Principles of Experimental Design for Art Conservation Research

Terry J. Reedy

Chandra L. Reedy

GCI SCIENTIFIC PROGRAM REPORT

JANUARY 1992

Stat/Consul
3 Stage Road
Newark, Delaware 19711

University of Delaware
Art Conservation Department
303 Old College
Newark, Delaware 19716

The Getty Conservation Institute
4503 Glencoe Avenue
Marina del Rey, California 90292

Principles of Experimental Design for Art Conservation Research

Terry J. Reedy

Chandra L. Reedy

GCI SCIENTIFIC PROGRAM REPORT
JANUARY 1992

Stat/Consul
3 Stage Road
Newark, Delaware 19711

University of Delaware
Art Conservation Department
303 Old College
Newark, Delaware 19716

The Getty Conservation Institute
4503 Glencoe Avenue
Marina del Rey, California 90292

Cover printed on Simpson Quest,
a recycled stock using 100%
post-consumer waste

CONTENTS

THE AUTHORS

Terry J. Reedy

Dr. Terry J. Reedy has degrees in mathematics, operations research, and ecology. He was a consulting statistician from 1979 to 1989 in the Biomathematics Unit of the Center for Ulcer Research and Education in the School of Medicine, University of California at Los Angeles. This work gave him broad experience with the practical problems of data analysis in scientific research. He also has experience in the problems of statistics in archaeometry, art history and conservation research. He currently works as an independent consultant and writer.

Chandra L. Reedy

Dr. Chandra L. Reedy received her Ph.D. in Archaeology from the University of California at Los Angeles in 1986, where her areas of specialization were materials analysis and South Asian art and archaeology. She did conservation research at the Los Angeles County Museum of Art from 1986 to 1989. She currently coordinates the Ph.D. program and teaches in the Art Conservation Department of the University of Delaware. Her particular interests are regional provenance studies and the role of experimental design and statistical analysis in the study of art and art conservation.

PREFACE

The subject of this report is experimental design for art conservation research. It covers both practical and statistical aspects of design, and both laboratory experiments on art materials and clinical experiments with art objects. General principles developed in other fields are applied to concrete research problems in conservation. Problems encountered in conservation practice illustrate the points discussed. Most of the material should be comprehensible to working conservators and conservation scientists alike.

Chapters 1 and 2 set the scene by discussing conservation research and the scientific method. In broad terms, an experiment is planned, performed, analyzed, and reported. The remaining chapters generally focus on the design phase of research, which comes before the performance and analysis phases. They explore possibilities for designing experiments in art conservation but do not evaluate current practice.

Chapter 3 presents designs for experiments with single objects and the corresponding treatment randomization tests. Both subjects are ignored in most books on experimental design and statistical analysis, but they are especially pertinent to conservation research. The designs formalize and extend the current practice of testing treatments on small patches before treating a whole object. The corresponding statistical tests are comprehensible without esoteric assumptions and mathematical derivations. This material should be especially accessible to conservators without training in research and statistics.

Chapter 4 systematically examines the major aspects of design: goals, objects, measurements, and treatments. In chapter 7, this scheme is used to discuss a particular design: treatment trials directly comparing two or more treatments on art objects. The authors believe that more extensive and more rigorous use of this design would benefit conservation practice.

More traditional material on experiments with groups of objects and statistical tests based on normal distributions and ranks is presented in chapters 5 and 6. These chapters do not duplicate the detail found in books on these subjects. However, the preponderance of scientific experiments use the simpler designs given the most attention in chapter 6. This chapter ends with a work sheet that readers can use to design similar studies (section 6.5).

In spite of the focus on design, there is some material on statistical data analysis. One reason is that an initial plan of analysis is part of the design. Chapters 6 and 7, especially, identify for each design statistical techniques that might be part of such a plan. Readers without previous statistical experience or training will need additional help from a statistical expert to actually do the analysis after performing the experiment. Knowing the name of a test or procedure will facilitate finding the relevant section of a statistics book or statistical program manual and may also be useful when consulting with a statistician (see section 5.4).

Another reason for including statistical material is that other parts of the design are governed by statistical considerations. It is the post-experiment analysis that will actually answer the question that motivates the design. Some statistical discussions, especially in Chapter 5, give optional background for understanding. Some require previous statistical knowledge and can be skipped by those without. The STATISTICAL GLOSSARY & INDEX briefly defines many statistical terms and gives page references to their use in the text.

Our previous study and technical report, *Statistical Analysis in Art Conservation Research* (Reedy and Reedy 1988), reviewed 320 papers from the conservation literature, covering five years of four publications. It critically evaluated presentation of designs, data, and statistical analyses in this body of published art conservation research. Though written first, it is logically a sequel to this work. It gives several examples of statistical analysis of art conservation data and discusses issues of data organization and plotting that are not covered here.

This report results from a conceptual investigation combined with experience teaching this material to art conservation graduate students. It builds on and is colored by the senior author's statistical experience, which is primarily laboratory and clinical research at a university medical school. An agricultural or industrial statistician would certainly emphasize different topics. By its size, it is more of an overview than a detailed compendium. It is an opening statement rather than a closing argument. The authors invite reasoned reader responses.

Acknowledgments

This work was sponsored by the Getty Conservation Institute scientific research program. We thank the students of the Art Conservation Program of the University of Delaware and Winterthur Museum for listening to earlier drafts of this material. We also thank Laurie German, Hope Gumprecht, Eric Hansen, Jim Druzik, and Mary Striegel for their review and comments.

CHAPTER 1

CONSERVATION SCIENCE AND PRACTICE

1.1 CONSERVATION RESEARCH

Art conservation practice combines philosophy, skill, and knowledge. Philosophy covers the goals and evaluation of conservation practice, including aesthetics, questions of restoration versus preservation, and questions of reversibility versus permanence. Skill comes from hands-on practice obtained in apprenticeships and training programs. Knowledge can be communicated but is originally derived from experience, either accidental or deliberate. This book discusses the design of intentional research projects aimed at increasing the knowledge of objects, materials, and processes that forms the basis of art conservation practice.

The purpose of research is to answer questions. In the field of conservation, theoretical questions involve some aspect of conservation science. Their answers add to a developing body of scientific knowledge. Practical questions deal with some phase of a conservation project. Such technological questions and their answers help achieve the goal of the project. Research can be directed at either technological or scientific questions. The principles used in research study design and experimental design can be applied to questions of both types. Both types of questions are covered in this book.

To put a boundary on the subject of conservation research, we require that it involve some art object or material. If several adhesives are studied "in the pot" by measuring variables such as viscosity, density, shelf life, and fume toxicity, then the subject is adhesive chemistry. Even if such work is done by a conservation scientist and is relevant to conservation practice, it is not conservation science as defined here. If one applies the same adhesives to some art material and then measures properties such as strength and discoloration, then the study falls within conservation science.

1.2 CLASSIFICATION OF CONSERVATION RESEARCH AND KNOWLEDGE

1.2.1 Phase

Conservation of an object has three possible phases:

Composition Determining the composition of the object.

Deterioration Determining how the object might or is most likely to deteriorate (for prevention) or how it has deteriorated (for remediation).

Treatment Determining and applying a satisfactory treatment based upon the information derived from the two previous phases.

Composition, deterioration, and treatment are three major divisions or aspects of conservation. This scheme covers most studies published in the conservation literature. It is also useful as one basis for organizing research questions in conservation science.

This three-part categorization has direct analogies in medical science and medical practice. The study of *composition* can be likened to anatomy and physiology, which cover the structure and function of the human body. Pathology and etiology include the nature and cause of disease (*deterioration*). The corresponding phase of medical practice is diagnosis. Pharmacology (with the attendant practice of drug prescription) and surgery are two of the subareas of medical therapeutics (*treatment*). Medical students study all these areas of medical science as a basis for medical practice. Research in medical science continues in all these areas.

1.2.2 Study Types

Conservation questions can be divided or categorized on another basis -- the subject or type of object studied:

Method A particular technique, process, or type of equipment.

Object An art object, or a few objects, considered individually.

Class A generic class of art objects, with particular objects possibly used as illustrative or experimental material.

Surrogate Art materials or objects made from art materials especially for the study.

A class of objects can include all the objects in a large collection or a group selected by any reasonable criterion. The dividing line between an object study of a few objects and a class study of a small group is sometimes a bit fuzzy. When a group is so large that one cannot study each object with individual attention, then one is doing a class study.

The key to surrogate studies is the use of art materials in place of actual art objects. Surrogate objects may be artificially aged and given possibly dangerous treatments. Occasionally, old materials that have aged naturally are available. Other advantages of surrogates are the availability of as much material as needed and the freedom to do destructive testing. The disadvantage is that results must be extrapolated to real objects, sometimes with questionable validity.

1.2.3 Conservation Research Problems

To make these classifications more concrete, we give an example from the literature for each of the twelve combinations of phase and type:

Composition	Method	Quantitative determination of red anthraquinone dyes on ancient textile fibers (Wouters 1985).
	Object	Identification of the pigments in the "Peacock Room" by Whistler (Winter and FitzHugh 1985).
	Class	Presence of manganese black pigment on Etruscan terracottas (Schweizer and Rinuy 1982).
	Surrogate	Discrimination between re-creations of three nineteenth-century blue textile dyes (Cordy and Yeh 1984).
Deterioration	Method	Rapid determination of exhibition and storage materials likely to tarnish silver objects (Daniels and Ward 1982).
	Object	Nature and cause of sandstone deterioration on a reconstructed eighteenth-century building in Bern, Switzerland (Zehnder and Arnold 1984).
	Class	Contribution of various weathering processes to deterioration of rock art inside caves (Dragovich 1981).
	Surrogate	Effect of temperature and humidity change on cracking, splitting, and anisotropic movement of ivory (Lafontaine and Wood 1982).
Conservation	Method	Rapid and easy testing of polyethylene glycol (PEG) penetration in wood stabilization (Hoffmann 1983).
	Object	The most efficacious and safe solvent-reagent system for cleaning the damaged gilding on the "Door of Paradise" panels by L. Ghiberti (Fiorentino et al 1982).
	Class	The best method for eliminating chlorides from highly corroded excavated iron objects (Rinuy and Schweizer 1981).
	Surrogate	Effectiveness of benzotriazole (BTA) for inhibiting the corrosive behavior of stripping reagents used on bronze while still allowing the patina to be removed (Merk 1981).

1.2.4 Art Objects and Materials

A third basis for categorizing conservation projects and research is by art material and object class. Within the American Institute for Conservation of Historic and Artistic Works (AIC), there are specialty groups for book and paper, paintings, textiles, photographic materials, objects, wooden artifacts, and architecture. This is one possible categorization on this basis. Objects can be further subdivided into metal, stone, and ceramic, among others. Each poses different problems of deterioration and conservation.

Books, works of art on paper, paintings, textiles, and photographs are legitimate areas of specialization, each with their own problems and practices. Their common feature is that they are basically two-dimensional, with a thin third dimension consisting of image material superimposed on a mechanical substrate. Image materials -- such as ink, pigment, dye, and silver nitrate -- all have the unfortunate property that they can fade and change colors. Substrate materials, such as canvas, paper, textile, wood, and plaster, have the common problem that they can lose strength and fall apart.

Here is one possible classification of art materials and objects:

Two-Dimensional Image	substrate	wood
		canvas
		textile
		paper
		glass
		plastic
		leather
	image	paint
		dye
		silver grains
		ink
		gilding
Three-Dimensional Object		wood
		stone
		metal
		glass
		plastic

1.3 EXPERIMENTAL DESIGN

1.3.1 Literature Survey

In *Statistical Analysis in Art Conservation Research* (Reedy and Reedy 1988, hereafter referred to as *Statistical Analysis*), we reported the results of a survey of over 300 articles in the English-language art conservation literature from 1980 to 1985. It used the phase and type categories described above, since they relate to many aspects of study design and analysis. In that survey, research questions and the experimental designs used to answer them were taken as given, without evaluation. We did look at and evaluate (a) statistical analysis and

(b) the reporting of design, data, and analysis. Our primary purpose at that time was to quantify, categorize, and evaluate the statistical methods used in conservation research.

We discovered that few papers in the conservation literature had any formal statistical content. For only 10% of all papers could we say that the experimental design was such that the statistical technique of hypothesis testing could and should have been applied. For nearly half of the studies none of our evaluation categories, not even description-of-treatment-number, was applicable, because the experimental design did not allow for it.

In *Statistical Analysis* we said that many of the studies reviewed could have been designed to make numerical techniques and statistical analysis applicable and useful, but that design questions were beyond the scope of that volume. Here we examine the appropriateness and tradeoffs of various experimental designs in relation to conservation research questions.

1.3.2 Planning

What is experimental design? In one sense, it is whatever one does in an experiment. It also refers to the action of planning. An experimental design is analogous to an architectural plan. One could start with the goal of building a two-story house with four bedrooms, three bathrooms, and a fireplace, but one usually would not immediately start to build. An architectural plan would be drawn first. The degree of detail for experimental designs, as with architectural plans, ranges from a rough sketch to a complete finished plan.

A plan requires a goal and consideration of context. Chapter 2, which reviews the scientific method, discusses the scientific aspects of both goals and contexts. An important point is that a scientific experiment has to have the possibility of more than one result. The hypotheses behind the experiment should each be specific enough to be contradicted by one of the possible results. In other words, scientific hypotheses have to be falsifiable.

A major differentiation in study design is between experiments, which manipulate or treat objects with the intention of inducing a change, and observational studies or surveys, which do not. Survey measurements may cause unintentional and even undesirable changes, but these are not the focus of the study. An experiment concerning the effect of treatments must involve a new treatment that would not have been applied if the study were not done. Observation of the outcome of standard treatments that were or would be applied anyway, without the study, is a survey.

Since conservators take the composition of art materials and objects as given, composition studies of actual objects are surveys. Composition method studies may involve active experiments. Deterioration studies of real art objects are usually surveys. One can experiment by placing objects of a class in an environment hypothesized to retard deterioration, but definitive results may take too long. Deterioration experiments with surrogate objects are more feasible because of the freedom to use accelerated aging. Surveys of treatment outcomes have uses in conservation, as in medicine, such as indicating areas where improvement is most needed. Treatment experiments are also possible with real objects. Chapter 3 shows how planned experimentation can be done with single objects that are to be treated. It is therefore particularly relevant for practicing art conservators.

1.3.3 Major Aspects of Experimental Design

Given the goal or question to be answered by a study, the major aspects of study design, discussed in detail in Chapter 4, are:

Objects Number, source, and method of selection or sampling; grouping of experimental units into homogeneous blocks to reduce the experimental error in the comparison of treatments or populations.

Variables Number, measurement protocol, repetition, units, recording, and reduction.

Treatments Number, organization, and replication; assignment to experimental units in the design blocks; application protocol.

The answers to the questions implied by the list above (such as "How many objects?") largely make up a design. In this sense, Chapter 4 is a step-by-step guide to experimental design.

Unfortunately, cookbook answers are not possible. For example, there is no simple answer to the question of how many objects to study. It depends on the variability of objects, treatments, and measurements and on the desired precision of the answer. Workers in various fields develop rules of thumb. Such rules should be based upon relevant experience and statistical consideration of that experience. A general answer is to use as many objects or samples as required to get an answer that is both statistically and scientifically meaningful. Statistically, more is better. Scientific considerations usually give an upper limit, because the detection of very small differences is often unimportant. For instance, conservators would typically not consider it worthwhile to do enough samples to detect a 1% difference in decay half-life between two treatments. At some point, it is scientifically more worthwhile to finish and publish the experiment and go on to the next. Economic and career considerations also come into play.

Chapter 5 gathers discussions of some particular points of statistical analysis related to the major aspects of experimental design. These include problems of measurement repetitions, estimation of curve parameters, inference and hypothesis testing, tests based on normal distributions, tests based on ranks, and working with statisticians. It is not a full discussion of statistical analysis. It can easily be supplemented by any of the numerous books on statistical analysis available in libraries. There is also additional material in *Statistical Analysis*.

Chapter 6 presents designs for multiple object and multiple group studies. It incorporates the major aspects of design described in Chapter 4 into a variety of experimental designs for specific situations. The chapter ends with schematic outlines of several real examples and a work sheet that readers can copy or modify for their own use.

Chapter 7 applies the material of Chapters 4 and 6 to a class of designs directly comparing treatments on real or closely simulated objects. Treatment trials were developed for medical research and should also be useful for improving conservation practice. They examine both the safety and effectiveness of new treatments compared to standard or control treatments. A treatment trial begins with a clearly written protocol detailing object selection criteria, measurement and treatment procedures, and an initial plan of analysis. Randomization and the masking of treatment identities are important measures for eliminating bias.

CHAPTER 2

THE SCIENTIFIC METHOD AND CONSERVATION RESEARCH

A science is a systematized body of knowledge usually comprising both facts and principles. It is derived from a sophisticated coordination of thought and action -- thought in the form of hypothesis and inference, action in the form of experiment and observation. Thought, though starting with prior knowledge, is modified on the basis of experience, and is used as a guide to action, which in turn leads to new experience, which is used to generate new thought, and so on. When coupled with publication and peer review of results, this iterative process is called the scientific method. It has been extraordinarily successful. It is as applicable to conservation and conservation research as to other practical applied endeavors.

Modern science and scientific method dates from about the year 1600. Two primary innovators were Francis Bacon, the philosopher, and Galileo Galilei, the experimentalist. The number of scientist practitioners has since grown by several orders of magnitude, while a far smaller group continues developing the philosophy and history of science and the scientific method. Among the many recent discussions of the scientific method, our favorites include Chamberlin (1965), Platt (1964), Hempel (1966), and Harré (1981). These were the basis for the following discussion.

2.1 STEPS IN THE SCIENTIFIC METHOD

The basic scientific method, which leads to rapid advance and progress in a scientific field, consists of several steps:

Observe Record the observation or phenomenon stimulating the study.

Specify Choose a research question, problem, or area of study as the basis for a research program. The question or problem should be explicit, clear, and answerable.

Hypothesize Explicitly construct alternative hypotheses. Write several possible answers to the research question or several possible explanations of the initial phenomenon.

Infer Work out the implications of the hypotheses. Write some concrete predictions. If a hypothesis is true, what therefore has to happen or be observable?

Design Devise tests to support or eliminate one or more of the hypotheses. Think of experiments that should generate the predicted phenomena.

Explain If the relationship among hypotheses, implications, tests, and design is not obvious, give a rationale to explain the relationship.

Experiment Select objects, measure variables, and apply treatments according to the design. Observe what happens.

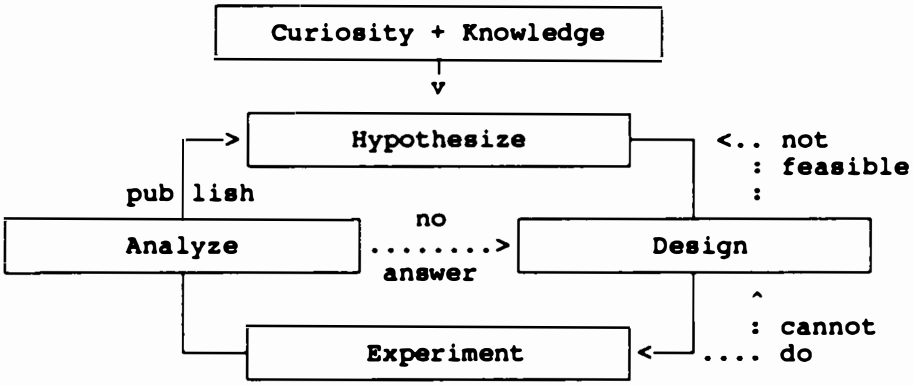
Analyze Analyze and interpret the results of the tests. Reduce the data, do statistical analysis, and compare actual results to expectations for the different hypotheses.

Publish When appropriate, make results public for discussion and use by others.

Build Repeat the procedure after constructing sub-hypotheses or sequential hypotheses to refine the possibilities that remain. If appropriate, refine the research question.

Figure 2.1 illustrates the research cycle, with backtracking for lack of success at any step.

Figure 2.1 Research Cycle



Major questions usually require many studies and experiments. Progress begins by choosing experimental projects that are appropriate given current knowledge and feasible given current resources. Do not try to do everything in one research cycle. Progress continues by combining results from several studies. The new discipline of meta-analysis does this in a quantitative way. Glass, McGaw, and Smith (1981), Light and Pillemer (1984), and Hunter and Schmidt (1990) are recent presentations of the newest methods.

2.2 RESEARCH QUESTIONS

Research questions fall along a spectrum from *technological* (what happens, what results treatments have) to *scientific* (why). Between these two extremes are questions of when, under what conditions, to what degree, and with what patterns things happen. Additional types of questions concern variability, estimation error, and importance. Is something different enough from some neutral value to be worth additional attention?

A technological or phenomenological study is often the necessary first step in a new research project. We need to know that something is present or occurs before we can usefully ask why. Once we establish a factual base, organization and unifying principles are pragmatically useful and intellectually satisfying. A modern trend in most applied disciplines is the development of a scientific base so that solutions to problems can be designed in light of predicted behavior rather than merely discovered through trial and error.

For example, an initial experiment with colorant C determines that C fades over time at a moderate rate. Further experiments investigate the relationship among light exposure, temperature, humidity, air pollutants, and fading rate. The answers allow interpolations that predict the outcome under conditions not yet tested. Experiments closer to the scientific end of the spectrum modify the composition of both the substrate for C and the air in the experimental chamber. Suppose ozone accelerates the fading of C regardless of the value of other variables. The next few experiments might then investigate the mechanism of this effect. This hypothetical example study started by identifying which factors had which effect and continued by finding out why.

For another example, a study of several colorants starts with ranking their degree of fading under a particular set of environmental conditions. The scientific study of these materials continues by developing an explanation for the observed ranking. This initially takes the form of partial correlations between fading and other properties of the colorants, such as structure and composition. These hint at some of the underlying relationships. These scientific answers lead to non-obvious predictions of the fading of previously untested colorants with better-than-random accuracy.

Discovery of structure-function relationships has become of intense interest, for instance, in the pharmaceutical industry, where trial-and-error searches may involve thousands of compounds. Science involves generalizations that enable us to make predictive jumps.

It may be hard work to refine and simplify a research question until it is unambiguous and clearly answerable in one particular experiment. A major question will often need to be split into subquestions. For example, "What causes deterioration of stone sculpture?" is too large a question to address in one experiment. It requires a major research program with a series of experiments each addressing subquestions.

2.3 HYPOTHESES

A scientist might address the question of why colorant C changes color by hypothesizing: (1) hydration reactions, (2) photo reactions, (3) chemical reactions with its substrate, or (4) inherent instability. He or she needs an understanding of the chemistry of C to refine these hypotheses. The specific set of hypotheses developed should build on previous work by that experimenter and by others.

Hypotheses and their implications should be explicit. Write them down or discuss them with others, even if they seem very obvious. This will clarify what you are doing and why, both to you yourself and to others. That which is obvious to you while immersed in your work is often not obvious to others or to you at a later time. If it is not obvious how the hypotheses might answer the research question or how experimental outcomes might eliminate or support them, provide a rationale explaining the relationship.

If these steps are easy there is a tendency to think that they are not worthwhile. If they are difficult, there is a tendency to think that they are too much trouble. However, vague thinking and planning at the beginning of a study can lead, after months or years of work, to vague results with little or no value.

2.3.1 Contingency

Several authors have stated that a scientific hypothesis must be falsifiable. More exactly, it must be logically and psychologically contingent on experience so that there is some real possibility of refuting it. The evidence gathered from observation and experience must matter in order for the judgement rendered to be within the realm of evidential reasoning (Lett 1990).

A scientific hypothesis should not be contradictory and necessarily false regardless of the evidence. One is most likely to commit this error if one has a complicated hypothesis compounded from many simpler statements or if one combines into a single hypothesis competing hypotheses that should be kept separate. An example would be, "This chemical consolidates stone *and* this chemical does not consolidate stone."

A scientific hypothesis should also not be a logical tautology (true regardless of evidence) or a subject-matter platitude. "This chemical consolidates stone *or* this chemical does not consolidate stone" is a logical tautology. A subject matter platitude is a statement so weak that no practical amount of evidence could disprove it. An example is "This treatment might be valuable under some circumstance." A hypothesis should say something worthwhile but should not try to say everything.

Similarly, a scientific hypothesis must not be a value judgement, philosophical stance, or religious belief that is outside the influence of evidence or considered to be so by the individual involved.

Lett (1990) gives two ways to violate the rule of falsifiability. He summarizes these as "whatever will be, will be" (a platitude) and "heads I win, tails you lose" (philosophical slipperiness). The first violation is to make a broad, vague claim with little content and blank spaces that are filled in, if ever, only after evidence is collected. The second violation is to make a claim with some content, but then to generate excuses as needed to explain away contradictory evidence.

To illustrate with an only-somewhat-ridiculous example, suppose Dr. Weasel claims that colorant fading is due to enteation by lingrems. Without a definition for "enteate" and "lingrem," he has only said that colorant fading is attributable to some action involving some agent other than the colorant itself. If allowed to be this vague, he can claim any evidence as support for his hypothesis. When Dr. Skeptic tries to pin him down, Dr. Weasel tells her that lingrems are a hitherto unknown microscopic quasi-organism and that enteation is a surface reaction. When she says that light accelerates fading, he says that lingrems absorb light and use the energy to enteate. She adds that ozone does the same. He adds that lingrems use it as an alternative energy source. When trying to replicate his experiment, she cannot destroy them with heat or cold. He claims that they have a super-resistent spore-like stage. If she cannot see them with light microscopy, he says that they are either transparent or too small, and she cannot see them with scanning-electron microscopy because they are destroyed by preparation for SEM. And so the dialogue continues without resolution.

2.3.2 Multiplicity

The scientific reason for having multiple hypotheses is that there may be multiple explanations or facets to the problem involved. A single hypothesis may misleadingly simplify complex relationships.

A psychological reason for having multiple working hypotheses, even if one of the hypotheses seems trivial, is to prevent over-attachment to any one of them (Chamberlin 1965). If everything depends upon supporting one hypothesis, the mind excels at seeking out or only noticing the evidence that tends to do so. One becomes remarkably blind to all evidence refuting it. Platt (1964) also discussed the subconscious tendency to try to force results to fit a pet hypothesis.

Another psychological reason is to prevent under-commitment, in the form of having no hypothesis. Being vague about what might happen is another trick the mind uses to avoid being wrong. This can be combatted by bearing in mind that it is hypotheses and not scientists that are wrong. Clearly stated multiple hypotheses combine the virtues of explicitness and flexibility.

If one can only think of or only has interest in one hypothesis, then it is always possible to derive a second by negating the first. This may result in a null hypothesis, as discussed in 2.5.1. Section 2.6 has several examples.

2.3.3 Two Cautionary Tales

The first, personal, is a tale of under-definition. One of the senior author's first statistical assignments was to help organize and analyze 7 years and 5000 lines of clinical data comprising 4 tables with over 200 variables. Unfortunately, the medical investigator who started the project had died without leaving a written set of questions, hypotheses, and rationales for the data collected. From this mountain of data we extracted a mole of a result that generated one paper with little impact. The study continued, slightly revised, with a vague purpose that can be paraphrased, not unfairly, as "to see what we can see." The vague hypothesis was "we might see something worthwhile." Five years later, after a couple person-years of work distributed among several people and some frayed relationships, the data remained mute and the project abandoned with no scientific result.

The second, public, is a tale of over-zealous definiteness. Two chemists measured more thermal energy coming out of a room-temperature electro-chemical cell than electrical energy going in. They hypothesized that the surplus heat was generated by a hitherto-unknown cold fusion reaction. Apparently blinded by visions of a Nobel Prize, patent royalty millions, and a new future for humanity, they announced their hypothesis as a conclusion at a news conference. They did not properly consider and test the alternative hypotheses that nearly everyone else considered more likely. Only later did they submit a paper for peer review, which they withdrew when asked for certain methodological details. Although definite about cold fusion, they were vague about their methods. They explained negative results by others attempting to duplicate their results as due to undisclosed but essential differences in technique.

2.4 EXPERIMENTS AND OBSERVATIONS

The details of experiments and measurements are specific to the research questions and hypotheses. Later chapters discuss many of the general principles. An important question is, what is the simplest experiment that can exclude one of the hypotheses? A simple, short, elegant experiment that clearly eliminates one possibility is intellectually and financially preferable to a long and complicated one that produces that same result.

2.5 ANALYSIS AND INTERPRETATION

When the outcome predicted by a hypothesis does not occur, one should question that hypothesis. If the outcome is accepted as correct, the hypothesis must be rejected.

When the predicted outcome does occur, the hypothesis is not necessarily true. Another hypothesis may also predict and explain the same observation. Suppose Reverend Ego hypothesizes "I am the Chosen of God" and infers "It will snow tonight because I want to go skiing tomorrow." Snow that night does not prove his divine favor when it has another explanation.

Successful prediction of observations or events, while not proving a hypothesis, *supports* it. The strength of the test depends upon how unusual the observation is. Snow on a winter night is a weak prediction, though stronger than predicting that the sun will rise in the morning. Snow on a summer day is a much stronger prediction.

Sometimes two people doing the same experiment get different results. Experimental results are not always as clear as we would like. The sample sizes often used for experiments allow substantial random variations.

Reporting hypotheses and their tests clearly enough that other researchers can repeat the process helps substantiate one's work. It is important to distinguish between raw data and interpretations. Vague experimental designs, inappropriate statistical tests, and incorrectly applied tests give results more likely to be wrong and thus not reproducible.

2.5.1 Statistical Hypothesis Testing

Hypothesis testing has a specific sense in statistics. The focus is on rejection of null hypotheses. A null hypothesis is a hypothesis that some groups are the same with respect to some characteristic, or that several treatments are the same with respect to some outcome, or that certain variables have no inherent relationship. In other words, a null hypothesis is a hypothesis of zero difference or zero correlation. This definition encompasses the proposition that a treatment has no effect, since there is an implicit comparison with the alternative "treatment" of doing nothing. Doing nothing is a type of control treatment as discussed more thoroughly in later chapters.

Statistical hypothesis testing is a tool used in scientific hypothesis testing. Because of the difference of focus between numerical relationships and subject matter content, making the two mesh usually requires some thought. If we use neutral hypothesis as a synonym for null hypothesis, the relationship may be a little clearer. A neutral hypothesis that all treatments are equally good is similar to saying that all theories or hypotheses are equally valid. The common theme is to not prejudge and to keep an open mind until one has clear evidence, and to keep open to further evidence and development of a new and even better theory or treatment.

When one of the hypotheses regarding the outcome of an experiment is itself a neutral hypothesis, the situation is a little more complicated. The usual procedure is to stick with the null hypothesis unless and until there is reason to reject it and consider an alternative. This follows the principle enunciated in the early fourteenth century by William of Occam (or Ockham) and known as Occam's Razor -- do not multiply entities beyond necessity. In other words, do not declare the existence of phenomena and causes unless there is a good reason. Yet another wording of this principle is to stick with the simplest hypothesis that accounts for the most facts.

Our law courts follow a similar procedure with the presumption that a defendant is innocent until proven guilty ("one of the most important null hypotheses in Western Civilization" as Jim Druzik remarked). There are millions of possible suspects for any crime, and investigators sometimes select the wrong person. Prosecutors are therefore required to publicly present both the evidence and reasoning supporting a hypothesis of guilt. The jury must be convinced beyond a reasonable doubt before rejecting the null hypothesis of no connection with the crime. Only then is (punitive) action taken.

In medicine, there are millions of possible causes for every disease, and millions of possible treatments. Usually only one or a few are of major importance. The goal of medical research is to find these few that are worth acting on.

From the viewpoint that everything is connected with everything else, the null hypothesis should never be true in an absolute sense. If carried to enough decimal places, the correlation between any human attribute and a particular disease should be different from 0. Similarly, there might be millions of people remotely connected with any crime by virtue of some interaction with either criminal or victim, or with someone who later had such an interaction, or through some lengthier chain of interactions. But this is not the issue in either courts of law or scientific judgment. Is the correlation, effect, or connection strong enough to be worth paying attention to? The practical null hypothesis for any question is that it is not.

2.6 EXAMPLES

2.6.1 Technological Example

Observation Some adhesives used to conserve paper type P subsequently discolor.

Question Which of three specific adhesives will discolor least when used on paper P?

Hypotheses (0) There is no difference in degree of discoloration that appears over time with the three adhesives on paper P.

(A) Adhesive A applied to paper P discolors less over time than do adhesives B and C.

(B) Adhesive B applied to paper P discolors less over time than do adhesives A and C.

(C) Adhesive C applied to paper P discolors less over time than do adhesives A and B.

Hypothesis A refines to ABC and ACB, where hypothesis ABC is that adhesive A discolors less than B which discolors less than C. Hypotheses B and C have similar refinements.

Method Measure the color of the adhesives applied to paper substrate P before and after thermal artificial aging and calculate the color difference.

Rationale There are problems interpreting artificial aging data. The correlation of the effect of short-term severe environments with the effect of long-term natural aging conditions needs more assessment. However, the conservation literature accepts the theoretical use of thermal aging tests to rank the performance of materials.

Variables In addition to color change one might include peel strength, reversibility, and change in pH.

2.6.2 Scientific Example

Observation Some adhesives used in the conservation of paper type P discolor noticeably more than do other adhesives.

Question What factors determine how much an adhesive will discolor on paper type P?

Hypothesis (1) Adhesives with certain chemical bonds susceptible to hydration will react with water, causing discoloration.

Implication Color measurements of some adhesives subjected to varying humidity will show greater discoloration after exposure to high humidity.

Rationale Excess moisture allows hydration reactions to occur.

Hypothesis (2) Adhesives containing the impurity phenol formaldehyde will discolor over time.

Implication Color measurements made before and after thermal artificial aging on adhesives with and without the impurity will show greater discoloration on the ones with the impurity.

Rationale Phenol formaldehyde is a highly reactive impurity that can be introduced into an adhesive during synthesis or processing.

2.6.3 Observational Example

Observation Medieval-period copper-based statues from Kashmir are often difficult to distinguish stylistically from those produced in West Tibet (Reedy 1986; Reedy and Meyers 1987; Reedy 1988; Reedy 1991).

Question Can we distinguish statues from the two regions on technical grounds?

Hypothesis (1a) Statues from the two regions are distinguishable from each other in casting and decorating technology.

Rationale The casting and decorating technology employed in statue production involves many steps, with alternative choices available at each. Statues originating from a set of workshops located within one specific region might vary in casting and decorating methods from statues originating in another regional set of workshops.

Hypothesis (1b) Statues from the two regions are indistinguishable in casting and decorating technology.

Rationale Historical texts say that medieval Himalayan craftsman were often mobile, and artists from Kashmir went to West Tibet to produce statues for monasteries there. They might have used the same techniques in both places, resulting in overlapping technologies between the two regions.

Method Identify and record surface and internal casting and decorating features visible by eye, binocular microscope, and X ray. Use stepwise discriminant analysis to show how well these data distinguish between the two regions.

Hypothesis (2a) Statues from the two regions are distinguishable from each other in elemental composition of the metals.

Rationale Differences in major and minor elements will result if different alloying practices exist from one region to another. Trace element differences may result from the exploitation of different copper ore sources by different regions.

Hypothesis (2b) There are no significant differences between the two regions in metal composition.

Rationale There may be no significant alloying differences between regions since there are limited variations in copper-based alloys. Metal was probably imported into West Tibet since it had limited ore resources, and also might have been imported into Kashmir at times. If the imports into the two regions came from the same source(s), we would find no significant trace element differences in metals.

Method Sample 25 mg of metal from each statue and analyze by inductively-coupled plasma emission spectrometry (ICP-ES) to measure major, minor, and trace element concentrations. Use stepwise discriminant analysis as before.

Hypothesis (3) Statues from the two regions are distinguishable from each other in clay core composition.

Rationale Since clay core materials are unlikely to have been imported from an outside region to the manufacturing workshop, this should be an ideal material for a regional characterization. Since Kashmir and West Tibet are separate geographic areas, the sandy clay used in core manufacture should show considerably more mineralogical and elemental variation between two regions than it does within a single region.

Method Sample core material when available and do petrographic analysis to identify the mineralogical variables and instrumental neutron activation analysis to quantify the trace elements. Use stepwise discriminant analysis as before.

Hypothesis (4) Combining all datasets will result in an increased ability to distinguish between the two regions.

Rationale The more information we have about each regional group, the more likely we are to be able to characterize it and distinguish it from the other regional groups.

Method Perform stepwise discriminant analysis with various combinations of the datasets to see which gives the best results.

2.6.4 Summary Example

Here is a final example of multiple hypotheses. Implications, methods, and rationales are omitted.

Observation The surface of a bronze grizzly bear statue in Yellowstone Park is deteriorating, resulting in severe discoloration.

Question What is causing the deterioration?

Hypotheses

(1) Rain beating on the surface or components within rainwater such as acid.
(2) Excess sunlight exposure.
(3) Fluctuating temperatures.
(4) Fire ash and smoke (from the 1988 forest fires).
(5) Fluctuating humidity.
(6) Condensing moisture that sits on the surface of the statue and works its way inside through small cracks.
(7) Wind abrasion.
(8) Finger oil from tourists who touch the statue.
(9) Exposure to animals and their urine and droppings.
(10) Sulphur exposure from adjacent hot springs.
(11) Unstable metal composition and casting methods resulting in "inherent vice."
(12) Surface patina applied by the artist with an unusual component.

Devising multiple hypotheses may increase creativity and the ability to come up with many possible answers to a question. Some hypotheses may be eliminated without tests by using logic and library research. Implications are needed for the rest. Their tests can be as simple as looking by eye or with a magnifying lens. High-tech analytical equipment and long-term experiments are not always needed. After eliminating more hypotheses by experiment, more than one may remain supported. Follow-up research can assess the degree to which each probable agent contributes to the problem, and thus where the conservation priorities lie.

CHAPTER 3

SINGLE-OBJECT STUDIES

3.1 ADVANTAGES AND USE

The scope of conservation projects and conservation research ranges from one object to thousands of objects and generic classes of objects. The same is true of the healing, training, and education disciplines that work with human beings. This chapter concentrates on studies of single objects. Such studies are perhaps most important to practicing conservators, who spend much time with a single objects and need to justify particular treatments without always having similar objects available for experimentation. Single-object studies also may aid researchers who are more interested in conservation science than in a particular object but who are studying actual works of art rather than surrogate samples of art materials.

Historically, applied disciplines began with a focus on individuals, with the primary method of research being the case study. This century has seen a shift of emphasis in research and experimental design toward group comparison methods. The latter is the primary content of most books on experimental design and analysis.

In recent years, there has been a revival of interest in single-subject studies for gaining knowledge (Barlow and Hersen 1984). This does not mean a return to subjective case studies. It instead means applying the principles used in group studies to derive statistically valid conclusions from single-subject studies. The key idea is that an experimenter can sometimes make an object serve as its own control by treating it more than once in a formal experimental design. It is therefore possible to make formal statistical statements comparing treatments applied to only one object. Such designs are alternatives to case-study trial-and-error. They should result in a more rapid accumulation of knowledge.

Strictly speaking, the results of a single object study are applicable only to that object. However, the conservator who must treat a particular object, just like the doctor who must treat a live patient, can use objective substantiated conclusions to plan a treatment for that object and to develop knowledge for future projects.

Single-object studies also may interest researchers who would otherwise prefer more general results. Material may be in short supply, especially for destructive experiments. Resources may be limited or uncertain. A complete study on one or two objects is better than an incomplete study on six that is prematurely terminated as resources run out. The treatment protocol for an experiment may be in flux and subject to change after the experience with each object. A series of completed single-object studies may lead to the desired generalizations.

Compared to human beings, art objects have advantages for individual study. Most treatments of human subjects are systemic and affect the entire body or at least an entire organ. Multiple treatments of one patient must nearly always be given at different times.

Over-the-counter headache remedies are a typical example. A limited exception is the simultaneous application of skin treatments that have only local effects. Art objects can more often receive different treatments at different places within one treatment session. Conservation researchers therefore have added flexibility for designing rapid studies.

A second advantage of art objects is that the effect of time is more stable and predictable. Unlike human medical subjects, objects do not spontaneously heal themselves. We can more plausibly conclude that observed improvement results from treatment. Bronze disease does not cure itself, whereas organisms are active entities and tumors sometimes disappear without a doctor's intervention. The art object might get worse, but it rarely cures itself.

Art conservation has the additional advantage of greater stability and dependability of treatment outcome. We use this in everyday conservation practice to make cause-and-effect conclusions. However, people can be fooled. Coincidences do occur and are constantly subject to misinterpretation. People remember dreams that come true much better than they recall those that do not. But it is harder to apply a treatment and not notice that there has been no effect.

In part, this chapter builds upon the everyday art conservation practice of testing patches on an art object. It develops that practice into formal, statistically analyzable experiments. Once learned, the techniques often require relatively little extra time, perhaps an hour or so. This is not much compared with the total time involved in some treatments.

We are not suggesting that every treatment on every object be subject to a rigorous test. Instead, we discuss when and how one justifies making routine conclusions that a treatment works, give suggestions on how to proceed when its value is not so clear, and show how to test an object's treatment statistically by applying it to test patches within test intervals.

3.2 MEASUREMENTS ON SINGLE OBJECTS

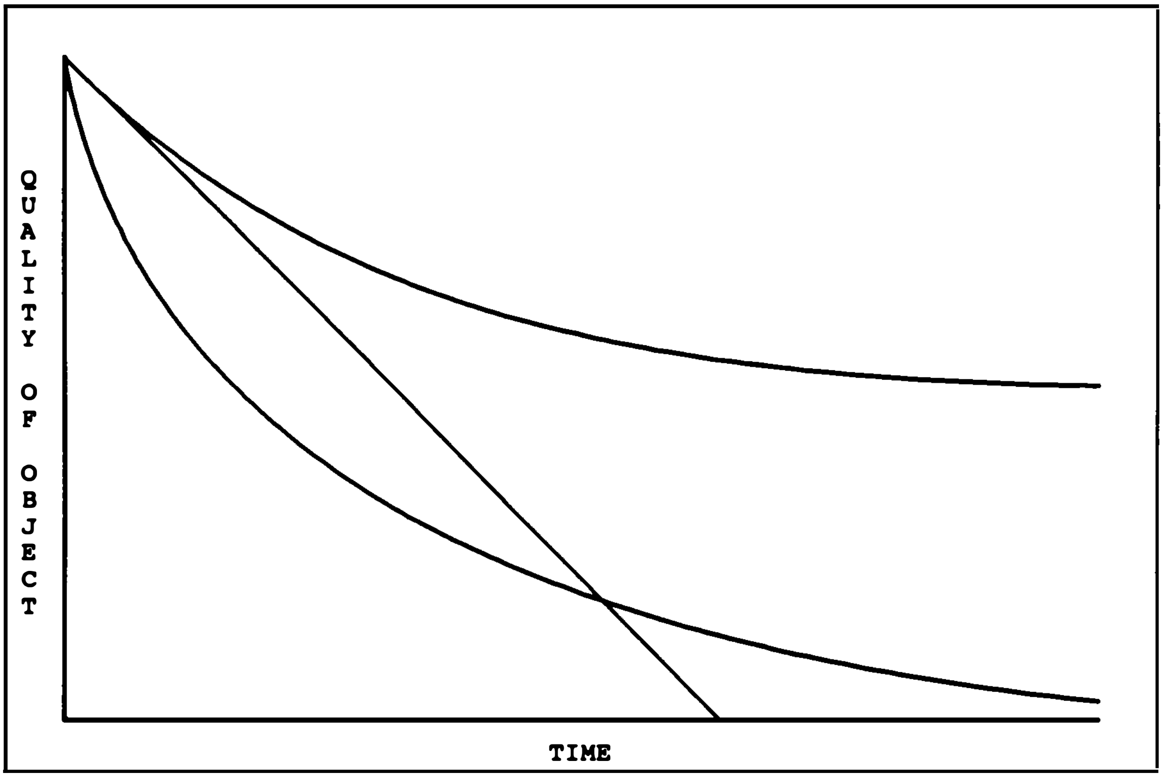
3.2.1 One Measurement

A single measurement reveals the state of the object at the time of measurement. It may be useful for choosing a treatment but does not in itself constitute a study.

3.2.2 Simultaneous Measurement of Multiple Variables

Composition studies typically measure multiple constituents in each sample. An example is an elemental analysis for copper, zinc, tin, lead, iron, and arsenic in a sample from a copper-alloy statue. Treatment studies often measure multiple outcomes for evaluating the treatment. A study of adhesives might use peel strength, color change, and reversibility. The most common statistical analysis techniques work with one outcome variable. We need

3.2.4A Example time trends of quality decay for an object



special multivariate statistical techniques to analyze multiple outcomes as a group. To avoid complications, we assume that only one outcome variable is of immediate interest even if there are others to be analyzed later in the same study.

3.2.3 Two Measurements of One Variable

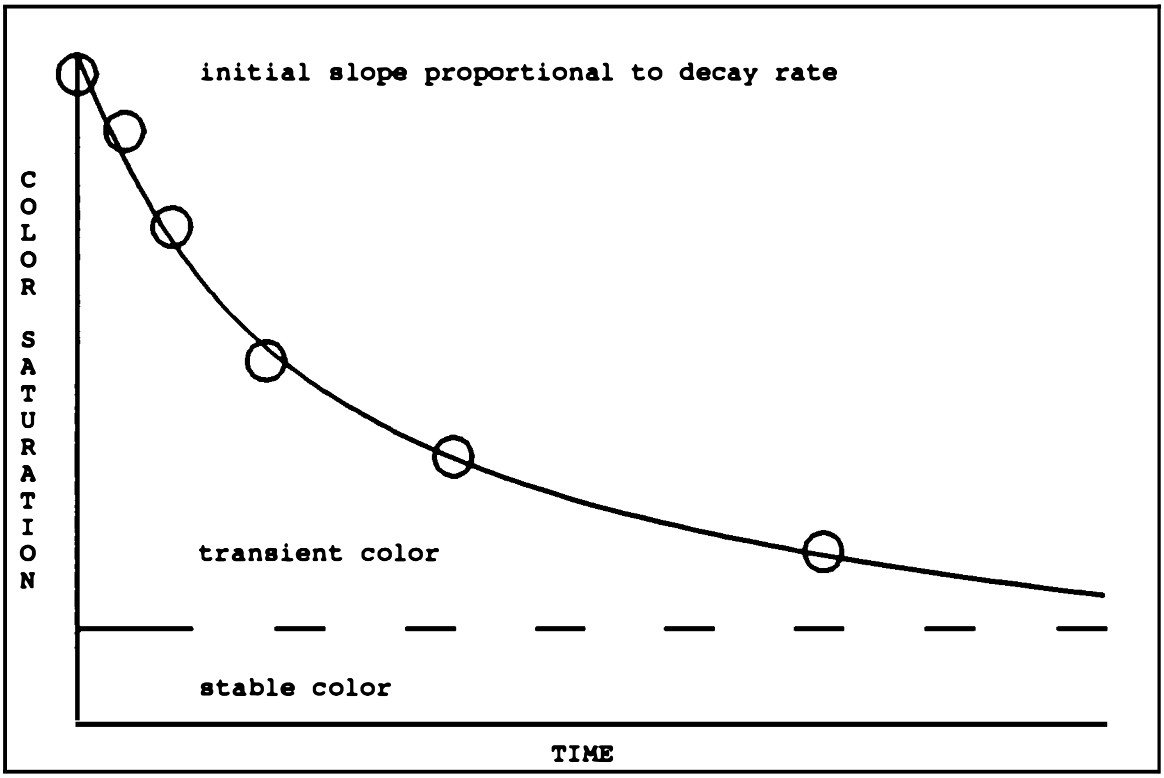
The difference between two measurements estimates how much the object changed during the interval between the measurements. However, real measurements are always subject to variation. Concluding that there has been a change in the state of the object requires information about the size of measurement errors. We must assume that this information is applicable to the current measurements.

Treating or otherwise manipulating the object between the two measurements generates a minimal before-and-after study. Concluding that the action affected the outcome requires additional assumptions about the change expected if there were no intervention. The validity of the conclusion is no greater than the validity of the assumptions.

3.2.4 Repeated Measurements of One Variable

Two measurements are absolutely minimal for a kinetic study of deterioration from aging and exposure. Multiple measurements should be obtained when possible in order to get

Figure 3.2.4B Negative exponential fit to simulated data



information about the shape of the decay curve. Does the tensile strength of a fiber decay in a straight line to zero, exponentially decline to 50% of its original value, or something else? Figure 3.2.4A shows some graphical examples.

Three or more sequential measurements under constant conditions give additional kinetic information. Least-squares regression and other curve-fitting techniques fit straight lines and curves to such data. These summarize and condense the data and estimate characteristics of the object in that particular condition that we cannot measure directly. If the number of measurements is greater than the number of parameters estimated, then we can judge the statistical significance of the slope or change parameter. With enough measurements we can estimate both the time effect and the measurement error without making assumptions about the relevance of previous information.

An example of a time series is a sequence of color measurements at predefined times or after predefined exposures to a hypothesized fading agent. In the units usually used to measure color intensity, the relationship between exposure and remaining intensity is often approximately a declining (negative) exponential. We can then summarize several repeated measurements by a single decay constant, which we could call "fading rate." Figure 3.2.4B shows a declining exponential fit to hypothetical color saturation measurements. The fit estimates the long-term residual color, the portion that disappears, and the rate. Expressed as numbers, these three aspects of the system are called parameters.

Simultaneous measurements at different places, usually over a two-dimensional surface, also have application in conservation studies. The difference between measurements on the edge and interior of a two-dimensional surface estimates edge effects and may give an initial indication of the presence of decay.

3.2.5 Treatment Effects

Statistical tests of treatment effects compare observed differences or changes to the amount of difference one would expect to see from random experimental error. The expected size of the random variation must be known from prior data or estimated from the current data. The key to experimental design for one object is that there are several times or places at which one might apply the treatment. Random assignment of treatments to times or places gives an internal experimental basis for the calculation of the differences expected from random error. One can then do a formal statistical test and evaluate the outcome even though only a single object is involved.

3.3 DESIGN 1: ONE TREATMENT INTERVAL

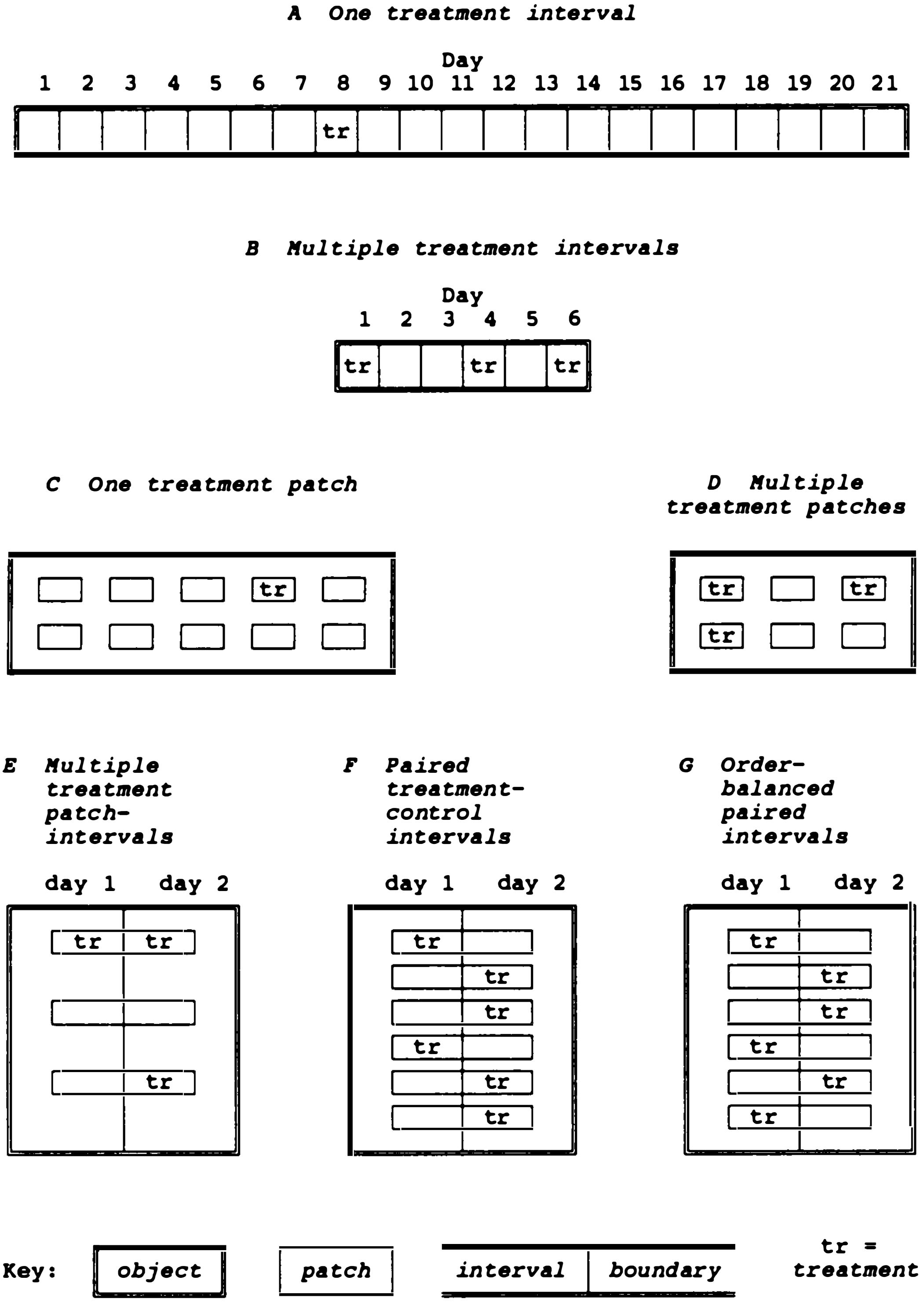
The following three sections describe seven different single object designs. Figure 3.3 gives a schematic example of each. The text and the figures are meant to work together to reinforce understanding of these designs.

3.3.1 Random Selection from Multiple Intervals

Subsection 3.2.3 introduced the minimal before-and-after study. Such a measure-treat-measure design does not tell us that an observed change results from the treatment unless we have external information that the difference is much larger than both normal measurement variation and the changes induced by other conditions of the experiment. Consider this design to be a selection of one treatment interval from one possibility. Design 1 extends this minimal design to a random selection of one treatment interval from several possibilities. Adding the element of choice allows a statistically valid conclusion for one object without using previous data or making assumptions.

To introduce design 1, consider the following situation. A museum has several limestone sculptures that are dirty, partly from being touched by visitors. The curator asks C, an objects conservator, to clean them. C proposes to treat the sculptures, after cleaning, with an agent that will seal the pores and prevent the penetration of dirt -- especially hand oil. Museum official M objects, claiming that the agent would alter the appearance of the limestone. C says that there would be no visible difference in normal museum viewing conditions. After further discussion, a statistician proposes an experiment to gather some evidence with regard to the two competing hypotheses.

Figure 3.3 Single Object Designs



A One treatment interval

B Multiple treatment intervals

C One treatment patch

D Multiple treatment patches

E Multiple treatment patch-intervals

F Paired treatment-control intervals

G Order-balanced paired intervals

C cleans one sculpture and puts it in an empty gallery, along with some other work projects. At 9 a.m. for 21 successive working days, M enters the gallery, looks at the sculpture from a distance of three feet under normal gallery lighting, makes notes, and leaves. C enters the gallery and goes to work with no one else present. On one of the first 20 days, C applies the quick-sealing agent to the sculpture (a small piece). Thus, M observes the piece at least once before the treatment and at least once after. C selects or has a third party select the one treatment day from the 20 possibilities by a random process that makes each equally likely. (See 4.4.2 for a discussion of how to do randomization.)

C does not tell M when the object is going to be treated, or afterward that it has been treated. The visual measurements are, in this sense, done blindly, without the influence of knowledge of when the treatment has been applied. On the morning of the 21st day, after the final observation, M consults memory and notes and selects the one interval, out of the 20 observed, during which a change was most noticeable. The identity of the treatment done each day (something or nothing) is kept from M until after M's selection of the apparent treatment day.

At the end of the experiment, M and C give analyst A their respective data. In the simplest version of design 1, the data are only the apparent and actual treatment days and the number of possible treatment days. A can be a third person or either of C or M.

The null hypothesis for this experiment is that the treatment does not affect any characteristic observed by M. If this is true, then M's selection is just a guess. Saying that the treatment does not affect M's observations means that M makes the same sequence of observations and the same selection regardless of when the treatment is done, or if it is done at all.

Because of the treatment randomization procedure, there is, with 20 possible treatment days, a 1/20 or 5% chance of a match between the day selected for treatment and the day selected by M. If C just "picks a number" between 1 and 20, without a proper randomization procedure, the match probability would almost certainly be an unknown value different from .05. It is very difficult to mentally pick numbers with equal probability even when one intends to. In this example, the match probability would depend upon the psychology of both C and M, including their second-guessing of each other. Random selection of the treatment day eliminates all such difficulties and considerations.

Selection of the correct day by M has a 1 in 20 chance by accident. This is sufficiently low that M and C might reject the hypothesis of accident and accept the alternate hypothesis of visible difference. However, the choice whether to make such a judgment at this level of chance is beyond the realm of statistics.

3.3.2 Test Statistics

Now suppose that before finding out the correct treatment day, M also indicates a second choice, a third choice, and so on, all the way to a twentieth or last choice. M gives A a data table that looks like one of the two sets of columns in table 3.3.2. Whichever form M gives

A, A can do the rank analysis described next. A might also re-sort the data to produce the other form of data table for additional insight into the experiment and its results.

Table 3.3.2 Example rank data for design 1

Choice (rank)	Potential treatment day
1	11
2	12
3	13*
4	10
5	5
6	6
7	4
8	14
9	15
10	9
11	8
12	7
13	3
14	2
15	1
16	16
17	17
18	18
19	19
20	20

* actual day

or

Potential treatment day	Choice (rank)
1	15
2	14
3	13
4	7
5	5
6	6
7	12
8	11
9	10
10	4
11	1
12	2
13*	3
14	8
15	9
16	16
17	17
18	18
19	19
20	20

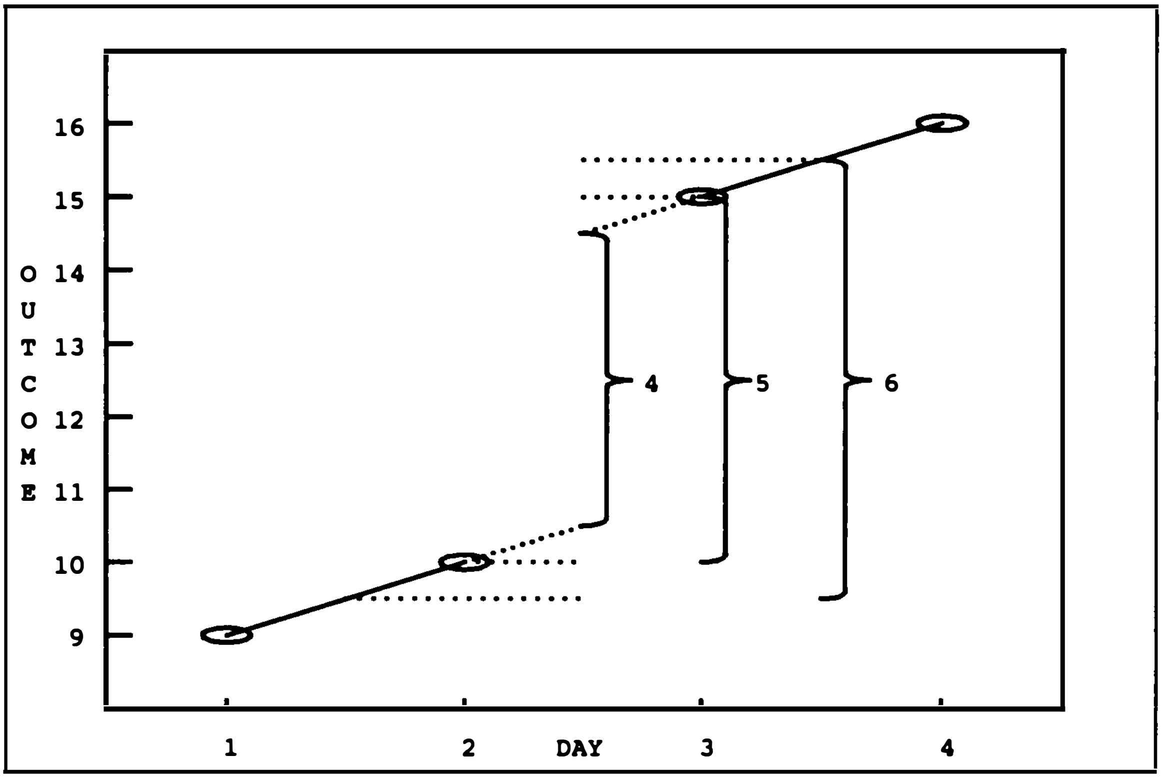
* actual day

A numerical summary of data designed to reveal a particular aspect of an object, population, or measurement process is a statistic. A test statistic, in particular, is a number that summarizes data for the purpose of deciding between competing hypotheses. If M only indicates a top choice, the test statistic is 0 or 1 for a miss or a match. Given a complete ranking, a better test statistic is the rank of the actual treatment day. For the hypothetical data table given in table 3.3.2, the summary test statistic is 3 for 3rd choice.

To interpret the test statistic, A calculates the probability of getting particular values when the null hypothesis is true. For miss or match, the probabilities were .95 and .05. For any particular rank, the probability is .05. A next sums the probabilities for the potential values that are as extreme or more so than the actual test statistic. This gives a cumulative probability known as a "p value." Extreme is defined as being in the direction predicted by the alternate hypothesis. For the ranks, M's hypothesis of easily visible treatment effect predicts a value of 1. For an actual value of 3, the potential values of 1, 2, and 3 are as much in the direction of 1 as 3 is. A reports a p value of .15 for this experiment. There is a 15% chance that the actual day randomly chosen by C will be among the top three picked by M. In general, the p value for this particular experiment and outcome measure is k times .05, where k is the rank of the actual treatment day.

Finally, suppose that M gives A 21 photometer measurements from the 21 morning observation periods. A initially estimates the treatment effect by subtracting the measurement taken just before the treatment from the measurement taken just after.

Figure 3.3.2 Treatment effect estimates for example data



A can try to improve the estimate of treatment effect by using the complete set of measurements rather than just the ones immediately before and after the treatment. If the measurements are noisy but otherwise stable, A can subtract the average of the measurements before the treatment from the average of the measurements after the treatment. If there is a consistent trend in the measurements before and after treatment, a more sophisticated adjustment for slope might be used.

To be concrete, limit the experiment to 4 days instead of 21. Suppose that M measures values of 9, 10, 15, and 16 on days 1, 2, 3, and 4, respectively. Suppose further that C treats the object on day 2 after the measurement of 10 and before the measurement of 15. Then the simple estimate of treatment effect is 15 - 10 = 5. The mean-adjusted estimate is (15+16)/2 - (9+10)/2 = 6, and a slope-adjusted estimate is 5 - (1+1)/2 = 4. Each of the three estimates of treatment effect is a possible test statistic. These are illustrated in figure 3.3.2.

The null hypothesis for this experiment is that the treatment has no effect. For the 21-day version, this means that the difference for the 1 treatment interval should be about the same as the differences for the 19 non-treatment intervals. The presence of 21 measurements instead of just 2 (20 differences instead of just 1) enables A to do a statistical significance test of whether the observed treatment effect is larger than expected from random variation alone.

Using the simple difference as a test statistic, A does a randomization test by calculating the 20 differences that correspond to the 20 possible treatment day choices. This produces a derived data table similar to the second set of columns in table 3.3.2, with the differences replacing the rank choices. A then sorts this table from largest difference to smallest difference to produce a second derived data table similar to the first set of columns in table 3.3.2. A then counts down from the top to get the rank of the actual treatment effect in the set of possible treatment effects. This determines the p value as before. If the observed treatment difference is the largest of the 20 possible, A declares it significant at the .05 level. Subsection 3.3.3 more thoroughly describes the nature and logic of randomization tests.

Design 1 constitutes a legitimate experiment and statistical analysis without using any other object as a control for comparison. The before-and-after measurement pairs for the intervals without the treatment show what happens when C leaves the object alone. The measurement intervals can be any length as appropriate for the particular treatment without changing the logic of the design or analysis. They also can be unequal, to accommodate work schedules, although there could be complications if there is substantial drift in either the object characteristic or measurement apparatus.

This first design is simple but illustrates some key points of experimental design for single objects. However, intervals of a week or more, as might be needed to apply the treatment or to allow for drying time, would make the procedure a bit tedious. The number of intervals needed can be decreased by accepting a greater uncertainty in the outcome than 5%. This reduces the amount of information that needs to be collected, as discussed in the next subsection. An alternative is using one of the other designs discussed in 3.4 and 3.5.

3.3.3 Treatment Randomization Tests

The logic of the test described for design 1 is the following: We hypothesize that the treatment has no effect on the object. We give it an equal probability of being applied during any of 20 days. If both are true, then the measured difference for the treatment day has an equal 1-in-20 (or 5%) chance of being in any of the 20 possible positions (excluding ties) in the sorted set of 20 differences for the 20 possible treatment days. For example, it has a 5% chance of being 8th largest purely by random happenstance. The same applies to being the largest or smallest.

If the observed difference is the largest, we may choose to reject the null hypothesis of no treatment effect because the data make it seem unlikely. We say that the p value or significance level is less than or equal to .05, or that we are rejecting the null hypothesis at the 5% level. A p value is the probability of observing a result at least as extreme as the actual result if the null hypothesis is true. Learning these catch phrases helps one read or write about statistical tests of experimental results.

This randomization test is a statistical test that depends upon treatment randomization. It has nothing to do with random selection or sampling of objects from a population. Rather, C randomly samples one treatment time from the 20 possible treatment times.

The test does not depend on the nature or distribution of the measurements as long as there is a summary statistic that we can rank within the reference set of possible outcomes. One can be rigorous even with subjective scores. They may be less precise than instrumental measurements, but this is a only a reason to be more careful with the design and analysis. A's numerical calculations based on instrumental measurements are an objective, public alternative to M's private subjective rankings of the likelihood of each possible treatment protocol. But the logic of the test is the same.

It is crucial that the ranking of the possible treatment intervals be based only on M's observations without knowledge of C's actual random choice. This is obvious if M makes the rankings directly but just as true if A calculates and sorts a set of test statistics. A may choose a test statistic from many possibilities after looking at the data but not after knowing the actual treatment. Under the null hypothesis, the data contain no information about the actual treatment, so there is no harm in looking, and there may be some benefit in making the test statistic more sensitive to situations where the alternative hypothesis is true. If A chooses a test statistic after knowing the correct answer, then it is possible to do so to manipulate and therefore invalidate the apparent p value and the judgement of the null hypothesis.

With n possible treatment intervals, there is a 1/n probability that the actual randomly selected treatment interval will have the largest value of the test statistic purely by accident. The p value resulting from the experiment will be 1/n times the rank of the actual test statistic among the n possible. If we reduce design 1 to 10 or 5 possible treatment intervals, then we can calculate the p value only to the nearest 10 or 20%. With 100 intervals, the minimum p value would be .01 and the treatment statistic would only have to be in the largest 5 of the 100 possible to be significant at the .05 level.

Section 5.3 has more material on hypothesis testing. Some warnings on p values are given in 5.3.5.

3.3.4 One-Sided and Two-Sided Tests

In 3.3.2, the calculated test statistics are the differences between successive measurements. The p value is lowest when the actual treatment difference is the largest, with negative numbers being less than positive numbers. This corresponds to an alternative hypothesis that claims that the treatment will increase the measured attribute of the object. If the treatment is expected to decrease the measurement, the test is reversed by reversing the signs of the differences. In either case, the test is called a one-sided test, since it is sensitive to differences in only one direction. A treatment difference that is large but in the opposite direction will have a high p value that supports the null hypothesis. If the alternative hypothesis is a positive treatment effect, then the null hypothesis is effectively that the treatment has no effect or maybe a negative one.

A does a two-sided test that is sensitive to changes in either direction by ignoring the signs of the differences and sorting them according to their absolute values. The null hypothesis is that the treatment has no effect, and the alternative is that it has some effect, in either direction. This is especially appropriate for treatments such as cleaning or consolidation that could make the object look worse instead of better. Using the signed difference as the test statistic increases the probability of detecting beneficial treatments but at the cost of giving up the possibility of detecting harmful treatments. Using two-sided tests is therefore standard practice.

This distinction only applies when there is one treatment and a control of no treatment or two treatments and no control. With more treatments, there are several ways to violate a null hypothesis of treatment equivalence. Most standard tests are multi-sided: they are sensitive to any difference among the treatments.

3.4 DESIGN 2: MULTIPLE TREATMENT INTERVALS

If the treatment has an effect that is either temporary or cumulative, then it can be applied many times to the same object and make a visible difference after each application. Reagan (1982) microwaved wool fabric samples several times and counted the number of live and dead insects after each successive treatment. This design can reduce drastically the number of intervals in the experiment without increasing the uncertainty in the null hypothesis p value. A balance of treatment and control intervals is usually best.

3.4.1 Example with Calculations

In particular, let M make seven measurements and C randomly select three of the six intervals for treatment. In the design 1 example, there were 20 ways to select one interval from the 20 available. Here there are 20 ways to select three intervals from the six available. Therefore this example version of design 2 also has a minimum p value of 5%, nicely matching the example version of design 1.

Table 3.4.1A Original and derived data tables for design 2 example

Day	Measure
1*	1
2	8
3	6
4*	9
5	12
6*	11
7	16

Interval	Difference
1*	7
2	-2
3	3
4*	3
5	-1
6*	5

* treatment

A calculates the test statistic as the mean of the three treatment differences minus the mean of the three control differences. Suppose the original measure and differences are as in table 3.4.1A. Then the treatment effect is estimated as (7+3+5)/3 - (-2+3-1)/3 = 15/3 -

0/3 = 5 - 0 = 5. A repeats this calculation for each of the other 19 possible sets of three treatment intervals selected from the six available. There is again a 1 in 20 chance that the test statistic will be largest for the actual treatment assignment. A would ignore signs to make a two-sided test.

Table 3.4.1B shows the test statistic for each of the 20 possible treatment assignments. Since both treatment and control effects have a divisor of 3, the division has been omitted. The test statistic for the actual treatment assignment is therefore listed as 15 -- instead of 5, as calculated in the last paragraph. Doing the division or not has no effect on the relation of different values. To properly estimate the treatment effect, the division is necessary. To calculate the p value, it is not, so it is omitted to save effort.

Table 3.4.1B Randomization calculations for design 2 example

Treatment intervals	Statistic	Treatment intervals	Statistic
1 2 3	8- 7= 1	4 5 6	-1
1 2 4	8- 7= 1	3 5 6	-1
1 2 5	4-11= -7	3 4 6	7
1 2 6	10- 5= 5	3 4 5	-5
1 3 4	13- 2= 11	2 5 6	-11
1 3 5	9- 6= 3	2 4 6	-3
1 3 6	15- 0= 15	2 4 5	-15
1 4 5	9- 6= 3	2 3 6	-3
* 1 4 6	15- 0= 15	2 3 5	-15
1 5 6	11- 4= 7	2 3 4	-7

* actual treatment assignment
Interval values are 7, -2, 3, 3, -1, and 5.

The symmetry of the design means that switching treatment and control intervals merely changes the sign of the statistic, so only half must be calculated. Those in the second column were derived from those in the first by reversing the signs. Calculation is further simplified by remembering that the sum of the differences (15) is the last measurement minus the first (16 - 1) and that the sum for the control intervals is the total sum minus the treatment sum. In the last line of the table, 4 can be calculated as 15 - 11 instead of -2 + 3 +3.

Because intervals 3 and 4 have the same difference of 3, there are duplicate values. The actual assignment is tied with another for top position, and so p is .10. This value suggests that the treatment in question has an effect but is not usually considered conclusive. If the differences for intervals 3 and 4 were 2.9 and 3.1, there would be no tie and p would be .05, which is stronger evidence of an effect. Randomization tests are more powerful when measurements are sufficiently accurate to avoid ties. With an odd number of intervals, avoidance of changes calculated as 0 is also better.

If signs are ignored to do a two-sided test, then ties are inevitable with an even number of intervals, and the minimum p value for six intervals is .10. In this example, the additional ties would make the minimum .20. If the minimum p value is too high for rejection of the null hypothesis, then the experiment is too weak. It should be modified or not done. For two-sided tests with design 2, an odd number of intervals will break the symmetry and result in a minimum p value the same as for a one-sided test.

Design 2 is an improvement over design 1 in that it is faster and requires fewer measurements, reducing the cost of both waiting for an answer and of making measurements. However, it requires more treatments and thus increases that cost. Averaging across three treatments somewhat improves our confidence in the generality of the results. The multiple treatments also give an estimate of the variability of treatment effect. But making these gains requires a possibly dubious assumption, which adds an uncertainty cost. If the treatment has a permanent effect that is not enhanced by additional applications, then treatment intervals after the first will have a difference similar to the control intervals. The estimated treatment effect will be less than its true value. This dilution increases the danger that we falsely conclude that an effect is not significant. The reduction of control intervals from 19 to 3 reduces our information about inherent measurement variation, but this is usually of lesser importance. Finally, the increased calculation in the analysis suggests a need for computer programs.

3.4.2 Selecting Combinations

The number of ways to select k intervals for treatment from among n intervals in the experiment is n!/[(n-k)!k!]. In this expression, n! is the factorial function n x (n-1) x (n-2) x ... x 3 x 2 x 1 for positive integer n and 0! = 1. For example, the number of ways to select 3 intervals from 6 is 6x5x4x3x2x1 / (3x2x1 x 3x2x1) = 20. Table 3.4.2 gives the number of possible treatment sets for balanced experiments for n from 1 to 10.

Table 3.4.2 Number of ways to select k items from n total

N	K	Subsets	Ratio
1	0,1	1	
2	1	2	2
3	1,2	3	1 1/2
4	2	6	2
5	2,3	10	1 2/3
6	3	20	2
7	3,4	35	1 3/4
8	4	70	2
9	4,5	126	1 4/5
10	5	252	2

The ratio in the last column is the ratio of the number of subsets on the current line and the line above. The pattern of ratios can be used to extend the table. The number of ways to select 5 or 6 items from 11 is 252 x 1 5/6 = 252 x 11/6 = 42 x 11 = 462. The number of ways to select 6 from 12 is twice that, or 924.

Even for large N, random selection of a subset for treatment can be done by the manual methods given in 4.4.2. However, when N reaches 7 or 8, a computer becomes necessary for analysis by calculation of the test statistic for each possible subset. At present, programs are not as easy to find as they should be, but we expect this situation to improve in the future.

3.5 OTHER DESIGNS FOR ONE TREATMENT ON A SINGLE OBJECT

3.5.1 Design 3: Single Treatment Patch

We can often treat a small test patch on an object instead of treating the entire object. A modification of design 1 is to treat one of many possible patches instead of one of many possible intervals. An experiment that leaves most of the object untreated is necessary if the experiment is undertaken to determine how best to treat the object.

An example version of design 3 has 20 patches on the object. They do not have to be marked on the object but must be dependably relocatable by both M and C using marked photographs, recorded coordinates, or other descriptions. M measures all 20, C randomly selects 1 and treats it, and M remeasures all 20. M gives A a data table with the form of table 3.5.1, where the Ms indicate data filled in by M. A calculates the 20 differences marked A (the fourth column) and proceeds as in design 1. C indicates where the asterisk (*) for the treated patch goes (the example has patch 8 so marked).

Table 3.5.1 Data table for design 3

Patch	Before	After	Difference
1	M	M	A
2	M	M	A
3	M	M	A
4	M	M	A
5	M	M	A
6	M	M	A
7	M	M	A
8*	M	M	A
9	M	M	A
10	M	M	A
11	M	M	A
12	M	M	A
13	M	M	A
14	M	M	A
15	M	M	A
16	M	M	A
17	M	M	A
18	M	M	A
19	M	M	A
20	M	M	A

* patch actually treated

Design 3 is much faster than design 1, but takes 40 measurements instead of 21. There is also the additional effort of defining 20 patches instead of just 1 and the concomitant multiplying of any deleterious effects of the measurement process.

3.5.2 Design 4: Multiple Treatment Patches

We combine the ideas of balance and patch treatment, which produced designs 2 and 3 from design 1, to produce design 4. An example version consists of treating three patches out of a possible six that we measure at the beginning and end of one treatment interval. This only requires the assumption that multiple patches can be treated without affecting each other. Again, there are 20 possible treatment arrangements and a 1 in 20 chance that the three patches treated have the three largest values of the test statistic. The data table is similar to 3.5.1 but with fewer lines and more asterisks.

The comparison of designs 1 and 2 given above also applies to designs 3 and 4, with the modification that the repeated treatment assumption is replaced by a neighboring treatment assumption. Averages and deviations then apply to spatial rather than temporal differences. Similarly, the comparison of designs 1 and 3 also applies to designs 2 and 4.

3.5.3 Design 5: Multiple Treatment Patch-Intervals

Designs 1 to 4 show that experimental design involves tradeoffs among several cost factors:
1. spatial extent -- number of patches;
2. temporal extent -- number of intervals;
3. number of measurements;
4. number of treatments;
5. various types of uncertainty about the results.

These tradeoffs continue to appear in additional designs that have both multiple patches and multiple time intervals.

An example of design 5 has three spatial patches and two time intervals. There are six possible patch/interval treatment units from which three can be randomly selected, as in designs 2 and 4. Table 3.5.3 outlines the data structure. Three of the six patch-interval differences calculated by A would be marked as treated. Once the patch-intervals are arbitrarily numbered 1 to 6, the randomization analysis table would be similar to 3.4.1B.

The number of measurements (9, including the before measurements) is between the numbers (7 and 12) needed for designs 2 and 4. A disadvantage is the need for the often-dubious assumption about the additive effect of multiple treatments. Even if true, the likely imbalance of treating one patch twice and another never is undesirable. Such imbalance will confuse spatial differences with treatment differences when the data are analyzed. This phenomenon of one type of effect masquerading as another because of the design is called *confounding*.

Table 3.5.3 Data table for design 5 example

				Inverval 1	Inverval 2
Patch	Before	Middle	After	Mid-Bef	Aft-Mid
1	M	M	M	A	A
2	M	M	M	A	A
3	M	M	M	A	A

M: measurement supplied by measurer M
A: calculated by analyst A

3.5.4 Design 6: Paired Treatment-Control Intervals

We can remove the need for the additive treatment assumption and the possibility of treatment-patch imbalance in a second patch-interval design. The cost is additional patches, measurements, and treatments. We restrict the randomization so that we treat each patch exactly once, either in the first interval or in the second, and obtain design 6. With two choices for each of n patches, there are 2-to-the-nth-power sets of treatment assignments. This gives 16 sets of treatment assignments with 4 patches and 32 with 5. With this design, A calculates the difference between treatment and control for each patch separately and takes the average across patches to estimate the treatment effect as indicated in table 3.5.4.

Table 3.5.4 Data table for design 5 example

	Treatment				Effect		
Patch	Interval	Before	Middle	After	Treatment	Control	T - C
1	C	M	M	M	A	A	A
2	C	M	M	M	A	A	A
3	C	M	M	M	A	A	A
4	C	M	M	M	A	A	A
5	C	M	M	M	A	A	A

C: treatment interval 1 or 2 randomly selected by conservator C
M: measurement supplied by measurer M
A: calculated by analyst A according to the treatment interval

Design 4 with six patches and design 6 with four patches both require 12 measurements. The latter requires one more treatment and has four fewer sets of assignments, giving a minimum p value of .07 instead of .05. It should only be used if the patches are sufficiently heterogenous in responses to make it desirable to compare treatment and control within patches instead of between patches. This pulls patch differences out of the estimate of treatment effect. However, subtraction of pre-treatment values is usually sufficient. Design 6 is included here partly for completeness and partly to introduce ideas that reappear in chapter 6 (6.1.3).

3.5.5 Design 7: Order-Balanced Paired Intervals

In designs 1, 2, 5, and 6, we assumed that taking differences between the measurements at the beginning and end of an interval is sufficient to eliminate any interaction or carryover between treatment and control effects. In other words, we have assumed that the control effect is the same whether a control interval comes before or after treatment.

If we suspect an order effect, we can modify design 6 by further restricting the randomization so that the treatment is applied to an equal number of patches in the first and second interval. An experiment balanced for treatment order facilitates analysis for carryover. An example version of design 7 has six patches and two intervals with three patches treated in the first interval and three in the second. There are 20 possible treatment assignments, as in design 4, versus the 64 in design 6 with six patches.

This design is more costly in time, measurements, and treatments than design 4 with the same number of patches and possible assignments. Design 6 would only be used if thought necessary to control for both patch heterogeneity and order effects. Carryover is more likely if, to anticipate the next section, we compare two treatments instead of treatment and control.

3.6 MULTIPLE TREATMENTS AND CONDITIONS

With one treatment under consideration, the question is -- does this treatment have a good effect, or does it have a deleterious effect, or no effect? Especially with multiple patches, we can rephrase the question as follows: Is our proposed treatment better or worse than the control treatment of doing nothing?

In these terms, we can view and use the one-treatment designs 1 to 7 as two-treatment designs. However, we should eliminate the unbalanced designs (1,3) where one treatment is given once and the other several times. They are inefficient in discriminating between two treatments that have anywhere near the same cost of application. Two-treatment designs with multiple intervals (1,2,5,6,7) and one or more applications of each treatment to each patch are problematical. There may be occasions where applying each treatment once to each patch may be tolerable (6,7), especially with order-balanced (7). Generally, however, the balanced multiple-patch design (4) is the choice for conservation research.

Human studies are different. With humans, patch treatments are usually not possible and many treatments have temporary or cumulative effects on the whole organism. Design 2 is thus more common than design 4 for one-subject human experiments.

If there is a standard treatment that is going to be applied without indication of anything better, then one can directly compare the new treatment to the standard treatment and leave out the control of no treatment. Medical researchers often do this when the benefit of a drug is well enough established to make it unethical to give a placebo treatment. (A placebo is composed of substances known to have no effect except the possible psychological effect of contact with a healer and the apparent action of doing something.)

We can easily extend design 4 to three or more treatments while maintaining the principles of balance and random assignment. One of the treatments may still be a control or standard treatment that is a benchmark of comparison for the new treatments being tested. We can also apply the multi-factor designs discussed in chapter 6 to single-object studies. Additional complications arise from multiple patch types. Blue areas and red areas in a painting might need different cleaning or treatment procedures. At this point, we are approaching multiple object designs, since we would effectively be treating pigment type as the experimental unit, rather than the painting as a whole.

CHAPTER 4

EXPERIMENTAL DESIGN

4.1 GOALS

4.1.1 Strategy

The single-object designs in the last chapter show that we must consider several items when planning an experiment. For discussion, we group them into five areas -- goals, objects, measurements, treatments, and analysis. This chapter has a subsection to each of the first four areas. The next chapter focuses on certain aspects of statistical analysis that are especially pertinent to the discussion of designs in chapters 6 and 7.

While we have to discuss these topics in some particular order, the process of design requires some skipping around and backtracking. It involves various tradeoffs between different costs and benefits and adjustment of interconnected factors. It should involve consideration of alternatives for the aspects that are not completely determined by the purposes, goals, and hypotheses of the study.

At times, the number of possibilities is overwhelming. It helps to identify which aspects of the experiment are fixed and which are still subject to adjustment. If there are still too many variables to keep in mind at one time, fix some more, even if only temporarily. In particular, design involves minimizing costs and maximizing benefits, but it is not possible to do both simultaneously. Holding one of these two major properties of a study constant while optimizing the other leads to two basic strategies of design.

4.1.2 Questions

The first step in designing a study is to formulate the goals, questions, and hypotheses, as discussed in the first two chapters. For the most part in the rest of this book, the goals and hypotheses of a study will be taken as fixed. However, knowledge of what information one can gain with limited resources will influence the initial choice of a goal.

Art conservation research questions should initially be expressed in art conservation terms. While designing a study, add detail about time, place, and circumstance. Turn questions about general classes of objects into questions about specific objects. Translate questions and hypotheses into technical and even statistical terms as needed. Alternative designs represent alternative specific goals and hypotheses.

4.1.3 Research Programs

Thinking in terms of a research program influences the design of a particular experiment.

A basic strategy is "divide and succeed": Start with a broad question and work towards the details as appropriate. For instance, to study air pollution as a cause of watercolor discoloration, we might first compare pure air (a nitrogen + oxygen mixture) to typical city air with a mixture of pollutants. If we find air to be important, then we continue with further experiments with specific atmospheric components to isolate the guilty pollutant(s).

Another example is a study assessing the performance of an adhesive on ceramics of varying composition. The first experiment starts with porcelain (pure clay), glass (pure silica), and a heterogeneous mixture of clay, sand, and organic material. If ceramic body type affects adhesive performance, follow-up experiments would use other combinations of components to identify further the mechanisms involved.

Research programs use the results of one experiment to design the next. Important results usually appear from such progressive development. Trying to do everything in one experiment may confuse both the experimenter and everyone else, and the results may not be clear. Most crucial experiments are in the middle of a long chain of experiments. The basic pattern is to do a little bit and then see where to go with more detailed experiments.

A single experiment typically lasts about a month, at most three. It might even be shorter. Aim for the simple yet elegant experiment.

4.1.4 Observational Studies

In an observational study the researcher does not manipulate variables but observes the outcome of natural processes. Careful formulation of the research problem, selection of objects, and measurement of variables are just as important for this type of study as for experiments. The outcome is a correlation rather than a causal inference. The basis for hypothesis testing has to be random selection from a population (see section 5.3) rather than random assignment of treatments, since there are no treatments.

To clarify the relationship between the two types of studies, consider the hypothesis that exposure to chemicals used in conservation treatments contaminates the blood of conservators. An observational study might randomly sample conservators and an unexposed control population, sample and analyze the blood of each participant, and then correlate occupation with blood level of various chemicals. An obvious problem is selection of an appropriate control group. A negative result would not necessarily be conclusive since those who cannot detoxify themselves may self-select themselves out of conservation work.

An experimental version of the study might start with a blood sample, randomly assign half the members of each group to a week of either office work or laboratory work with known exposures, and finish with a second blood sample. The problem here is the difficulty of enlisting subject cooperation and the possible inadequacy of only a week of exposure. Because different types of experiments have different advantages and disadvantages, medical researchers study human disease in laboratories, clinics, and natural populations rather than focusing on just one type of study. Conservation scientists can do the same with art object deterioration.

4.2 OBJECTS

The goal of an art conservation study is to gain knowledge about art objects and their conservation. The investigator must first decide whether to observe and experiment with actual art objects or to use simulated or surrogate objects composed of art materials. This is similar to the basic decision for a biological investigator between working with live organisms (*in vivo*) or cell or tissue cultures in glass dishes (*in vitro*). This decision has a major impact on the character of an experiment, which is why we have used it to classify studies by type. Working with surrogate entities eliminates many ethical constraints and adds flexibility. However, the results obtained need verification with actual art objects.

The next decision is whether to study one or a few objects intensively, or many objects more superficially. This also has a major impact on the type and design of the study. With this determined, the remaining questions are how many and which objects (or materials and sources thereof) to use. How will we select, group, and characterize the objects? Will we group them by existing characteristics or those we impose by manipulation?

Clinical trials in medicine have explicit entry and exclusion criteria based on patient characteristics. Examples are sick enough to need treatment but not so hopeless that cannot benefit, competent to give informed consent, available for the entire study, and no history of interfering prior treatments. Similar criteria apply to conservation studies on real objects. One difference is that the owner of the object rather than the object itself must give consent and promise availability until the end of the study.

4.2.1 Study Units

This book makes frequent use of the concept of "unit" as a generalization of the concept of "object." The generalization includes collections of objects considered as a whole, portions of an object, and batches or portions of an art material. The unit concept also carries with it an implicit contrast to the concepts of group and class. Units of various types are subject to measurement and manipulation as part of a study.

The primary unit of a study should be the type of unit one wants to learn about. It may be called an experimental or observational unit, depending upon the type of study. In the study of the works of a particular artist, the study units are usually the art objects attributed to that artist. In a survey of a museum collection, the works of an artist at that museum could be one unit characterized by number, value, average condition, and priority for expansion. In single object studies, the primary unit for measurement and treatment is sometimes a portion of the object. In a treatment study, bottles of glue from various manufacturers might be study units.

The related word "sample" is somewhat ambiguous. It sometimes refers to the portion of one unit used for analysis. It may also refer to several units selected from a population for a study. The conservator blood contamination example above used "sample" in both senses. The word "specimen" frequently refers to individual study units.

4.2.2 Replicates

Replicates are multiple objects or experimental units of a given type that we measure under the same particular set of relevant treatment conditions. We expect replicates to behave about the same. However, because of unavoidable object, treatment, and measurement variations, replicates vary in the exact measured value of their variables.

Replicates therefore have a dual purpose. The first is to estimate treatment effects more accurately than is possible with one object by canceling errors when calculating their average behavior. The second purpose is to get an estimate of variability that is unobtainable from a single experimental unit. The latter is important if we intend replicates to represent a larger class or population of objects. In particular, statistical hypothesis testing uses such estimates of variability.

Splitting a single experimental unit into pieces does not create replicates of that unit. For example, suppose we experiment with adhesive A as a conservation material for paper type P. Suppose we want the results to apply to the class of batches of adhesive A rather than to the class of aliquots from a single batch of adhesive A. Then we gain more information about the population of batches if we take one sample from each of several different batches to form a group of replicates.

If we take many samples from one batch, the multiple analyses are usually repeated readings of that one batch (4.3.1). The conclusions of the experiment would strictly apply only to that one batch of adhesive. Similarly, if we want our conclusions to apply to the class of paper P and not just to one particular sheet of paper P, we should use multiple sheets.

Conservation research is somewhat different from research in chemistry or physics. Except for minor impurities, reagent-grade chemicals should be the same from batch to batch. In contrast, most art materials receiving a conservation treatment are heterogeneous, structured, mixtures. For example, differences in raw material, processing, and aging result in differences in the "paper" of museum and library items. If all samples in a study derive from one roll of paper or one bolt of cloth, the experiment itself gives no idea of how well one can generalize to other rolls of paper or other bolts of cloth.

Conservation studies frequently commend or condemn conservation treatment methods and materials for use on art objects. Such experiments should include replication in order to assess the potential variability of treatment results and to safeguard against errors leading to fluke outcomes. Even studies of real art objects can usually have replicates. Treating and analyzing fewer types of objects but including replicates often improves the reliability of a study.

4.2.3 How Many Experimental Units?

The question of how many replicates to use in each group is important but hard and fast rules are hard to come by. The answer depends on several factors. Rules such as ASTM testing standards may or may not give appropriate numbers for a particular experiment.

The standard approach given in statistics and experimental design books is to fix the goal of the experiment and calculate the required number of units from formulas and tables. This method requires the following: an estimate of the variability of the outcome measure and the assumption that the variability will remain about the same after treatment; a decision about the statistical analysis that will be done, including the p value that will be considered significant; and a decision about the size of effect that one wants to see and with what probability.

The estimate of variability requires prior experience with the material and measurements of the study. The decision about effect size is usually difficult for people to make. Answers to the question "What size of effect should be declared significant with a probability of 80%?" and the same question for 90% tend not to be consistent in terms of the derived study size. Perhaps asking the question for a 50% probability would be more successful. This approach also requires the ability to calculate statistical power from non-central distributions that are specific to particular experimental designs and statistics and that are usually based on an assumption of normal errors.

A more realistic approach is to fix the resources of the experiment, consider how many objects are possible or convenient, and decide if the potential answer is worth the cost.

The optimum number of study units depends upon the magnitude of differences that you want to see, the uniformity of the material you are studying, and the precision of your measurement techniques. If you have no prior information on how much variation to expect, start with two or three replicates. Experimenting with one object gives you no information on variability. However, using an arbitrarily large number of replicates at the initial stage of research can waste time and money.

Increasing the number of treatments tested in one experiment tends to reduce the number of replicates needed. With many treatments, two replicates for each may be enough to show the variability of outcome. If the variability is large, follow-up experiments can have more replicates. With one treatment, six or eight replicates give a good idea of the variation. It will rarely be necessary to use extremely large numbers of replicates, and it is not necessarily the case that "the more the better." Biological experiments are often successful with six to eight replicates when two to four treatments are tested with quantitative outcomes. This is true in spite of the high amount of variation between living organisms. Increasing the number of replicates up to 20 or 30 will improve the determination of the variation present in the population, but after that there is little gain. The more replicates you use the more likely you are to see significant differences, but there is a limit to what is useful.

Experiments involving binary outcome measures generally need many more replicates than experiments with quantitative measures. Suppose the current standard treatment for a certain class of badly damaged objects has a 50% success rate as measured by some criterion such as being acceptable or not for museum display. There is a new treatment that is more expensive but might, for plausible reasons, be more successful. We propose to test the new treatment against the standard treatment in a randomized trial. The treatments will be declared to have significantly different success rates if the p value for the chi square statistic is less than or equal to .05. (Chi square is one standard test statistic used for designs with two treatments and two outcomes.) If we want an 80% chance of detecting a 20% treatment difference, then we need 73 objects in each group, as calculated by a standard formula found in statistical texts. This number is likely to be impractical and suggests that we look for an alternative.

The required study size can be reduced by accepting p values above .05 as significant. This increases the chance of thinking that the new, more expensive treatment is an improvement, when in fact it is not. Or we can accept more than a 20% (100% minus 80%) chance of missing a true improvement of 20%. The best choice when possible is to change the outcome measure to one that makes finer distinctions.

Experiments without replicates require assumptions, extra care in statistical analysis, and qualification or avoidance of significance statements. A common design without replicates is a screening experiment. Industrial chemists, for example, may have hundreds of combinations of factors to test but insufficient time and money to do replicates or even to test all combinations once. One of their solutions is to do fractional factorial experiments (see 6.3.6), which only look at main effects and assume away interactions. They give a general idea of what to expect, what seems to work, and what does not. They are not used to discover laws of nature but are used only as a "quick and dirty" way to improve upon current practice. Without replicates there is no protection against experimental blunder. Treatments passing a screening experiment should be studied more or carefully monitored if immediately put in use.

4.2.4 Random Selection

Random sampling of experimental units or objects from a population of interest is one method to eliminate bias. It makes probability calculations possible for statistical tests. However, it is often difficult or impossible to do when working with real art objects. Stratified random sampling -- random sampling of selected strata such as the edge of a painting -- is one alternative. Spread sampling, in which one explicitly attempts to encompass as much of the variation present as possible, is another strategy. An example is an authenticity study to exclude the possibility that a piece dates from a particular period by showing that it has a characteristic never found in that period. This requires samples encompassing all the possibilities of the period. Sometimes when working with art objects, the only practical method is to take what you can get. In research reports, describing the method and rationale for choosing objects is especially important when the selection is haphazard instead of a random sample from a population of interest. Any generalization of the results to objects other than those studied will then have to be made on non-statistical grounds.

4.3 MEASUREMENTS

The investigator must decide what variables to measure; how, when, and where to measure them; and what units and method of recording to use. This forms a measurement protocol. Some variables may be set by the investigator while others are observed in their natural state. Study variables may be divided into those describing the objects, those describing the treatments, and those describing the result of applying the treatments. Some result variables may be included to detect treatment side-effects that the investigator hopes will never occur.

4.3.1 Dataset Structure

All variables must be recorded and organized, which usually involves some combination of paper data forms and computer files. The measurements form a dataset that has a structure that needs to be well-understood for its correct analysis and interpretation. Structural outlines of data tables for several example designs are given in chapters 3 and 6.

An important part of dataset structure is the presence of missing values or holes in the dataset. They should usually be avoided if possible. This should be easier in conservation research than in medical research, where cultures and organisms sometimes die prematurely.

An occasional planned exception is a block of variables applicable to only a subgroup of objects. These variables require subanalyses with just those objects. When studying copper-alloy statues, for instance, clay core mineralogy and elemental composition can be important. However, solid-cast statues have no clay core to analyze.

Table 4.3.1 Example variables of different types

Categorical	artist
	place (identified by name)
	pigment type
Ordered	scales such as awful, bad, fair, good, and superb
	ranks of objects sorted by some quality
	mineral hardness scale from 0 to 10 (with diamond = 10)
Arithmetic	linear time
	strength, reflectance, color intensity
	most instrument readings
Binary	chemical or pigment present or absent
	treatment applied or not
Spatial	x,y coordinates on a painting
	color coordinates in various systems
Cyclic	time of day
	season of year
	angle around the axis of a statue

Another crucial aspect of the structure of a dataset is the set of values allowed for each variable. Possibilities include unordered categories, ordered categories or scales, true arithmetic numbers, and specialized forms, such as time of day or two-dimensional location. Binary variables are variables with only two possible values, such as present or absent, yes or no, or 0 or 1. They can be treated as either categorical or arithmetic, as convenient. Some art conservation examples are given in table 4.3.1.

Replicate measurements are measurements of the same variable on replicate units. Multiple measurements of the same variable on one unit can be either repeated readings or repeated measurements. It is useful to differentiate multiple measurements by our reason for making them and our expectation for their outcome. These factors affect what we do with them. Our treatment of a set of numbers depends upon our viewpoint, which can change during analysis.

4.3.2 Repeated Readings

Repeated readings are multiple measurements of one variable on a single experimental unit. They are separated by time, space, portion of the unit, instrument, or some combination of these, but not by a factor that is part of the study. Repeated readings monitor and average out measurement errors, such as instrument noise and reading error, for increased accuracy. We expect them ideally to be the same. We are not consciously varying any factors that should cause them to be different. They generally are taken one right after the other, with three being a typical number. The multiple readings are summarized by their average or median to give one number for use in the study. They could be treated as measures on a micro-experimental unit. However, their variation is not of much interest and is usually suppressed, since it has no direct bearing on the analysis of the experiment. Variability does appear indirectly, since replicate variance will include measurement reading variance.

Computerized readings taken directly from a machine with no significant drift or noise may not need repeated readings. It is also possible that the averaging process is built into the equipment or computer program. Standardized procedures usually exist for each instrument in a given laboratory. To decide how many repeated readings to do, compare their variability to the difference between experimental units or replicate samples. If the variation is relatively small, multiple readings may not be worth their time and cost. More replicates or another treatment group may be more valuable.

4.3.3 Repeated Measurements

Repeated measurements on a single object assess changes within that object under different conditions. The experimental unit or specimen is measured more than once, either at different times or different places or both. We expect repeated measurements to possibly vary because the time or space separation is sufficient for natural change or because each has a different treatment condition. Variation is the object of the study of repeated measurements. We therefore do not immediately average these numbers as with repeated readings, although there may be later analyses in which we do so.

The changing factor is time when we measure color before and after artificial aging or take several sequential measurements for a decay curve. Time effects may be driven by internal dynamics of the object under constant conditions, natural variation of environmental conditions, or experimentally imposed variations. Constant conditions may be natural or artificial. Ensuring control of all relevant conditions usually requires an enclosed chamber. Natural or imposed variations include light, temperature, and humidity.

The changing factor is space when we test four varnish removal agents on four patches of each of two pigments on a painting. The painting has eight test patches and eight corresponding measures characterized by treatment and pigment. There are two repeated measures for each of the four treatments and four repeated measurements for each of the two pigments. If each patch were measured twice, before and after varnish removal, there would be 16 repeated measures characterized by the three dimensions of time, pigment, and treatment.

The experiment tests both treatment and pigment type as spatially separated repeated measures factors. Patches are often different because of actions by artist or conservator that have no inherent relationship to position. They may also differ for positional reasons such as gradient and edge effects.

It is usually not productive to make repeated measurements simply for the sake of making them without a clear reason to expect variation. Repeated measurements should be tied to the purpose of the study. For example, kinetic studies of change with time can use repeated non-destructive measurements of one sample after several periods of time.

4.3.4 Avoidance of Bias

Loosely speaking, a biased experiment is one expected to give a wrong answer because of its design. Avoid object bias by random selection and care in making generalizations, as previously discussed. Avoid treatment bias by random treatment assignment, as discussed in the next section. Avoid measurement bias by careful organization of measurement procedures.

Normally, all measurements of a given variable should be made with the same combination of people and instruments. Measuring replicates in treatment group A with one setup and those in treatment group B with another will add the deviation between setups to the apparent treatment difference. Recalibrating an instrument, as is sometimes done with each run of a batch-oriented process, effectively creates a new setup. Measurements with different setups should be assessed and possibly corrected for systematic differences. Treatment groups should be mixed among setups.

Similarly, objects and patches should be measured in a time sequence that avoids having instrumental drift show up as a spurious treatment effect. Measuring all the replicates in treatment group A first and then all the replicates in treatment group B is a bad procedure. One solution is to measure objects from the two groups alternately (ABAB...). This will work as long as the instrument does not have the quirk of alternately giving high and low measurements. The most reliable solution is to randomize the order of measurement.

Repeated measures require additional care. If the measurement unit is a patch or spot on the treatment or study unit, then each measurement must be located. For example, color measurements on 1" x 3" strips of colored paper coated with adhesive measure only a small portion of the specimen. Again, the solutions are to be either systematic or random. To be systematic, measure each sample in the same place, such as its center. Another set of measurements after a period of artificial aging should be taken in the same spot for each specimen as before. When measuring each specimen at the center and along the edge to check for edge effects, be consistent in selecting the edge position. While random selection of measurement position is statistically useful, so is systematic selection of positions that reduce the variation between measurements. In addition, the physical mechanics of locating and measuring a particular random spot is sometimes difficult. Generally, randomization is more useful for measurement order than for measurement location.

Conscious or subconscious preconceptions can affect any step of the measurement process done by human beings. People knowing the treatment applied to each unit sometimes tend to make systematic errors and see and record the results they expect. The conscious effort required to avoid such error can be bothersome. Whenever possible, measure without knowing which object received which treatment. If possible, have one person apply the treatments and another person record the results. If that is not possible, have an assistant shuffle and add random code numbers to the specimens between treatment and measurement. Automated recording of measurements removes both bias and transcription errors.

4.4 TREATMENTS

The basic questions are how many treatments, which treatments, the reason for their choice, their time and place of application, and their assignment to objects. The treatment protocol for a study should be explicit and put into writing.

It is important to identify at the outset whether a project is a survey or an experiment. An experiment requires some treatment or manipulation of variables by the scientist. The word treatment is used here in the broad sense of action designed to have an effect, as opposed to the narrower sense of action designed to improve a condition.

Studies of the composition of works of art, although requiring scientific analyses, are observational studies. Studies of deterioration can be either observational or experimental. Studies of conservation materials and methods and their usefulness in conservation practice should be true experiments when possible.

We can apply different treatments to different parts of a single object. It is then best if each object in the experiment receives all treatments tested, so that they are compared on the same group of objects. This permits the use of statistical computer programs for the analysis of repeated measures, which require a complete design with no holes or missing values. A good example is Barger *el al* (1984), who studied two coatings and a control treatment on 17 nineteenth-century daguerreotypes. They applied each treatment to one-third of each daguerreotype for a complete repeated measures design.

4.4.1 Controls

Controls are objects that receive a null or neutral treatment. A simple experiment varies only one factor at a time, holding constant all other variables. If we observe an effect, then we can attribute it to that one factor. In practice there are often variables that we cannot control. There will be others that we do not even measure. We therefore need controls to be more confident that the measured effect was indeed due to the factor we intentionally varied. Controls are objects subjected to the same protocol as others except that the treatment is replaced by either a sham or dummy treatment or by no treatment at all.

For example, we measure the color and pH of adhesive-coated papers on day one, thermally age each preparation for 30 days, and remeasure to determine the difference. Control specimens are paper units prepared, measured, and aged exactly as all the others were except for the adhesive. Some have nothing applied. Others receive dummy treatments consisting of an application of the solvent that serves as a vehicle for one of the adhesives. A third set of controls gets adhesive but no thermal aging.

A similar experiment exposes to ozone canvas swatches painted various colors. Demonstrating that observed changes are caused by ozone is aided by controls that are not exposed to ozone. It is also useful to know the effect of ozone on the color of unpainted control swatches.

4.4.2 Randomization

One purpose of randomizing the treatment of experimental units is to prevent the bias that can occur when an investigator subjectively chooses which units get which treatment. Another is to evenly distribute among the treatment groups uncontrolled factors that might affect treatment outcome. Except in special situations, each treatment unit should be equally likely to get any specific treatment. Moreover, the assignment of treatments to different units should be independent except for randomization constraints, such as having equal numbers of replicates for each treatment.

Violation of these conditions impairs the validity of statistical analyses and the reliability of conclusions. Randomization is assumed in mathematical methods of estimating error and testing the significance of observed differences. A corollary is that the results of an experiment done without randomization may be mistaken and not be reproducible.

Two examples appear in Wilson (1952). The first experiment tested the accuracy of a particular analytical procedure. Each sample analyzed was immediately reanalyzed. Then the next sample was analyzed twice. Agreement between the two analyses of each sample was good. So the experimenter confidently accepted the usefulness of the analytical procedure. Later, another laboratory did the same experiment but randomized the order of analyses. The two analyses of each sample were no longer in direct sequence but were intermixed with all the rest. The new results showed wide discrepancies within pairs of analyses. Additional work showed that a zinc reductor gradually lost its effectiveness because of certain elements present in the samples. From one analysis to the next the effect was small, but by the end of the day the absolute values were very much in error. Thus the analytical procedure was not very useful.

Another experiment tried to relate the length of time a plastic part is pressed in its mold to its subsequent strength. Hot plastic was injected into the mold, pressed for 10 seconds, then removed. The next pressing was 20 seconds, then 30 seconds, and so on. A plot of strength against pressing time showed a strong linear dependence. However, the research supervisor criticized the lack of randomization of pressing time. When the experiment was repeated with randomization, the linear dependence of strength on pressing duration disappeared. It turned out that it was the order, not the duration, that had an effect. As the mold grew warmer and warmer during the experiment, the strength of the plastic part increased.

An example of treatment randomization follows. We want to test metal coatings A and B for discoloration over time when applied to five metal plates of different compositions. We want to test each plate twice with each coating and a control treatment C of no coating, so we cut six 1" squares from each of the five plates. We uniquely label one side of each square with a letter indicating the type of metal and a number from 1 to 6. The coating will go on the other side of each square.

Starting with any group of six squares, we roll a die and assign the square of that metal with the number on the top of the die to treatment A. We roll the die again until a new number is shown and assign the corresponding square to treatment A also. The next two squares selected get treatment B. The two remaining squares of that metal get treatment C. We repeat the process on each of the other four metals. For each metal, we have two replicates for each of three treatments. Within each metal, one square is as likely to be assigned to a particular treatment as any other. This process chooses a particular set of treatment assignments from all those possible as illustrated in Table 4.4.2A.

When rolling the die, we used the random symbol on top to select an object for a particular treatment. We can equally well use random symbols to select a treatment for a particular object. Suppose we eliminate control treatment C, rename the two coatings as H and T, and decide to apply each in triplicate to each metal. We cut and label six squares as before. Then we flip a coin and assign the treatment corresponding the upper side to square 1 in the first metal group. We repeat for square 2, square 3, and so on, until we have assigned one of the two treatments to three of the squares. The remaining squares get the

Table 4.4.2A Randomization of 1" squares from five metal plates to treatment A, B, or C by rolling a die

	Metal Plate				
Square	1	2	3	4	5
1	B	B	C	C	C
2	A	A	A	B	B
3	C	A	B	B	A
4	A	B	C	A	A
5	C	C	B	C	B
6	B	C	A	A	C

other treatment. We apply this alternative randomization method to each of the other metals as well to obtain an assignment table such as 4.4.2B.

Table 4.4.2B Randomization of treatments H and T to 1" squares from five metal plates by flipping a coin

	Metal Plate				
Square	1	2	3	4	5
1	H	H	H	H	H
2	T	H	H	T	T
3	H	H	T	T	H
4	T	T	H	T	T
5	T	T	T	H	H
6	H	T	T	H	T

Randomization means using a process with known probabilities, nearly always meant to be even, rather than haphazardly and subjectively picking units or treatments with unknown probabilities and serial correlations. Flipping a coin and tossing a die give sufficiently even probabilities when done properly. Two dice of different colors can generate a random number from 1 to 36: multiply the value of one die by 6, subtract 6, and add the value of the other. Do not simply add the two numbers to get a sum from 2 to 12, since 7 is six times as probable as 2 or 12. To do the same with coins, assign heads the value 0 and tails the values 1, 2, 4, 8, and 16 for a penny, nickel, dime, quarter, and half dollar respectively. Simultaneously toss one coin of each denomination and sum their values to generate a random number from 0 to 31.

An ancient method of randomization, drawing lots, is to pick unseen and well-mixed items labeled with numbers or other identifiers out of a container or pile. A pile such as a deck of playing cards should be shuffled at least seven times to be well mixed. Items in containers also need much more mixing than people commonly think is necessary.

Another method is a random number table. A sample table and a description of its use are given below. Pseudorandom number functions are deterministic functions whose successive values are chaotic and evenly distributed in some range, giving the appearance of a random series. They are widely used for computerized randomization.

For small numbers (up to 10), mixing and selecting numbered objects may be easiest. For intermediate numbers (up to 100), a random number table can be used, as described in the last paragraph of this section. For very large numbers, computerized methods are useful. Whatever procedure used should be briefly described in any research report. Just saying that the treatments were randomized is not sufficient, since that has too often been used as a synonym for haphazard assignment.

Randomization can be restricted to avoid sets of assignments that have undesirable features and to increase statistical power. A set of assignments is chosen with equal probability from among those that are allowed. Chapter 3 discussed some common restrictions, such as equal number of replicates for each treatment and treatment order. The metal-coating example above exemplifies independent randomization within groups of homogeneous objects.

One should clearly state the rationale for any other type of restriction. For example, Peacock (1983) examined whether deacidification agents successfully used in paper conservation can also reduce the rate of deterioration of a cellulose fiber textile (flax linen) during accelerated aging tests. The experiment included three deacidification agents and two application methods. She applied each of the six combinations to ten samples. The assignment was done so that, "Within each group of ten specimens no two samples had warp or weft threads in common. Therefore, samples were structurally independent of one another."

Table 4.4.2C contains 4200 effectively random digits produced by a computer program. The following five steps use this table to match objects and treatments. Letters indicate mutually exclusive versions of the corresponding steps. In other words, do 1a or 1b and one of 5a, 5b, and 5c. The example experiment used for steps 1 and 5 has 3 treatments with 6 replicates each for a total of 18 objects.

1a. To assign treatments to objects, number the K treatments from 0 to K-1. For the example, the three treatments are numbered 0, 1, and 2. Arrange or list the N objects in any arbitrary order. Use this method if objects become available and are treated one at a time. M, the maximum random number needed, is K-1.

1b. To assign objects to treatments, number the N objects from 0 to N-1. For the example, the 18 objects are numbered 0 to 17. Arrange or list the K treatments in any arbitrary order. M is N-1.

2. Pick a starting digit in the random number table by closing your eyes and hitting the page with a pencil. Circle the nearest digit as the starting point.

3. Pick a direction by closing your eyes, hitting the page with a pencil, and selecting the nearest digit in the range 0 to 7. Let 0, 2, 4, and 6 stand for up, right, down, and left

and 1, 3, 5, and 7 for the diagonal directions in between.

4. Lay a ruler against the starting digit to aid reading successive digits in the direction chosen. Upon reaching the edge of the table, continue at the opposite edge without moving the ruler. When reaching the starting digit, slide the ruler so that the next line marked is to the right of the previous line, as viewed in the direction of reading.

5a. If $M < 9$, use successive random digits to make assignments. If $M \geq 5$, ignore the digits from M+1 to 9. If $M \leq 4$, optionally use some of the digits in the range from M+1 to 9 by using the procedure illustrated in the following example and explained in 5b. Example with K=3, N=18, and M=2: if the next random digit is 0, 3, or 6, assign treatment 0 to the next object; if the next digit is 1, 4, or 7, assign treatment 2; if the next digit is 2, 5, or 8, assign treatment 3. Ignore digit 9, since its use would unbalance the assignment probabilities. Once a treatment has been assigned to six objects (its quota for this example), ignore its corresponding digits.

5b. If $10 \leq M \leq 99$, use a pair of digits to make a two-digit number for each assignment. In this context, 00 to 09 are two-digit numbers even though they have single digit values. If $M \geq 50$, ignore the numbers from M+1 to 99. If $M \leq 49$, then optionally make more efficient use of the random numbers. Find the largest multiple of M+1 that is less than or equal to 100, subtract 1, and label the result M'. Ignore random numbers in the range from M'+1 to 99. Divide other random numbers by M+1 and use the remainder as a random number from 0 to M for the assignment. Example with K=3, N=18, M=17: M'=(18 x 5) - 1 = 90 - 1 = 89, so ignore numbers from 90 to 99. If the first random number is 0, 18, 36, 54, or 72, assign object 0 to the first treatment. If the first number is 17, 35, 53, 71, or 89, assign object 17. Follow the same pattern for other numbers and objects. Once you assign an object to a treatment, ignore its number when it reoccurs. Assign the first six objects selected to the first treatment and the second six objects selected to the second treatment. Unless treatment order is also being randomized, assign the remaining six objects to the third treatment without further use of the table.

5c. If $100 \leq M \leq 999$, use the table in a similar fashion to provide random three-digit numbers from 000 to 999. It probably would be easier to use a computer.

As the randomization progresses, more and more digits and numbers become useless. At some point, it may be worthwhile to jump back to step 1 and renumber the items and then continue with step 5.

If the treatments are to be applied to one object at a time and there is any plausible possibility of drift or improvement in the treatment procedure and all objects are available at the beginning of the study, then formally randomize the order in which the objects are treated. If only one object is treated at a time, use 1a and randomly shuffle the objects to get their treatment order. If treatments are done in parallel, use 1b and keep track of the order in which objects are assigned to each treatment.

Table 4.4.2C 4200 Random Digits

```
7161266253748183949789442169759372263372403532247333081761137099601023
0933083184729825154082818486501358836664304696836916252223062498069535
7313415760719312473013084445770828930249893709544908447847785717033183
7747516689039089208012007945209245017622093608909774253679742979210861
4818284724061326313816935971964947086641382019374091361166978889257987
1339121230752917072954818704588585906664285634643802030961688642439942
5137978607713667584916032844560835752831204121334593231218021897841903
0732088392480466135057899057635054714370946916209729413502916403433965
4333044439608267673779506659637336975471908301334353043666656247763101
5916598250124487245713721656558554545251969713420557490621481131358476
7193494168105382456224491448605717181154650746806047125275805524972735
4253302268349872943418481446941826472561680353697791479129940574742853
8231726609682320311218881689545365502406121079544812297261892010080649
6147917970422304081694356348414721350405960231995210747961621643497427
4467277884704387303903764818986606889956662256669868616589582680841258
9131651713153207028888065392440733212019041344078821547733754937437470
4768484927860044287636760988781826340511916432918482216021963295979734
4646952921917771667687842695452012247941043827335870118349746500465381
4381299297539224627826142719776988558688361805822015663921113691662587
6729737351608198569471655696983617065279709726464515566071460594124369
0852363929724282123850242920439636764748585124188445422768871292937752
7047650976554944106550328568735130181059653805625012458622152891429988
4674343329625204850861121697157954324476815145076526730928245911513776
3830411279319535064557740135693068587900520881262313729131818706307169
7343842743290596558562352724757498935929376476526400620008226406033037
9982193680439224385908888749092190753643538429190486325931254009725939
0166715728342553252525727723643385687979995463873407626679089744986045
9246512988558923355735149016546668167510475288199912769617020959616443
5198922068838526842079142523088462010844827292620108293414259117256365
7600833261825592022772222534425095477800194946751135551058581904484743
4291978899307051191417188020301038951123766805286906663672208322340014
5115066980662587736057162403957092790993421636152733762113030531423381
4977019222463162903634835416407001791236746636005797433834971695508935
7244740999142575476420252487478704164741326588197722263854460528053750
0141698002536297811775150877543604922981087028820915183396274706462489
1439565349348430177856569135479019605263272434258436654916470869429229
3063382158033294979999594643921051501826621355027514906362014658159834
0195348226170847493332634607429399088737520951582609757286654181937128
2056602375844295314337241817220704251790555289638483269393349458874324
3871740929082820115459803129231533914605436370342590315587839359008043
4963697736051141004901358199963882618397569384368203664523403391555005
3227011663625487623643981731697726947899283980001419912539460438021525
9239319704049382458962641890521742942651840123918594161214816434554384
5743707208505772376084010980982679666579268179063651880990680161863092
8848481748002485736426181614268460558004935983438899967420928522166232
4546632285205940127411864446232857049254098194454696704445444142137656
5966588021304148483152220078949803108788493814339083360002828536202436
3600196862077561815616471630725007609473403292202468719426756167337245
3869616393816764420067938031104984745263643255859585185720729721087607
1601112308434431492086652563535559868519165477021910814662186125981313
9368939686175502509083611298098077432084598581171571281552763202442601
2293218328915938947745673545698948980745301588600345436225823914662575
2750598260652959695886519108934978208293647406936277687168119896531746
2459524825303621117541094137754000638833657133556841449323963105197041
4849956660516960132304553547600694015735415962661708276444130027857709
7529792015355512925946254307227643455403942616391933202258599491683832
8009803925535849231043820574000670368979707989063345331059212523365145
7718958428457116677813577292678373151881247640394897195632329639677692
9495642741356912644642983094234074710450702126305542319429365260575327
3414765879839854820145897291014179977720208750864409484544399390963041
```

CHAPTER 5

STATISTICAL ANALYSIS

While this book covers experimental design and not statistical analysis as such, the two fields are tightly interrelated. Analysis of experimental data should both answer the target questions and be appropriate for the design. This chapter discusses some statistical points particularly related to other material in this book. Section 5.1 should be read by all. Sections 5.2 and 5.3 give background that is used in chapters 6 and 7. Some points in these sections will be of greatest interest to those researchers who already have some familiarity with statistical methods. Those without such familiarity may want to refer to the STATISTICAL GLOSSARY & INDEX, *Statistical Analysis*, and introductory texts on statistics. Actually analyzing the data from an experiment may be facilitated by consultation or collaboration with a professional statistician (5.4).

5.1 DATASET STRUCTURE

5.1.1 Missing Values

We said in chapter 4 that the structure of a dataset determines the type of analyses that can be done. In particular, we recommended the avoidance of missing values. The reason is that most statistical techniques require a complete set of data. Either the object or the variable with a missing value will have to be left out of any particular calculation.

It is sometimes possible to fill a hole with a guess as to the most likely value by using other data. The easiest guess is the mean or median of the variable that has a missing value. A more sophisticated method is to use the values of other variables that are present for a particular object and the correlations between variables in the other objects to calculate a likely value for the one that is missing. This latter process of imputing values is probably most used in survey work, such as that done by the United States Census Bureau. However, either process has obvious limitations for the analysis of experimental data. Values filled in by the experimenter cannot substitute for real measurements except as a matter of convenience. With too many fill-ins the statistical results become dubious.

5.1.2 Repetitions

A correct statistical analysis that adequately addresses the questions of interest requires clear distinctions among replicates, repeated measures, and repeated readings. Lack of correct differentiation for statistical analysis often leads to claims that effects are significant when they are not. Occasionally one misses real effects instead of seeing spurious effects.

The classification of multiple measurements of a single variable sometimes depends upon the population of interest and the desired scope of the results. For example, measurements of many samples taken from one batch of adhesive are repeated readings, not replicates, with respect to the population of batches. Generalizing results to that population requires samples from multiple batches or an assumption that the one batch is representative. However, a conservator interested only in a batch on hand as material to use or discard would see multiple samples from that batch as replicates. The general procedure is to identify a population or class of interest, sample multiple units of that population, and then generalize the results to that population.

5.1.3 Data Values and Analytical Methods

Categories and numbers are the two main types of values for variables. Section 4.3 discussed a few more. With care, ordered scales can sometimes be treated as if they were arithmetic.

Most statistical analyses separate the variables considered into two categories: stimulus and response, predictor and outcome, or, more abstractly, independent and dependent. The first group are considered to affect the values of the second. Table 5.1.3 gives example analyses that are appropriate for different combinations of the types of variables in the two groups.

Table 5.1.3 Statistical analyses for predictor-outcome combinations

	Outcome	
Predictor	Categorical	Arithmetic
Categorical	contingency table	variance
Arithmetic	cluster	correlation
	discriminant	regression

Whether a variable is a predictor or an outcome depends upon the context. In a study of infrared, ultraviolet, or X-ray analysis of paintings, ground pigment, pigment binder, paint thickness, and surface coating might all be predictor variables. In a psychosocial study of artists' techniques, they might all be outcome measures.

One can replace a categorical variable by a set of binary variables representing the presence or absence of each possible category. For each object, one indicator must be 1 while all the others are 0. Since binary variables are arithmetic as well as categorical, this transformation extends the scope of arithmetic variable techniques. In particular, analysis of variance usually uses least-squares regression to estimate treatment effects by transforming categorical variables to sets of binary indicator variables. On the other hand, regression uses variance analysis to assess the significance of the parameters estimated. Analysis of covariance and some repeated measures analysis of variance use mixtures of categorical and numerical predictor variables.

Numerous statistical texts give full descriptions of the statistical methods listed above. The variables in contingency-table analysis and correlation analysis do not necessarily have to fall into separate predictor and response categories. Cluster analysis finds or creates groups or classes. Discriminant analysis characterizes known groups and assigns unknown objects to the most likely group. Variance analysis, construed broadly, includes parametric analysis of variance and the randomization and rank tests discussed in chapter 3 and section 5.3. Correlation analysis includes principal component, factor, and correspondence analysis. Regression is discussed in section 5.2. Specialized techniques for spatial and cyclic variables are the subject of entire books (Ripley 1981; Batschelet 1981) but are currently of little importance in conservation research.

Statistical Analysis provides conservation examples of all the methods in the table above except cluster and discriminant analysis. Its Chapter 5 and Appendix 9 give a correlation and contingency table analysis of the survey of conservation literature. Analysis of covariance, analysis of variance, correlation, regression, and repeated measures are listed in the index. Actual analyses are in the appendixes. The regression examples (appendixes 6 and 8) are done with a repeated-measures analysis of variance program because the numerical predictor is the time of repeated measurements on each experimental unit. Reedy (1991) gives discriminant analyses from two provenance studies done on bronze statues.

5.2 ESTIMATION OF PARAMETERS

5.2.1 Data Models

Many statistical analyses begin with a model of the data. First consider a single measurement of one arithmetic variable on one object. We regard the observed value to be the true value plus a measurement error. However, we must use the observed value as an estimate of the true value and can say nothing about the measurement error. We expect that it could have been different and will be different if we repeat the measurement. The measurement error has a distribution of possible outcomes, rather being a fixed number.

To improve the estimate of the true value and estimate the measurement error distribution, we measure the object twice or more. The most common estimate of the true value is the average of the repeated readings. For some error distributions, a median or some other statistic is more likely to be close to the true value. Subtracting the summary estimate from each observed value gives a set of estimated deviations. Subtracting the mean gives deviations with a mean of 0. Subtracting the median gives deviations with a median of 0. We summarize the deviations in turn by a measure of spread such as their standard deviation or interquartile range.

Measurements of replicate objects in a group have the following model:

observed value = group mean + object deviation

If the objects sample a larger population, then the population mean can replace the group mean in the model. Object deviations, which have their own probability distribution, usually include measurement errors.

The idea of measurement repetition is sometimes a bit slippery. Measuring the reflectance of a silver plate at a particular wavelength requires positioning the plate in the light beam, setting the wavelength, and reading a reflectance. The reflectance might be read to the nearest .01 from a dial marked 0 to 1 in increments of .05. Repositioning the plate, resetting the wavelength, and waiting an hour or a day give larger deviations than merely rereading the dial after a 10 second delay. Saltzman and Keay (1965) isolated and estimated several components of color measurement error. Their data model includes effects of dye lot, milling day, sample, swatch, spot, measurement day, and immediate repetition. They estimated the data variance due to each of these possible sources of error.

Data models have the following abstract form:

observation = deterministic part + random part

The deterministic and random parts are sometimes called the fit and residual, or explained and unexplained. These formulations are similar but not exactly equivalent, since the fitted part of the model may include random components with known explanations. For experimental data, the most important components are treatments. The most important analyses estimate the magnitude of their effects and test their significance.

5.2.2 Curve Fitting

A mathematical function specified by a few parameters often approximates the empirical relationship between two continuous variables. The model is that one variable equals some member of a family of functions of the other variable plus a random component. The most common example in conservation is a declining exponential relationship between desired quantities, such as substrate strength or image color intensity, and decay factors, such as time or cumulative exposure to light, heat, or moisture.

Fitting a curve reduces several measurements to a few standard values, such as initial level, decay rate, and final level. It also isolates standard aspects of the relationship for further analysis. In a fading study, the initial difference of color intensity of replicate specimens is a nuisance factor that does not depend on the subsequent treatment. Exponential fading rates that do depend on the treatment are the primary target of experimental manipulation and statistical analysis.

In the exceptional case of linear relationships, simple linear regression estimates the parameters (Draper and Smith 1981). In the general case of nonlinear relationships, there are two approaches to estimating parameters (Bates and Watts 1988).

The old approach transforms the data points to make them lie more or less along a line. A straight line is drawn by eye on graph paper or fitted with a linear regression program. There is a different mathematical trick for straightening each type of curve. If there are more than two parameters, all but two must be estimated by eye or some other ad hoc method before doing the transformation. Analysts developed these methods before the widespread availability of computers.

Figure 3.2.4B shows an example fit of a negative exponential declining to a non-zero baseline. The data model with three parameters is as follows:

value = stable part + transient part * exp(-rate*time) + residual

where exp() is the exponential function and residual is anything left over. To linearize this model, estimate the stable part (the non-zero baseline) by eye, subtract it from the original values, and take logarithms to get the following:

new value = log(value-base) = log(transient) + rate * time + new residual

Different persons, and the same person at different times, will make different estimates of the baseline and thereby calculate different estimates of the other two parameters. The result will often be worse than the fit shown in the figure.

The modern approach is to fit the curve to the data on a computer with direct nonlinear regression, with appropriate weights for the data points. This nearly always gives more accurate estimates of the true parameter value. This approach uses the repetitive calculation skill of computers, instead of merely automating clever pencil-and-paper methods designed to avoid as much calculation as possible. Any decent statistical package now includes some method for true nonlinear regression. Fast, flexible, and easy-to-use nonlinear curve fitting is now available even for microcomputers. This removes constraints on experimental design that existed even in the early 1980s.

Logarithms of the dependent variable linearize simple negative exponential curves that decline to 0. If measurements are always positive and measurement errors are large for large measurements and proportionally small for small measurements, then this transformation is acceptable. If measurements can have negative values, if errors remain large for small measurements, if there is a non-zero minimum level, if the decay pattern is other than a simple exponential, or if alternative models are to be tested for a better fit, then one should fit a negative exponential without linearization, as in figure 3.2.4B.

5.2.3 More Models

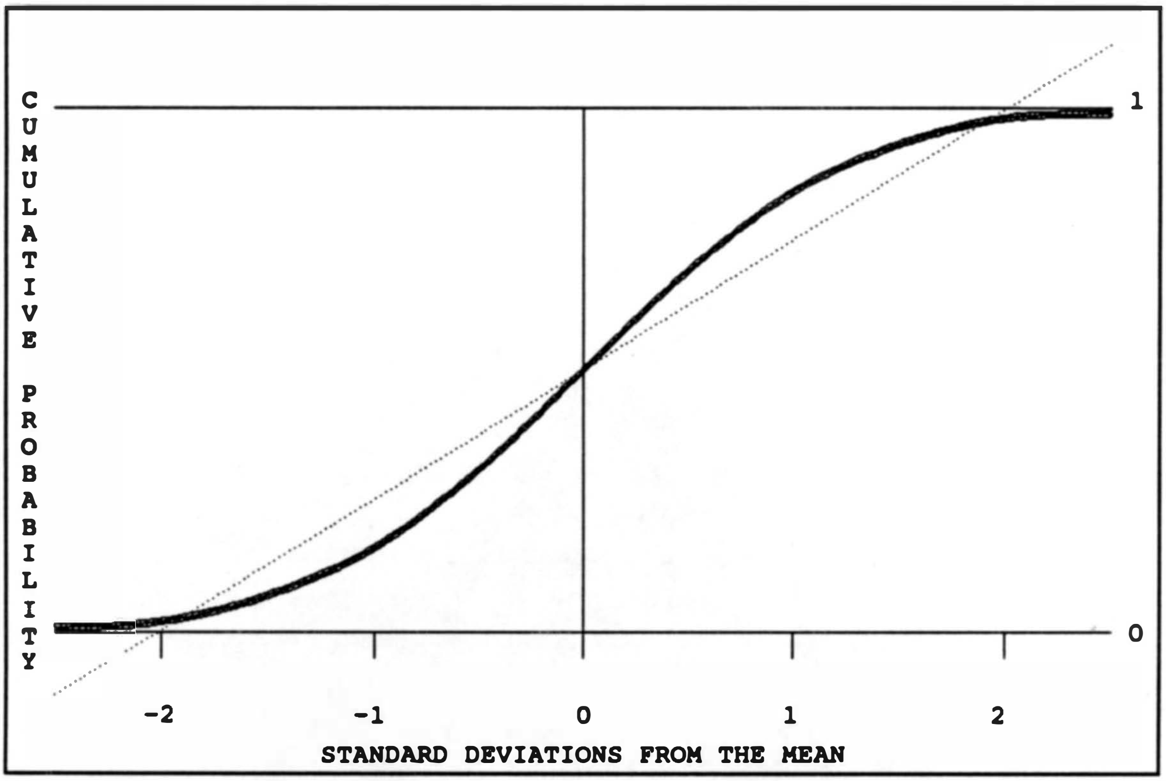
Imagine figure 3.2.4B with time replaced by treatment dose and the curve flipped so that it rises instead of falls. It would then show object quality improving with dose, but with diminishing effect, up to a saturation level. This type of dose-response relationship is fairly common in chemistry and biology.

If dose is replaced by log(dose), the relationship may become sigmoid (S-shaped), similar to the one shown in figure 5.2.3. (For the moment, ignore the fact that the curve happens to represent the cumulative normal distribution.) One can think of this as measuring dose on a different scale, much as pH is a logarithmic hydrogen ion concentration (activity) scale. A common model used for sigmoid dose-response relationships is the logistic:

response = baseline + range / (1 + exp(logdose - halfdose))

where halfdose is the dose of half (50%) response. In figure 5.2.3, baseline = 0, range = 1, and halfdose = 0. Ratkowsky (1983, 1990) gives a hundred more models.

Figure 5.2.3 Normal cumulative distribution curve



Polynomials are often used as empirical models. Assume that response to a particular dose is measured to within 5% of the full range and that figure 5.2.3 is the true but unknown model. If doses are restricted to the range -.7 to +.7, the response curve is linear. If the range of doses is expanded to -2.5 to +1, the faint dotted line might represent a linear fit. The systematic variation of residuals from positive to negative to positive suggests that a quadratic component also be fit. If the doses vary from -2.5 to +2.5 and generate the entire range of responses, the linear fit might still be the dotted line. Now, however, the residuals switch sign once more and suggest the fitting of a cubic component instead of a quadratic component.

This thought experiment shows two principles: (1) Empirical models depend upon the doses applied, and (2) extrapolation of response curve beyond the observed range may lead to great error.

5.3 INFERENCE AND HYPOTHESIS TESTING

5.3.1 One-Sided and Two-Sided Tests

Comparison of two treatments may use either signed or unsigned differences for the test statistic and corresponding reference distribution (3.3.3). A one-sided test uses signed differences and is only significant if the treatment is sufficiently better than the control. However, if a treatment has a bad effect and is worse than doing nothing, we usually want to know. This requires designs that have enough measurements and are therefore powerful enough for two-sided rejection of the null hypothesis when using absolute differences (3.3.4).

5.3.2 Randomization, Ranks, and Distributions

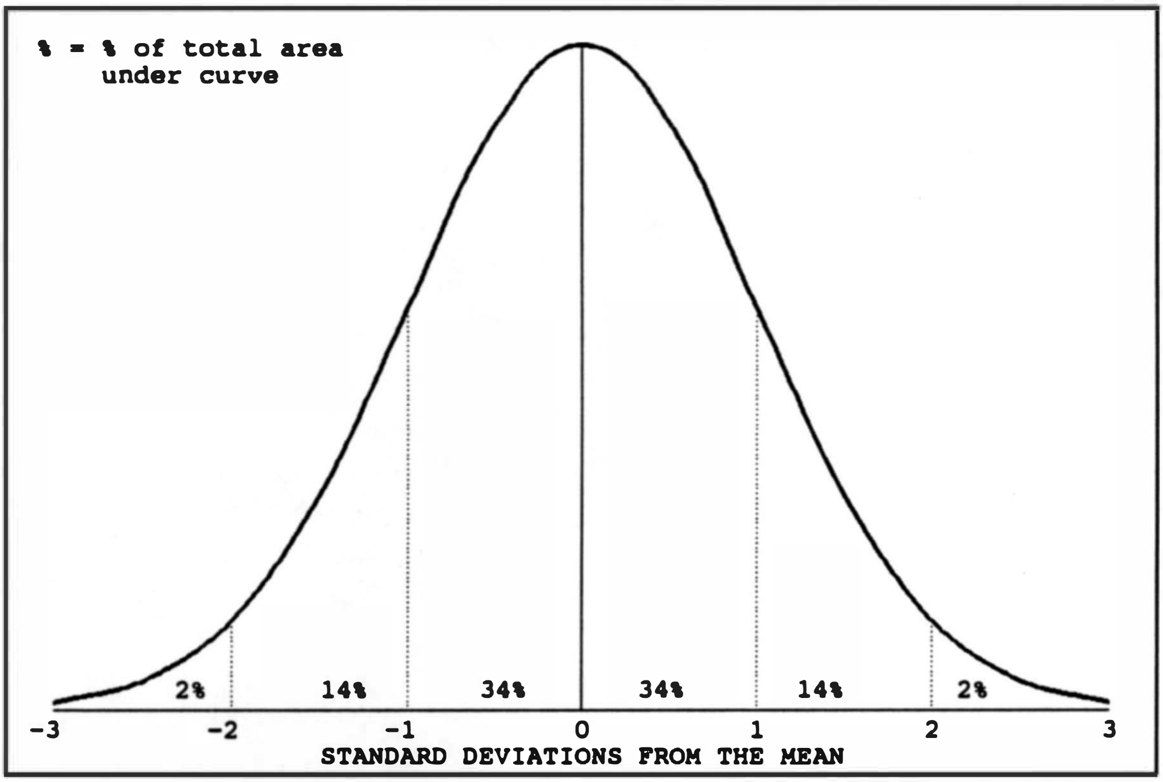
Randomization tests first appeared in the 1930s (Fisher 1935 and Pitman 1937). Only Fisher's Exact test for 2-by-2 contingency tables gained wide acceptance. For practical computational reasons, the rest disappeared. Tests based on normal distribution theory and ranks, discussed in the next two subsections, still dominate statistics books and programs. However, the wide availability of low-cost computing has recently increased interest in tests that use all the information available, avoid assumptions known to be approximate, and are clearly intelligible (Howard 1981, Edgington 1987). Random assignment of treatments is as effective as random sampling of populations as a basis for statistical inference, especially with proper analysis.

Each family of tests uses slightly different test statistics. Each calculates different reference distributions of the results expected from random variation when the null hypothesis is true. Rank tests approximate randomization tests by replacing measured values by their ranks when sorted. These ranks of measured values are different from the rank of the test statistic in the reference distribution. Normal-theory tests approximate randomization tests by replacing the empirical reference distribution by a theoretical distribution based on calculated means and standard deviations and the assumption that the object deviations in a data model are independent selections from a normal distribution. When the assumption is sufficiently accurate or when there are enough experimental units, the match between the two distributions may be quite good.

Without a computer, normal theory and rank tests are easier than randomization tests. For a particular design, the reference distribution for the test statistic only depends on the number of objects and not on the observed data. Ranks and normal theory test statistics (such as chi square, F, and t) are designed to be independent of the scale used for the outcome measure. Once calculated for each sample size, the critical values can be printed in tables. Some tables list the p value for given values of the statistic calculated from the data. Others list the values of the statistic needed to attain given p values, such as .1, .05, .01, and .001. The rows or columns of these tables are sorted by either sample size or degrees of freedom.

Degrees-of-freedom is a technical concept that statistics has borrowed from mechanics. In general, it is the sample size minus the number of parameters estimated, not counting the variance of the errors. An exact understanding is not important here.

Figure 5.3.3 Normal density curve



5.3.3 Tests Based on the Normal Distribution

The first set of alternatives to randomization tests is the family of normal distribution theory procedures. Tests in this family assume that the observed data values are independent random samples from a normal distribution (figure 5.3.3). This bell-shaped probability distribution is sometimes called the Gaussian or error distribution. The integral of this probability density function is the cumulative distribution function shown in figure 5.2.3.

Equivalently, these tests assume that deviations from the true, unknown and unknowable mean value of the distribution sampled are independent and normally distributed with a mean of 0. This assumption makes possible the calculation and tabulation of a standard distribution for a particular experimental design and sample size. The t, F, chi-square, and normal distribution itself are some of the standard distributions in common use.

To do a t test in design 1 (section 3.3), A calculates the mean and sample standard deviation of the 19 control differences, subtracts the mean from the treatment difference, and divides the result by the standard deviation and then by the square root of 19. A then compares the resulting t statistic to a t distribution with 18 degrees of freedom either by direct calculation of a p value or by reference to a table. (The 18 degrees of freedom are the 20 data values minus 2 for the treatment and mean control differences.)

In design 1, the successive measurements might have a drift component, but regression can estimate and eliminate it. Even if the original values are stable and independent, derived successive differences tend to be negatively correlated. A high measurement usually causes a positive difference followed by a negative difference, and vice versa for low measurements. The degrees of freedom should be reduced according to the amount of serial correlation, but we do not know of any formula applicable to this particular design.

The assumption of independent normal errors adds information to the observed data. When this information is correct, it adds power to the test. The power of a test is the likelihood that a true difference will be detected and declared significant. If the information is incorrect, it may make the test both invalid, by giving an incorrect p value, and less powerful.

The normal distribution assumption can be examined by looking at the data, histograms, or normal probability plots. There are also formal statistical tests. Transformation of skewed data by square roots or logarithms often makes the distribution of values for each group more symmetric. Non-normal symmetric distributions are more troublesome. Proper experimental design minimizes correlation, but it is inherent in repeated measures.

Statisticians have modified various normal-theory tests in several ways to make them less vulnerable to violations of the error assumption. Unfortunately, the best modification depends on the true error distribution, which is hard to estimate from one experiment, so none is particularly dominant or widespread. It is fortunate that normal theory methods usually give roughly correct answers when the residual assumption is only roughly true. It is somewhat ironic, however, that a dependable test of the normality assumption requires enough data to estimate the true error distribution for comparison to the assumed distribution. The assumption then adds very little in the way of statistical power.

5.3.4 Tests based on ranks

The second set of alternatives to randomization tests is the family of rank-based procedures. Other names for this category of tests are distribution-free or non-parametric. These tests sort the n measured data values and replace them by their ranks from 1 to n. Tied data values are replaced by their mean rank. Reference distributions are calculated by the same permutation procedure used for treatment randomization tests. In other words, the commonly used rank tests are randomization tests with the measured data replaced by ranks.

The advantage of replacement is that all experiments with the same number of samples ideally have the same set of substitute measurements, 1 to n. Calculation of the reference set is done once and the result published in tables for general use. Such tables are, however, limited to small sample sizes, and end with a formula for a normal distribution approximation for sample sizes larger than those tabulated. Rank data with mean ranks due to ties in the original data have a different reference distribution. To do a rank test anyway, the problem of ties is ignored and the standard table is used. If there are too many ties the test cannot be done. The disadvantage of rank replacement is that it reduces the precision of measurement and hence the power of the experiment to detect real differences. When randomization tests are widely available on microcomputers, much of the advantage of tables in books and the need for normal approximations will disappear.

The important point for experimental design is that we can analyze ranks. Ranks can be original data based on human visual assessment rather than replacement data for quantitative instrument measures. For instance, Wharton, Lansing, and Ginell (1988) ranked about 80 1" x 3" silver sample plates, after artificial tarnishing and polishing by various compounds, according to visible residual tarnish and scratching.

5.3.5 Points, Intervals, and Significance

Thus far, we have discussed estimation of treatment effects and other parameters by a single number. This is called point estimation. We have also discussed attachment of a p value to a point estimate of treatment effect by comparison with a reference distribution of possible outcomes expected under the null distribution.

We have intentionally downplayed the labeling of p values as significant or not according to arbitrary cutoffs such as .05 or .01 and have avoided the term "significance test" in favor of "hypothesis test." The lower the p value, the stronger the evidence against the null hypothesis and for some alternative. In writing this book and *Statistical Analysis*, we are NOT recommending that conservation adopt the ritual search for arbitrarily significant p values that has infected many fields of science despite the disapproval of many statisticians (Salsburg 1985).

An alternative or complement to a point estimate and a p value is an interval estimate. An interval estimate of an aspect of the world is a pair of numbers that hopefully brackets the true value. When the hope is quantified by a confidence factor between 0 and 100%, the interval is called an X% confidence interval. Higher confidence requires a wider interval. The width of the interval for a given confidence level is a measure of the uncertainly of the point estimate.

For instance, when a statistic can be said to have a normal distribution, the interval from the mean minus the standard deviation to the mean plus the standard deviation is a 68% confidence interval (see figure 5.3.3). The mean and standard deviation here refer to the distribution of the calculated statistic, not of measured data. The mean is usually the same as the point estimate. If the standard deviation is multiplied by 1.96, then the wider interval has a confidence of 95%. In this simple case, the point estimate has a p value of X when 0 is one endpoint of an X% confidence interval.

The distribution of a statistic can be based on either hypothetical repetition of an experiment or subjective assessment. It depends on the number, quality, and distribution of measurements. The confidence factor X can be interpreted as an estimated probability that the interval (variable from experiment to experiment) encloses the fixed true value. The process is something like tossing an elastic horseshoe at a fixed target. A larger horseshoe makes a ringer more likely.

Because of the abuse of p values, some statisticians go so far as to suggest that they be abolished and replaced by confidence intervals (Hunter and Schmidt 1990). Of course, both can be calculated, especially when computers are available. Because confidence intervals belong to analysis rather than design, further discussion is beyond the scope of this book.

5.4 WORKING WITH STATISTICIANS

Experimental designs may be complex and often depend upon a thorough knowledge of the appropriate statistical techniques. In many other fields, such as biomedicine, researchers routinely collaborate with a statistician from the outset of a project. This is especially useful when planning a major project, and the cost can be included in grant proposals. Hiring a specialist to determine the fewest number of samples and measurements required to adequately examine the effects of interest may ultimately save time and money and ensure correct statistical analysis.

Statistics is a large field. Statisticians specialize and have different areas of knowledge and expertise, just as do doctors and conservators. They also have different views about how to apply statistical knowledge to specific problems. More easily than patients and objects, a set of data can be manipulated and analyzed several times. It is possible to get a second statistical analysis, just as one can get a second expert opinion.

Statistics is an actively expanding field. There are general statistics journals, such as the *Journal of the American Statistical Association* (over 1000 pages a year) and *Applied Statistics*. There are also specialized journals such as *Biometrics*, *Technometrics*, *Chemometrics*, and *Envirometrics* for applications of statistics to specific fields. A current trend is the development of techniques that make use of the computing power sitting idle on millions of desks (Efron and Tibshirani 1991). Another is the formalization of methods of exploratory data analysis (Hoaglin, Mosteller, and Tukey 1983 and 1985).

The following are some specific suggestions for working with statisticians.

- Initiate the relationship before beginning the experiment instead of after finishing.
- Bring a written outline of the experiment similar to the work sheet in 6.5.
- Explain the background and details of the experiment when asked. They may be necessary to do a proper analysis and solve the problem.
- Remember that conservation can be as strange to a statistician as statistics can be to a conservator.
- Allow more than five minutes for a consultation.

CHAPTER 6

MULTIPLE-OBJECT (GROUP) STUDIES

This chapter presents several basic designs for group-comparison studies. They are illustrated in figure 6. The mention of more complicated designs indicates some possible directions for further study. As stated in the preface, knowing the names of standard statistical procedures for most of the designs should help one talk with a statistician or find the appropriate section of a statistics book or program manual.

A small sampling of the many books that cover topics in this chapter includes Bethea, Duran, and Boullion (1975); Campbell and Stanley (1963); Cox (1958); Keppel (1982); Milliken and Johnson (1984); Milliken and Johnson (1989); and Rasch and Herrendörfer (1986). Any good library or technical bookstore should have several more.

6.1 ONE GROUP

6.1.1 Comparison to a Standard

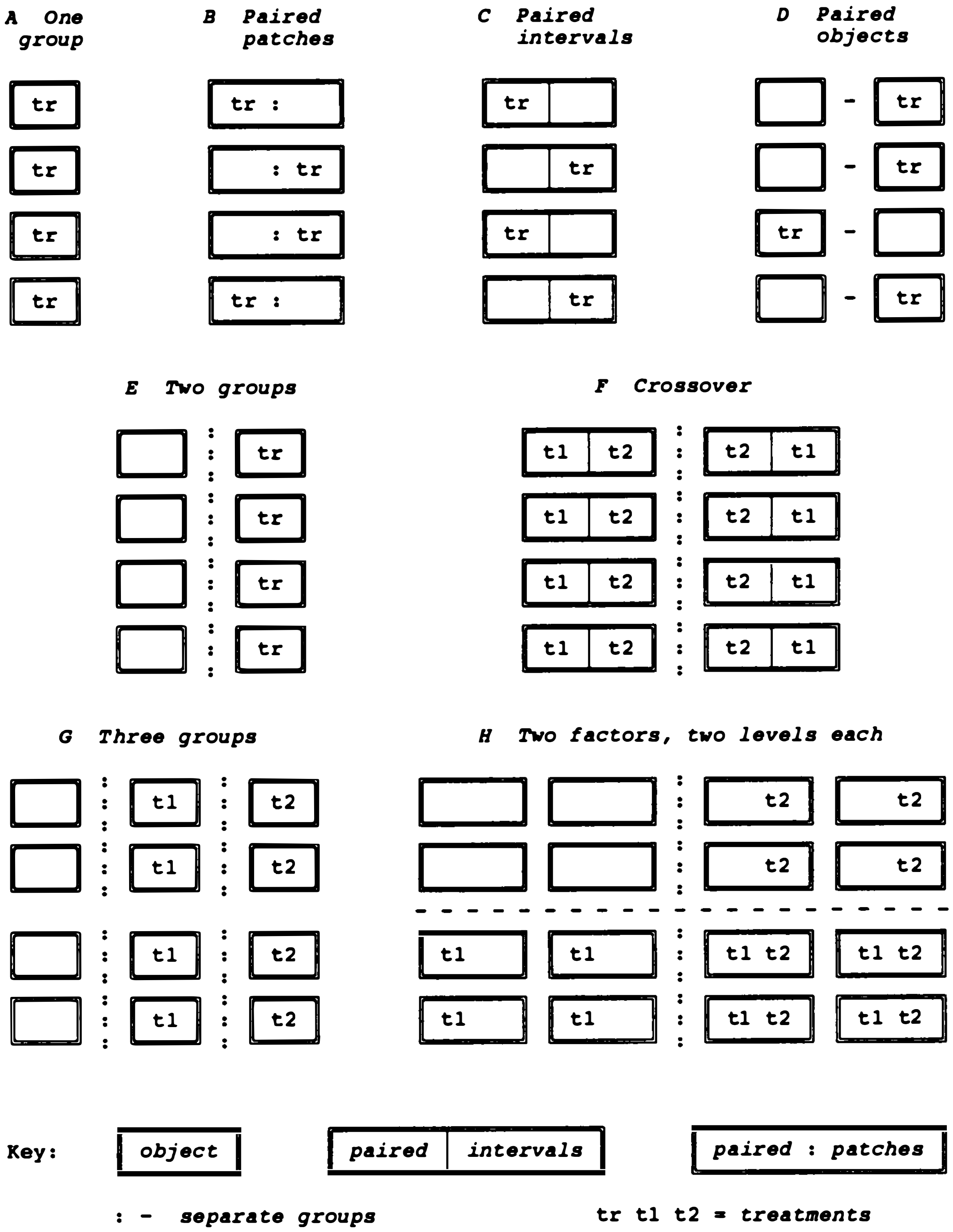
For the simplest multiple-object design, we study one group of objects and measure each once (figure 6A). The measurement follows some treatment, possibly nothing, that is applied to each object. In this chapter, we assume that measured outcomes have arithmetic values. We also assume that averages and standard deviations adequately summarize the central tendency and variability of a batch of measurements.

This design characterizes the group of objects at a particular time and in a particular condition. This condition includes having received whatever treatment was given. We can say little about the effect of the treatment, but can compare the measurements to some standard value. This might help us decide whether the objects or some population they represent are suitable for some particular purpose.

Assume that the value for comparison is 0. If it is not, subtract it from each measurement. Either way, we take "positive" to mean "greater than the standard value" and "negative" to mean "less than the standard value." In practice, immediate subtraction is not necessary for the summary statistics. Averaging and then subtracting the standard value from the result gives the same mean as subtracting the standard from each value and then averaging. The subtraction does not affect the standard deviation. However, the individual values adjusted to the standard are sometimes useful for visual examination.

If the question is whether each object exceeds the standard, the answer is easy (leaving aside questions of measurement error, as we will do for now). It is also simple to determine whether the calculated group mean exceeds the standard value. More difficult to answer is the question of whether any particular measurement is unusual in some sense. The

Figure 6 Multiple Object Designs



real difficultly comes when one asks about what would happen if we were to measure more objects that meet the criteria for belonging to the group. Would the mean change sign? Is its current sign or difference from 0 in some sense just an accident due to the particular choice of objects? The goal is to use the data at hand to make an inference about other objects or possible future measurements.

Since there has been no treatment randomization, no randomization test is possible. An alternative basis for statistical inference is an assumption about the process generating the data or the distribution of the resulting values. The general model is that each measured value is the sum of two components -- a fixed or deterministic part and a random part. The random part is usually the residual or error term. For this simple design, assume the fixed component to be a constant that is the same for each object in the group.

For a test based on the theory of normal distributions, assume that the random part has a normal distribution that is the same for each observed value. The mean of the distribution is 0 and the standard deviation is known, assumed, or estimated from the data itself. The last is the most usual case. An additional assumption is that the random parts of different measurements are independent of each other.

The test statistic is a t statistic, the group mean divided by its standard error. (The standard error of the mean is the standard deviation of the data divided by the square root of the number of objects.) If we want to know whether the mean is, within the resolution of the experiment, different from 0, we assume that it is 0 until demonstrated otherwise by a two-sided test. If we want to know whether it is positive (a one-sided test) we have a choice between being optimistic and assuming that it is positive unless shown to be negative or being pessimistic and assuming that it is negative unless shown to be positive.

From a sampling viewpoint, the t test assumes that the objects are a random sample from an effectively infinite population with normally distributed values for the variable measured. More exactly, random sampling from such a population is one means of guaranteeing that the data have the characteristics outlined above. We also can generate such data by constructing, treating, and measuring simulated objects. The resulting inferences and conclusions apply to the population or process that generated the data.

6.1.2 Color Match Example

To make the discussion in 6.1.1 more concrete, suppose a paintings conservation instructor assigns five new students the task of visually color matching a red spot on a dark background in an old painting. They are to paint a 1" spot on a 3" x 5" white card. The six spots are then instrumentally measured for color in random order.

Table 6.1.2 shows the results. DeltaC is the measure of a student sample (s#) minus the target measure (30). S.d. and s.e. are the standard deviation of the s# samples and the standard error of the mean.

The color of all five student samples is positive relative to the target, which is the standard for this experiment. The instructor announces that one purpose of the exercise was

Table 6.1.2 Paintings conservation student color match (hypothetical data, arbitrary units)

Object	Color	DeltaC
target	30	
s1	32	2
s2	33	3
s3	31	1
s4	36	6
s5	33	3
mean	33	3
s.d.		1.9
s.e.		.8

to show that visually apparent color is modified by the background color. S1, a skeptical student, asks whether these particular results might just be an accident. S2, who has taken one semester of statistics, looks at a t table in a statistics book and finds that a t statistic of 3.75 (3/.8) with 4 degrees of freedom has a two-sided p value of .02. The class decides that, assuming that they are representative, it is highly unlikely that naive conservation students can make unbiased visual color matches under the conditions of the exercise.

S5 wonders how their class compares to others. The instructor says that 68 undergraduate art majors averaged 34.7 in a similar exercise. S2 calculates that using 34.7 as a standard gives a t statistic of -2.13 ((33-34.7)/.8) with a p value a bit above .10. There is plausible evidence that, as graduate trainees, they were fooled less than undergraduates.

6.1.3 One Treatment

A conclusion about the effect of the treatment applied to each object before measurement requires information about what would have been the state of the objects if they had received some other treatment or none at all. We get this information from either the same group of objects (the remainder of this section) or another group of objects (section 6.2). Additional information from the same group comes from more measurements (this subsection) or another treatment (6.1.5).

The simplest source of additional information is to measure the objects before the treatment. We often can reasonably assume that the objects would have remained the same or deteriorated further without the treatment. The treatment effect for each object is then the difference between its two measurements. The differences form a new group of derived measurements that we usually analyze with a t test as described in 6.1.1 and illustrated in 6.1.2.

This procedure effectively assumes that the treatment makes about the same change for all objects, regardless of prior condition. This is reasonable, with a proper scale of measurement, for a small dose of a treatment such as cleaning or polishing.

Other treatments tend to bring the object to a particular state regardless of the initial condition. Paintings coated with a high-gloss varnish will be glossy regardless of their initial surface condition. The change in gloss will be negatively correlated with the initial gloss. The arbitrary variation in initial surface condition of the paintings in the group will become part of the empirical error distribution. This distribution will be highly dependent upon the particular set of objects used for the experiment. There is no reason that it should be a normal distribution.

The sign test is an alternative to the t test. The null hypothesis is that positive and negative changes are equally likely. A sign test makes no other assumption about the distribution of changes. The test statistic is the number of positive differences in a group of n objects. Its p value is calculated from the binomial distribution, just as for flips of unbiased coins. For instance, six positive values out of six differences has a two-sided probability under the null hypothesis of .03. Other values are widely available in tables.

Another alternative is valuable when the treatment difference is only partly dependent upon the initial value. Subtracting the before value from the after value is only one possible adjustment of the post-treatment observation. More general is subtraction of any fraction of the pre-treatment value. The best choice is the fraction whose subtraction most reduces the dispersion of the resulting adjusted values. Linear regression, available on hand calculators and explained in every beginning statistics book, gives this value. We then test the new adjusted values as described in 6.1.1.

6.1.4 Color Match, Part 2

The next day, the paintings conservation instructor asks the five students to repeat the color match exercise. S2 realizes that the class is in the second part of a measure-treat-measure study, and that the treatment was their examination and discussion of the feedback from the instrumental measurements. To avoid influencing the response of the other students, S2 keeps silent.

Table 6.1.4A Paintings conservation student color matches (hypothetical data, arbitrary units)

	Color of red spot		
Student	Before	After	Change
s1	32	28.5	-3.5
s2	33	30	-3
s3	31	30.5	-0.5
s4	36	25	-11
s5	33	31	-2
mean	33	29	-4
s.d.			4
s.e.			1.8
t			-2.2

After the second set of measurements, S2 presents the rest of the class with table 6.1.4A and announces the p value as a bit under .10. S5 complains that it somehow should be lower, to more conclusively show that they had learned something after the first set of measurements. S2 replies that S4 is an atypical outlier. With S4 eliminated, t is -3.4 (-2.3/.66) for a two-sided p value of about .03. S5 is happier. S4 angrily objects that she is as much a part of the group as any of them. S3 wonders why eliminating the largest change differences across the intervals. As discussed in chapter 3, this design is less likely to be useful with two active treatments.

In both of these paired designs, the treatment and measurement unit is a subdivision of the object. If we change our viewpoint, the pair of randomized treatments is a composite treatment that we apply to the object as a whole. Correspondingly, we can consider the calculated contrast between treatment and control, based on two, three, or four measurements, to be a composite measurement characterizing the object as a whole. This is why the one-group tests are applicable to the paired-treatment designs. Other designs resulting in one composite measurement per object, such as multiple-object extensions of those in chapter 3, can be similarly analyzed.

This viewpoint also suggests a third paired treatment design in which the composite treatment/measurement unit is a pair of objects matched by their initial similarity (figure 6D). The objects serve the same role in this design as the patches did in the first, and the analysis is otherwise the same.

6.1.6 Color Match, Paired

A year later, the paintings conservation instructor has six students. The instructor displays the same painting as before and hands each student two 3" x 5" cards -- one white and one black. On successive days, the students are to paint a red spot on one card matching the spot on the painting. Each flips a coin to determine which card to use on the first day: heads for white, tails for black. All cards are measured on the third day. Table 6.1.6 outlines the data table. Readers are free to imagine the results.

Table 6.1.6 Color matching on black and white cards (data structure outline only)

	Card		black -
Order	black	white	white
R	M	M	C
	etc		

R: random order, bw or wb
M: measured value
C: calculated value

This experiment is very similar to design 6, chapter 3. One can regard the two measurements as two measures of each student on successive days (the second paired design). Alternatively, one can regard them as measures of two objects that are paired by their origin (the third paired design). It is a matter of viewpoint whether the experimental unit is the student or the card; the analysis is the same.

Recording the order in which the samples are prepared is not strictly necessary. It does, however, allow a check on whether the black-white difference depends on their preparation order. This effectively divides the students and their pair of samples into two groups. This is formally included in the design by restricting the randomization so that equal numbers of pairs are prepared in each order (chapter 3, design 7 and 6.2.3).

6.2 TWO GROUPS

Besides pairing, another way to get control information is to measure a second group of untreated objects.

6.2.1 Historical Controls

A two-group design with a non-experimental control group uses an existing set of measurements on a similar group of objects that did not get the treatment under question. If only the control group mean is available, we use it as the standard value for the one group test. However, this mean is only an estimate of the true value. When the individual measurements are available, we do a two-group test. If we know the mean, standard deviation, and number of objects in the control group, we can do a two-group t test.

The data model for this design is as follows:

value = control value + treatment effect (if treated) + random component

The control value is estimated by the control group mean. The treatment effect is estimated as the treatment group mean minus the control group mean. The random component for each object is estimated as its measured value minus the value predicted by the rest of the model (the deterministic part).

Using a historical control group requires the often dubious assumption that the control value and distribution of random deviations are the same for both groups. The difference of group means may then be a valid estimate of the treatment effect. It also requires that the measuring instruments for the current treatment group and the historical control have the same calibration. Because of the need for this assumption, designs with historical control groups are better for generating hypotheses than for confirming them.

6.2.2 Two Treatment Groups

A two-group experiment starts with one group of objects that is randomly split into two treatment groups (figure 6E). The randomization is usually restricted to put equal numbers of objects into each treatment group. One of the treatments may be a control treatment. If not, the wording of the data model in the previous paragraph must changed to:

value = mean + treatment effect + random component

This design is the same as designs 2, 4, and 5 in chapter 3, with the experimental units being whole objects instead of intervals, patches, or patch-intervals.

When doing an analysis by hand, it is common to list the data for each group in a separate column. When preparing data for computer analysis, it is usually necessary to put the measured values in one column and an indicator of group membership in a second. Table 6.2.2 outlines both types of data tables.

Table 6.2.2 Data structure for two treatments

Treatment group	
1	2
M	M
etc	

M: measured value

or

Treatment group	Measured value
T	M
etc	

T: treatment id such as 1,2

This design is similar to the paired-object design (the third paired-treatment design, 6.1.5), but without the pairing. In each design, half the objects get each treatment (usually). The difference is in the treatment randomization procedure and the subsequent analysis. Selecting 4 objects from 8 is different from selecting 1 object from each of 4 pairs.

If the variables used to match objects predict treatment outcome, then pairing objects reduces the variability of treatment differences. This increases the precision of the treatment-effect estimate. On the other hand, pairing halves the number of analysis units, tending to decrease the precision of the error-variability estimate. Pairing may increase the t statistic by decreasing the denominator. It simultaneously decreases the number of degrees of freedom, thereby requiring a larger t statistic for a given p value. There is a similar tradeoff with randomization tests.

In the paired-patch design of 6.1.5, we could unpair the patches and independently randomize each into two treatment groups. The two patches on an object would sometimes get the same treatment and the objects themselves would not figure in the analysis. This might occasionally be appropriate, especially with multiple patches per object. However, in most biological experiments and probably in most art conservation experiments, patches from one organism or object are much more similar than patches from different objects. Keeping the patches paired usually results in a net gain in precision. In addition, it is usually objects and not patches that constitutes the desired unit of inference and conclusion.

6.2.3 Crossover and Related Two-Factor Designs

The paired-interval design restricts randomization to the two intervals for each object. The objects fall into two groups according to the order of the treatments, as mentioned in 6.1.6. If one of the two treatments is a control of no treatment, this is not important. If both treatments are active treatments, there may be an important carry-over effect. The effect of the first-period treatment carries over to affect the second-period outcome. In other words, the apparent effect of a treatment may depend upon whether we give it first or second. We can make the groups defined by treatment order equal in size by further restricting the randomization as in design 7 of chapter 3 (figure 6F). Estimating the order effect and testing it for significance is discussed in many experimental design books.

Suppose we have 10 insect-infested wool textiles. As an experiment, we fumigate each by low doses of two fumigants. The outcome measure for each fumigation is the number of dead insects counted the next day. To reduce carry-over effects without over-prolonging the experiment, we wait a week between fumigations. Carry-over may still be present and is a nuisance. To estimate its magnitude and minimize its interference with the fumigant comparison, we randomly assign half the objects to each of the two possible treatment orders.

In art conservation, this particular design is likely to be less useful relative to other designs than it is in human studies. The important point for is that the treatment of each unit is defined by two factors -- what and when -- that are independently variable. Furthermore, treatments are compared within each object, while treatment orders are compared between objects. Other types of experiments using mixed within- and between-object comparisons should be useful in conservation research.

Paired patches within an object have no intrinsic order. We can divide the objects into two groups by some other property or treatment. Suppose we want to compare two oil painting varnishes. We also suspect that application temperature might influence the outcome. There may also be an interaction between varnish and temperature. There is an interaction if the difference between varnishes is different at different temperatures, or, equivalently, if the temperature effect is different for different varnishes.

We start with 12 disposable paintings and randomly select 6 for the low temperature room (at 50°F, say). The other 6 go into a high temperature room at 90°F. Extreme temperatures maximize the probability of seeing a temperature effect if there is one. They also maximize the usefulness of a negative result since it would apply to nearly all temperature conditions encountered in the United States. We select two patches on each painting and randomly apply one of the varnishes to each patch.

Table 6.2.3A Varnish and temperature I

	Outcome measure	
Temperature	Varnish A	Varnish B
50	M	M
	(+ 5 more lines)	
90	M	M
	(+ 5 more lines)	

Table 6.2.3A outlines the data table. We analyze it with the help of a statistics program designed to work with such mixed within-and-between designs, which are also called repeated-measures designs. The analysis estimates and tests the significance of the varnish difference, the temperature difference, and their interaction.

To understand interaction, suppose the means of the four groups of measurements defined by temperature and varnish are 0, 0, 0, and 1, with the 1 belonging to varnish B at 90°. Assume that the number of replicates and the size of measurement variations is such that the difference between 0 and 1 is noticeable. Then temperature has no effect on the outcome with varnish A but does affect the outcome with varnish B. Similarly, varnish A and B have the same outcome at 50° but different outcomes at 90°. The effect of one treatment factor depends on the value of the other treatment factor.

Comparisons within objects are usually more precise than comparisons between objects. One guide to choosing your design is to decide which effect is more important.

Practicality is another. To reverse the design given above, we would randomize each painting to treatment by one of the two varnishes. We would then delineate two patches on each painting and varnish one at the high temperature and the other at the low temperature, using the same varnish for each patch. This would probably be logistically more difficult, requiring painting movement or drastic room temperature modification. It also introduces some concern about the order of application. The data table would then have the form of table 6.2.3B

Table 6.2.3B Varnish and temperature II

	Outcome measure	
Varnish	50°	90°
A	M	M
	(+ 5 more lines)	
B	M	M
	(+ 5 more lines)	

6.3 MULTIPLE GROUPS

6.3.1 Multiple Treatments

The simple two-group, two-treatment design generalizes to multiple-group, multiple-treatment designs in which each group of objects gets a different treatment. The simplest multiple-group designs have only one treatment factor with several values or levels (figure 6G). All the treatments belong to one class or category. Examples are multiple varnishes, multiple polishes, multiple cleaners, or multiple pigments. There may be one or more control treatments.

Each treatment group usually gets the same number of objects. The major exception is that control groups are sometimes larger for specialized reasons. If the primary purpose of the experiment is hypothesis testing rather than estimation of treatment effects, then as the number of treatment groups increases, the number of replicates in each can decrease.

The data matrix for this design is only a slight modification of table 6.2.2. For hand analysis, there are three or more columns, one for each treatment, instead of two. For machine analysis, the treatment variable has three or more values delineating three or more blocks of outcome measures, instead of two.

The most common analysis for this design is one-way analysis of variance. Nearly every statistics book discusses this procedure; nearly every statistics program package includes it. There are also rank and randomization procedures for this design. The usual null hypothesis for this design is that all group means are equal. Rejecting this hypothesis does not say which treatment effects are unequal, just that they are not all the same. Rejection may be followed by comparison of particular pairs of treatments.

If the multiple treatments are multiple doses of one treatment, then use regression analysis or curve fitting instead of analysis of variance. This is discussed in 3.2.4 and 5.2.

Examples of quantitatively differentiated treatments are varnish coat thickness, cleaning time, polishing pressure, and pigment concentration.

The fifth example study outlined at the end of this chapter surveyed books from the Yale University libraries. The classification factor is major library subunit: Law-library-building and main-library-floor are levels of this factor. Use of "level" to refer to unordered categories in addition to ordered and quantitative variables is initially confusing but unfortunately well entrenched. For some analyses, subunits of a library were combined to produce a library factor with levels such as main-library. Since the outcome measures were all categorical, contingency tables with chi-square statistics were used instead of analysis of variance and F statistics.

6.3.2 Multiple Factors

In designs with two groups and paired treatments within each group (6.2.3), there are two treatment factors -- one within objects and one between the groups. The last two examples in 6.2.3 used varnish type and application temperature as the two factors. Factorial multiple group designs are similar, with one group for each combination of factors. An alternative to the 6.2.3 examples is to use four groups of paintings. The four treatments are varnish A at low temperature, varnish A at high temperature, varnish B at low temperature, and varnish B at high temperature. Each group might have 6 objects for a total of 24 experimental units. The data structure for this is shown in 6.3.2.

Table 6.3.2 Varnish and temperature III

Varnish	Temperature	Outcome
A	50°	M
A	90°	M
B	50°	M
B	90°	M
(+ 5 more in each group)		

In a factorial design, there are two or more treatment factors, each with two or more values or levels. The total set of treatments includes all possible combinations of all values of each factor (figure 6H). The total number is the product of the number of levels of each factor. In the example above, two varnish types times two application temperatures equals four combinations and therefore four composite treatment groups. There are preferably at least two replicates in each group.

Standard statistical programs analyze factorial experiments by multi-way analysis of variance. This process tests the main effect of each factor, the interaction of each pair of factors, the triple interaction of each triplet of factors, and so on up to the k-way interaction of all k factors. These are also called the first-order to kth-order effects. Third-order and higher-order effects tend to be difficult to interpret when significant, and in most experiments, we prefer that they be negligible.

The development of factorial experiments and their analysis is one of the major contributions that statisticians have made to scientific and industrial experimentation. They have two advantages over multiple one-way experiments with the same total number of experimental units: They estimate main effects more precisely because all data are used for each estimate, instead of a subset of data from one of the one-way experiments. They also estimate and test interaction effects, which one-way designs cannot do at all. Art conservation research would benefit from their increased use.

A pair of one-way designs corresponding to the four-group example above is to test the two varnishes at one temperature (which?) with 12 objects and test the two temperatures with one varnish (which?) with another 12, giving the same total of 24 objects. However, each main effect would be estimated from 12 objects instead of all 24, and there would be no estimate of the interaction effect.

Some of the factors in a factorial design may be classifications rather than treatments, but they are generally analyzed in the same way. Sex and age are common human classifications. Time, place, and artist classify art objects. The interpretation of such classifications must be in terms of populations rather than processes.

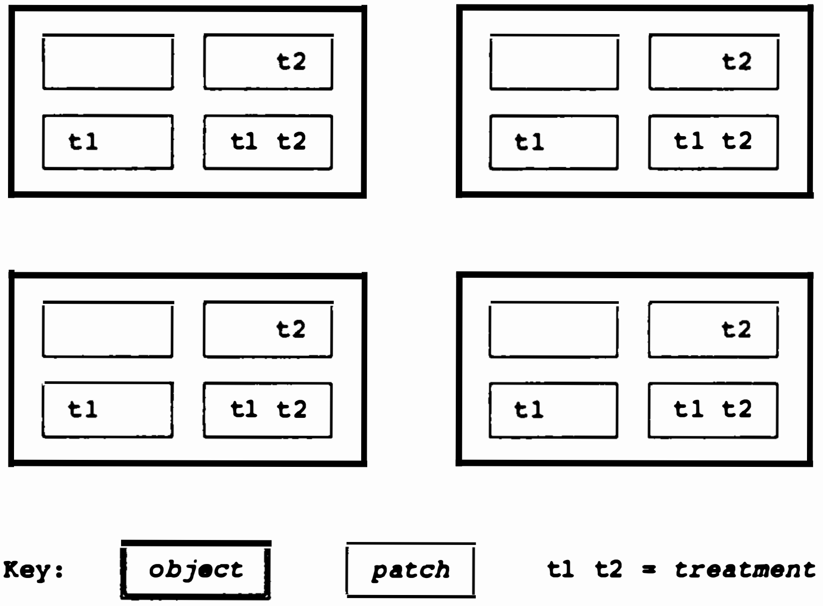
The four conservation experiments summarized at the end of this chapter all had two or three treatment factors. Four brands of cobalt blue were applied at four concentrations, giving 16 treatment combinations for 16 glass plates. Seventeen dyes were combined with five mordants; with two replicates, there were 17 x 5 x 2 = 170 surrogate textile units. Three dyes were exposed to three lamp types with and without a filter, making 18 combinations. Test patches of canvas were impregnated or not, shaded or not, and enclosed or not (eight combinations) for 24 years.

6.3.3 Random Effects

Up to this point we have regarded the treatments in our designs as being fixed. Repeating an experiment has meant using the same values for each treatment factor. The result of an analysis of variance of fixed effects is an estimate and comparison of the effect of those particular values.

Suppose we want to test the color constancy of acrylic paints under particular conditions. A supplier offers 97 choices other than white. We are interested in this whole set but only have resources to test 6 colors. So, from the 97 colors available, we randomly select 6, which are of interest as representatives of the 97 and not as 6 particular values. Repeating this experiment means randomly selecting another set of 6. A color would be repeated only by accident. We want the analysis of the observed changes of the 6 colors chosen to estimate what the average change of the 97 colors would be. We also want a measure of their variability and a test of whether the average and the variability are different from 0. Since the observed color effect (difference between color changes) depends upon the particular set of randomly chosen colors, it is a random effect rather than a fixed effect. Randomly chosen treatments require a slightly modified analysis. We will continue to regard all treatments as being fixed.

Figure 6.3.4 Repeated-measures factorial



6.3.4 Repeated-Measures Factorials

Factorial designs also can be carried out within objects (figure 6.3.4). The 24 treatment units in another version of the varnish/temperature experiment are 4 patches on each of 6 objects. The four treatment combinations are randomized to the four patches in each object. For each object, there are four repeated measurements with two factors -- varnish and temperature. Table 6.3.4 shows the structure of this data. An experiment also can have multiple factors both within and between objects.

Table 6.3.4 Varnish and temperature IV

	Outcome			
Object	A-50°	A-90°	B-50°	B-90°
1	M	M	M	M
	(+ 5 more objects)			

With only two repeated measures per object, one can analyze the difference between the two as an outcome measure, as if there were not a repeated-measures factor. Any other repeated-measures experiment should be analyzed with software designed for repeated measures. The required adjustments to analysis of variance calculations are best done automatically by a tested program (see the third paragraph in 6.3.5).

The surrogate experimental units for the first three examples at the end of this chapter were measured several times after their initial preparation. Such a series of values usually constitutes a quantitative repeated-measures or within-object factor. In the first experiment, however, they were reduced to the time of the highest, which was the outcome measure for analysis. In the second experiment, three human beings rated each unit at the end of its exposure. Treated as repeated readings, the three ratings would immediately be combined into one composite rating for analysis. Kept separate as repeated measures, one could investigate consistency among the raters.

6.3.5 Single-Replicate Factorials

There are advantages, as outlined above, to simultaneously testing several factors with multiple possible values. However, the total number of combinations rapidly becomes unmanageable. One may be tempted to expand an experiment until only one replicate per treatment combination is possible. The data for such an experiment looks like table 6.3.2 with only one line for each unique combination of treatment values.

Analysis of such an experiment by standard analysis of variance techniques requires the assumption that one or more of the highest order interactions are negligible. These are used to estimate the variance of the random part of the data model. The first experiment at the end of the chapter is an example of this design.

If we shift our viewpoint slightly, the repeated-measures experiment in 6.3.4 can be regarded as a three-factor design with no replicates. The objects can be considered to define a third factor, with six values, that classifies the patches. Then there is one patch for each object-varnish-temperature combination. This particular type of single replicate design does not require the no-interaction assumption. Instead, the interactions between object and other patch factors are used as replicate error terms. This is part of the adjustment to standard analysis of variance referred to in 6.3.4.

6.3.6 Fractional Factorials

Investigators sometimes expand the number of factors and levels until the number of combinations is far more than the feasible number of experimental units. If the number of experimental units is an appropriate fraction of the number of factorial combinations, it is possible to proceed with a fractional factorial experiment. An actual but extreme example is a design with nine factors at two levels each and 16 objects. One can partition the 512 (2^9) possible treatment combinations into 32 non-overlapping subsets of 16 treatment combinations each (32 x 16 = 512) so that each subset has an equivalent structure of treatment combinations. Because of this equivalence, any of the 32 subset of 16 treatments can be used for random assignment to the 16 objects. The nine first-order main effects are estimated and tested with an error term having six degrees of freedom (16 - 9 - 1). Experimental design books collectively give several possible fractional factorial designs.

A fractional factorial design is different from a haphazard collection of a subset of possible treatment combinations. The relationship between the treatment combinations in each fractional factorial subset is specifically chosen to allow independent estimation and testing of the main effects.

The problem with fractional factorials is that the formulas for computing groups of effects becomes the same. The example above requires the assumption that there are *no* interactions among the nine factors. Such an experiment is obviously only a screening experiment. Its results constitute hypotheses that need further testing, rather than conclusions to be immediately acted upon. Industries use fractional factorial designs to improve the efficiency and outcome of various processes. This is one of the techniques used in successful quality-improvement programs in Japan. It might also be useful for the improvement of art conservation treatment processes.

A fractional factorial experiment for varnishing paintings could have the following ten factors: paint medium, pre-varnish cleaning agent, varnish type, varnish concentration in solvent, temperature, brush, brushing speed, coat thickness, number of coats, and conservator doing the work.

6.3.7 Blocks and Latin Squares

Several of the designs considered thus far have within-object factors. These are examples of randomized-block designs, with the "block" being an object. A block also can comprise a group of similar objects or a batch of material. Some treatments may be applied to whole blocks instead of to the individual treatment units contained within the blocks. Then the blocks become higher-level treatment units. This design is sometimes called a split-plot design from the original use of agricultural plots as blocks. This idea can be extended to more than two levels of treatment units.

Blocking can be combined with fractional factorials. If there are multiple blocks (usually objects), then each gets a different fraction of the complete factorial design. Enough blocks make a complete factorial with one replicate. For example, 32 blocks with 16 patches each would be enough for the 512 combinations of the nine-factor, two-level design. Better would be 64 blocks or 32 patches in each block, allowing two replicates of each treatment combination.

Latin squares are another specialized design extensively tabulated in experimental design books, with instructions for use. They can be thought of as designs with two blocking factors and fractional replication. For example, let object and day be blocking factors. Randomly label four objects 1, 2, 3, and 4, and randomly label four treatments A, B, C, and D. Then table 6.3.7 indicates which treatment to give each object on each day.

Table 6.3.7 Order-balanced Latin square

	Day			
Object	1	2	3	4
1	A	B	C	D
2	B	D	A	C
3	C	A	D	B
4	D	C	B	A

All four treatments appear in each row and column. Since object is one of the blocking factors, we choose a Latin square that is balanced for treatment order (most are not). A appears once on the first day and once after each of B, C, and D. It appears once on the last day and once before each of B, C, and D. The same order balance is also true of B, C, and D. As well as possible, carry-over from one treatment to another is canceled in the analysis of the main effects. The analysis of a Latin square estimates and tests the main effect of the three factors (object, day, and treatment), combining interactions with the error term. Experimental design books give additional variations and developments.

6.3.8 Quality Control and Improvement

In an agricultural field trial, the usual goal is to find the combination of inputs that gives the highest harvestable output, possibly with adjustment for costs. In many conservation processes, the goal is to attain a particular value or result, rather than the highest possible value. A finished and dried varnish should have the gloss appropriate for the piece, rather than one as close to mirror-like as possible. Infill paint should have, and keep after aging, the appropriately matching hue and tint.

Analysis and evaluation of experiments with such agents and processes is slightly more complicated than simply optimizing the output. Instead, one may subtract the target value, average the difference for replicates of each treatment, take the absolute value of each mean, and compare to find the minimum.

If, as is often the case, there are several treatments that give a result close enough to the target, then another criterion is needed to choose among them. Another aspect of output quality from particular inputs is the variability of results when used in day-to-day studio work. Minimal sensitivity to minor treatment variations is another criteria for choice. One possible strategy is the following, developed in Japan (Box 1988; Dehnad 1989):

1. Select process factors and environmental factors that are economically feasible to control and modify.

2. On the basis of experiment, find settings of sensitivity-controlling factors that minimize output sensitivity to variations in other factors, especially including those that are too expensive to control.

3. Also on the basis of experiment, find settings of the other factors that give results acceptably close to the target.

We do not know of any quality control experiments in art conservation, but believe that these techniques, including others not mentioned here, are potentially useful to the field.

All four treatments appear in each row and column. Since object is one of the blocking factors, we choose a Latin square that is balanced for treatment order (most are not). A appears once on the first day and once after each of B, C, and D. It appears once on the last day and once before each of B, C, and D. The same order balance is also true of B, C, and D. As well as possible, carry-over from one treatment to another is canceled in the analysis of the main effects. The analysis of a Latin square estimates and tests the main effect of the three factors (object, day, and treatment), combining interactions with the error term. Experimental design books give additional variations and developments.

6.3.8 Quality Control and Improvement

In an agricultural field trial, the usual goal is to find the combination of inputs that gives the highest harvestable output, possibly with adjustment for costs. In many conservation processes, the goal is to attain a particular value or result, rather than the highest possible value. A finished and dried varnish should have the gloss appropriate for the piece, rather than one as close to mirror-like as possible. Infill paint should have, and keep after aging, the appropriately matching hue and tint.

Analysis and evaluation of experiments with such agents and processes is slightly more complicated than simply optimizing the output. Instead, one may subtract the target value, average the difference for replicates of each treatment, take the absolute value of each mean, and compare to find the minimum.

If, as is often the case, there are several treatments that give a result close enough to the target, then another criterion is needed to choose among them. Another aspect of output quality from particular inputs is the variability of results when used in day-to-day studio work. Minimal sensitivity to minor treatment variations is another criteria for choice. One possible strategy is the following, developed in Japan (Box 1988; Dehnad 1989):

1. Select process factors and environmental factors that are economically feasible to control and modify.

2. On the basis of experiment, find settings of sensitivity-controlling factors that minimize output sensitivity to variations in other factors, especially including those that are too expensive to control.

3. Also on the basis of experiment, find settings of the other factors that give results acceptably close to the target.

We do not know of any quality control experiments in art conservation, but believe that these techniques, including others not mentioned here, are potentially useful to the field.

6.4 CONSERVATION RESEARCH EXAMPLES

This section outlines the design aspects of four actual conservation research experiments and one conservation survey of a large collection (observational study). These are the examples referred to earlier in the chapter. The four experiments are discussed in much more detail in *Statistical Analysis* (pp 39-46 and appendix), which also gives our suggested analyses, statistical program setups, and outputs. For each of these experiments, the listing under "Analysis" is our suggestion rather than what the original authors did. "ANOVA" is an abbreviation for "analysis of variance."

6.4.1 Cobalt Blue and Linseed Oil Fading

Reference:
 Simunkova, Brothankova-Bucifalova, and Zelinger (1985)
Research problem:
 effect of cobalt blue pigment on linseed oil drying
Experimental units:
 glass plates coated with cobalt blue pigments in linseed oil (32)
Between-unit factors:
 pigment type (4 commercial brands of cobalt blue)
 pigment concentration (4: 5, 10, 20, and 30 units)
Replicates:
 2
Within-unit factors:
 none, but weighed at several time intervals
Outcome measure:
 time to maximum dryness from plotted curve
Analysis:
 analysis of covariance with interaction used for error term

6.4.2 Mordant and Yellow Dye Fading

Reference:
 Crews (1982)
Research question:
 effect of mordant on natural yellow dye fading
Experimental units:
 worsted-wool flannel samples (170)
Between-unit factors:
 dye (17 natural yellows from American plant materials)
 mordant (5 in common use: tin, alum, chrome, iron, and copper)

Replicates:
  2
Within-unit factor:
  exposure (5: 5, 10, 20, 40, 80 AATCC exposure units)
    or (at end of exposure and therefore not crossed)
  rater (3 trained human beings)
Outcome measures:
  instrumental delta E color change (5, after each exposure)
  visual assessment after 80 units of exposure
Analysis:
  repeated measures ANOVA for each within factor and its corresponding outcome

6.4.3 Light and Textile Dye Fading

Reference:
  Bowman and Reagan (1983)
Research problem:
  textile dye fading by infrared and ultraviolet light
Experimental units:
  bleached cotton cloth specimens, 5 x 8.5 cm (18 x ?)
Between-unit factors:
  dye (3: tumeric, madder, indigo)
  lamp (3: incandescent, fluorescent, tungsten halogen quartz)
  filter (2: none or type appropriate for lamp)
Replicates:
  at least 6
Within-unit factor:
  exposure time (4: 100, 200, 300, and 400 hours of light)
Outcome measure:
  reflectance readings
Analysis:
  ANOVA with linear, quadratic, and cubic time effects

6.4.4 Linen Canvas Weakening

Reference:
  Hackney and Hedley (1981)
Research problem:
  reduction of linen canvas weakening by impregnation, shading, and enclosure
Experimental units:
  linen canvas samples (20)

Between-unit factors:
impregnations (2: none or beeswax-and-resin lining mixture)
shading (2: open to normal room light or covered)
enclosure (2: open to normal room air or sealed in case)
Replicates:
2 (if unimpregnated) or 3 (if impregnated)
Within-unit factor:
none, but aged naturally on one of three boards for 24 years
Outcome measures:
tensile strength averaged for 30 to 40 yarns
pH from cold water extract
Analysis:
analysis of variance for each outcome

6.4.5 Book Deterioration Survey

Reference:
Walker, Greenfield, Fox, and Simonoff (1985)
Research problem:
condition and preservation needs of Yale Library collections
Observation units:
books selected from 8 million in Yale University Library system (36,500)
Between-unit factors:
library (15) or major subunit (36)
areas, bookcases, and shelves were used for random sampling
Replicates:
about 1,000 per subunit (222 minimum)
Identification item:
call number
Classification factors:
country of publication
date of publication
circulation (yes or no)
primary protection (rigid, limp, acidic pamphlet)
outer hinge or joint cover (cloth, paper, leather)
leaf attachment (sewn fold, oversewn, stabbed, adhesive)
gutter (inner margin) width (cm)
Outcome measures:
primary protection (intact or not)
leaf attachment (intact or not)
brittleness (breaks with two folds, four folds, or not)

pH (above or below 5.4 as determined by bromocresol green)
printed area (all pages intact or not)
mutilation (by human or animal, yes or no)
environmental damage (fading or water damage, yes or no)
immediate treatment needed (replace, reproduce, repair, rebind; yes or no)

Analysis:
one-, two-, and some three-way tabulations of percent and standard error

6.5 WORK SHEET FOR EXPERIMENTAL DESIGN

A one-page summary of an experiment or study is useful in reviewing and understanding the work of others and in planning and communicating one's own designs. The following page contains a blank work sheet that combines the summary headings used for the examples in chapter 2 and those immediately above. Readers are encouraged to copy it for their own use or to make modified versions more appropriate to their particular circumstances.

Sketching the design in a fashion similar to figures 3.3, 6, and 6.3.4 is often useful. So is explicitly listing the treatment combination for each object.

Art Conservation Experiment Summary

Experimenter

Date

Research problem

Brief background, previous work, and known facts

Hypotheses

Implication and rationale for each hypothesis

Experimental units

Selection of units

Grouping, classification, or between-unit treatment factors

Replicates

Within-unit measurement and treatment factors

Outcome measures

Analysis

Other crucial information

7. CONSERVATION TREATMENT TRIALS

7.1 CLINICAL TRIALS IN MEDICINE

7.1.1 History

In traditional medicine, practices such as bloodletting sometimes continued for centuries in spite of being unsafe, ineffective, or both. Likewise, safe and effective practices were sometimes ignored for far too long. The prevention of scurvy by citrus juice was discovered by Englishmen as early as 1600 but was not uniformly practiced by the British Navy until 1795 -- 195 years later (Drummond and Wilbraham 1940).

The success of modern medicine is due in part to the development of randomized controlled clinical trials. Planned medical experiments date back at least to 1721 when Lady Wortley-Montague and Maitland evaluated experimental smallpox inoculation in the light of previous experience (historical controls). In 1747, Lind directly compared 6 treatment regimes in 12 scurvy patients (2 replicates each) and discovered (again) that citrus gave the best result. In the 1920s, Fisher introduced treatment randomization in the context of agricultural trials. A randomized, placebo-controlled, double-blinded trial of a common cold treatment was published in 1938 (Meinert 1986:4-6 and references therein).

Planned experiments testing treatment efficacy thus developed over two centuries. In simplest form, a clinical trial compares the outcome of patients receiving the test treatment with the outcome of patients receiving a control treatment. Treatments are randomly assigned to patients to reduce bias from uncontrolled factors. Treatment assignments (who got which treatment) are masked from patients and study personnel as far as possible. Both groups are enrolled, treated, and followed over the same time period. The control group may receive a placebo or a standard treatment that is the logical alternative to the new treatment. In a drug trial, a placebo treatment is an inert pill that mimics the treatment pill in appearance and taste.

Today, such trials are required for the approval of new drugs and medical devices in the United States and elsewhere. There is a Society for Clinical Trials and a specialized journal, *Controlled Clinical Trials*. Improvements in clinical trial design, conduct, and analysis continue.

7.1.2 Trial Personnel

The principal investigators in medical treatment trials always include one or more physicians, who as "treaters" are analogous to conservators. Physicians have the expertise to select treatments for testing, diagnose patients for entry into a trial, apply treatments, and evaluate outcomes. The physician determines if a patient should be removed from the trial early due to adverse effects of the treatment and makes the final decision about how to proceed in his or her medical practice based upon the results of the trial.

Scientists and engineers often develop and conduct laboratory testing of drugs and devices later tested on human beings in clinical trials. They may or may not be involved at the clinical stage. Biostatisticians use their knowledge and experience to design the statistical aspects of a trial and conduct the subsequent analysis. They sometimes join doctors as principal investigators. The conduct of a trial is usually assisted by a study nurse. Various people do different types of tests and laboratory analyses. A trial conducted at multiple sites, to increase the number or diversity of patients, requires a trial coordinator.

Of great importance is the study sponsor. Studies of patented drugs and devices are usually paid for by the corporation that will benefit from approval and sale. Other studies are sometimes funded by one of the National Institutes for Health (NIH). Small studies can be conducted, within ethical guidelines, by individual physicians as part of their practice.

7.2 GOALS AND PLANS

7.2.1 Better Treatments

The AIC Code of Ethics says that "the conservator should honestly and sincerely advise what he considers the proper course of treatment" (AIC 1990:23). But it gives no guidance about how to demonstrate that a course of treatment is proper or how to evaluate multiple treatment choices.

Medical and conservation research share the problem that laboratory experimentation is incomplete. Eventually, treatments must be tested on real subjects -- human patients for doctors and works of art for conservators. This almost always introduces additional uncontrollable variables. Yet the results of careful experimentation with real subjects are usually more applicable to treatment practices in general than are the results of less realistic laboratory tests alone. Medicine and conservation share the ethical concern that treatments should be adequately tested and evaluated for whether or not they (1) have harmful side effects, immediate or delayed, (2) really work as intended, and (3) are at least as good as alternatives.

These similarities between medicine and conservation suggest that conservation could benefit from increased use of treatment trials modeled on medical clinical trials. Conservation should be able to learn from decades of treatment-trial development and refinement. This chapter is an attempt to show how and give some modifications.

The general goal of a treatment trial is to improve treatment practice. The specific goal is to directly compare two or more treatments in a controlled situation close to actual practice. Sometimes the specific question is whether a new treatment is definitely better than either doing nothing or applying current treatments that are considered unsatisfactory. In other trials the specific question is whether a new treatment is at least as good as a successful existing treatment.

7.2.2 Drug Testing Phases

The standard program for testing drugs consists of five phases stretching over several years. The initial laboratory phase is followed by four possible phases of testing in human beings. Aspects of both safety and efficacy are a concern at each phase. The process can stop at any phase if the drug fails to pass the corresponding test.

Testing for effectiveness in the laboratory phase depends on the target disease. A cancer drug is tested on human cancer cell lines and animals with cancer. A diabetic drug is tested on animals with induced diabetes. An attempt is made to determine the physiological mode of action. Safety is tested on both human cell or tissue cultures and various species of animals.

The first phase of human testing uses healthy volunteers to determine basic pharmokinetics: How fast does the drug appear in the blood and what is the peak concentration for a given administered dose? How long does it stay around and where does it go? Any possible symptoms or side effects due to the drug are carefully noted.

The next phase uses patients with the target disease in uncontrolled trials to adjust dosage, determine apparent effectiveness compared to historical experience, and further assess safety. If successful, these are followed by human phase III clinical trials, the randomized, controlled trials that are the focus of this chapter.

The final phase is continued surveillance after the drug is approved and marketed for general use. One purpose is to catch rare side effects that do not show up in the earlier trials, which usually involve at most a few thousand persons. This phase depends on good communications and computer data management.

7.2.3 Conservation Research

The five-phase drug testing model is modified by several factors when applied to conservation. The ability to create and age simulated art objects that are much closer to the real thing than animals are to human beings blurs the distinction between laboratory and clinical experiments. The expendability of simulated and low-value objects widens the range of allowable treatments and measurements. Both factors may speed the testing process. The application of poisonous volatile compounds to large areas dictates more attention to safety for the practitioner, as opposed to the patient.

The most successful treatment trials are those that are part of a research program involving a series of trials that build an accumulation of knowledge. Trials are not a substitute for laboratory research with surrogate objects. Treatments are not included in a trial until laboratory testing shows that they are likely to work and that they are probably safe to use on valuable objects.

Conservators must be principal investigators along with scientists, since they have the experience to know which treatments need testing, the skills to apply treatments to real works of art, and the final responsibility for the objects.

The product of the design phase of a treatment trial should be a fairly detailed written protocol. It should start with a rationale for the study, a review of previous research, and a listing of the main goal and any subsidiary goals. Procedures for selecting, measuring, and treating the objects are discussed in 7.3, 7.4, and 7.5. The next part of the protocol, the initial plan for analysis of the data, is discussed in 7.6. The last part of a protocol is a list of references. A good protocol makes it a straightforward matter to conduct, analyze, and report the results of a treatment trial.

7.3 OBJECTS

It is important to clearly define the class of objects eligible for the study so that other researchers can assess whether or not the results apply to their objects. A balance is necessary between homogeneity of objects and conditions on the one hand and practicality and applicability of results on the other. Restricting trial entry to similar objects with similar conditions tends to increase the similarity of treatment outcome within each treatment group, making the difference between treatments, if any, more obvious. Restricting eligibility too much makes it difficult to obtain enough objects to do a trial. It also restricts the applicability of results.

For example, a photograph conservator might conduct a research program to find better treatments for mold. For the first trial, "any photograph with any mold" would probably be too broad an entry criterion. "Albumen prints from the 1890s with mold species X" might be too narrow. The mold on a specific print should be severe enough for treatment to show definite improvement. Each print should be in good enough condition to plausibly tolerate any of the treatments without falling into pieces. Other criteria such as being in an accessible collection may be necessary for practical reasons. "Severe enough," "good enough," and any other criteria should be defined clearly enough in the protocol so that two conservators have a reasonable chance of agreeing on the eligibility of most candidate objects.

The entry criteria for objects must be broad enough to allow a sufficient number of objects into the study. Some minimum number is required to have a reasonable chance of answering the question that motivated the trial. Medical trial investigators usually consult with a statistician on this point, since adequate sample size varies depending upon the number of treatment groups and the number and type of outcome measures. This issue was discussed in 4.2.3. Meinert (1986) and Friedman, Furberg, and DeMets (1982) each devote a chapter to this subject.

7.4 MEASUREMENTS

The measurement protocol for a treatment trial includes both baseline (pre-treatment) and outcome (post-treatment) variables. Reasons for taking baseline measurements include:

1. verify object eligibility for entry into the trial;
2. characterize the study population for post-trial reports;
3. facilitate administration of the trial;
4. control variables that are part of the treatment protocol;
5. adjust outcome measures, as by subtraction;
6. explore relationships between object characteristics and treatment outcome.

One outcome variable should be designated as primary for judging the success of the treatments. This may, however, be a composite variable derived from several different measurements. A common outcome is the status of an object at the end of the trial procedure. An alternative outcome is the time until treatment failure. Treatment failure can be determined by any consistent criterion. Either type of outcome measure can be used with an accelerated-aging protocol.

Data forms are part of the protocol and should be drawn up before the trial begins. Those collecting data or measuring outcomes for the trial must demonstrate that they do so correctly and consistently with each other.

Blinding or masking prevents subjective bias from affecting the results. Whenever possible, the person evaluating treatment outcome should not know which object received which treatment. If possible, the same should be true of the person applying the treatment, in which case that same person can do the evaluation. In medical trials there is the additional concern of blinding the patient but this is not a concern for conservation. A subtle point is that the person evaluating an object for entry into the trial should not know which treatment will be applied to the object if it is entered.

For some medical conditions, such as cancer, heart disease, and recurrent gastric ulcers, the concept of a cure has been nebulous or difficult to measure. The outcome of treatment trials for such diseases has often been time of survival or lack of recurrence up to a cutoff of, say, five years. The time from conception to publication of results could easily be a decade. In the meanwhile, patients have been treated and additional trials may have been started without benefit of the new information.

A current area of research for clinical trials is the development of valid surrogate measures that are available within months instead of years. For cancer, improved imaging techniques and detection of cancer-specific biochemical markers present in the bloodstream at parts per trillion show promise. For heart disease, removal of risk factors such as smoking, hypertension, and high cholesterol are increasingly accepted as endpoints. The problem is that a drug that reduces cholesterol may not actually improve the survival of heart patients. There is controversy over how much risk of error is acceptable in return for quicker results.

In conservation, the true outcome of object treatments may not be evident for decades. One possible speedup is artificially aging expendable objects, but there is sometimes disagreement about extrapolation of results to real conditions. The other speedup technique is to find microstructural, chemical, or other measures that plausibly correlate with long-term outcome.

7.5 TREATMENTS

7.5.1 Selection

Most clinical trials compare two treatments, but more than two are possible. The treatments compared in a trial can come from several sources:

1. folk practice,
2. untested professional practice,
3. mass screening,
4. design from physical-chemical principles,
5. variation of known successes.

Mass screening means applying a large number of candidates to a simple laboratory test system. Thomas Edison's discovery of a usable light-bulb filament material after testing hundreds of possibilities is the classic example. Antibacterial drugs extracted from fungi have been discovered by sprinkling dirt from around the world on bacteria cultures and looking for clear zones a few days later. The recent discovery of taxol extracted from Pacific yew trees comes after tens of thousands of plant extracts have been applied to cancer cell cultures. Other names for this process are trial-and-error (test-and-discard), exhaustive search, and searching for a needle in a haystack or a diamond in the rough.

7.5.2 Randomization and Masking

Randomization of objects to treatment groups (4.4.2) prevents bias from entering into the selection process. The surface reason for requiring randomization is that all statistical hypothesis tests assume in their mathematical probability calculations that randomization was done. If that assumption is violated, statistical results are not valid. The deeper reason is that it serves to distribute between treatment groups, in a manner with known probabilities, the uncontrolled differences in objects. This is especially important for real art objects.

For example, in the study of photographs with mold, we rarely know all details of the history of each object. If too many objects that happen to have one history, such as being displayed in a smoke-filled room, were all put into one group, apparent treatment differences might actually be due to past variations in smoke exposure. Randomly distributing the exposed and unexposed objects between treatment groups will mitigate that problem.

Random assignment produces a schedule of object identifiers and treatments. It is desirable to mask the assignment schedule from as many persons involved in object evaluation and treatment as possible. If the different treatments involve different chemicals identically applied, an assistant can fill coded containers for each object. Additional measures may be required to make the treatment doses indistinguishable. This will be more difficult in conservation than in medicine, where doses are usually packed in tablets or syringes.

If ingenuity will not mask chemical differences or if there are procedural differences, then the treatment applier will know which is which. If this person evaluates objects for entry to the trial before each treatment, or must perform some preliminary procedure on all objects, then treatment names or codes should be written on slips of paper that are sealed inside opaque envelopes that are only opened as needed to treat an object. It is also useful to have someone other the treatment applier evaluate the treatment outcome. When the objects are identical surrogates created for the study, identifying labels can be kept out of sight and objects shuffled before evaluation.

7.5.3 Protocol

The treatment protocol has two parts, the uniform and the differential. The first consists of the procedures and conditions applied to all objects in the trial. These include any initial preparation before application of one of the test treatments and any post-treatment stress or aging before measurement of the outcome. Post-treatment follow-up of an individual object may terminate when the object reaches a certain status (such as failure), after a fixed time interval, or when the entire trial is stopped.

A trial stops when all objects finish the uniform protocol. It may stop sooner -- on a fixed date, when a fixed number of failures have accumulated, or when sufficient data have accumulated to accomplish the goal of the study. If so, different objects may have different follow-up times and the outcome measure may be incomplete for some. If the intended outcome is time to failure, then we only know that it is longer than the observed non-failure time and the true answer is censored. Stopping a trial due to sufficient accumulated data is a specialized topic that is the subject of current research (Elashoff and Reedy 1984).

The second part of the treatment protocol consists of the treatments being compared, or more specifically, the difference among them. Statistical analysis considers whether the actual difference between treatment groups affects the outcome. Treatment blinding or masking, controls, and randomization all serve to make the actual difference as close as possible to the intended difference.

Consider the opposite extreme. Conservator A cleans five French impressionist paintings in a New York museum with agent X. Curator B evaluates the results on a scale of 1 to 10. A few years later, conservator C cleans five German expressionist paintings in a London museum with agent Y. Curator D judges this second batch. Blind application of a statistical test to the 10 outcome measures gives a p value of .03, which is declared significant. There are numerous factors that could explain the difference between the two groups of numbers. No useful interpretation is validly possible. If, however, all factors other than cleaning agent are either held constant or randomized between the two cleaning groups, then a p value of .03 indicates that the observed difference between the two groups of numbers is unlikely to be due to the random effects and is therefore plausibly ascribed to an actual difference between the cleaning agents.

The treatment protocol should be clearly written before a trial so that all objects are treated consistently. This is critical if there are multiple experimenters. This section of the trial protocol is the first draft of the methods section of the research report that will enable others to evaluate, use, reproduce, and extend the results of the trial.

7.6 ANALYSIS

The first factor that determines the specific analysis of a treatment trial is the type of variable that defines the difference between treatments. If there are two treatments (or a treatment and control), then the treatment variable is binary. If more, the treatment variable is usually either categorical, for different types of treatments, or numerical, for different doses of one treatment. Untreated controls can be regarded as a dose of 0 of any treatments. This works fine as the endpoint of a numerical series. More care is required when the treatments other than control are categorically different. A mixture such as two doses of three agents for six treatment combinations defined by two treatment variables (dose and agent) will likely work. An unbalanced mixture such as one dose of treatment A and two doses of treatment B causes more difficulty. Grouping treatment units into pairs or blocks for the randomization generates another treatment factor.

Equally important for analysis is the outcome variable. Binary outcomes are relatively easy to analyze. Numerical outcomes are more sensitive. Ordered categories can sometimes be converted into numbers. The judgements poor, fair, good, excellent might be scored 1, 2, 3, 4, or something else, such as 1, 4, 5, 9. To prevent bias, the choice should be made before the treatment assignments are unmasked. Even so, the results may be more controversial than with conventional numerical measurements. Unordered categories are rare as the outcome of a treatment trial.

An important subtype of numerical outcome is the time to an event. This type of outcome results in partial information when an object is withdrawn from the trial or the trial is stopped before the event happens. There are actually a pair of outcome variables that have to be analyzed together: the first is a time, the second an indicator of whether the time is the time of an event or merely a censored lower limit on the actual time. Lee (1980), Nelson (1990), and others discuss the special procedures of survival or life-table analysis.

Analyses with observed predictor or prognostic variables are optional. These variables are usually measured when an object enters the trial (baseline variables) but may be measured at any time before the final outcome. The general purpose is to find variables other than the treatment that are correlated with and possibly affect the outcome. One specific purpose is to increase the sensitivity of the treatment comparison by adjusting outcomes values to remove the effect of other variables. The adjusted outcome may be less variable within treatment groups. Another purpose is to look for interactions between treatment and other variables that define subgroups of objects for which a treatment is particularly valuable or useless. By itself, this type of analysis should be regarded as generating hypotheses rather than proving facts. If one looks at enough variables, one is sure to find apparently significant but accidental correlations that will disappear with further experiments.

Table 7.6 lists some possible combinations of treatment variable, outcome variable, and analysis. Matching or blocking is included where appropriate. Where baseline number is indicated, categorical baseline variables must be replaced by a set of binary indicator variables that can be used as numerical variables. A categorical treatment variable must also be so converted to do a survival+hazard analysis.

Table 7.6 Analyses for different combinations of treatment and outcome variables

Treatment	Outcome	Analysis
binary	binary	2x2 table, Fisher, Chi-square
binary	number	two-group, t-test
binary, matched	number	pairs, paired t-test
binary	time	life table, survival
binary + baseline number	time	survival+hazard
category	binary	proportions
category	number	analysis of variance
category, blocked	number	two-way ANOVA
category + baseline	number	analysis of covariance
category	time	life table, survival
number	binary	logit, d50, others
number + baseline number	binary	logistic regression, discriminant
number	number	regression
number + baseline number	number	multiple regression

Meinert (1986) examines many practical aspects of clinical-trial data analysis. Fleiss (1986) focuses more on statistical issues.

STATISTICAL GLOSSARY & INDEX

In the entry for each term, italics indicate a cross reference to another entry. The italicized numbers at the end of each entry refer to pages in the text where the term is used or further defined.

analysis of variance A technique for measuring the effect of *categorical variables* on a *continuous variable*. It is based on dividing (analyzing) the observed variation of the continuous variable into components, which are assigned to the effects. *58-59,79,80,83,86-88,99*

arithmetic variable A variable whose values are numbers rather than categories. *47,58, 98-99*

average See **mean.**

binary variable A variable whose values represent the presence or absence of a single category. *47,58,98-99*

carry-over effect In a treatment experiment with more than one treatment interval, the effect of *treatment* in one interval on results of later intervals. *77-78*

categorical variable A variable whose possible values are categories (e.g., the variable "blue pigment" with categories azurite, lazurite, and cobalt blue). *47, 58,80,98-99*

chi square A *test statistic* measuring the extent to which the numbers in a *frequency table* differ from those expected on the basis of some *hypothesis*. Numerically, the sum for all observed/expected pairs of the difference squared divided by the expected value. *46,63,99*

cluster analysis A multivariate technique for dividing objects into groups, or clusters. *58-59*

confidence interval An *interval estimate* with a confidence attached. The confidence is an estimate of the probability that the interval encloses the fixed target. *66*

confounding An inability to separate two types of effects because of the way an experiment is designed. *38,84*

contingency table A *frequency table* with at least two dimensions. *58-59,80*

continuous variable A variable representing a physical quantity considered capable of continuous change. Its values are usually considered to be real numbers or decimals rather than integers. See *arithmetic variable*.

control A sample subjected to the same protocol as the other samples in an experiment except that it does not receive a *treatment*. The purpose is to increase confidence that any effect measured during the experiment is in fact due to the *factor* that was intentionally varied and not due to another, uncontrolled, factor. *51,76,91*

correlation An observed relationship between two ordered variables such that low and high values of one tend to occur with low and high values of the other respectively (positive correlation) or vice versa (negative correlation). *57-59*

curve fitting See *regression*. *60-61*

data model A model describing how a set of data is generated, what it should look like, and how it should be understood. There are usually deterministic or fixed components involving other variables and random portions. *59-62,71,76*

degrees of freedom In mechanics, the number of values required to specify the position and orientation of an object, taking into account the contraints on the system. A train on a track has one; a bead on a wire has two. In statistics, the number of freely variable values, allowing for constraints. Three variables that must sum to 10 have two degrees of freedom. *72*

discriminant analysis A technique for determining the best way to combine *arithmetic variables* to derive a discriminant function that assigns objects to one of several possible groups or categories. The stepwise version selects a parsimonious subset of the variables. *58-59,99*

error term The random component of a variable in a *data model* (which see).

experimental unit The primary unit that is treated and measured for change in an experiment. *43,44,59,82,94*

exponential decay Decay in which the rate at which a measure decreases is proportional to the value remaining. *26-27,60-61*

F A *test statistic* calculated as the ratio of two *variances*. Named after Fisher, who developed *analysis of variance*. *63,79*

factor A variable characterizing one dimension of a total treatment applied to an object. The differences along this dimension may be either quantitative or qualitative. For instance, a pigment factor may represent differences in either concentration or color. *46,48-49,51, 60,70,77-89*

factorial experiment An experiment involving the simultaneous testing of two or more treatment *factors*, each with two or more values or levels. The total set of treatments includes all possible combinations of all values of each factor. *80-82*

fractional factorial An experiment in which the number of *experimental units* is an appropriate fraction of the number of factorial combinations. *46,83-84*

frequency table A table whose columns represent the categories of a particular variable. If there are multiple lines, each row represents the categories of another variable. The entries in the body of the table are the frequency of occurrence (number of occurrences) of a particular category or combination of categories. See *contingency table*. *89,99*

hypothesis A possible answer to a research question; a possible explanation of a phenomena under investigation. *13-15*

hypothesis test A decision as to whether observed experimental data are consistent with a particular hypothesis about the system being investigated. *17-18,62-66,99*

integral A calculus term for a generalization of sums to continuous curves that gives the area under the curve between two endpoints. *64*

interaction A situation when the effect of one variable or treatment depends upon the value of another variable. *78,80*

interquartile range The upper *quartile* minus the lower quartile. *59*

interval estimate An estimate consisting of two endpoints that are intended to enclose some parameter of the population or data model. The size of the interval is related to the uncertainty of the estimate. *66*

least squares An abbreviation for least sum of squared deviations, a criterion for chosing one estimate as best from a plethora of possibilities. The *mean* of a batch of numbers is the least squares estimate of its location. Also used for choosing parameter estimates in *regression*.

level A possible value of a treatment *factor*. This term is used even when the possible values are unordered categories. *70,80*

logarithm The power of a base number required to get a target number. The base 10 logarithms of 100, 500, and 1000 are 2, 2.7, and 3. *61*

mean A *statistic* describing a batch of numbers, calculated as their sum divided by their count. The mean of 3, 5, and 8 is (3+5+8)/3 = 16/3 = 5 1/3. *31,34,57,59-60,62-64,66,69,71-74,76,78-79,85*

measurement protocol A set of decisions regarding what variables to measure in an experiment; how, when, and where to measure them; and what units and method of recording to use. *47-48,94-95*

median The middle value of a batch of numbers. If the count is even, the mean of the two middle values. The median of 3, 5, and 8 is 5 and the median of 3, 5, 6, and 8 is 5 1/2. *57,59*

mode The most frequent value in a batch of numbers.

normal distribution A probability distribution in which most of the values fall near the *mean* with a few higher and a few lower, producing a bell-shaped curve when a histogram is plotted. *62,63-65,71*

null hypothesis A hypothesis that some groups are the same with respect to some characteristic, or that several treatments are the same with respect to some outcome, or that certain variables have no inherent relationship, or that a treatment has the same effect as some *control* treatment. *17,29-31,32,73*

observational study A study in which the researcher does not manipulate variables, but simply observes the outcome of natural processes; a survey. *19-20,42*

one-sided test A treatment-control or treatment-treatment test sensitive to differences in only one direction. *33-34,62-63*

p value The probability of observing a result at least as extreme as the actual result if a certain *null hypothesis* is true. *30,32-35,46,63-65,66-67,73-74*

parameter A number that partly determines the exact form of a mathematical formula that describes a situation. The parameters of a *normal distribution* are its *mean* and *standard deviation*. The parameters of a straight line are an intercept and a slope. A given mathematical model can have several equivalent sets of parameters. *26*

parametric test A *hypothesis test* that requires the estimation of the *parameters* of a *probability* distribution, such as the *normal distribution*. The *p value* is sometimes looked up in a table giving the distribution of the parametric *test statistic*, such as *chi square*, *F*, or *t*.

point estimate A single number or statistic that estimates a parameter of a data model or characteristic of an object or population of objects. *66*

polynomial A sum of multiples of powers, such as x^3+2x^2-3x+5, which has a cubic, quadratic, linear, and constant term. *62,87*

probability Mathematically, one of a set of positive fractions that sum to 1. Interpreted as either the relative frequency of occurrence of an event or the relative degree of confidence in an event or hypothesis. *29-30,32-34,45-46,54,60,64,66*

quartile The lower quartile of a batch of numbers is a number above 1/4 and below 3/4. The upper quartile is above 3/4 and below 1/4. The middle quartile is the *median*. *59*

random sample A sample from a population of objects that are the subject of study. The sample is done by a random process with known *probabilities*. *46*

randomization Mutual assignment of *treatments* and *experimental units* to each other on the basis of a random process which is nearly always designed to give equal probability to each possible set of assignments. The purpose is to eliminate possible bias by the investigator in the choice of which samples should undergo a certain treatment, and to distribute uncontrolled factors in a controlled way. *27, 50,51-56,76,96*

randomization test A *hypothesis test* based on *treatment randomization* in which a *test statistic* is calculated for all possible treatment assignments. The *p value* is the rank of the statistic for the actual assignment in the sorted set of possibilities, divided by their number. *29, 32-33,63*

range The difference of two numbers. Without modification, the largest value of a batch of numbers minus the smallest. Also see *interquartile range*. *54-55,59,61-62*

rank test A *randomization test* in which data values are sorted and replaced by ranks before the calculation of test statistics. In practice, the *p value* is looked up in a pre-calculated table. *63,65*

regression The estimation of a functional relationship between one or more variables, often called dose, stimulus, predictor, or independent variables, and a response or dependent variable. Called linear regression when the geometric representation of the function is a straight line, flat plane, etc. *58-59,60-61,73,79,99*

repeated measurements Multiple measurements of one variable on a single experimental unit intended to assess that variable's change under different conditions at different times or different locations on the unit. *48-49,50,58,59,82-83,87*

repeated readings Multiple measurements of one variable on a single measurement unit not separated by any factor that is part of the study and intended to monitor and average out measurement errors for increased accuracy. *48,58*

replicates Multiple objects or experimental units of a given type that are measured under a particular set of treatment conditions. Replicates estimate treatment effects more accurately than a measurement on one object can by canceling errors when calculating an average. They also estimate treatment effect variability, which cannot be done with one unit without additional assumptions, and which is the basis of statistical inference. *44-46*

residual The portion of an observation that is not explained by a functional relationship between the observation and other variables. See *data model.*

significance level See *p value.*

standard deviation A measure of the spread of a batch of numbers. Calculated as the square root of the *variance. 59,62-64,66,69,71,76*

standard error A measure of the variability of a *parameter* or *statistic* that is estimated from a *standard deviation* by an appropriate formula. *71,89*

statistic A number, such as a *test statistic*, calculated from and summarizing raw data. *30*

statistics In a general sense, the theory and techniques developed for calculating and using numbers calculated from raw research data. In a narrow sense, statistics are used to describe objects, estimate the characteristics of a population from a sample, and test hypotheses or ideas about the subject of a study.

t A *test statistic* that is the ratio between a difference and its standard error. *63-64,71-72*

test statistic A *statistic* calculated as part of a *hypothesis test* to decide amoung competing *hypotheses*. Most often used to decide whether to reject a *null hypothesis*. Usually related to an estimate of treatment effect. *29-32,63,66,71*

treatment In the context of a scientific experiment, a treatment refers to some manipulation of variables on the part of the investigator, designed to have an effect on the object or experimental unit. *29,72,97*

treatment effect The effect of a particular *treatment* on an outcome variable that makes the outcome different from some standard value. The standard value is usually either the value resulting from a control treatment or an experiment-specific mean of the outcomes from several treatments. *27,45-46,77*

treatment protocol A set of decisions regarding how many and which treatments to test, when and where they will be applied, and how they are assigned to objects or experimental units. *50-51,96-98*

two-sided test A treatment-control or treatment-treatment test sensitive to differences in either direction. *33-34,35,62-63,73*

variance A measure of variation; the average square of the deviations from the *mean* for a batch of number or *probability* distribution. *58,83*

REFERENCES

The italicized numbers after the entries are the text pages referring to these sources.

CONSERVATION

AIC

1990 Code of ethics and standards of practice, in *AIC 1990-91 Directory*: 21-31. Washington, D.C.: American Institute for Conservation of Historic and Artistic Works. *92*

Barger, Susan M., A.P. Giri, William White, William Ginell, and Frank Preusser

1984 Protective surface coatings for daguerreotypes. *Journal of the American Institute for Conservation* 24(1):40-52. *51*

Bowman, Janet Gilliland and Barbara M. Reagan

1983 Filtered and unfiltered lights and their effects on selected dyed textiles. *Studies in Conservation* 28(1):36-44. *87*

Cordy, Ann and Kwan-nan Yeh

1984 Blue dye identification of cellulosic fibers: indigo, logwood, and Prussian blue. *Journal of the American Institute for Conservation* 24(1):33-39. *5*

Crews, Patricia Cox

1982 The influence of mordant on the lightfastness of yellow natural dyes. *Journal of the American Institute for Conservation* 21(2):43-58. *86*

Daniels, V. and S. Ward

1982 A rapid test for the detection of substances which will tarnish silver. *Studies in Conservation* 27(1):58-60. *5*

Dragovich, D.

1981 Cavern microclimates in relation to preservation of rock art. *Studies in Conservation* 26(4):143-149. *5*

Fiorentino, P., M. Marabelli, M. Matteini, and A. Moles

1982 The condition of the "Door of Paradise" by L. Ghiberti: tests and proposals for cleaning. *Studies in Conservation* 27(4):145-153. *5*

Hackney, S. and G. Hedley

1981 Measurements of the ageing of linen canvas. *Studies in Conservation* 26(1):1-14. *87*

Hoffmann, Per

1983 A rapid method for the detection of polyethylene glycols (PEG) in wood. *Studies in Conservation* 28(4):189-193. *5*

Lafontaine, Raymond H. and Patricia A. Wood

1982 The stabilization of ivory against relative humidity fluctuations. *Studies in Conservation* 27(3):109-117. *5*

Merk, Linda E.

1981 The effectiveness of benzotriazole in the inhibition of the corrosive behaviour of stripping reagents in bronzes. *Studies in Conservation* 26(2):73-76. *5*

Peacock, Elizabeth E.

1983 Deacidification of degraded linen. *Studies in Conservation* 28(1):8-14. *54*

Reagan, Barbara

1982 Eradication of insects from wool textiles. *Journal of the American Institute of Conservation* 21(2):1-34. *34*

Reedy, Chandra L.

1986 *Technical Analysis of Medieval Himalayan Copper Alloy Statues for Provenance Determination*. Ph.D. dissertation, UCLA Archaeology Program. UMI No. 8606473. *19*

Reedy, Chandra L.

1988 Determining the region of origin of Himalayan copper alloy statues through technical analysis. In *A Pot-Pourri of Indian Art*, edited by Pratapaditya Pal:75-98. Bombay: Marg Publications. *19*

Reedy, Chandra L.

1991 Petrographic analysis of casting core materials for provenance studies of copper alloy sculptures. *Archeomaterials* 5(2):121-163. *19,59*

Reedy, Chandra and Pieter Meyers

1987 An interdisciplinary method for employing technical data to determine regional provenance of copper alloy statues. In *Recent Advances in the Conservation and Analysis of Artifacts*, James Black, ed.:173-178. London: Summer Schools Press. *19*

Rinuy, Anne and François Schweizer

1981 Methodes de conservation d'objets de fouilles en fer: etude quantiative comparee de l'elimination des chlorures. *Studies in Conservation* 26(1):29-41. *5*

Schweizer, François and Anne Rinuy
1982 Manganese black as an Etruscan pigment. *Studies in Conservation* 27(3):118-123. *5*

Simunkova, E., J. Brothankova-Bucifalova, and J. Zelinger
1985 The influence of cobalt blue pigments on the drying of linseed oil. *Studies in Conservation* 30(4):161-166. *86*

Walker, Gay, Jane Greenfield, John Fox, and Jeffrey S. Simonoff
1985 The Yale survey: a large-scale study of book deterioration in the Yale University Library. *College and Research Libraries* 46:111-132 (March 1985). *88*

Wharton, Glenn, Susan Lansing, and William S. Ginell
1988 *Silver Polish Project Technical Report*. Marina del Rey: Getty Conservation Institute. *66*

Winter, John and Elisabeth West FitzHugh
1985 Some technical notes on Whistler's "Peacock Room." *Studies in Conservation* 30(4):149-154. *5*

Wouters, Jan
1985 High performance liquid chromatography of Anthraquinones--analysis of plant and insect extracts and dyed textiles. *Studies in Conservation* 26(3):89-101. *5*

Zehnder, K. and A. Arnold
1984 Stone damage due to formate salts. *Studies in Conservation* 29(1):32-34. *5*

SCIENCE AND STATISTICS

Barlow, David H. and Michel Hersen
1984 *Single Case Expperimental Designs*. New York: Pergamon Press. *23*

Bates, Douglas M. and Donald G. Watts
1988 *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley. *60*

Batschelet, Edward
1981 *Circular Statistics in Biology*. London and New York: Academic Press. *59*

Bethea, Robert M., Benjamin S. Duran, and Thomas L. Boullion
1975 *Statistical Methods for Engineers and Scientists*. New York: Marcel Dekker. *69*

Box, George

1988 Signal-to-noise ratios, performance criteria and transformations. *Technometrics* 30(1): 1-17. *85*

Campbell, Donald T. and Julian C. Stanley

1963 *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally. *69*

Chamberlin, T.C.

1965 Method of multiple working hypotheses. *Science* 148:754-759. *11,15*

Cox, D.R.

1958 *Planning of Experiments*. New York and London: John Wiley and Chapman & Hall. *69*

Dehnad, Khosrow

1989 *Quality Control, Robust Design, and the Taguchi Method*. Pacific Grove: Wadsworth & Brooks/Cole Advanced Books & Software. *85*

Draper, Norman and Harry Smith

1981 *Applied Regression Analysis*, 2nd ed. New York: John Wiley. *60*

Drummond, J.C. and A. Wilbraham

1940 *The Englishman's Food: A History of Five Centuries of English Diet*. London: Jonathan Cape. *91*

Edgington, Eugene S.

1987 *Randomization Tests*. Second Edition. Statistics: Textbooks and Monographs, 77. New York and Basel: Marcel Dekker. *63*

Efron, Bradley and Robert Tibshirani

1991 Statistical data analysis in the computer age. *Science* 253(5018): 390-395. *67*

Elashoff, Janet D. and Terry J. Reedy

1984 Two-stage clinical trial stopping rules. *Biometrics* 40: 791-795. *97*

Fisher, Ronald A.

1935 *The Design of Experiments*. New York: Hafner Publishing. *63*

Fleiss, Joseph L.

1986 *The Design and Analysis of Clinical Experiments*. New York: John Wiley. *99*

Friedman, Lawrence M., Curt D. Furberg, and David L. DeMets

1982 *Fundamentals of Clinical Trials*. Boston: John Wright, PSG. *94*

Glass, G.V., B. McGaw, and M.L. Smith
1981 *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications. *12*

Harré, Rom
1981 *Great Scientfic Experiments*. Oxford: Phaidon. *11*

Hemple, Carl G.
1966 *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall. *11*

Hoaglin, David C., Frederick Mosteller, and John W. Tukey (editors)
1983 *Understanding Robust and Exploratory Data Analysis*. New York and Toronto: John Wiley & Sons. *67*

Hoaglin, David C., Frederick Mosteller, and John W. Tukey (editors)
1985 *Exploring Data Tables, Trends, and Shapes*. New York and Toronto: John Wiley & Sons. *67*

Howard, Margret
1981 Randomization in the analysis of experiments and clinical trials. *Scientific American*. 245(3, September):98-102. *63*

Hunter, John E. and Frank L Schmidt
1990 *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage Publications. *12,67*

Keppel, Geoffrey
1982 *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs: Prentice-Hall. *69*

Lee, Elisa T.
1980 *Statistical Methods for Survival Data Analysis*. Belmont, CA: Wadsworth. *98*

Lett, James
1990 A field guide to critical thinking. *Skeptical Inquirer* 14(2):153-160. *14*

Light, Richard L. and David B. Pillemer
1984 *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press. *12*

Meinert, Clifford
1986 *Clinical Trials: Design, Conduct, and Analysis*. Oxford: Oxford University Press. *91,94,99*

Milliken, George A. and Dallas E. Johnson
1984 *Analysis of Messy Data. Vol. 1: Designed Experiments*. Belmont: Lifetime

Learning Publications. *69*

Milliken, George A. and Dallas E. Johnson
1989 *Analysis of Messy Data. Vol. 2: Nonreplicated Experiments*. New York: Van Nostrand Reinhold. *69*

Nelson, Wayne
1990 *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*. New York: John Wiley. *98*

Pitman, E.J.G.
1937 Significance tests which may be applied to samples from populations. *Journal of the Royal Statistical Society Series B* 4:119-130; 225-232. *63*

Platt, John R.
1964 Strong inference. *Science* 146(3642):347-353. *11,15*

Rasch, Dieter and Gunter Herrendörfer
1986 *Experimental Design: Sample Size Determinations and Block Designs*. Dordrecht: D. Reidel Publishing. *69*

Ratkowsky, David A.
1983 *Nonlinear Regression Modeling*. New York: Marcel Dekker. *61*

Ratkowsky, David A.
1990 *Handbook of Nonlinear Regression Models*. New York: Marcel Dekker. *61*

Reedy, Terry J. and Chandra L. Reedy
1988 *Statistical Analysis in Art Conservation Research*. Research in Conservation, 1. Marina del Rey: Getty Conservation Institute. *6,7,8,57,59,66,86*

Ripley, Brian D.
1981 *Spatial Statistics*. New York and Toronto: John Wiley & Sons. *59*

Salsburg, David S.
1985 The religion of statistics as practiced in medical journals. *American Statistician*. 39(3):220-223. *66*

Saltzman, Max and A.M. Keay
1965 Variables in the measurement of colored samples. *Color Engineering* 3(5):1-6. *60*

Wilson, E. Bright
1952 *An Introduction to Scientific Research*. New York: McGraw-Hill. *52*

THE GETTY
CONSERVATION
INSTITUTE