Topic 4

Producing Data

Statistics has been the handmaid of science, and has poured a flood of light upon the dark questions of famine and pestilence, ignorance and crime, disease and death. - James A. Garfield, December 16, 1867

Our health care is too costly; our schools fail too many; and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet.

These are the indicators of crisis, subject to data and statistics. Less measurable but no less profound is a sapping of confidence across our land a nagging fear that America's decline is inevitable, and that the next generation must lower its sights. - Barack Obama, January 20, 2009

4.1 Preliminary Steps

Many questions begin with an anecdote or an unexplained occurence in the lab or in the field. This can lead to fact-finding interviews or easy to perform experimental assays. The next step will be to review the literature and begin an **exploratory data analysis**. At this stage, we are looking, on the one hand, for patterns and associations, and, on the other hand, apparent inconsistencies occurring in the scientific literature. Next we will examine the data using quantitative methods - summary statistics for quantitative variables, tables for categorical variables - and graphical methods - boxplots, histograms, scatterplots, time plots for quantitative data - bar charts for categorical data.

The strategy of these investigations is frequently the same - look at a **sample** in order to learn something about a **population** or to take a **census** or the total population.

Designs for producing data begin with some basic questions:

- What can I measure?
- What shall I measure?
- How shall I measure it?
- How frequently shall I measure it?
- What obstacles do I face in obtaining a reliable measure?

The frequent goal of a statistical study is to investigate the nature of **causality**. In this way we try to explain the values of some **response variables** based on knowing the values of one or more **explanatory variables**. The major issue is that the associated phenomena could be caused by a third, previously unconsidered factor, called a **lurking variable** or **confounding variable**.

Two approaches are generally used to mitigate the impact of confounding. The first, primarily statistical, involves subdividing the population under study into smaller groups that are more similar. This subdivision is called **cross tabulation** or **stratification**. For human studies, this could mean subdivision by gender, by age, by economic class,

by geographic region, or by level of education. For laboratory, this could mean subdivision by temperature, by pH, by length of incubation, or by concentration of certain compounds (e.g. ATP). For field studies, this could mean subdivision by soil type, by average winter temperature or by total rainfall. Naturally, as the number of subgroups increase, the size of these groups can decrease to the point that chance effects dominate the data.

The second is mathematical or probabilistic modeling. These models often take the form of a mechanistic model that takes into an account the variables in the cross tabulation and builds a **parametric model**.

The best methodologies, of course, make a comprehensive use of both of these types of approaches.

4.2 Formal Statistical Procedures

As a citizen, we should participate in public discourse. Those with particular training have a special obligation to bring to the public their special knowledge. Such public statements can take several forms. We can speak out as a member of society with no particular basis in our area of expertise. We can speak out based on the wisdom that comes with this specialized knowledge. Finally, we can speak out based on a formal procedure of gathering information and reporting carefully the results of our analysis. In each case, it is our obligation to be clear about the nature of that communication and that the our statements follow the highest ethical standards. In the same vein, as consumers of information, we should have a clear understanding of the perspective in any document that presents statistical information.

Professional statistical societies have provided documents that provide guidance on what can be sometimes be difficult judgements and decisions. Two sources of guidance are the *Ethical Guidelines for Statistical Practice* from the American Statistical Society.

`http://www.amstat.org/about/ethicalguidelines.cfm`

and the International Statistical Institute *Declaration on Professional Ethics*

`http://www.isi-web.org/about-isi/professional-ethics`

The formal procedures that will be described in this section presume that we will have a sufficiently well understood mathematical model to support the analysis of data obtained under a given procedure. Thus, this section anticipates some of the concepts in probability theory like independence, conditional probability, distributions under different sampling protocols and expected values. It also will rely fundamentally on some of the consequences of this theory as seen, for example, in the law of large numbers and the central limit theorem. These are topics that we shall soon explore in greater detail.

4.2.1 Observational Studies

The goal is to learn about a population by observing a sample with as little disturbance as possible to the sample.

Sometimes the selection of treatments is not under the control of the researcher. For example, if we suspect that a certain mutation would render a virus more or less virulent, we cannot ethically perform the genetic engineering and infect humans with the viral strains.

For an observational study, effects are often confounded and thus causation is difficult to assert. The link between smoking and a variety of diseases is one very well known example. We have seen the data set relating student smoking habits in Tucson to their parents. We can see that children of smokers are more likely to smoke. This is more easily described if we look at **conditional distributions**.

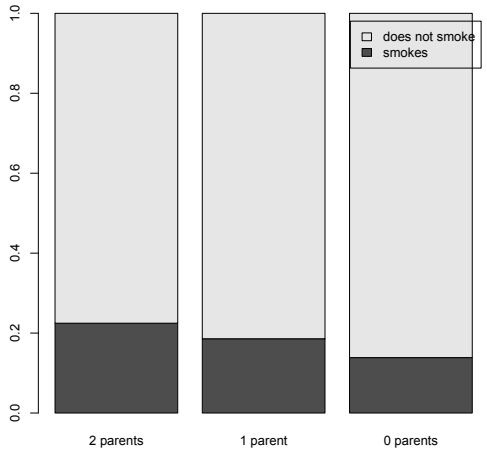
0 parents smoke	
student smokes	student does not smoke
0.1386	0.8614

1 parent smoke	
student smokes	student does not smoke
0.1858	0.8142

2 parents smoke	
student smokes	student does not smoke
0.2247	0.7753

To display these conditional distributions in R:

```
> smoking<-matrix(c(400,1380,416,1823,188,1168),ncol=3)
> smoking
     [,1] [,2] [,3]
[1,]  400  416  188
[2,] 1380 1823 1168
> condsmoke<-matrix(rep(0,6),ncol=3)
> for (i in 1:3)
  {condsmoke[,i]=smoking[,i]/sum(smoking[,i])}
> colnames(condsmoke)
  <-c("2 parents","1 parent", "0 parents")
> rownames(condsmoke)
  <-c("smokes","does not smoke")
> condsmoke
               2 parents  1 parent 0 parents
smokes         0.2247191 0.1857972 0.1386431
does not smoke 0.7752809 0.8142028 0.8613569
> barplot(condsmoke,legend=rownames(condsmoke))
```



Even though we see a trend - children are more likely to smoke in households with parents who smoke, we cannot assert causation, i.e., children smoke because their parents smoke. An alternative explanation might be, for example, people may have a genetic predisposition to smoking.

4.2.2 Randomized Controlled Experiments

In a controlled experiment, the researcher imposes a treatment on the **experimental units** or **subjects** in order to observe a response. Great care and knowledge must be given to the design of an effect experiment. A University of Arizona study on the impact of diet on cancers in women had as its goal specific recommendations on diet. Such recommendations were set to encourage lifestyle changes for millions of American women. Thus, enormous effort was taken in the design of the experiment so that the research team was confident in its results.

A good experimental design is one that is based on a solid understanding of both the science behind the study and the probabilistic tools that will lead to the inferential techniques used for the study. This study is often set to assess some hypothesis - *Do parents smoking habits influence their children?* or estimate some value - *What is the mean length of a given strain of bacteria?*

Principles of Experimental Design

1. **Control** for the effects of lurking variables by comparing several treatments.

2. **Randomize** the assignment of subjects to treatments to eliminate bias due to systematic differences among categories.
3. **Replicate** the experiment on many subjects to reduce the impact of chance variation on the results.

Issues with Control

The desired control can sometimes be quite difficult to achieve. For example;

- In medical trials, some individuals may display a **placebo effect**, the favorable response to any treatment.
- Overlooking or introducing a lurking variable can introduce a **hidden bias**.
- The time and money invested can lead to a subconscious effect by the experimenter. Use an appropriate **blind** or **double blind** procedure. In this case, neither the experimenter nor the subject are aware of which treatment is being used.
- Changes in the wording of questions can lead to different outcomes.
- Transferring discoveries from the laboratory to a genuine living situation can be difficult to make.
- The data may suffer from undercoverage of difficult to find groups. For example, mobile phone users are less accessible to pollsters.
- Some individuals leave the experimental group, especially in longitudinal studies.
- In some instances, a control is not possible. The outcomes of the absence of the enactment of an economic policy, for example, a tax cut or economic stimulus plan, cannot be directly measured. Thus, economists are likely to use a mathematical model of different policies and examine the outcomes of computer simulations as a proxy for control.
- Some subjects may lie. The **Bradley effect** is a theory proposed to explain observed discrepancies between voter opinion polls and election outcomes in some US government elections where a white candidate and a non-white candidate run against each other. The theory proposes that some voters tend to tell pollsters that they are undecided or likely to vote for a black candidate, and yet, on election day, vote for his white opponent. It was named after Tom Bradley, an African-American who lost the 1982 California governor's race despite being ahead in voter polls going into the elections.

Setting a Design

Before data are collected, we must consider some basic questions:

- Decide on the number of explanatory variables or **factors**.
- Decide on the values or **levels** that will be used in the treatment.

Example 4.1. *For over a century, beekeepers have attempted to breed honey bees belonging to different races to take advantage of the effects of hybrid vigor to create a better honey producer. No less a figure than Gregor Mendel failed in this endeavor because he could not control the matings of queens and drones.*

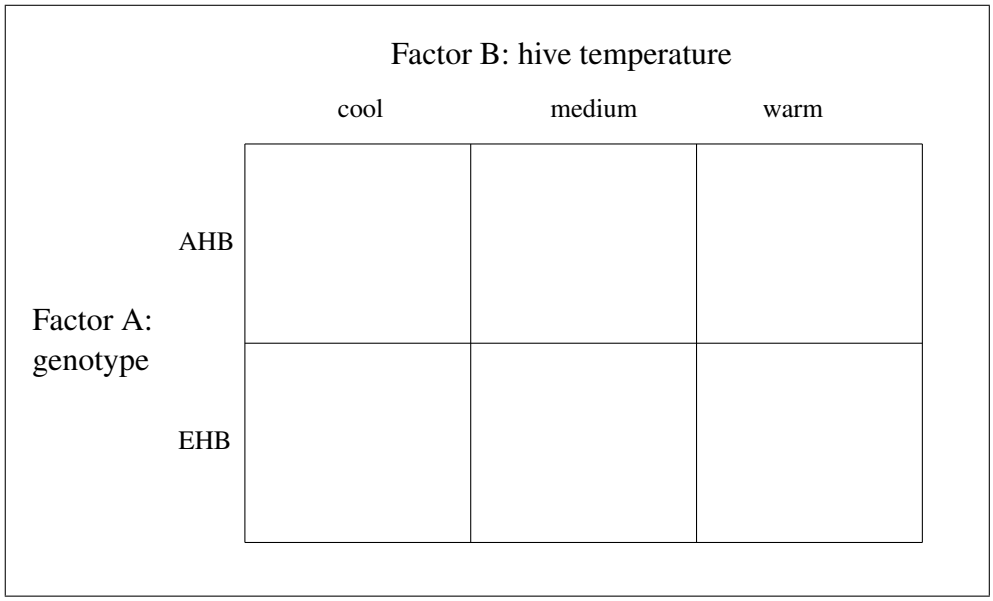
A more recent failure, a breeding experiment using African and European bees, occurred in 1956 in an apiary in the southeast of Brazil. The hybrid Africanized honey bees escaped, and today, in the western hemisphere, all Africanized honey bees are descended from the 26 Tanzanian queen bees that resided in this apiary. By the mid-1990s, Africanized bees have spread to Texas, Arizona, New Mexico, Florida and southern California.

When the time arrives for replacing the mother queen in a colony (a process known as **supercedure***), the queen will lay about ten queen eggs. The first queen that completes her development and emerges from her cell is likely to become the next queen. Suppose we have chosen to investigate the question of whether a shorter time for development*

for Africanized bee queens than for the resident European bee queens is the mechanism behind the replacement by Africanized subspecies in South and Central American and in the southwestern United States. The development time will depend upon hive temperature, so we will determine a range of hive temperatures by looking through the literature and making a few of our own measurements. From this, we will set a cool, medium, and warm hive temperature. We will use European honey bee (EHB) queens as a control. Thus, we have two factors.

- *Queen type - European or Africanized*
- *Hive temperature - cool, medium, or warm.*

Thus, this experiment has **6 treatment groups**.



The response variable is the queen development time - the length of time from the depositing of the egg from the mother queen to the time that the daughter queen emerges from the hive. The immature queen is kept in the hive to be fed during the egg and larval stages. At that point the cell containing the larval queen is capped by the worker bees. We then transfer the cell to an incubator for the pupal stage. The hive where the egg is laid and the incubator that houses the queen is checked hourly. A few queens are chosen and their genotypes are determined to verify the genetic designations of the groups. To reduce hidden biases, the queens in the incubator are labeled in such a way that their genotype is unknown.

We will attempt to rear 120 queens altogether and use 20 in each treatment group. The determination how the number of samples in the study is necessary to have the desired confidence in our results is called a **power analysis**. *We will investigate this aspect of experimental design when we study hypothesis testing.*

Random Samples

A **simple random sample (SRS)** of size n consists of n individuals chosen in such a way that every set of n individuals has an equal chance to be in the sample actually selected. This is easy to accomplish in R. First, give labels to the individuals in the population and then use the command `sample` to make the random choice. For the experiment above, we rear 90 Africanized queens and choose a sample of 60. (Placing the command in parenthesis calls on R to print the output.)

```
> population<-c(1:90)
> (subjects<-sample(population,60))
```

```
 [1] 61 16 65 73 13 25 10 82 24 62 28 66 55  8 26 72 67 17 58 69  6 27 41 20
[25] 87 68 22 11  5 48 33 63 50 88 35 37 84 12  4 59 90 86  2 60 19 18 74 23
[49] 78 49 45  7 64  3 42 57 81 56 46 32
```

If your experimental design call for grouping similar individuals, called **strata**, then a **stratified random sample** from the full sample by choosing a separate random sample from each stratum. If one or more of the groups forms a small fraction of the population, then a stratified random sample ensures the desired number of sample from these groups is included in the sample.

If we mark the 180 queens 1 through 180 with 1 through 90 being Africanized bees and 91 through 180 being European, then we can enter

```
> population<-c(1:180)
> subjectsAHB<-sample(population[1:90],60)
> subjectsEHB<-sample(population[91:180],60)
```

to ensure that 60 come from each group.

For the example above, we divide the sampled Africanized queens into 3 treatment groups based on hive temperature. Here `dim=c(3,20)` signifies that the array has 3 rows and 20 columns. Let the first row be the choice of queen bees for the cool hive, the second row for the medium temperature hive, and row three for the warm hive.

```
> groups<-array(subjects,dim=c(3,20))
> groups
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]   61   73   10   62   55   72   58   27   87    11    33    88    84
[2,]   16   13   82   28    8   67   69   41   68     5    63    35    12
[3,]   65   25   24   66   26   17    6   20   22    48    50    37     4
     [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,]    59     2    18    78     7    42    56
[2,]    90    60    74    49    64    57    46
[3,]    86    19    23    45     3    81    32
```

Most of the data sets that we shall encounter in this book have a modest size with hundreds and perhaps thousands of observations based on a small number of variables. In these situation, we can be careful in assuring that the experimental design was followed. we can make the necessary visual and numerical summaries of the data set to assess its quality and make appropriate corrections to ethically clean the data from issues of mislabeling and poorly collected observations. This will prepare us for the more formal procedures that are the central issues of the second half of this book.

We are now in a world of massive datasets, collected, for example, from genomic, astronomical observations or social media. Data collection, management and analysis require new and more sophisticated approaches that maintain data integrity and security. These considerations form a central issue in modern statistics.

4.2.3 Natural experiments

In this situation, a naturally occurring instance of the observable phenomena under study approximates the situation found in a controlled experiment. For example, during the oil crisis of the mid 1970s, President Nixon imposed a 55 mile per hour speed limit as a strategy to reduce gasoline consumption. This action had a variety of consequences from reduced car accidents to the economic impact of longer times for the transportation of goods. In this case, the *status quo ante* served as the control and the imposition of new highway laws became the natural experiment.

Helena, Montana during the six-month period from June 2002 to December 2002 banned smoking ban in all public spaces including bars and restaurants. This becomes the natural experiment with the control groups being Helena before and after the ban or other Montana cities during the ban.

4.3 Case Studies

4.3.1 Observational Studies

Governments and private consortia maintain databases to assist the public and researchers obtain data both for exploratory data analysis and for formal statistical procedures. We present several examples below.

United States Census

The official United States Census is described in Article I, Section 2 of the Constitution of the United States.

> *The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of 10 years, in such manner as they shall by Law direct.*

It calls for an actual enumeration to be used for apportionment of seats in the House of Representatives among the states and is taken in years that are multiples of 10 years.

`http://2010.census.gov/2010census/`

U.S. Census figures are based on actual counts of persons dwelling in U.S. residential structures. They include citizens, non-citizen legal residents, non-citizen long-term visitors, and undocumented immigrants. In recent censuses, estimates of uncounted housed, homeless, and migratory persons have been added to the directly reported figures.

In addition, the Censsu Bureau provides a variety of interactive internet data tools:

`https://www.census.gov/main/www/access.html`

Current Population Survey

The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

`http://www.census.gov/cps/`

Selecting a random sample requires a current database of every household. The random sample is mutistage.

1. Take a sample from the 3000 counties in the United States.
2. Take a sample of townships from each county.
3. Take a sample of blocks from each township.
4. Take a sample of households from each block.

A household is interviewed for 4 successive months, then not interviewed for 8 months, then returned to the sample for 4 months after that. An adult member of each household provides information for all members of the household.

World Heatlh Organization Global Health Observatory (GHO)

The Global Health Observatory is the World Heatlh Organization's internet gateway to health-related statistics. The GHO compiles and verifies major sources of health data to provide easy access to scientifically sound information. GHO covers global health priorities such as the health-related Millennium Development Goals, women and health, mortality and burden of disease, disease outbreaks, and health equity and health systems.

`http://www.who.int/gho/en/`

The Women's Health Initiative

The Women's Health Initiative (WHI) was a major 15-year research program to address the most common causes of death, disability and poor quality of life in postmenopausal women.

`http://www.nhlbi.nih.gov/whi/`

The WHI observational study had several goals. These goals included:

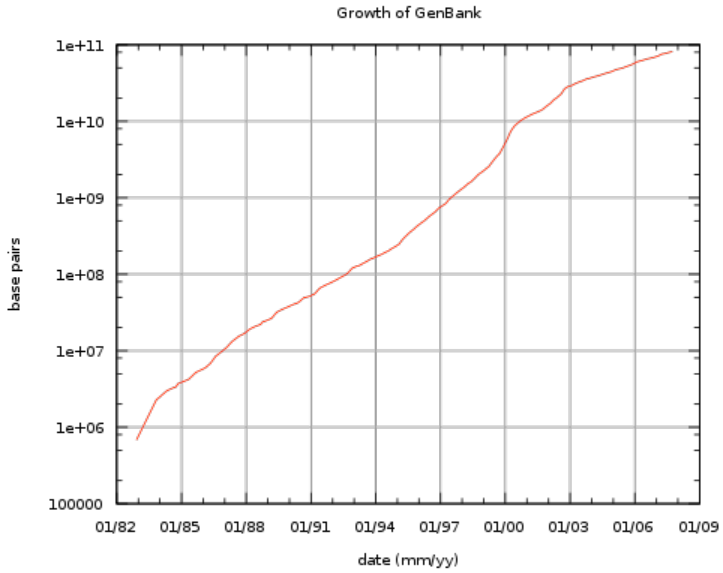
- To give reliable estimates of the extent to which known risk factors to predict heart disease, cancers and fractures.
- To identify "new" risk factors for these and other diseases in women.
- To compare risk factors, presence of disease at the start of the study, and new occurrences of disease during the WHI across all study components.
- To create a future resource to identify biological indicators of disease, especially substances and factors found in blood.

The observational study enlisted 93,676 postmenopausal women between the ages of 50 to 79. The health of participants was tracked over an average of eight years. Women who joined this study filled out periodic health forms and also visited the clinic three years after enrollment. Participants were not required to take any medication or change their health habits.

GenBank

The GenBank sequence database is an open access of nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC. GenBank has approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the whole genome sequence (WGS) division as of April, 2011.

`http://www.ncbi.nlm.nih.gov/genbank/`



4.3.2 Experiments

The history of science has many examples of experiments whose results strongly changed our view of the nature of things. Here we highlight two very important examples.

Light: Its Speed and Medium of Propagation

For many centuries before the seventeenth, a debate continued as to whether light travelled instantaneously or at a finite speed. In ancient Greece, Empedocles maintained that light was something in motion, and therefore must take some time to travel. Aristotle argued, to the contrary, that "light is due to the presence of something, but it is not a movement." Euclid and Ptolemy advanced the emission theory of vision, where light is emitted from the eye. Consequently, Heron of Alexandria argued, the speed of light must be infinite because distant objects such as stars appear immediately upon opening the eyes.

In 1021, Islamic physicist Alhazen (Ibn al-Haytham) published the Book of Optics, in which he used experiments related to the camera obscura to support the now accepted intromission theory of vision, in which light moves from an object into the eye. This led Alhazen to propose that light must therefore have a finite speed. In 1574, the Ottoman astronomer and physicist Taqi al-Din also concluded that the speed of light is finite, correctly explained refraction as the result of light traveling more slowly in denser bodies, and suggested that it would take a long time for light from distant stars to reach the Earth. In the early 17th century, Johannes Kepler believed that the speed of light was infinite since empty space presents no obstacle to it.

In 1638, Galileo Galilei finally proposed an *experiment* to measure the speed of light by observing the delay between uncovering a lantern and its perception some distance away. In 1667, Galileo's experiment was carried out by the Accademia del Cimento of Florence with the lanterns separated by about one mile. No delay was observed. The experiment was not well designed and led to the conclusion that if light travel is not instantaneous, it is very fast. A more powerful experimental design to estimate of the speed of light was made in 1676 by Ole Christensen Romer, one of a group of astronomers of the French Royal Academy of Sciences. From his observations, the periods of Jupiter's innermost moon Io appeared to be shorter when the earth was approaching Jupiter than when receding from it, Romer concluded that light travels at a finite speed, and was able to estimate that would it take light 22 minutes to cross the diameter of Earth's orbit. Christiaan Huygens combined this estimate with an estimate for the diameter of the Earth's orbit to obtain an estimate of speed of light of 220,000 km/s, 26% lower than the actual value.

With the finite speed of light established, nineteenth century physicists, noting that both water and sound waves required a medium for propagation, postulated that the vacuum possessed a "luminiferous aether", the medium for light waves. Because the Earth is in motion, the flow of aether across the Earth should produce a detectable "aether wind". In addition, because the Earth is in orbit about the Sun and the Sun is in motion relative to the center of the Milky Way, the Earth cannot remain at rest with respect to the aether at all times. Thus, by analysing the speed of light in different directions at various times, scientists could measure the motion of the Earth relative to the aether.

In order to detect aether flow, Albert Michelson designed a light interferometer sending a single source of white light through a half-silvered mirror that split the light into two beams travelling at right

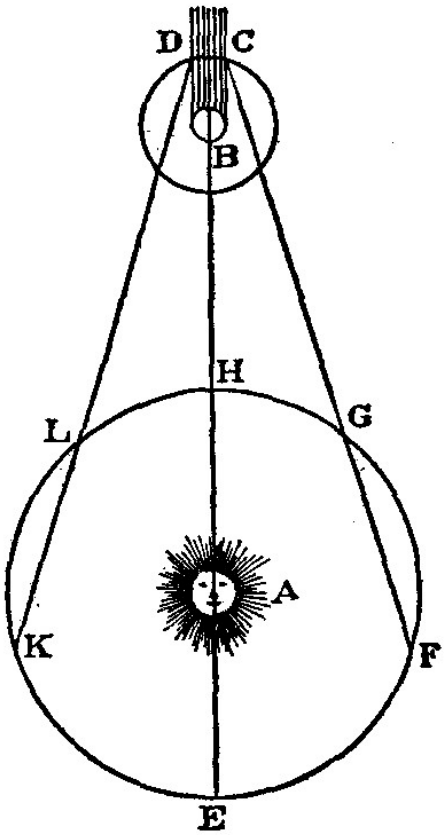


FIG. 70.

Figure 4.1: Romer's diagram of Jupiter (B) eclipsing its moon Io (DC) as viewed from different points in earth's orbit around the sun

angles to one another. The split beams were recombined producing
a pattern of constructive and destructive interference based on the travel time in transit. If the Earth is traveling through aether, a beam reflecting back and forth parallel to the flow of ether would take longer than a beam reflecting perpendicular to the aether because the time gained from traveling with the aether is less than that lost traveling against the ether. The result would be a delay in one of the light beams that could be detected by their interference patterns resulting for the recombined beams. Any slight change in the travel time would then be observed as a shift in the positions of the interference fringes. While Michaelson's prototype apparatus showed promise, it produced far too large experimental errors.

In 1887, Edward Morley joined the effort to create a new device with enough accuracy to detect the aether wind. The new apparatus had a longer path length, it was built on a block of marble, floated in a pool of mercury, and located in a closed room in the basement of a stone building to eliminate most thermal and vibrational effects. The mercury pool allowed the device to be turned, so that it could be rotated through the entire range of possible angles to the hypothesized aether wind. Their results were the first strong evidence against the aether theory and formed a basic contribution to the foundation of the theory of relativity. Thus, two natural questions - how fast does light travel and does it need a medium - awaited elegant and powerful experiments to achieve the understanding we have today and set the stage for the theory of relatively, one of the two great theories of modern physics.

Principles of Inheritance and Genetic Material

Patterns of inheritance have been noticed for millenia. Because of the needs for food, domesticated plants and animals have been bred according to deliberate patterns for at least 5000 years. Progress towards the discovery of the laws for inheritance began with a good set of model organisms. For example, annual flowering plants had certainly been used successfully in the 18th century by Josef Gottlieb Kölreuter. His experimental protocols took the advantage of the fact that these plants are easy to grow, have short generation times, have individuals that possess both male and female reproductive organs, and have easily controlled mating through artificial pollination. Kölreuter established a principle of equal parental contribution. The nature of inheritance remained unknown with a law of blending becoming a leading hypothesis.

In the 1850s and 1860s, the Austrian monk Gregor Mendel used pea plants to work out the basic principles of genetics as we understand them today. Through careful inbreeding, Mendel found 7 true-breeding traits - traits that remained present through many generations and persisted from parent to offspring. By this process, Mendel was sure that potential parent plants were from a true-breeding strain. Mendel's explanatory variables were the traits of the **parental generation**, G. His response variables were the traits of the individual plants in the **first filial generation,** F_1 and **second filial generation,** F_2.

Mendel noted that only one trait was ever expressed in the F_1 generation and called it **dominant**. The alternative trait was called **recessive**. The most striking result is that in the F_2 generation the fraction expressing the dominant trait was very close to 3/4 for each of the seven traits. (See the table below summarizing Mendel's data.) These results in showing *no* intermediate traits disprove the blending hypothesis. Also, the blending theory could not explain the appearance of a pea plant expressing the recessive trait that is the offspring of two plants each expressing the dominant trait. This lead to the hypothesis that each plant has two units of inheritance and transmits one of them to each of its offspring. Mendel could check this hypothesis by crossing, in modern terms, heterozygous plants with those that are dominant homozygous. Mendel went on to examine the situation in which two traits are examined simultaneously and showed that the two traits sort independently. We now use the squares devised in 1905 by Reginald Punnett to compute the probabilities of a particular cross or breeding experiment.

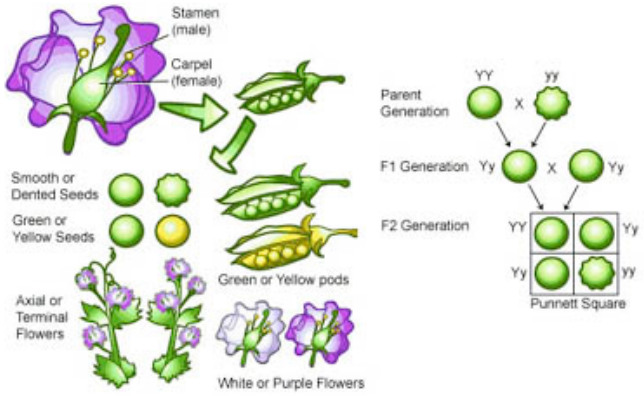


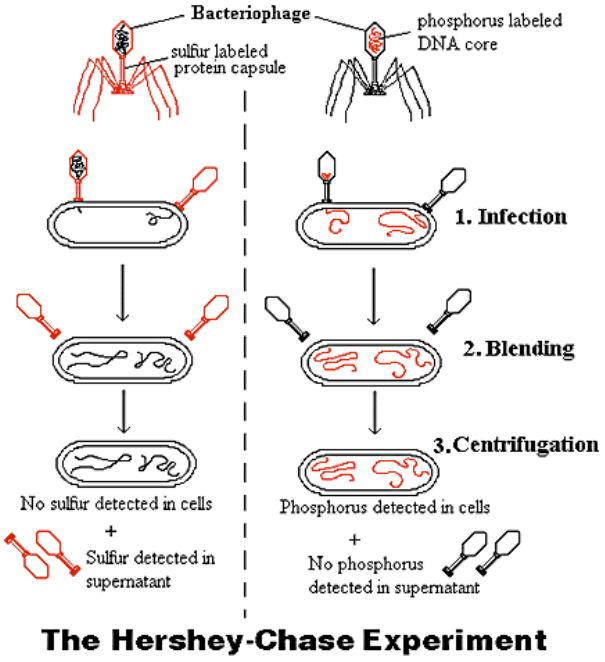
Figure 4.2: Mendel's traits and experiments.

parental phenotypes			F_2 generation phenotypes			
dominant		recessive	dominant	recessive	total	fraction dominant
spherical seeds	×	wrinkled seeds	5474	1850	7324	0.747
yellow seeds	×	green seeds	6022	2001	8023	0.751
purple flowers	×	white flowers	705	224	929	0.758
inflated pods	×	constricted pods	882	299	1181	0.747
green pods	×	yellow pods	428	152	580	0.738
axial flowers	×	terminal flowers	651	207	858	0.759
tall stems	×	dwarf stems	787	277	1064	0.740

We now know that many traits whose expression depends on environment can vary continuously. We can also see that some genes are linked by their position and do not sort independently. (A pea plant has 7 pairs of chromosomes.) The effects can sometimes look like blending. But thanks to Mendel's work, we can see how these expressions are built from the expression of several genes.

Now we know that inheritance is given in "packets". The next question is what material in the living cell is the source of inheritance. Theodor Boveri using sea urchins and Walter Sutton using grasshoppers independently developed the **chromosome theory of inheritance** in 1902. From their work, we know that all the chromosomes had to be present for proper embryonic development and that chromosomes occur in matched pairs of maternal and paternal chromosomes which separate during meiosis. Soon thereafter, Thomas Hunt Morgan, working with the fruit fly *Drosophila melanogaster* as a model system, noticed that a mutation resulting in white eyes was linked to sex - only males had white eyes. Microscopy revealed a dimorphism in the sex chromosome and with this information, Morgan could predict the inheritance of sex linked traits. Morgan continued to learn that genes must reside on a particular chromosomes.

We now think of chromosomes as composed of DNA, but it is in reality an organized structure of DNA and protein. Thus, which of the two formed the inheritance material was in doubt. Phoebus Levene, who identified the components of DNA, declared that it could not store the genetic code because it was chemically far too simple. At that time, DNA was wrongly thought to be made up of regularly repeated tetranucleotides and so could not be the carrier of genetic information. Indeed, in 1944 when Oswald Avery, Colin MacLeod, and Maclyn McCarty found that DNA to be the substance that causes bacterial transformation, the scientific community was reluctant to accept the result despite the care taken in the experiments. These rsearchers considered several organic molecules - proteins, nucleic acids, carbohydrates, and lipids. In each case, if the DNA was destroyed, the ability to continue heritability ended.



Alfred Hershey and Martha Chase continued the search for the genetic material with an experiment using bacteriophage. This virus that infects bacteria is made up of liitle more than DNA inside a protein shell. The virus introduces material into the bacterium that co-opts the host, producing dozens of viruses that emerge from the lysed bacterium. Their experiment begins with growing one culture of phage in a medium containing radioactive phosphorus (that appears in DNA but not in proteins) and another culture in a medium containing radioactive sulfur (that appears in proteins but not in DNA). Afterwards they agitated the bacteria in a blender to strip away the parts of the virus that did not enter the cell in a way that does minimal damage to the bacteria. They then isolated the bacteria finding that the sulfur separated from the bacteria and that the phosphorus had not. By 1952 when Hershey and Chase confirmed that DNA was the genetic material with

their experiment using bacteriophage, scientists were more prepared to accept the result. This, of course, set the stage for the importance of the dramatic discovery by Watson, Crick, and Franklin of the double helix structure of DNA.

Again, for both of these fundamental discoveries, the principles of inheritance and DNA as the carrier of inheritance information, the experimental design was key. In the second case, we learned that even though Avery, MacLeod, and McCarty had designed their experiment well, they did not, at that time, have a scientific community prepared to acknowledge their findings.

Salk Vaccine Field Trials


All Local Teams Are
Mother Objects To Son's Girl
Journal American
7 SPORTS WALL ST. SPECIAL
SALK'S VACCINE WORKS!
Official Count
Polio Vaccine Reported 80 to 90% Effective
All Local Teams Are Raised Out

Poliomyelitis, often called polio or infantile paralysis, is an acute viral infectious disease spread from person to person, primarily via the fecal-oral route. The overwhelming majority of polio infections have no symptoms. However, if the virus enters the central nervous system, it can infect motor neurons, leading to symptoms ranging from muscle weakness and paralysis. The effects of polio have been known since prehistory; Egyptian paintings and carvings depict otherwise healthy people with withered limbs, and children walking with canes at a young age. The first US epidemic was in 1916. By 1950, polio had claimed hundreds of thousands of victims, mostly children.

In 1950, the Public Health Service (PHS) organized a field trial of a vaccine developed by Jonas Salk.

Polio is an epidemic disease with

- 60,000 cases in 1952, and
- 30,000 cases in 1953.

So, a low incidence without control could mean

- the vaccine works, or
- no epidemic in 1954.

Some basic facts were known before the trial started:

- Higher income parents are more likely to consent to allow children to take the vaccine.
- Children of lower income parents are thought to be less susceptible to polio. The reasoning is that these children live in less hygienic surroundings and so are more likely to contract very mild polio and consequently more likely to have polio antibodies.

To reduce the role of chance variation dominating the results, the United States Public Health Service (PHS) decided on a study group of two million people. At the same time, a parents advocacy group, the National Foundation for Infantile Paralysis (NFIP) set out its own design. Here are the essential features of the NFIP design:

- Vaccinate all grade 2 children with parental consent.
- Use grades 1 and 3 as controls.

This design fails to have some of essential features of the principles of experimental design. Here is a critique:

- Polio spreads through contact, so infection of one child in a class can spread to the classmates.
- The treatment group is biased towards higher income.

Thus, the treatment group and the control group have several differences beyond the fact that the treatment group receives the vaccine and the control group does not. This leaves the design open to having lurking variables be the primary cause in the differences in outcomes between the treatment and control groups. The Public Health Service design is intended to take into account these shortcomings. Their design has the following features:

- Flip a coin for each child. (randomized control)
- Children in the control group were given an injection of salt water. (placebo)
- Diagnosticians were not told whether a child was in treatment or control group. (double blind)

The results:

	PHS		NFIP	
	Size	Rate	Size	Rate
Treatment	200,000	28	225,000	25
Control	200,000	71	725,000	54
No consent	350,000	46	125,000	44

Rates are per 100,000

We shall learn later that the evidence is overwhelming that the vaccine reduces the risk of contracting polio. As a consequence of the study, universal vaccination was undertaken in the United States in the early 1960s. A global effort to eradicate polio began in 1988, led by the World Health Organization, UNICEF, and The Rotary Foundation. These efforts have reduced the number of annual diagnosed from an estimated 350,000 cases in 1988 to 1,310 cases in 2007. Still, polio persists. The world now has four polio endemic countries - Nigeria, Afghanistan, Pakistan, and India. One goal of the Gates Foundation is to eliminate polio.

The National Foundation for Infantile Paralysis was founded in 1938 by Franklin D. Roosevelt. Roosevelt was diagnosed with polio in 1921, and left him unable to walk. The Foundation is now known at the March of Dimes. The expanded mission of the March of Dimes is to improve the health of babies by preventing birth defects, premature birth and infant mortality. Its initiatives include rubella (German measles) and pertussis (whooping cough) vaccination, maternal and neonatal care, folic acid and spin bifida, fetal alcohol syndrome, newborn screening, birth defects and prematurity.

The INCAP Study

The World Health Organization cites malnutrition as the gravest single threat to the world's public health. Improving nutrition is widely regarded as the most effective form of aid. According to Jean Ziegler (the United Nations Special Rapporteur on the Right to Food from 2000 to 2008) mortality due to malnutrition accounted for 58% of the total mortality in 2006. In that year, more than 36 million died of hunger or diseases due to deficiencies in micronutrients.

Malnutrition is by far the biggest contributor to child mortality, present in half of all cases. Underweight births and inter-uterine growth restrictions cause 2.2 million child deaths a year. Poor or non-existent breastfeeding causes another 1.4 million. Other deficiencies, such as lack of vitamins or minerals, for example, account for 1 million deaths. According to The Lancet, malnutrition in the first two years is irreversible. Malnourished children grow up with worse health and lower educational achievements.

Thus, understanding the root causes of malnutrition and designing remedies is a major global health care imperative. As the next example shows, not every design sufficiently considers the necessary aspects of human behavior to allow for a solid conclusion.

Figure 4.3: The orange ribbon is often used as a symbol to promote awareness of malnutrition,

The Instituto de Nutrición de Centro Americo y Panama (INCAP) conducted a study on the effects of malnutrition. This 1969 study took place in Guatemala and was administered by the World Health Organization, and supported by the United States National Institute of Health.

Growth deficiency is thought to be mainly due to protein deficiency. Here are some basic facts known in advance of the study:

- Guatemalan children eat 2/3 as much as children in the United States.
- Age 7 Guatemalan children are, on average, 5 inches shorter and 11 pounds lighter than children in the United States.

What are the confounding factors that might explain these differences?

- Genetics
- Prevalence of disease
- Standards of hygiene.
- Standards of medical care.

The experimental design: Measure the effects in four very similar Guatemalan villages. Here are the criterion used for the Guatemalan villages chosen for the study..

- The village size is 150 families, 700 inhabitants with 100 under 6 years of age.
- The village is culturally Latino and not Mayan
- Village life consists of raising corn and beans for food and tomatoes for cash.
- Income is approximately $200 for a family of five.
- The literacy rate is approximately 30% for individuals over age 7.

For the experiment:

- Two villages received the treatment, a drink called *atole*, rich in calories and protein.
- Two villages received the control, a drink called *fresca*, low in calories and no protein.
- Both drinks contain missing vitamins and trace elements. The drinks were served at special cafeterias. The amount consumed by each individual was recorded, but *the use of the drinks was unrestricted*.
- Free medical care was provided to compensate for the burden on the villagers.

The lack of control in the amount of the special drink consumed resulted in enormous variation in consumption. In particular, much more fresca was consumed. Consequently, the design fails in that differences beyond the specific treatment and control existed among the four villages.

The researchers were able to salvage some useful information from the data. They found a linear relationship between a child's growth and the amount of protein consumed:

$$\text{child's growth rate} = 0.04 \text{ inches/pound protein}$$

North American children consume an extra 100 pounds of protein by age 7. Thus, the protein accounts for 4 of the 5 inches in the average difference in heights between Latino Guatemalans and Americans.