of A
5	Analysis of A only data set
6	Journal accepts paper based on A only
7	Analysis of B data set gives Addendum

Designed Experiments are the topic of this course. In this type of study, the scientist has control over many important aspects of the study. This is accomplished by one of the following methods:

- Randomly assigning the experimental subjects to treatments: Randomly assign homogeneous plots of cotton to be planted with one of five amounts of fertilizer; randomly assign rabbits to receive one of four sources of fiber in their diet; randomly assign fifth grade math students to be evaluated by one of three methods of instruction.

- Randomly selecting subjects from different populations: randomly select 100 males and 100 females to complete a survey evaluating views on concealed weapon laws; randomly select 500 each of urban, suburban, and rural residents to participate in a virtual focus group to assess consumer demand for high speed internet service.

In the first case, the scientist controls the assignment of homogeneous experimental units to the treatment groups.
In the second type of designed experiment, random samples are selected from natural populations. The scientist controls the random sampling but not the assignment of experimental subjects to the treatment groups.

In designed experiments, there is a formal randomization over factors not of interest to the researcher (these factors may influence the measured response) while maintaining control over the assignment of the experimental subjects to the treatment groups. Thus, on the average, any major differences in the measured responses between the treatment groups can be attributed to the treatment groups rather than to potential factors which were not controlled. In observational studies or surveys, great care must be taken in making inferences beyond the observed data due to many uncontrolled sources of variation. In those situations where the researcher does not have control over important factors that may influence the measured response, a detailed discussion of the many limitations of any inferences beyond the observed data should be a crucial component in the report detailing the experiment results. In studies involving subjects which survive well beyond the time period in which the study is conducted, there is a great concern about spurious association between the observed (measured) response and the treatment groups. For example, in prospective and retrospective studies, experimental subjects are followed forward or backward, respectively, in time, to examine health differences between the treatment groups. There are many uncontrolled factors which may impact on the measured response. A study is designed to determine the impact of smoking on the degree of blockage in retinal veins. However, people are not randomly assigned to the treatment groups, smoker vs non-smoker. Instead there is a self-selection mechanism which is based on a variety of genetic, behavioral, environmental, occupational, and many other factors. Most of these factors will be completely unknown to the researcher but many of the factors may have an important influence on the degree of retinal vein blockage.

EXPERIMENTAL DESIGN PRINCIPLES

Selected Comments from *Experimental Design* by W. Federer

I. "All fields of research have at least one feature in common:

The variability of experimental responses."

II. When there is considerable variation (large σ) from observation to observation on the same experimental material due to uncontrolled factors and it is not feasible to run a large number of experiments (which would reduce the variation in the mean response: $\hat{\mu} = \bar{Y} \quad \Rightarrow SE(\hat{\mu}) = \sigma/\sqrt{n}$), THEN the experimenter must:

1. Refine the experimental design in order to obtain a specified degree of precision

   Blocking Designs: Group units based on Time of year, Age of subjects, Occupation, Genetics, Fertility of Soil, Environmental conditions, etc.
2. In order to attach a probability statement to the observed treatment mean differences (a measure of the degree of confidence in the observed results), it is necessary that proper Randomization and Replication occur.

III. Certain Principles of Scientific Experimentation should always be followed: (Many are non statistical, however, the analysis of the data resulting from improperly designed and conducted experiments may complicate the analysis to the point at which a VALID analysis of the data can not be conducted.)

P1. Formulation of Questions to be Asked and Research Hypotheses to be Tested:

Clearly stating and precisely formulating questions and hypotheses prior to the running of the experiments will help to

1. Minimize the number of replications required
2. Make sure all necessary measurements are taken.

P2. A Critical and Logical Analysis of the Stated Research Hypotheses:

1. Review the relevant literature
2. Evaluate the reasonableness and utility of the aim of the experiment as reflected in the Research Hypotheses. (May need to reformulate the Research Hypotheses.)
3. Forecast the possible outcomes of the experiment in order to determine the proper statistical methodology to analyze the resulting data: For example,
   a. A large percentage of the data values are 0's: Overdispersed Poisson or Discrete-Continuous mixture
   b. The data is categorical rather than numerical: Log-linear models
   c. Too few replications for projected variability: Bootstraping or increase number of reps
   d. Correlated (nonindependent observations): Spatially or Temporally correlation

In an experiment involving correlated observations, it may be necessary to increase the number of reps in order to achieve the level of precision in estimating treatment effects or the power in test of differences. In these types of experiments, $SE(\hat{\mu})$ can be considerably larger than the standard error in a comparable experiment with independent observations.

For example, if the data can be adequately modeled by an AR(1) model:

$y_i = \theta + \rho y_{i-1} + e_i$ with $0 < \rho < 1$ then $\quad SE(\hat{\mu}) \approx \frac{\sigma}{\sqrt{n}}\sqrt{\frac{1+\rho}{1-\rho}} > \frac{\sigma}{\sqrt{n}}$

If the data was assumed to be independent when determining the appropriate sample size to achieve a given level of precision in estimating the population mean, μ, then the resulting sample size will be too small to achieve this goal if the data is positively correlated.

P3. Selection of Procedures for Conducting Research

1. What Treatments to be included in experiment?
2. What Measurements should be made on the experimental units?
3. How should experimental units be selected?
4. How many experimental units should be used?
5. What sampling or experimental design should be used?
6. What is the effect of adjacent experimental units on each other? How can this effect be controlled? (Competition between experimental units leads to dependent data.)

   Feeding animals out of the same container

   Applying various herbicides too close to adjacent plots
7. Outline of pertinent summary tables for recording data.
8. Experimental procedures outlined and documented.
9. Statement of costs in terms of materials, personnel, equipment.

- Consideration of the above items may often result in a restricted experiment, rather than an experiment in which the results are highly incomplete and not very useful. It may be necessary to reduce the number of treatments or increase the number of replications per treatment.

P4. Selection of suitable Measuring Devices and Elimination of Personal Biases and Favoritisms:

1. Never observe 3 samples and discard "most discrepant" observation
2. Never place "Favorite Treatment" under the best experimental conditions
3. Discard data values from abnormal experimental units only after a **critical examination** of the experimental units and a determination of the degree of unsuitability of the results in reference to standard experimental conditions. **Always** report the data values and explain why they were excluded from the analysis.

P5. Carefully evaluate the statistical tests and the necessary conditions needed to apply these tests with respect to experimental procedures and underlying distributional requirements. (Residual analysis to check that assumptions hold.)

P6. Quality of the Final Report:

1. Include well designed graphics
2. Include description of statistical procedures and data collection methodology so that the reader of the report can determine the validity of your experiment and analysis.
3. A report should be prepared whether or not the research hypotheses have been supported by the data; otherwise Type I errors alone may produce misleading conclusions. Many experiments result in the acceptance of the null hypothesis but no report is written. Thus, even when the research hypothesis is in fact false, there may be a few experiments (5% Type I Errors) that support this research hypothesis incorrectly whereas the vast majority of the experiments (95%) in fact find that the research hypothesis is not supported by the data. However, because a report is not written when the null is not rejected, the research hypothesis may incorrectly be supported in the literature due simply to Type I errors in the tests of hypotheses.
4. A second approach to addressing the Type I error problem, is to report the size of the **treatment effect**, for example an estimate of the difference in two treatment means: $\mu_i - \mu_k$. The effect size should be reported and not just the p-value of the test statistic. This will often demonstrate that although the p-value was slightly less than .05 the difference in the treatment means is not of practical significance. Therefore, we should always include both point estimates and confidence intervals on the effect size. Thus, a distinction is being made between **Statistically Significant Results** (small p-value) and **Practically Significant Results** (small p-value with large Treatment effect).

IV. Statistically Designed Experiments are

- Economical
- Allow the measurement of the influence of several factors on a response
- Allow the estimation of the magnitude of experimental variability
- Allow the proper application of statistical inference procedures

EXPERIMENTAL DESIGN TERMINOLOGY

I. Designed Experiment Consists of Three Components:

C1. Method of Randomization:

a. Completely Randomized Design (CRD)
b. Randomized Complete Block Design (RCBD)
c. Balanced Incomplete Block Design (BIBD)
d. Latin Square Design (LSD)
e. Crossover Design
f. Split Plot Design
g. Many others

C2. Treatment Structure

a. One Way Classification
b. Factorial
c. Nested Factors
d. Fractional Factorial
e. Fixed, Random, Mixed factor levels

C3. Measurement Structure

a. Single measurement on experimental unit
b. Repeated measurements on same experimental unit:

   Crossover Design - Different Treatments for each Measurement

   Each EU is observed under all of the t treatments

   Comparing t drugs, each EU is observed under each of the drugs with a washout period between applications of the treatment
c. Repeated measurements on same experimental unit:

   Repeated Measures Design - Longitudinally or Spatially

   Mice are injected with toxic chemical, blood samples taken every 10 hours to determine how long chemical stays in blood

   1-acre plots of land are randomly assigned to be sprayed with specified amounts of herbicide. Measurements are taken at specified locations on the plots to determine the saturation level of the herbicide.
d. Subsampling of experimental unit:

   Measurements on a portion of EU

   1-acre plots of land are randomly assigned to be planted with one of 5 varieties of wheat. Measurements are taken at 10 randomly selected locations on the plots to determine the variation in nitrate levels across the plots.

II. Specific Terms Used to Describe Designed Experiment:

1. **Experimental Unit:** Physical Object to which a treatment is randomly assigned

   Measurement Unit: Physical Object on which measurements or observations are made
2. **Homogeneous Experimental Unit:** Units that are as uniform as possible on all characteristics that could affect the response
3. **Block:** Groups of experimental units which, prior to receiving the treatment, have units within the same block more alike than units in different blocks
4. **Factor:** Aspects of an experiment
   a. which may be under the control of the researcher, such as types of drugs or amount of fertilizer
   b. which occur naturally, such as fertility of soil or slope of land
   c. which may not be of direct interest to the researcher, such as gender, or species or location.

   The levels of the factor are varied in order to determine if the average responses from the EU's differ across the levels of the factor.
5. **Level:** Specific value of a factor
6. **Experimental Region (Factor Space):** All possible factor-level combinations for which experimentation is possible
7. **Treatment:** A single level of a factor or a combination of the levels of two or more factors.
8. **Replication:** Observations on two or more experimental units which have been randomly assigned to the same treatment and are observed under the SAME experimental conditions.

9.A **Sub sampling:** Multiple measurements at randomly selected times or locations on the same experimental unit under the same treatment

9.B **Repeated Measurements:** Multiple measurements at specified locations or times on the same experimental unit under the same treatment

10. **Response:** Outcome or result of an experiment; Either measured or observed
11. **Effect:** Change in the average response between two factor-level combinations or between two experimental conditions: $\mu_{ij} - \mu_{kh}$
12. **Interaction:** Existence of joint factor effects in which the effect of each factor depends on the levels of the other factors
13. **Confounding:** One or more effects that cannot unambiguously be attributed to a single factor or interaction using the observed data. It may be possible to eliminate the confounding in future experiments by using a more complex design.
14. **Covariate:** An uncontrollable variable that influences the response but is unaffected by any other experimental factors. For example Weather conditions - temperature, humidity, wind, amount of sunlight, etc. ; Differences in EU prior to assigning treatments - health, age, size, ability, pretest scores, etc. ; Differences in plots of land - record soil conditions prior to applying treatments

EXAMPLE #1

A semi-conductor manufacturer is having problems with scratching on their silicon wafers. They propose applying a protective coating to the wafers, however, the wafer engineers are concerned about the diminished performance of the wafer. An experiment is designed to evaluate several types and thicknesses of coatings on the conductivity of the wafer. Two types of coatings and three thicknesses of the coating are selected for experimentation. A random sample of 54 wafers are selected for use in the experiment with 9 wafers randomly assigned to each combination of a type of coating (C_1, C_2) and a thickness of coating (T_1, T_2, T_3). Only 24 wafers can be evaluated on a given day. Thus, the engineers each day test either 2, 3, or 4 wafers under each of the coating types-thicknesses combinations. On each wafer, the conductivity is recorded before and after applying the coating to the wafer. Furthermore, to assess the variability in conductivity across the wafer surface, conductivity readings are taken at five locations on each wafer.

Treatment	Day 1				Day 2			Day 3	
C_1T_1	W_1	W_2	W_3	W_4	W_{25}	W_{26}	W_{27}	W_{43}	W_{44}
C_1T_2	W_5	W_6	W_7	W_8	W_{28}	W_{29}	W_{30}	W_{45}	W_{46}
C_1T_3	W_9	W_{10}	W_{11}	W_{12}	W_{31}	W_{32}	W_{33}	W_{47}	W_{48}
C_2T_1	W_{13}	W_{14}	W_{15}	W_{16}	W_{34}	W_{35}	W_{36}	W_{49}	W_{50}
C_2T_2	W_{17}	W_{18}	W_{19}	W_{20}	W_{37}	W_{38}	W_{39}	W_{51}	W_{52}
C_2T_3	W_{21}	W_{22}	W_{23}	W_{24}	W_{40}	W_{41}	W_{42}	W_{53}	W_{54}

For the above experiment, identify all the components and definitions:

I. Designed Experiment Consists of Three Components:

C1. Method of Randomization:

C2. Treatment Structure:

C3. Measurement Structure:

II. Specific Terms Used to Describe Designed Experiment:

1a. **Experimental Unit:**

1b. **Measurement Unit:**

2. **Block:**

3. **Homogeneous Experimental Unit:**

4. **Factor:**

5. **Level:**

6. **Experimental Region (Factor Space):**

7. **Treatment:**

8. **Replication:**

9. **Subsampling:**

10. **Response:**

11. **Effect:**

12. **Interaction:**

12. **Interaction: (continued)**

13. **Confounding:**

14. **Covariate:**

OTHER POSSIBLE WAYS OF CONDUCTING THE WAFER EXPERIMENT

Scenario I: All 54 wafers are evaluated on the same day. Nine wafers are randomly assigned to each of the 6 treatments $((C_i, T_j), i = 1, 2; j = 1, 2, 3)$. The conductivity readings are all done in the same lab under essentially identical conditions. The application of the coatings to the 54 wafers would be done in a random fashion by randomly permuting the numbers 1 to 54 and then applying the coatings in the order of the permutation. The type of randomization is denoted as a completely randomized design (CRD).

Treatment	Day 1								
C_1T_1	W_1	W_2	W_3	W_4	W_{25}	W_{26}	W_{27}	W_{43}	W_{44}
C_1T_2	W_5	W_6	W_7	W_8	W_{28}	W_{29}	W_{30}	W_{45}	W_{46}
C_1T_3	W_9	W_{10}	W_{11}	W_{12}	W_{31}	W_{32}	W_{33}	W_{47}	W_{48}
C_2T_1	W_{13}	W_{14}	W_{15}	W_{16}	W_{34}	W_{35}	W_{36}	W_{49}	W_{50}
C_2T_2	W_{17}	W_{18}	W_{19}	W_{20}	W_{37}	W_{38}	W_{39}	W_{51}	W_{52}
C_2T_3	W_{21}	W_{22}	W_{23}	W_{24}	W_{40}	W_{41}	W_{42}	W_{53}	W_{54}

Scenario II: A maximum of 24 wafers can be evaluated on the same day thus requiring 3 days to complete the experiment. On Day 1, 4 wafers are randomly assigned to each of the 6 treatments: $(C_i, T_j), i = 1, 2; j = 1, 2, 3$, on Day 2, 3 wafers are randomly assigned to each of the 6 treatments, and on Day 3, 2 wafers are randomly assigned to each of the 6 treatments. The conductivity readings are all done in the same lab under essentially identical conditions. On each of the 3 Days, the application of the coatings to the wafers would be done in a random fashion by randomly permuting the numbers 1 to 24 for Day 1, 25 to 42 for Day 2, 43 to 54 for Day 3 and then applying the coatings in the order of the permutation. The type of randomization is denoted as a randomized block design (RCBD), with Day being the Blocking Factor.

Treatment	Day 1				Day 2			Day 3	
C_1T_1	W_1	W_2	W_3	W_4	W_{25}	W_{26}	W_{27}	W_{43}	W_{44}
C_1T_2	W_5	W_6	W_7	W_8	W_{28}	W_{29}	W_{30}	W_{45}	W_{46}
C_1T_3	W_9	W_{10}	W_{11}	W_{12}	W_{31}	W_{32}	W_{33}	W_{47}	W_{48}
C_2T_1	W_{13}	W_{14}	W_{15}	W_{16}	W_{34}	W_{35}	W_{36}	W_{49}	W_{50}
C_2T_2	W_{17}	W_{18}	W_{19}	W_{20}	W_{37}	W_{38}	W_{39}	W_{51}	W_{52}
C_2T_3	W_{21}	W_{22}	W_{23}	W_{24}	W_{40}	W_{41}	W_{42}	W_{53}	W_{54}

There is a restriction on the randomization in that the 54 wafers are first randomly assigned to the 3 days with 24 assigned to Day 1, 18 to Day 2, and 12 to Day 3. Then, within each day, a randomization of wafers to the 6 treatments is implemented.

Scenario III: Suppose only 2 wafers can be evaluated on the same day in a given lab. Thus to reduce the number of days to complete the experiment, 6 different labs are used. Two wafers are randomly assigned to each of the 6 treatments in each of the 6 labs on each of the 6 days of the experiment resulting in the use of 72 wafers in the experiment. The randomization is such that 2 wafers are observed under each treatment on every Day in every Lab. The type of randomization is denoted as a Latin Square Design (LSD) with 2 replications for each of the 36 Day-Lab combinations.

Lab	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
1	$C_1T_1 - W_1, W_2$	$C_1T_2 - W_3, W_4$	$C_1T_3 - W_5, W_6$	$C_2T_1 - W_7, W_8$	$C_2T_2 - W_9, W_{10}$	$C_2T_3 - W_{11}, W_{12}$
2	$C_1T_2 - W_{13}, W_{14}$	$C_1T_3 - W_{15}, W_{16}$	$C_2T_1 - W_{17}, W_{18}$	$C_2T_2 - W_{19}, W_{20}$	$C_2T_3 - W_{21}, W_{22}$	$C_1T_1 - W_{23}, W_{24}$
3	$C_1T_3 - W_{25}, W_{26}$	$C_2T_1 - W_{27}, W_{28}$	$C_2T_2 - W_{29}, W_{30}$	$C_2T_3 - W_{31}, W_{32}$	$C_1T_1 - W_{33}, W_{34}$	$C_1T_2 - W_{35}, W_{36}$
4	$C_2T_1 - W_{37}, W_{38}$	$C_2T_2 - W_{39}, W_{40}$	$C_2T_3 - W_{41}, W_{42}$	$C_1T_1 - W_{43}, W_{44}$	$C_1T_2 - W_{45}, W_{46}$	$C_1T_3 - W_{47}, W_{48}$
5	$C_2T_2 - W_{49}, W_{50}$	$C_2T_3 - W_{51}, W_{52}$	$C_1T_1 - W_{53}, W_{54}$	$C_1T_2 - W_{55}, W_{56}$	$C_1T_3 - W_{57}, W_{58}$	$C_2T_1 - W_{59}, W_{60}$
6	$C_2T_3 - W_{61}, W_{62}$	$C_1T_1 - W_{63}, W_{64}$	$C_1T_2 - W_{65}, W_{66}$	$C_1T_3 - W_{67}, W_{68}$	$C_2T_1 - W_{69}, W_{70}$	$C_2T_2 - W_{71}, W_{72}$

Scenario IV: A new machine used to apply the coating to the wafers has recently been purchased. This machine requires a considerable amount of time in order to change from applying coating type C_1 to C_2 but almost no set-up time for changing from one thickness to another thickness. Therefore, the engineers want to apply all three thicknesses of coating C_1 and then apply all three thicknesses of coating C_2 rather than doing the applications in a random fashion. This will save them considerable amount of set-up time. Furthermore, only 24 wafers can be coated in a given day and only 1 lab is available for the experiment. Therefore, the following randomization was conducted. On a given day, 12 wafers were randomly assigned to each of the two coatings. Then, 4 of these 12 wafers were randomly assigned to each of the three thicknesses. The randomization was repeated on each of the three days needed to complete the experiment. The type of randomization is a Randomized Complete Block Design with a Split-Plot treatment assignment.

Coating	Thickness	Day 1	Day 2	Day 3
C_1	T_1	W_1,W_2,W_3,W_4	W_{25},W_{26},W_{27},W_{28}	W_{49},W_{50},W_{51},W_{52}
	T_2	W_5,W_6,W_7,W_8	W_{29},W_{30},W_{31},W_{32}	W_{53},W_{54},W_{55},W_{56}
	T_3	W_9,W_{10},W_{11},W_{12}	W_{33},W_{34},W_{35},W_{36}	W_{57},W_{58},W_{59},W_{60}
C_2	T_1	W_{13},W_{14},W_{15},W_{16}	W_{37},W_{38},W_{39},W_{40}	W_{61},W_{62},W_{63},W_{64}
	T_2	W_{17},W_{18},W_{19},W_{20}	W_{41},W_{42},W_{43},W_{44}	W_{65},W_{66},W_{67},W_{68}
	T_3	W_{21},W_{22},W_{23},W_{24}	W_{45},W_{46},W_{47},W_{48}	W_{69},W_{70},W_{71},W_{72}

COMMON PROBLEMS IN EXPERIMENTAL DESIGNS

I. Masking of Factor Effects

When the variation in the responses are as large as the differences in the treatment means, the treatment differences will not be detected in the experiment. For example, σ_e is large relative to $\mu_i - \mu_k$ in a completely randomized design. In this situation, the experiment must be redesigned to reduce

$\text{StDev}(\hat{\mu}_i - \hat{\mu}_k) = \sigma_e\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}$ by

1. Increasing the sample sizes n_i, n_k to reduce $\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}$
2. Blocking the experimental units to reduce the size of σ_e

- Create k blocks of EU's where the variance in responses within the blocks, σ_e^* is less than the variance in the responses without the blocking, $\sigma_e^* < \sigma_e$

3. Using Covariates to reduce the size of σ_e

II. Uncontrolled Factors

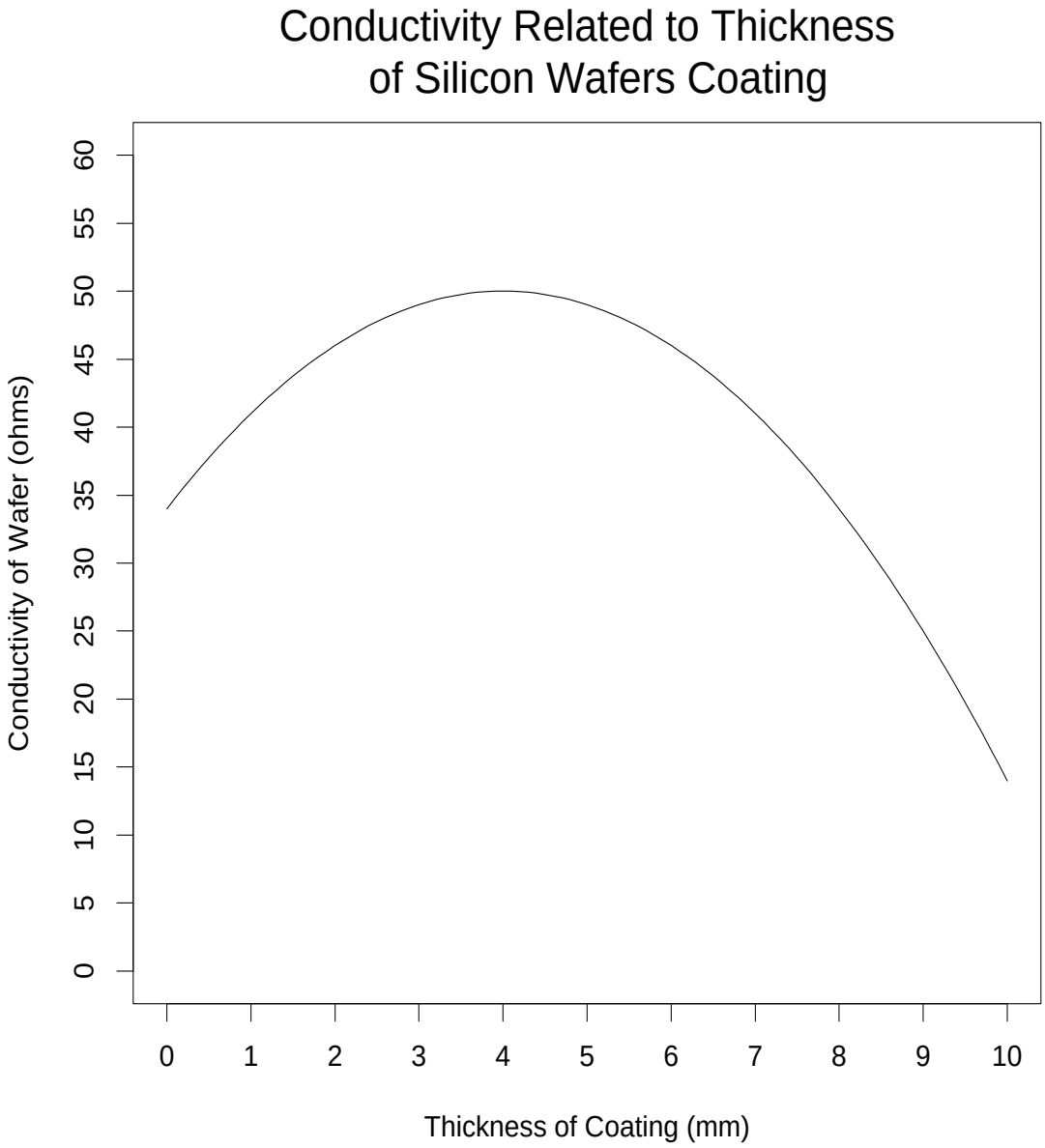
If factors are known to have an effect on the response variable, then these factors should be included in the experiment as either treatment factors or blocking variables. Failure to carefully consider all factors of importance can greatly compromise the extent to which conclusions can be drawn from the experimental outcomes.

1. Differences between experimental plots in terms of soil fertility, drainage, exposure to sun, exclusion of wildlife, etc.
2. Position of experimental units on greenhouse benches
3. Position of experimental units on trays or in ovens
4. Time of day or lab in which experiment is run
5. Varying health condition, breed, age, size of experimental animals

III. Erroneous Principles of Efficiency

If the time to run experiments or the cost to run experiments place restrictions on the number of factors and the number of levels of the factors that can be included in the experiment, then the overall goals of the experiment must be reevaluated since

1. Important factors may be ignored or left uncontrolled
2. Non-linear effects may not be determined since the number of levels may be too few or not broad enough to detect higher order effects.



Consider the following 3 Designs: (the experimenter does not know the true shape of the curve relating Conductivity to Thickness)

Design 1: Select thicknesses $T = 1, 10$ for the experiment

Design 2: Select thicknesses $T = 3, 4, 5$ for the experiment

Design 3: Select thicknesses $T = 1, 2, 4, 9$ for the experiment

SELECTING AN APPROPRIATE EXPERIMENTAL DESIGN

I. Consideration of Objectives

1. Nature of anticipated results helps to determine what factors need to be included in the experiment:

   Suppose an experiment is designed to determine which of 6 fuel blends used in automobiles produce the lowest CO emissions. The 6 blends include a standard commercial gasoline and 5 different methanol blends. After determining that blend number 5 has the lowest CO emission, the question arises what properties of the blends (distillation temperature, specific gravity, oxygen content, etc.) made the major contributions to the reduced CO level in emissions using the selected blend. A problem that may arise is that the fuel properties may be confounded across the 5 blends and it may not be possible to sort them out with the given experimental runs. This problem could have been avoided if this question was raised prior to running the experiments.

2. Definition of concepts (Can the goals of the experiment be achieved) :

   Suppose we want to study the effects of radiation exposure on the life length of humans

   - Design 1: Subject randomly selected homogeneous groups of humans to various levels of radiation (unethical experiment)
   - Design 2: Use laboratory rats in place of humans (extrapolation problem)
   - Design 3: Use observational or historical data on groups that were exposed to radiation
     (Many uncontrolled factors, genetic differences, amount of exposure, length of exposure, occupational differences, daily habits)

3. Determination of observable variables

   - What covariates should be observed?
     Temperature, humidity, rainfall, weight, body temperature, blood pressure
   - How often?
     Every 15 minutes, hourly, daily, weekly
   - How accurately should they be measured?
     Nearest inch, cm, mm

II. Factor Effects

1. Inclusion of all relevant factors avoids uncontrolled systematic variation and possible confounding with factors not recorded in the study.

- A study of effectiveness of new treatment for heart worms is conducted using dogs from a kennel. The new drug and old drug are randomly assigned to 30 dogs each. New drug seems more effective until further study reveals that 20 of the 30 dogs receiving the new drug are a breed of dogs which has a natural resistance whereas only 5 of the 30 dogs receiving the old drug were of that breed.

2. Need to measure all important covariates to control heterogeneity of experimental units or environmental conditions in lab. For example, physiological differences in subjects; temperature, humidity, air quality in a greenhouse study; fertility of soil in an agricultural field trial.
3. Anticipated interrelationships between factor levels helps to determine type of design:
   a. No interactions between factor levels: Use simple screening design
   b. Interactions exist: Need full factorial design
   c. Higher order relationships between factor levels may require a greater number of levels of the factors in order to be able to fit high order polynomials to the responses.
4. Include a broad enough range of the factor levels so as not to miss important factor effects, include lowest and highest feasible values of factor.

III. Precision - Efficiency of Experiment

Degree of variability in response variable σ_e determines the number of replications required to obtain desired widths of confidence intervals and power of statistical tests. Determine variability through pilot studies or review literature for results from similar experiments.

IV. Randomization

In order to protect against unknown sources of biases and to be able to conduct valid statistical procedures:

1. The experimental units **MUST** be randomly assigned to the treatments or
2. The experimental units **MUST** be randomly selected from the treatment populations and
3. The time order in which experiments are run and/or spatial positioning of experimental units must be randomly assigned to the various treatments. This avoids the confounding of uncontrolled factor effects with the experimental factors. For example, drifts in instrumental readings, variation across the day in terms of temperature gradients, humidity or sunlight exposure, variation in performance of laboratory technicians (grad students), or various other conditions in the laboratory or field.

DESIGNING FOR QUALITY: INDUSTRIAL PROCESSES

Two Basic Types of Experiments

1. On-Line: Running experiments while process is in full production.

   EVOP - Evolutionary Operation

   Design strategy where 2 or more factors in an on-going production process are varied in order to determine an optimal operation level.

   Problem: Examining very narrow region of the factor space since only small deviations from *normal operations* are allowed by the company.

2. Off-Line: Running experiments in Laboratories or Pilot Plants

Two Basic Goals in Experiments Involving Quality Improvement

1. Bring product On Target

   Average measurement of product characteristic are equal to the target value

2. Uniformity - Consistency

   Measured product characteristics have a small variability about the target value

   Combining both of these criterions, we obtain

   $$\text{Minimize MSE} = (Bias)^2 + (StDev)^2 = (\text{Distance to Target})^2 + \text{Variance}$$

Taguchi Approach:

1. Emphasized the importance of using fractional factorial designs
2. His choice of designs were often highly inefficient
3. His analysis of experiments were often incorrect
4. He was successful in convincing engineers at large corporations to use designed experiments. The experiments were very successful even though there were not the best possible experiments that could have been run.

Proper Randomization for Valid Inferences

1. Randomization provides the justification for statistical inferences: Point Estimation, Confidence Intervals, Tests of Hypotheses

2. Proper replication provides sufficient data to estimate

   - the degree of experimental variation, σ_e
   - the mean response from each of the treatments, μ_is
   - the size of the treatment effects, $\mu_i - \mu_k$, across all pairs of means

3. The Independence Assumption can in most cases be justified through an examination of how the experiment was conducted. The physical proximity of EU's and relationships between EU's that existed prior to the start of the experiment may result in dependencies in the measured/observed responses. A proper randomization will often assist in overcoming such problems.

   - Plants placed too close together on greenhouse benches
   - Animals place in the same cages and pens
   - Children in the same classroom sharing resources

4. When the data values are not independent, as will occur when there are repeated measurements either temporally or spatially, the covariance structure of the data must be included as part of the model.

   - Record blood pressure of patient every 15 minutes for 4 hours after patient receives injection
   - Measure moisture depth every 10 cm from an experimental irrigation device
   - Treat 100 fire ant hills with chemical and then measure the weight of 20 randomly selected ants taken from each hill 5 days after chemical was applied.

5. Randomization also assists in overcoming biases due to a systematic assignment of EU's to treatments or a systematic order in which the treatments are observed in the experiment. The systematic assignment or ordering may provide an advantage to one treatment over another treatment.

   - Researcher decides which animals receive new feed supplement and which animals receive standard feed
   - Research Psychologist assigns 20 patients to new therapy and 20 patients to old therapy based on his judgement of which patients will best respond to new therapy.

Permutation Tests

If the randomization has been conducted properly, it is possible to conduct a statistical test for differences with almost no restrictions on the distributional properties of the responses. The conditions are that the population distributions are identical except for a possible shift in the location parameter. That is, the populations have cdf's from the same family of cdf's but the populations may have different location parameters.

Suppose we have t treatments: $A_i, i = 1, \ldots, t$. We wish to test if there is a difference in treatment mean responses:

$$H_o : \mu_{A_1} = \mu_{A_2} = \ldots = \mu_{A_t} \quad \text{versus} \quad H_1 : \quad \text{not all} \quad \mu_{A_i} \quad \text{are equal.}$$

Suppose we have $n = n_1 + n_2 + \ldots + n_t$ EU's with n_i EU's randomly assigned to treatment A_i.

The populations are identical except for except for possible location differences. Thus under H_o, no differences in location parameters, the n responses are independent and identical distributed. Thus, the treatments are just labels on the EU's.

Under H_o, we can then consider the n responses to be a random sample of size n from a single population (the t treatment populations are identical under H_o). Thus, which responses are labeled as from A_i or from A_j should have no effect on the sampling distribution of the test statistic if in fact H_o is true.

If H_1 is true, then certainly the switching of responses from one treatment to another would alter the sampling distribution of the test statistic because we would be assigning responses from a population having a small mean to a population having a large mean.

By considering all possible assignments of the n responses to t treatments, we can obtain the p-value of the test statistic as follows. Recall, the p-value is the probability, assuming that H_o is true, of obtaining a value of the test statistic which is as extreme or more extreme to H_o as is the value of the test statistic computed from the observed data.

Computation of p-value for Permutation Test:

1. Let N_p be the number of possible arrangements of the treatment labels to the responses

$$N_p = \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_{t-1}+n_t}{n_{t-1}}\binom{n_t}{n_t} = \frac{n!}{n_1!n_2!\cdots n_t!}$$

2. Compute the value of the test statistic for each of these assignments

3. Count the number, N_S, of assignments having a value of the test statistic as extreme or more extreme to H_o than the value of the test statistic computed from the original data.

4. p-value = N_S/N_p

Example with $t = 2$

Suppose we have two treatments A_1 and A_2 with $n_1 = 4$ and $n_2 = 3$ EU's randomly assigned to the two treatments, respectively. The following responses were obtained along with the value of the test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Unit	1	2	3	4	5	6	7		
Treatment	A_2	A_2	A_1	A_1	A_2	A_1	A_1	$\bar{y}_1 - \bar{y}_2$	t
Response	14	16	19	17	15	13	17	1.50	0.958

Next, we need to consider all the $N_P = \binom{7}{4} = 35$ arrangements of the 7 responses to the two treatment labels, A_1 and A_2.

	Response										Response								
AR	14	16	19	17	15	13	17	$\bar{y}_1 - \bar{y}_2$	t	AR	14	16	19	17	15	13	17	$\bar{y}_1 - \bar{y}_2$	t
1	A_1	A_1	A_2	A_2	A_1	A_1	A_2	-3.17	-3.348	19	A_1	A_1	A_1	A_2	A_1	A_2	A_2	0.33	0.196
2	A_1	A_2	A_2	A_1	A_1	A_1	A_2	-2.58	-2.065	20	A_1	A_1	A_2	A_1	A_2	A_2	A_1	0.33	0.196
3	A_1	A_2	A_2	A_2	A_1	A_1	A_1	-2.58	-2.065	21	A_2	A_2	A_1	A_1	A_1	A_1	A_2	0.33	0.196
4	A_1	A_1	A_2	A_1	A_2	A_1	A_2	-2.00	-1.380	22	A_2	A_2	A_1	A_2	A_1	A_1	A_1	0.33	0.196
5	A_1	A_1	A_2	A_2	A_2	A_1	A_1	-2.00	-1.380	23	A_1	A_2	A_1	A_1	A_1	A_2	A_2	0.92	0.555
6	A_1	A_2	A_1	A_2	A_1	A_1	A_2	-1.42	-0.896	24	A_1	A_2	A_1	A_2	A_1	A_2	A_1	0.92	0.555
7	A_1	A_2	A_2	A_1	A_2	A_1	A_1	-1.42	-0.896	25	A_2	A_1	A_1	A_1	A_2	A_1	A_2	0.92	0.555
8	A_2	A_1	A_2	A_1	A_1	A_1	A_2	-1.42	-0.896	26	A_2	A_1	A_1	A_2	A_2	A_1	A_1	0.92	0.555
9	A_2	A_1	A_2	A_2	A_1	A_1	A_1	-1.42	-0.896	27	A_2	A_1	A_2	A_1	A_1	A_2	A_1	0.92	0.555
10	A_1	A_1	A_1	A_2	A_2	A_1	A_2	-0.83	-0.502	28	A_1	A_1	A_1	A_1	A_2	A_2	A_2	1.50	0.958
11	A_1	A_1	A_2	A_1	A_1	A_2	A_2	-0.83	-0.502	29	A_1	A_1	A_1	A_2	A_2	A_2	A_1	1.50	0.958
12	A_1	A_1	A_2	A_2	A_1	A_2	A_1	-0.83	-0.502	30	A_2	A_2	A_1	A_1	A_2	A_1	A_1	1.50	0.958
13	A_2	A_2	A_2	A_1	A_1	A_1	A_1	-0.83	-0.502	31	A_1	A_2	A_1	A_1	A_2	A_2	A_1	2.08	1.462
14	A_1	A_2	A_1	A_1	A_2	A_1	A_2	-0.25	-0.147	32	A_2	A_1	A_1	A_1	A_1	A_2	A_2	2.08	1.462
15	A_1	A_2	A_1	A_2	A_2	A_1	A_1	-0.25	-0.147	33	A_2	A_1	A_1	A_2	A_1	A_2	A_1	2.08	1.462
16	A_1	A_2	A_2	A_1	A_1	A_2	A_1	-0.25	-0.147	34	A_2	A_2	A_1	A_1	A_1	A_2	A_1	2.67	2.194
17	A_2	A_1	A_1	A_2	A_1	A_1	A_2	-0.25	-0.147	35	A_2	A_1	A_1	A_1	A_2	A_2	A_1	3.25	3.662
18	A_2	A_1	A_2	A_1	A_2	A_1	A_1	-0.25	-0.147										

Note that arrangement #30 is the arrangement for the observed experiment.

Next, we count the number of arrangements, N_s having magnitude of the test statistic equal to or greater than the magnitude associated with the actual experiment:

$$N_S = \sum_{i=1}^{35} I(|t| \geq 0.958) = 13$$

Thus, the p-value $= \frac{N_S}{N_P} = \frac{13}{35} = 0.3714$

Compare this value to the value obtained from the t test: (which requires both treatment responses have a normal distribution and $\sigma_{A_1} = \sigma_{A_2}$)

$$p - value = 2P(t_5 \geq 0.958) = 2[1 - pt(.958, 5)] = 0.3831$$

In this case, the values are nearly identical.

Why are the t values identical for all arrangements having the save value for $\bar{y}_1 - \bar{y}_2$?

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\Delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{1}{n_1 + n_2 - 2}\left[C_2 - \left(n_1\left(\frac{n_2\Delta + C_1}{n_1 + n_2}\right)^2 + n_2\left(\frac{C_1 - n_1\Delta}{n_1 + n_2}\right)^2\right)\right]$$

Therefore, t is a function of $n_1, n_2, \Delta, C_1, C_2$ where

$\Delta = \bar{y}_1 - \bar{y}_2$

$C_1 = n_1\bar{y}_1 + n_2\bar{y}_2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}$

$C_2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}^2$

Note that C_1 and C_2 are the same value for all arrangements of the data values. Therefore, across the various arrangements of the data, the value of t depends only on the values of $\Delta = \bar{y}_1 - \bar{y}_2$.

I found the following article by John Timmer on line and thought you may enjoy it.

We're so good at medical studies that most of them are wrong

It's possible to get the mental equivalent of whiplash from the latest medical findings, as risk factors are identified one year and exonerated the next. According to a panel at the American Association for the Advancement of Science, this isn't a failure of medical research; it's a failure of statistics, and one that is becoming more common in fields ranging from genomics to astronomy. The problem is that our statistical tools for evaluating the probability of error haven't kept pace with our own successes, in the form of our ability to obtain massive data sets and perform multiple tests on them. Even given a low tolerance for error, the sheer number of tests performed ensures that some of them will produce erroneous results at random.

The panel consisted of Suresh Moolgavkar from the University of Washington, Berkeley's Juliet P. Shaffer, and Stanley Young from the National Institute of Statistical Sciences. The three gave talks that partially overlapped, at least when it came to describing the problem, so it's most informative to tackle the session at once, rather than by speaker.

Why we can't trust most medical studies Statistical validation of results, as Shaffer described it, simply involves testing the null hypothesis: that the pattern you detect in your data occurs at random. If you can reject the null hypothesisand science and medicine have settled on rejecting it when there's only a five percent or less chance that it occurred at randomthen you accept that your actual finding is significant.

The problem now is that we're rapidly expanding our ability to do tests. Various speakers pointed to data sources as diverse as gene expression chips and the Sloan Digital Sky Survey, which provide tens of thousands of individual data points to analyze. At the same time, the growth of computing power has meant that we can ask many questions of these large data sets at once, and each one of these tests increases the prospects than an error will occur in a study; as Shaffer put it, "every decision increases your error prospects." She pointed out that dividing data into subgroups, which can often identify susceptible subpopulations, is also a decision, and increases the chances of a spurious error. Smaller populations are also more prone to random associations.

In the end, Young noted, by the time you reach 61 tests, there's a 95 percent chance that you'll get a significant result at random. And, let's face itresearchers want to see a significant result, so there's a strong, unintentional bias towards trying different tests until something pops out.

Young went on to describe a study, published in JAMA, that was a multiple testing train wreck: exposures to 275 chemicals were considered, 32 health outcomes were tracked, and 10 demographic variables were used as controls. That was about 8,800 different tests, and as many as 9 million ways of looking at the data once the demographics were considered.

The problem with models Both Young and Moolgavkar then discussed the challenges of building a statistical model. Young focused on how the models are intended to help eliminate bias. Items like demographic information often correlate with risks of specific health outcomes, and researchers need to adjust for those when attempting to identify the residual risk associated with any other factors. As Young pointed out, however, you're never going to know all the possible risk factors, so there will always be error that ends up getting lumped in with whatever you're testing.

Moolgavkar pointed out a different challenge related to building the statistical models: even the same factor can be accounted for using different mathematical means. The models also make decisions on how best handle things like measuring exposures or health outcomes. The net result is that two models can be fed an identical dataset, and still produce a different answer.

At this point, Moolgavkar veered into precisely the issues we covered in our recent story on scientific

reproducibility: if you don't have access to the models themselves, you won't be able to find out why they produce different answers, and you won't fully appreciate the science behind what you're seeing.

Consequences and solutions It's pretty obvious that these factors create a host of potential problems, but Young provided the best measure of where the field stands. In a survey of the recent literature, he found that 95 percent of the results of observational studies on human health had failed replication when tested using a rigorous, double blind trial. So, how do we fix this?

The consensus seems to be that we simply can't rely on the researchers to do it. As Shaffer noted, experimentalists who produce the raw data want it to generate results, and the statisticians do what they can to help them find them. The problems with this are well recognized within the statistics community, but they're loath to engage in the sort of self-criticism that could make a difference. (The attitude, as Young described it, is "We're both living in glass houses, we both have bricks.")

Shaffer described how there were tools (the "family-wise error rate") that were once used for large studies, but they were so stringent that researchers couldn't use them and claim much in the way of positive results. The statistics community started working on developing alternatives about 15 years ago but, despite a few promising ideas, none of them gained significant traction within the community.

Both Moolgavkar and Young argued that the impetus for change had to come from funding agencies and the journals in which the results are published. These are the only groups that are in a position to force some corrections, such as compelling researchers to share both data sets and the code for statistical models.

Moolgavkar also made a forceful argument that journal editors and reviewers needed to hold studies to a minimal standard of biological plausibility. Focusing on studies of the health risks posed by particulates, he described studies that indicated the particulates in a city were as harmful as smoking 40 cigarettes daily; another concluded that particulates had a significant protective effect when it comes to cardiovascular disease. "Nobody is going to tell you that, for your health, you should go out and run behind a diesel bus," Moolgavkar said. "How did this get past the reviewers?"

But, in the mean time, Shaffer seemed to suggest that we simply have to recognize the problem and communicate it with the public, so that people don't leap to health conclusions each time a new population study gets published. Medical researchers recognize the value of replication, and they don't start writing prescriptions based on the latest gene expression studythey wait for the individual genes to be validated. As we wait for any sort of reform to arrive, caution, and explaining to the public the reasons for this caution, seems like the best we can do.