# 2

# Doing Science

> *"It is alleged to be found true by proof, that by the taking of Tobacco, divers[e] and very many do find themselves cured of divers[e] diseases; as on the other part, no man ever received harm thereby. In this argument there is first a great mistaking, and next a monstrous absurdity: . . .when a sick man has his disease at the height, he hath at that instant taken Tobacco, and afterward his disease taking the natural course of declining and consequently the patient of recovering his health, O then the Tobacco forsooth, was the worker of that miracle."*
>
> — King James I*, A Counterblaste to Tobacco.*

The previous chapter was fairly theoretical – it's time to get our hands dirty and think about *doing* science. In practice, this means doing experiments; either to explore new phenomena, or more often, to decide something. We define:

> An ***experiment*** is a controlled, reproducible examination of nature intended to arbitrate between competing hypotheses.

In spite of these intentions, an experiment can never settle an issue, it only adds to the evidence supporting (or refuting) a theory or model (reread Chap. 1 if this sounds odd). Of course, often the evidence becomes overwhelming and it becomes foolish to think that experiment has not proven a specific hypothesis (for example, that the Earth is round or that atoms exist).

The majority of a person's contact with science is through medicine. For that reason, this chapter will concentrate on medical science and its peculiar experimental methodology.


E.S. Swanson, *Science and Society*,
DOI 10.1007/978-3-319-21987-5_2

## 2.1 The Beginnings of Science

The notion that experimentation is a tool to understand nature is relatively new. Many of the ancient Greek philosophers, for example, distrusted earthly, fallible senses.

> "Did you ever reach [truth] with any bodily sense? and I speak not of these alone, but of absolute greatness, and health, and strength, and, in short, of the reality or true nature of everything."
>
> — Plato, *Phaedra*

It was the ***Scientific Revolution*** (1543–1727) that brought about a lasting change of attitude. An English lawyer, court intriguer, co-founder of the colonies of Virginia and Newfoundland, and member of parliament named Sir Francis Bacon (Fig. 2.1) was instrumental in bringing about this change. Although Bacon had a classical education, he became disenchanted with the diversity of ancient opinion and its search for teleological ultimate cause.

Figure 2.1: Sir Francis Bacon (1561–1626), 1st Viscount of St. Alban. English philosopher, statesman, and essayist.

> "For to what purpose are these brain-creations and idle displays of power .. All these invented systems of the universe, each according to his own fancy [are] like so many arguments of plays . . . every one philosophises out of the cells of his own imagination, as out of Plato's cave."

For Bacon the way out of the philosophical morass was clear: empty philosophizing must be replaced by ruthless ***empiricism***.[1] Natural philosophers should

[1] Empiricism is the idea that all knowledge is derived from the senses.

"cut nature to the quick" – only in this way could truth be obtained. Men should not seek final cause, said Bacon, rather they should be satisfied with what is knowable. He also stressed that science could be put to the use of the state, which of course has been taken up with enthusiasm by many modern governments. The reason is doubtlessly related to Bacon's most famous aphorism, "knowledge is power".

Bacon expressed other decidedly modern concepts, stressing the importance of the corpuscular theory,[2] for example, in understanding heat; care in experimentation, insisting that experimenters in different fields should communicate; and that the results of experiments should be meticulously recorded. In spite of these modern ideas, Bacon, like all men, was molded by his time and experiences: it appears that he imagined a form of experimentation that was modelled after legal proceedings.

The new methods of science soon made their way into medicine. One of the first studies made was by naval officer James Lind (1716–1794), who sought to alleviate the suffering of sailors due to scurvy. Lind thought that scurvy was caused by "putrefaction of the body" and that this could be remedied with acidic food. He tested his idea by giving six groups of sailors with scurvy identical diets but with differing supplements (barley water, cider, oranges, etc.) and remarked that those given citrus soon recovered.

In 1775, Sir Percivall Pott (1714–1788) found an association between scrotal cancer in chimney sweeps and exposure to soot, thereby demonstrating a link between occupation and cancer and the existence of environmental ***carcinogens***.[3]

Another watershed moment in the development of ***epidemiology***[4] came during an outbreak of cholera in London in 1854. John Snow, a local doctor, was skeptical of the miasma theory of disease and sought another explanation for the outbreak. By tracing the addresses of the sick he was able to identify a public water pump as the source of the disease. Later he used statistical methods to show there was a correlation between cholera incidence and water quality. Snow went on to become a founding member of the Epidemiological Society of London.

Finally, we consider the study of childbed fever conducted by the Hungarian physician, Ignaz Semmelweis (1818–1865). At the time, women contracted childbed fever with alarming regularity in European maternity clinics. Semmelweis noted that women attended by physicians tended to contract the disease more often than those attended by midwives. The death of a friend due to infection led him to guess it was physician contact with cadavers that caused childbed fever. Semmelweis tested his idea by requiring his doctors to wash their hands before treating women. The result was an immediate and dramatic drop in the infection rate. Later Semmelweis showed that the opening of a nearby pathological

[2] A corpuscle is a particle, so this refers to the idea that matter is made of atoms.

[3] A carcinogen is something that causes cancer.

[4] Epidemiology is the study of causes and effects of disease and other factors that impact health.

anatomy clinic was accompanied by an increase in fever rates – thereby establishing a correlation between the handling of corpses with the incidence of fever.

As bacteria were unknown at the time, Semmelweiss could not explain his findings. His results were ignored or derided by the medical establishment and he lost his job. Semmelweis did not deal with the rejection well and was eventually committed to an asylum, where he died of blood poisoning, probably as a result of being severely beaten by guards.

Important studies continue to be conducted. In the past 60 years it has been established that smoking is linked with lung cancer and that diet is associated with heart disease. These days science is big business. The United States government spends about $143 billion per year on research and development (about ½ of this is on military applications). It is estimated that the government and business spend $100 billion on medical studies every year, while world wide expenditure on biomedical research is around $270 billion per year. These budgets can be compared to the US Department of Energy outlay on particle physics of $¾ billion per year.

## 2.2 Studies

A ***clinical trial*** or ***study*** is an experiment that is typically performed on living things (like people). Because of this, studies have a different complexion than experiments in the physical sciences: people are not as reproducible as crystals or lasers, and the outcomes of experiments are often random (for example, smoking does not always cause cancer). Thus studies tend to seek average effects, require large numbers of subjects, and are prone to human biases.

The competing hypotheses that experiments and studies examine are usually of the type *A causes B* and *A does not cause B*. In the lingo, *A* is an ***exposure*** and *B* is an ***outcome***. Thus a typical study seeks to determine if there is a ***causal link*** between an exposure and an outcome. For example, one might hypothesize that smoking causes lung cancer in some way; here smoking is the exposure and having cancer is the outcome.

In spite of these goals, studies cannot determine whether an outcome is *caused* by an exposure – they can only measure the ***correlation*** between the exposure and the outcome. For example noting that smokers tend to have more lung cancer does not (necessarily) mean that smoking causes cancer. Perhaps the cancer is caused by something else that is in turn correlated with smoking.

| Correlation does not imply causation. |
|---|

This difficulty is captured in the phrase, “correlation does not imply causation”. Consider a study that finds a correlation between hot dog sales and drowning deaths. One might conclude that swimmers like hot dogs and when they eat

too many they tend to get cramps and drown. But a simpler explanation involves no causation at all: hot dog sales go up in the summer, which is when swimming happens. Similarly, one might examine divorce rates over the past 50 years and find a correlation with household television ownership. Maybe watching too much tv leads to divorce, but the correlation is probably due to an increase in both tv ownership and the divorce rate over the past few decades, presumably for independent reasons. A similar observation can be made about the sales of organic food and the rate of autism in the last decade (Fig. 2.2). Although the numbers track together remarkably well, it seems unlikely that this correlation is causal.

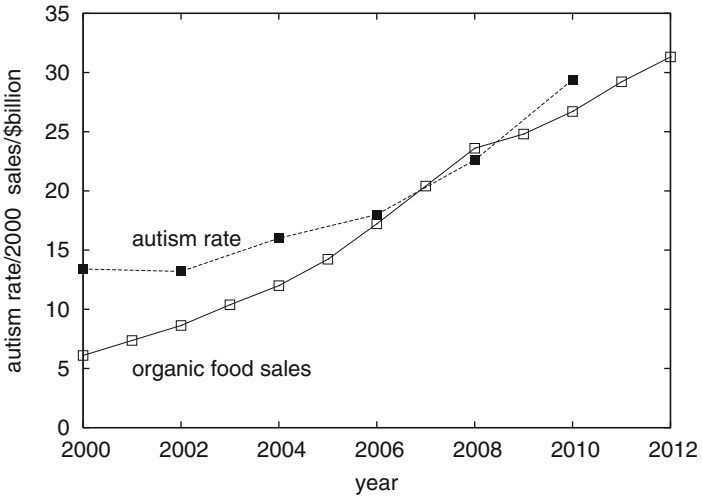


Figure 2.2: Autism rate in 8 year olds (per 2000) and sales of organic foods.

If experiment can never prove causality, how is one to make progress? The answer is that eventually so many experiments find tight correlation between exposure and outcome that it becomes nonsensical to assume a noncausal relationship between them (think of Occam's razor). In an effort to sharpen this, the British epidemiologist Sir Austin Bradford Hill (1897–1991) established a set of conditions necessary to argue for causality.

### Bradford Hill Criteria

**Analogy** Factors similar to a suspected causal agent should be investigated as possible causes.

**Biological Gradient** There should be a relationship between dose and patient response.

**Coherence** Controlled laboratory results should agree with epidemiological experience.

**Consistency** Results of studies should be consistent across a range of factors, such as who conducts the study, when it is conducted, or dosages employed.

**Direct Evidence** Direct controlled experimental demonstration of an outcome under the influence of an exposure is strong indication of causality.

**Plausibility** The apparent cause and effect must make sense in the light of current theories. If a causal relationship appears to be outside of current science then significant further testing must be done.

**Specificity** A specific relationship between outcome and exposure increases the likelihood of a causal relationship.

**Strength and Association** Strong correlation between outcome and exposure increases the likelihood of a causal relationship.

**Temporality** The outcome should occur after the exposure.

---

It is evident that the more of these common sense criteria that hold, the more likely it is that a true causal relationship has been established between exposure and outcome.

## 2.3 Study Design

### 2.3.1 Types of Study

Studies can vary widely in their design, depending on goals, financing, and other constraints. A fundamental distinction is between ***observational*** and ***randomized*** studies.

**observational** a study in which the investigator observes rather than influences exposure and disease among participants. Case-control and cohort studies are observational studies.

**case-control** an observational study that enrolls one group of persons with a certain disease, chronic condition, or type of injury (case-patients) and a group of persons without the health problem (control subjects) and compares differences in exposures, behaviors, and other characteristics to identify and quantify associations, test hypotheses, and identify causes.

**cohort** an observational study in which enrollment is based on status of exposure to a certain factor or membership in a certain group. Populations are followed, and disease, death, or other health-related outcomes are documented and compared. Cohort studies can be either prospective or retrospective.

The terms ***prospective*** and ***retrospective*** refer to whether trial subjects have or do not have the health outcome of interest at the beginning of a trial.

**prospective** a study in which participants are enrolled before the health outcome of interest has occurred.

**retrospective** a study in which participants are enrolled after the health outcome of interest has occurred. Case-control studies are inherently retrospective.

Ex. To test whether power lines cause cancer, researchers questioned cancer patients about how close they lived to power lines.

Ex. To test whether power lines cause cancer, researchers followed the health history of 1000 people living near power lines.

The first example is a case-control study, while the second is a prospective cohort study.

The other study category mentioned was "randomized", which refers to ***randomized controlled studies***. These are defined as

**randomized controlled** a study in which subjects are randomly placed in a two or more groups that receive different treatments,

and are regarded as the most reliable form of study.[5] The term ***control*** refers to a group that does not receive the test treatment and serves as a basis of comparison for the treatment group. The control group can receive treatment from a known drug or can be given a ***placebo***.

A placebo is a sham medicine (often a sugar pill) that is used to minimize the difference between the control and test groups. Specifically, there is a powerful psychological effect associated with the belief that one is being treated. It is therefore important that the control and the test groups both believe they are receiving treatment.

### 2.3.2 Bias

Researchers give placebos to control groups because they are trying to eliminate ***bias*** from their studies. There are many sources of bias in studies and researchers go to great efforts to eliminate or reduce them. For example, randomly assigning

[5] Many other types of studies exist, including cross-sectional (a survey), screening, and diagnostic.

persons to control and test groups removes ***selection bias***. To see the importance of this, consider a researcher with an interest in proving the efficacy of a new drug who selects members of the test group in a study. He could be tempted (unconsciously or otherwise) to choose healthier persons to be in the test group, thereby skewing the results of the study.

In a similar way, a researcher who assesses health outcomes during a trial could be tempted to be generous with people who he knows are in the test group. This is called ***interviewer bias***. There is a simple method to control for assessment bias called ***blinding***. A blinded study eliminates this bias by not revealing group membership to researchers. It can even be important to blind the study subjects so that they do not know if they are receiving treatment or a placebo. Such studies are called ***double blind***.

There are dozens of kinds of bias that can confound the most well intentioned research. Some of these are listed here.

**Study Biases**

**pre-study**

**study design** clear goals and criteria must be decided before a study is undertaken.

**selection bias** patients are not randomized or are not selected according to clear pre-set criteria.

**channeling bias** patients are not added to cohorts with clear pre-set criteria.

**in-study**

**interviewer bias** researcher interaction with subjects should be standardized and the researcher should be blinded to the exposure status of the subject.

**recall bias** patients who are asked to recount experience or results can introduce bias. It is preferable to find impartial methods to rate results.

**transfer bias** sometimes studies must follow-up with patients to obtain study results. A policy to deal with patients who cannot be found must be established before the study is made.

**dropout bias** people who leave studies can introduce bias if there is a common reason for withdrawal (such as being too sick to carry on).

**performance bias** studies that depend on procedures (such as surgery) can introduce bias due to time-dependence of ability (for example the surgeon gains experience, or the surgeon has a bad day).

**exposure misclassification** unclearly defined exposures can introduce bias.

**outcome misclassification** unclearly defined outcomes can introduce bias.

**post-study**

**citation bias** researchers who choose to refer to certain publications, but ignore others, introduce citation bias in their studies.

**salami slicing** researchers take the results from one study and slice the results into several reports without making clear that the reports are not independent. In this way a single positive trial can appear as many positive trials, giving a false impression.

**publication bias** researchers who decide not to publish results can skew the public record, leading to bias.

---

Let's consider examples of these biases.

Ex. one common example is the perceived association between autism and the MMR vaccine. This vaccine is given to children during a prominent period of language and social development. As a result, parents of children with autism are more likely to recall immunization administration during this developmental regression, and a causal relationship may be perceived.

Ex. in research on the effectiveness of batterers treatment programs, some researchers use conflictual couples seeking marriage counseling, and exclude court referred batterers, batterers with co-existing mental disorders, batterers who are severely violent, and batterers who are substance abusers . . . and then conduct the research in suburban university settings.

Ex. using psychology students in studies.

Ex. the Interphone study on cancer and cell phones determined usage by asking participants to estimate how many hours they used their phones per week.

Ex. paying subjects (procedural bias).

Ex. most medical studies have been done on white or black men (sampling bias).

Although these are just some of the ways that studies can be skewed, it is still a daunting list and illustrates just how careful an assiduous researcher must be. It is also the responsibility of readers to be aware of the limitations of any studies they are considering. To assist, a simple test called the ***Jadad scale*** has been devised by Alejandro Jadad Bechara (1963–) to assess the reliability of studies. Each affirmative answer earns one point: good studies should score 4 or 5, whereas studies scoring 0, 1, or 2 should not be relied on in forming opinion or courses of action.

**The Jadad Scale**

1. Is the study randomized?
2. Is the study double blind?
3. Were dropouts and withdrawals described?
4. Was the method of randomization described?
5. Was the method of blinding described?

Not all problems are associated with bias; simple methodology can lead to issues in interpreting studies.

> Ex. Members of the same research group went on to publish a comprehensive survey of the content and quality of randomized trials relevant to the treatment of schizophrenia in general. They looked at 2,000 trials and were disappointed in what they found. Over the years, drugs have certainly improved the prospects for people with schizophrenia in some respects. For example, most patients can now live at home or in the community. Yet, even in the 1990s (and still today), most drugs were tested on patients in hospital, so their relevance to outpatient treatment is uncertain. On top of that, the inconsistent way in which outcomes of treatment were assessed was astonishing. The researchers discovered that over 600 treatments – mainly drugs but also psychotherapy, for example – were tested in the trials, yet 640 different scales were used to rate the results and 369 of these were used only once. Comparing outcomes of different trials was therefore severely hampered and the results were virtually uninterpretable by doctors or patients. Among a catalogue of other problems, the researchers identified many studies that were too small or short term to give useful results. And new drug treatments were often compared with inappropriately large doses of a drug that was well known for its side-effects, even when better tolerated treatments were available –an obviously unfair test.
>
> I. Evans *et al.*, *Testing Treatments.*

## 2.4 Statistics and Studies

People are complex physical systems and do not respond in identical ways to external factors. Thus studies necessarily have an element of randomness to them and conclusions can only be expressed in terms of probabilities. For example, if 100 people are given an experimental drug, 20 may respond well, 30 may experience some benefits, 40 may remain indifferent, and 10 may have serious side effects. Doing the study again will yield different numbers. How is one to interpret the study data?

### 2.4.1 The Normal Distribution

The way to deal with randomness is with statistics. This can be an intimidating subject, so we are fortunate that a lot can be understood fairly simply.

You are probably familiar with the most basic and famous statistical quantity called the Gaussian or ***normal distribution***.

If one were to make a histogram of heights or IQ scores or weights of the American population they would look like normal distributions (Fig. 2.3). The average height for men is about 70 inches, which coincides with the peak of the distribution. The average is called the ***mean*** and is denoted $\mu$. The shape of the normal distribution (how narrow or fat it is) is given by another quantity called the ***standard deviation***, denoted $\sigma$. The standard deviation for height is about 3 inches.

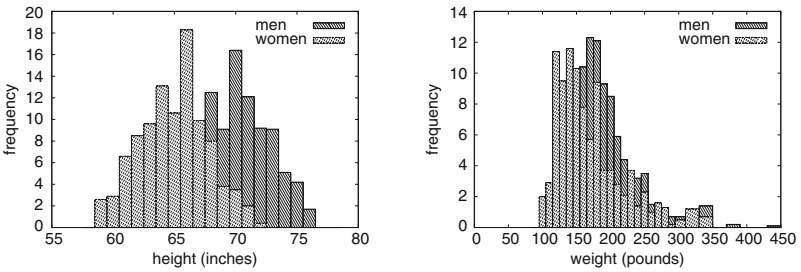


Figure 2.3: Height and weight distributions, US men and women, ages 20–29.
Source: CDC NHANES survey.

These experimental distributions are well approximated by the mathematical normal distribution,[6] shown in Fig. 2.4. The figure illustrates how the standard deviation is related to the shape of the curve. By definition the fraction of the curve between the mean ($\mu$) and the mean plus one standard deviation ($\mu + \sigma$) is 34.1 %. An additional 13.6 % is picked up between $\mu + \sigma$ and $\mu + 2\sigma$, and 2.1 % between $2\sigma$ and $3\sigma$. A final 0.1 % remains above $\mu + 3\sigma$.

[6]The formula for the normal distribution is $N(x) = \exp[-(x-\mu)^2/(2\sigma^2)]/(\sigma\sqrt{2\pi})$.

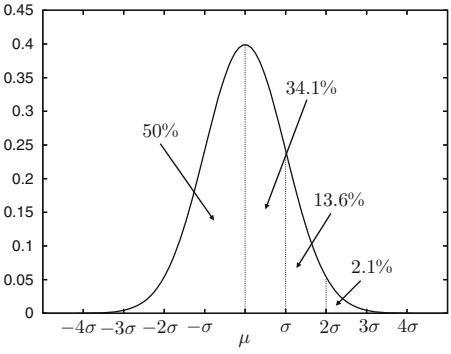


Figure 2.4: The normal distribution.

| Standard deviations | Cumulative fraction (%) |
|---|---|
| $\mu$ | 50.0 |
| $\mu + \sigma$ | 84.1 |
| $\mu + 2\sigma$ | 97.7 |
| $\mu + 3\sigma$ | 99.9 |
| $\mu + 4\sigma$ | $1 - 3.1 \cdot 10^{-5}$ |
| $\mu + 5\sigma$ | $1 - 2.9 \cdot 10^{-7}$ |
| $\mu + 6\sigma$ | $1 - 9.9 \cdot 10^{-10}$ |

Ex. The probability of landing within one standard deviation of the mean is 34.1% + 34.1% = 64.2%.

Ex. The mean for IQ tests is defined to be 100 and the standard deviation is about 15. If you have an IQ of 130, 97.7% people have a score lower than you. This is because a score of 130 is $2\sigma$ above the mean and the area under the normal curve up to $2\sigma$ is 50% + 34.1% + 13.6%.

The normal distribution seems to be everywhere and there is a good reason for this, called the ***central limit theorem***. The theorem states that the sum of many random numbers, no matter how they are distributed, approaches a normal distribution (Fig. 2.5). The theorem provides a clue about the pervasiveness of the normal distribution: it must be that height, IQ, etc. are net attributes due to many genetic and environmental factors, each one of them random.

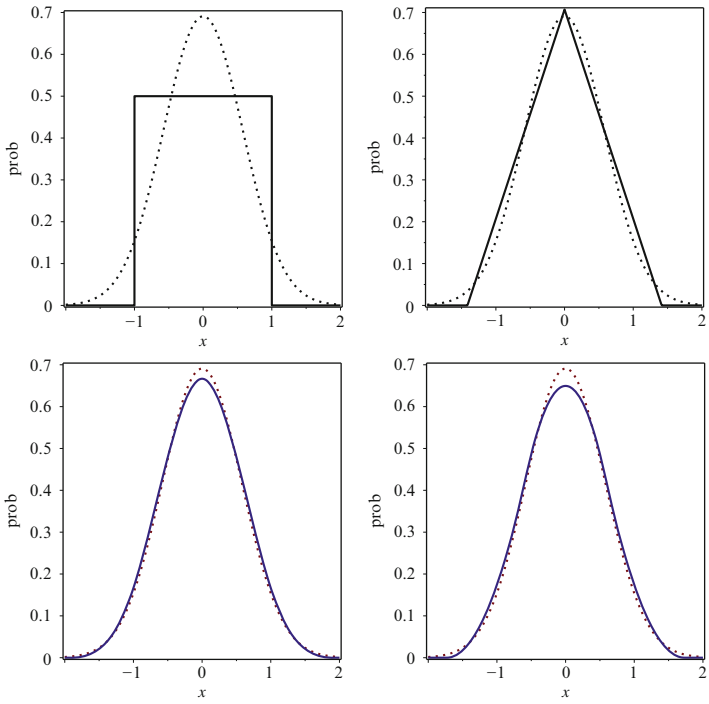


Figure 2.5: The central limit theorem. Clockwise from *top left*: the sum of 1, 2, 3, and 4 random variables.

### 2.4.2 Error Bars

A main use of the normal distribution is in estimating the reliability of conclusions. Say, for example, that you wish to determine how many Americans are in favor of capital punishment. In principle you could ask everyone in the country, but this would be expensive, and probably impossible to arrange. In practice, pollsters ask a random group of people, called a ***sample***. Say 1000 people are asked and the responses are 58 % in favor, 37 % not in favor, and 5 % have no opinion. How reliable is this result?

If you had a lot of money and time, you could ask another 1000 people and check. If you did this 47 times you could make a histogram of the results (Fig. 2.6). As usual, the histogram looks like a normal distribution and we can ask what the mean and standard deviation of the data is. The curve that best reproduces the survey results is shown as a dashed line and tells us that the mean is 59.2 % and the standard deviation ($\sigma$) is 0.9 %. This means that our best estimate is that 59.2 % of people agree with the survey question. Also, if we repeat the survey many times, we will find a result between 58.3 % and 60.1 % (59.2 % $\pm$ 0.9 %)

68.4 % of the time. People often say, the average is 59.2 % with an ***error bar*** of 0.9 % or the average is 59.2 % with 68 % ***confidence interval*** of 0.9 %.

There is no reason to run 47 different opinion surveys, we could simply make one survey of 47,000 different people and obtain the same agreement rate of 59.2 %. Then we can divide the results into 47 groups, recalculate the averages and make the histogram of Fig. 2.6. In fact this is the way statisticians obtain the confidence intervals that are reported in the media and in studies.

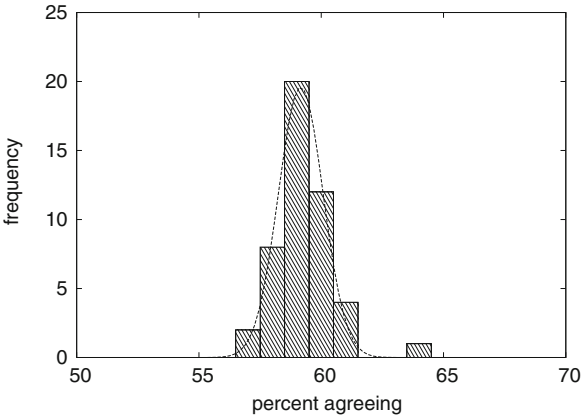


Figure 2.6: Results from your opinion survey.

> Ex. "The AP-GfK poll was conducted March 20–24, 2014. It involved online interviews with 1,012 adults and has a margin of sampling error of plus or minus 3.4 percentage points for all respondents."

What if your bosses wanted *really* accurate results? You might guess that the more people you ask, the more accurate the results. This is correct. The mathematics says that if the number of survey respondents is $N$ then the error bar goes like some number divided by $\sqrt{N}$. This means that if the error with $N = 15$ is $\sigma = 45$ then the error with $N = 1500$ is $\sigma = 4.5$ (an increase of a factor of 100 in survey size means a decrease in the error of a factor of 10).

| Data need error bars. |
|---|

It must be stressed that an experimental number without an error bar is meaningless. What do you care if a survey reports 80 % of people agree with you if you cannot estimate how reliable that number is? A result of 80 % with an error bar of 4 % is much more significant than one with an error bar of 60 %. As we have seen, the way to achieve this is to have a large ***sample size***, $N$. As a rough rule of

thumb $N = 1000$ corresponds to error bars of a few percent, while samples with 10s of subjects give errors of tens of percent.

> Ex. A survey of 400 people reveals that 32% prefer vanilla ice cream. The 68% confidence interval is 24%–40%. How many people would have to be interviewed to obtain a confidence interval of 28%–36%?
>
> A. The old error bar was 8%; the new one should be 4%, which means the sample size should go up by a factor of 4, so 1600 people are required.

While a sufficiently large sample is required to draw reliable conclusions, a researcher (and people who consult studies) must also pay attention to the quality of the sample. For example a survey of the popularity of dub step among 1000 college students would yield quite different results from a survey of 1000 retired people. This is an example of selection bias.

> Ex. Your grandmother and uncle both developed bunions while wearing Converse sneakers. You do *not* conclude that wearing Converses causes bunions because you realize that this is a survey of size $N = 2$ with strong selection bias.

| Anecdote is not data. |
|---|

### 2.4.3 Hypothesis Testing

The concepts of standard deviation and confidence intervals are central to ***hypothesis testing***. Recall that an experiment is an attempt to arbitrate between competing hypotheses. It is traditional to call one of these hypotheses the ***null hypothesis***. A null hypothesis is usually a statement that the thing being studied produces no effect or makes no difference. An example is "This diet has no effect on people's weight." Normally an experimenter frames a null hypothesis with the intent of rejecting it: that is, he seeks to show that the thing under study *does* make a difference.

Possible experimental outcomes are traditionally represented in a $2 \times 2$ grid as shown in Table 2.1.

The check marks indicate correct conclusions – either rejecting a false statement or accepting a true one.[7] A type I error is also called a false positive and means that a false statement has been accepted as true. A type II error implies that a true statement has been rejected as being false.

[7] The entry "accept null hypothesis" should more properly be called "fail to reject the null hypothesis". We stick with the first because it is less wordy.

Table 2.1: Experimental outcomes.

| | Null hypothesis true (no effect) | Null hypothesis false (effect exists) |
|---|---|---|
| Reject null hypothesis | Type I error (false positive) | ✓ |
| Accept null hypothesis | ✓ | Type II error (false negative) |

Ex. Type I error: a fire alarm goes off, yet there is no fire.

Ex. Type II error: you are pregnant but your test does not turn blue.

The probability of a type I error is called alpha ($\alpha$), while the probability for a type II error is called beta ($\beta$). Clearly one wants to minimize these probabilities so that the chance of making a correct conclusion is maximized. An example will illustrate the importance of these numbers.

Ex. About 1 in 10000 people have hepatitis C. An accurate test promises a false positive rate of 1.5%. What do you tell your cousin who tests positive? The actual probability for having the disease is 0.01%, which means that the odds of the test giving a false result is 150 times higher than the odds of actually having hepatitis C. Your cousin should not worry, although seeking another test is advisable.

***Power*** is defined as the probability of rejecting the null hypothesis given that the null hypothesis is false. Since beta is the probability of accepting a null hypothesis given that it is false, we derive

$$\text{power} = 1 - \beta. \tag{2.1}$$

It is desirable to have power as close as possible to 1.0 (values around 0.8 are a typical goal). Power depends on two quantities the researcher cannot control: the size of the effect being measured and the standard deviation of the sample data. In general, the larger the effect and the lower the standard deviation, the higher the power. It also depends on two quantities the researcher can control: the sample size and the desired statistical significance of the study. The higher the sample size and the lower the statistical significance, the higher the power.

Ex. With a power of 0.7, if 10 true hypotheses are examined 3 will be incorrectly rejected.

We have introduced the idea of ***statistical significance***. Informally, this is a measure of the chance of obtaining a given effect. It can be defined formally as a

***p-value***, which is the probability of getting the result you did (or a more extreme result) given that the null hypothesis is true. This is also known as a ***significance criterion***. It is desirable to have a low p-value so that one can be reasonably sure that the null hypothesis should be rejected. In practice, the researcher selects a desired value for alpha (recall this is the probability of accepting the null hypothesis if it false), computes the p-value, and rejects the null hypothesis if the p-value is less than alpha. A typical value for alpha is 0.05.

Unfortunately the interpretation of a p-value is regularly mangled in the media and by scientists themselves. Let us state clearly:

> There is no simple relationship between a p-value and the probability of a hypothesis.

Specifically, the p-value is a probability of observing something given a hypothesis (i.e., the null hypothesis). You are *not allowed* to turn it around and say that it is related to the probability of a hypothesis given your observation. To illustrate, consider the statements (a) the probability of being a woman given that you are in the House of Representatives[8] is 18 % (b) the probability of being in the House of Representatives given that you are a woman is 18 %. Clearly (a) makes sense while (b) is nonsense[9]

Making statements about beliefs in hypotheses is not the only way things can go wrong. A list of common mistakes includes:[10]

**Mistakes with p-value**

1. The p-value is not the probability that the null hypothesis is true.
2. The p-value is not the probability that a finding is a fluke (this error is very common).
3. The p-value is not the probability of falsely rejecting the null hypothesis.
4. The p-value is not the probability that a replicating experiment would not yield the same conclusion.
5. The (1-p)-value is not the probability of the alternative hypothesis being true.

[8] As of the 113th Congress there are 79 women out of 435 representatives.

[9] There is a way to calculate (b) called Bayes' Theorem.

[10] Source: M.J. Schervish, *P Values: What They Are and What They Are Not*, The American Statistician **50**, 203–206 (1996).

6. The significance level of the test is not determined by the p-value.

7. The p-value does not indicate the size or importance of the observed effect.

---

It is common for a study to seek to find a statistically significant difference between two outcomes. For example, drug A may be more effective than drug B, or women may score higher in IQ tests than men. Some mathematics that we need not go into allow us to apply the ideas of this section to this situation.

Assume that we have two data sets, one with average $\bar{A}$ and standard deviation $\sigma_A$, and the other with average $\bar{B}$ and standard deviation $\sigma_B$. We want to know how likely it is that the actual (as opposed to the measured) averages are equal. If these values are $\bar{A} = 1$, $\sigma_A = 0.1$ and $\bar{B} = 2.8$, $\sigma_B = 0.03$, then it is very unlikely that the actual averages are equal since the measured averages differ by 1.8 and the errors are quite small (Fig. 2.7 left). Alternatively, if $\sigma_A = 0.9$ and $\sigma_B = 2.2$ then it is much more likely that the actual averages are in fact equal (Fig. 2.7 right).

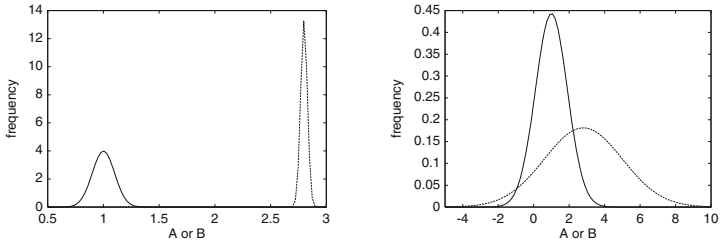


Figure 2.7: Two measured distributions.

The situation can be quantified with the following result: the probability distribution of the sum of two normal distributions is another normal distribution with mean $\bar{A} + \bar{B}$ and standard deviation $\sqrt{\sigma_A^2 + \sigma_B^2}$.[11]

Since this result applies to the difference of two normal distributions as well as to the sum, we can use it to quantify our example problems. In the first case the difference of the means is 1.8 while the new standard deviation is $\sigma = \sqrt{0.1^2 + 0.03^2} = 0.104$. In the second case the difference is still 1.8 but the standard deviation is $\sigma = \sqrt{0.9^2 + 2.2^2} = 2.38$. In the first case 1.8 is $17.3\sigma$ removed from zero, while in the second it is $0.76\sigma$ removed. Thus the probability that the first difference is consistent with zero is tiny (about $10^{-64}$), while the second is about 44 %.

[11] This result actually applies for the exact means and standard deviations. We will use it with the measured means and standard deviations, which are the best approximations we have to the (unknown) actual values.

If one assumes a null hypothesis that the two actual averages are identical, then the probability of obtaining a difference of means greater than 1.8 is $10^{-64}$ in the first case and 44 % in the second. Thus the p-value for the first experiment is tiny and for the second experiment is 0.44.

Assigning probabilities like this is called a ***hypothesis test***. More general hypothesis testing is possible. Consider, for example, the data in Fig. 2.8. The data are live birth sex odds (boys/girls) in the Russian Federation. The authors of the study from which these graphs were constructed hypothesized that the Chernobyl nuclear disaster in 1986 is the cause of the increase in the ratio seen in the figures. As proof they offered the right hand figure which shows a jump in the ratio in 1986. Their hypothesis is a step function like

$$\text{sex odds} = \begin{cases} a & \text{if year} < y \\ b & \text{if year} > y \end{cases}, \tag{2.2}$$

which is a three parameter model $(a, b, y)$ of the data. A fit yields $y = 1986.9 \pm 0.4$, which agrees with their contention about Chernobyl.

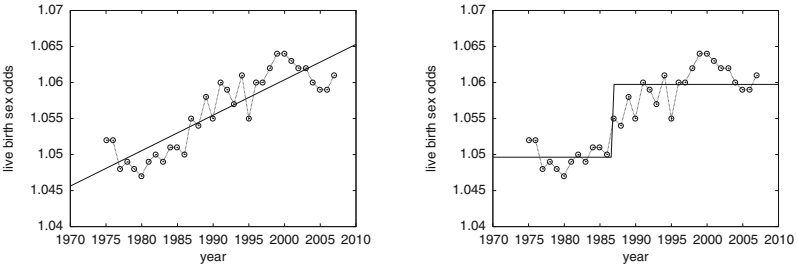


Figure 2.8: Live birth sex odds in the Russian Federation with two model fits.
Source: H. Scherb and K. Voigt, Environmental Science and Pollution Research, **18**, 697–707 (2011).

But there are problems: (i) the data have no error bars so it is difficult to see if the rise in sex odds is due to chance, (ii) even if there is a jump in 1986, there is no reason to associate this with Chernobyl; after all this was the year that *Walk like an Egyptian* was released by the Bangles, (iii) one should compare the model to alternative models to assess how reliable it is. To illustrate the last point I fitted the data to a straight line (left figure). The resulting quality of fit is nearly identical to the step function, and is achieved with one fewer adjustable parameter. Although there is an interesting question about the change seen in the sex odds, one cannot conclude that the Chernobyl disaster has anything to do with it.

Figure 2.9 illustrates another way in which data analysis can go wrong. In this case, the data have error bars and are fit nicely by a straight line. But something is wrong. Recall that an error bar usually represents the one-$\sigma$ variation in the data. Thus if the data really do follow the line indicated in the figure, 32 % of the data

points should lie more than an error bar away from the line. But *all* of the data lie closer (substantially closer) to the line than their error bars warrant.

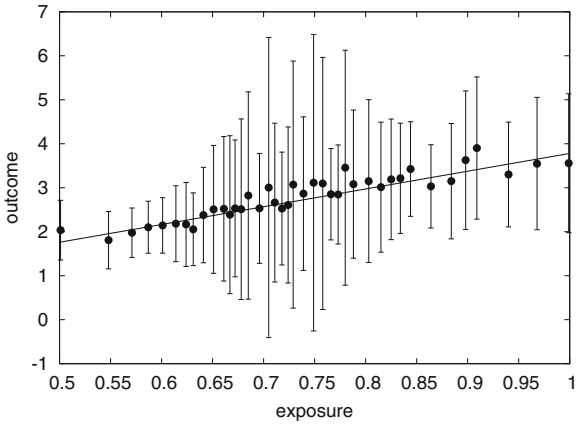


Figure 2.9: Outcome vs. exposure in a study.

### 2.4.4 The Statistics of Bias

In the past two sections we have examined statistical methods as applied to studies and a rather long list of the ways studies can go wrong. In this section we will use statistics to gain a quantitative understanding of how studies fail.

The simplest statistical concern is a sample size that is too small. The sample size refers to the number of subjects in the study – too few subjects means that conclusions are not statistically significant (we have already seen this in the discussion of the p-value). This is a serious, but common, problem, typically brought on by practical constraints, such as lack of funding or time.

The subtlest issue concerns psychological pressures on researchers. Scientists are under institutional and personal pressure to make important discoveries. When this is combined with financial pressure to publish frequently, it is easy for biases to subvert the accuracy of a study.

> "Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias."
>
> — John Ioannides.

John Ioannides (1965–), who studies clinical trial methodology has noted the following general features that carry negative impact on study fidelity[12]:

1. The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
2. The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
3. The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
4. The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
5. The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.
6. The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Much of what goes wrong is a simple consequence of false positives. To see this, consider a list of 1000 hypotheses that are deemed interesting enough to test. We will assume that 100 of these are actually true. We also assume that alpha = 0.05 and beta = 0.2. Thus when the 1000 studies are completed there will be $0.05 \times 900 = 45$ false positive results and $0.2 \times 100 = 20$ false negative results. This leaves 80 true hypotheses that have been confirmed (the power is 0.8). The net result is that the researcher believes she has $45 + 80 = 125$ true results, but $45/125 = 36\,\%$ of these are in fact false. Although the false positive probability is relatively small, it still skews the understanding of the scientific landscape because so few of the hypotheses are actually true (or so many are actually false).

Yet another simple statistical effect can confound the interpretation of large trials. Consider a study with alpha set to 0.05, so that the researcher seeks a p-value of 0.05 or less. Assume further that no effect exists, so that the null hypothesis is true. Since the p-value is the probability of obtaining the experimental result if the null hypothesis is true, on average one experiment in 20 will *not* find the result; i.e., will reject the null hypothesis and claim a discovery. This is called the ***multiple comparisons problem*** or the ***look elsewhere effect***.

| If you look often enough you will find something. |
|---|

Quite simply, if an experiment examines enough different outcomes for a given exposure, it is bound to observe a statistical fluctuation that will be interpreted

[12] PLoS Medicine, **2**, e124 (2005).

as a real rejection of the null hypothesis. The physicist Robert Adair (1924 –) has called studies that examine many possible outcomes ***hypothesis generating experiments*** because they can used to identify outcomes to be tested with subsequent studies. They cannot be used to claim real outcomes because of the look elsewhere effect.

> Ex. A 1992 Swedish study tried to determine whether power lines caused poor health effects. The researchers surveyed everyone living within 300 meters of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of over 800 (!) ailments. The study found that the incidence of childhood leukemia was four times higher among those that lived closest to the power lines, and it spurred calls to action by the Swedish government. The problem with the conclusion, however, was that they failed to compensate for the look-elsewhere effect; in any collection of 800 random samples, it is likely that at least one will be at least 3 standard deviations above the expected value, by chance alone. Subsequent studies failed to show any links between power lines and childhood leukemia.

A careful researcher can account for the look elsewhere effect. A simple, but approximate, way to do this is to replace the condition for significance, $p < \alpha$ with $p < \alpha/n$, where $n$ is the number of tested outcomes. This will make a dramatic difference in large-scale studies such as the Swedish study.

Notice that the look elsewhere effect applies equally well to a great many studies that consider the same exposure. Thus 800 studies with $\alpha = 0.05$ will find (on average) 40 false results just as surely as one large study. A current popular health concern and subject of many studies is the effect of electromagnetic radiation on health.[13] The look elsewhere effect *requires* that some of these studies will find an effect. Of course, subsequent studies (if they exist) will not find the same health outcome, but will "discover" a different, and random, outcome. How do you think that these findings will be reported in the media?

> Ex. In 2007 the BBC reported that the president of Lakehead University refused to install wifi on campus because he believes that "microwave radiation in the frequency range of wi-fi has been shown to increase permeability of the blood-brain barrier, cause behavioral changes, alter cognitive functions, activate a stress response, interfere with brain waves, cell growth, cell communication, calcium ion balance, etc., and cause single and double strand DNA breaks."

> Ex. A web site devoted to spreading alarm about electromagnetic radiation reports, "dozens of published papers have found links between living near power line electromagnetic radiation and a range of

[13] This topic will be revisited in Chap. 5.

> health woes including brain cancer and leukemia, breast cancer, birth defects and reproductive problems, decreased libido, fatigue, depression, blood diseases, hormonal imbalances, heart disease, sleeping disorders, and many others."

Yes, the media report *all* the random and false discoveries as real. Perhaps it goes without saying that when a single exposure can lead to so many deleterious outcomes one should immediately suspect the look elsewhere effect.

The net effect of all of these issues can be studied by examining published results and seeing how often they are confirmed by follow-up studies. The results are not encouraging. When the pharmaceutical firm Amgen attempted to replicate 53 landmark studies it was only able to reproduce 6 of them. Similarly, researchers at Bayer HealthCare were only able to reproduce one quarter of 67 seminal studies.

A less direct way to test study reliability is to check results as a function of who obtains them. The following table summarizes the results of trials testing the efficacy of acupuncture by country. It is not sensible that trials in North America find a favorable effect in 49 % of studies, while those in Asia find a favorable result 100 % of the time. Some difference, presumably cultural and historical, must be leading to significant bias in the Asian or North American trials (Table 2.2).

Table 2.2: Controlled clinical trials of acupuncture by country of research.

| Country | Trials | Trials favoring |
|---|---|---|
| USA | 47 | 25 |
| Canada | 11 | 3 |
| China | 36 | 36 |
| Taiwan | 6 | 6 |
| Japan | 5 | 5 |
| Hong Kong | 3 | 3 |

Source: A. Vickers *et al.*, Controlled Clinical Trials **19** 159 (1998).

Finally, studies reveal that 80 % of nonrandomized studies turn out to be wrong, along with 25 % of randomized studies, and 10 % of large scale randomized studies. These are not encouraging numbers and indicate the power of the forces arrayed against producing quality research and the difficulty in teasing results out of extraordinarily complex systems such as the human body.

## 2.5 Improving Study Reliability

Increasing awareness of the problem of unreliable studies has led to several efforts to improve methodology. One outcome of this effort was the CONSORT statement of minimum requirements for reporting randomized trials that was for-

mulated by an international group of medical journal editors, clinical researchers, and epidemiologists. The statement provides a standard way for authors to prepare reports of trial findings, facilitates complete and transparent reporting, and aids critical appraisal. You can find out more at `www.consort-statement.org`.

In 2000 the US National Institutes of Health (NIH) instituted a web site called `ClinicalTrials.gov` for tracking publicly funded clinical studies. The site is a Web-based resource that provides patients, their family members, health care professionals, researchers, and the public with access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions. Of course, it also addresses issues with publication bias and salami slicing.

The CONSORT and NIH web sites are tools meant for professionals. Fortunately, the *Cochrane Collaboration* was created to look after the rest of us. The collaboration is a non-profit global organization of independent health practitioners, researchers, and patient advocates that is dedicated to producing credible, accessible health care information that is free from conflict of interest. The chief product is a series of systematic reviews that address specific health care questions. These reviews are available online at `summaries.cochrane.org`.

> Ex: Entering "fish oils for the prevention of dementia in the elderly" into the Cochrane Summary search field yields a report with the following statement.
>
> "The results of the available studies show no benefit for cognitive function with omega-3 PUFA supplementation among cognitively healthy older people."

## REVIEW

**Important terminology:**

biases: publication, recall, citation, sampling, procedural [pg. 31]

blinding [pg. 32]

control group [pg. 31]

study types: cohort, randomized, observational, case controlled [pg. 31]

exposure and outcome [pg. 28]

mean and standard deviation [pg. 35]

placebo [pg. 31]

type I and type II errors [pg. 40]

alpha: the probability of rejecting the null hypothesis given that it is true. [pg. 40]

beta: the probability of accepting the null hypothesis given that it is false. [pg. 40]

p-value: the probability of getting the results found given that the null hypothesis is true. [pg. 41]

power: 1 - beta, or the probability of rejecting the null hypothesis given that it is false. [pg. 40]

**Important concepts:**

Correlation is not causation.

Anecdote is not data.

The Jadad scale.

The Bradford Hill criteria.

The Central Limit Theorem.

An experiment is a controlled, reproducible examination of nature intended to arbitrate between competing hypotheses.

The normal distribution is common because it is equivalent to the sum of many random variables.

Error bars permit assessing the reliability of conclusions.

There is no simple relationship between a p-value and the probability of a hypothesis.

When many hypotheses are false, truth can be overwhelmed by false positives.

The look elsewhere effect can give rise to false positives.

## FURTHER READING

Imogene Evans, Hazel Thornton, Iain Chalmers, and Paul Glasziou, *Testing Treatments*, Pinter and Martin, 2011.

Ben Goldacre, *Bad Science*, Faber and Faber, 2010.

## EXERCISES

1. It is common to hear that children at a birthday party are running wild because they are on a "sugar high". Suggest a causal mechanism for this. Suggest a noncausal mechanism. Suggest a way to test both ideas.

2. It is common to hear that someone caught a cold because they were cold. Suggest a causal mechanism for this. Suggest a noncausal mechanism.

3. In an effort to examine whether exposure to electromagnetic radiation is associated with cancer, a study examined cancer rates in power line workers and found significantly higher incidence of skin cancer. Suggest a causal link for this correlation. Suggest a noncausal link.

4. You wish to test the hypothesis that a coin is fair (i.e., the odds of coming up heads is ½). You flip the coin six times and obtain 5 heads.

   (a) Compute the probability of obtaining at least 5 heads.

   (b) Take the null hypothesis to be that the coin is fair. If your test criterion is alpha < 0.05, do you accept or reject the null hypothesis?

   (c) What p-value would you assign to the statement that the coin is fair?

5. Re-read the quotation at the beginning of this chapter. What point is King James making?

6. Suspicious Data.

   Look at Fig. 2.9 again. If the error bars represent a 90 % confidence interval, how many points do you expect (on average) to lie further from the fit line than their error bar?

7. A sample of crime-scene DNA is compared against a database of 10,000 people. A match is found and the accused person is brought to trial where it is stated that the odds of two DNA samples match is 1 in 5000. The prosecutor, judge, and jury all interpret this to means the odds the suspect is guilty is 4999 out of 5000. What do you say?

8. In 2006 the Times of London reported that "Cocaine floods the playground". The story noted that a government school-yard survey found that cocaine use in London schools had risen from 1 % in 2004 to 2 % in 2005, which they reported as "cocaine use doubles". The survey asked school children about their use of dozens of illegal substances. Why didn't government statisticians break this story?

9. To study the dangers of cellphones, researchers question 100 people with brain cancer to determine their rate of cellphone usage. What type of study is this?

10. Patients in a case controlled study on stomach cancer and antacid consumption are asked how many antacids they eat on average per month. What type of bias does this introduce to the study?

11. An experiment claims to find an effect with a p value of 0.04. If the experiment is repeated 100 times, about how many times will the effect not be seen?

12. A December 13, 2011 New York Times article, "Tantalizing Hints but No Direct Proof in Particle Search", reported

    > "The Atlas result has a chance of less than one part in 5000 of being due to lucky background noise, which is impressive but far short of the standard for a 'discovery', which requires one in 3.5 million odds of being a random fluctuation."

    What is wrong with this statement?

13. Combining Results. A manufacturer says that the length of his widgets is 1.3 m with an error of 1 mm, as measured by a sample of size $N = 10{,}000$. You buy 1000 widgets and measure an average length of 1.32 m with an error of 3 mm. Do you believe the manufacturer?

14. Consider the 1000 hypotheses scenario of Sect. 2.4.4 again, but this time assume alpha $= 0.05$ and beta $= 0.79$ (this corresponds to a power of 0.21, which is typical of neuroscience studies). What fraction of "true" hypotheses are actually true?

15. Height.

    You are 73 inches tall. What percentage of American men are taller than you?

16. Identify the types of studies mentioned in Sect. 2.1.

    (a) Lind's scurvy study

    (b) Snow's cholera study

    (c) Semmelweis's childbed fever study.

17. Mammograms and Cancer.

    Consider the following $2 \times 2$ table for a mammograms and cancer

    (a) What is the null hypothesis?

| | Cancer (1 %) | No cancer (99 %) |
|---|---|---|
| Test pos | 80 % | 9.6 % |
| Test neg | 20 % | 90.4 % |

(b) What are alpha and beta?

(c) Assume 1 % of people actually have cancer, what is the probability that you have cancer if you get a positive test result?

18. Weight distribution.

Look again at the distribution of weights in Fig. 2.3. It does not look very much like a normal distribution. Come up with possible reasons for this.

19. Scientists conducting the European Union INTERPHONE study of cancer and cellphone use disagreed about the validity of their study because patients were asked about their typical cellphone usage. What were the researchers worried about? How could the problem be circumvented?

20. Daniel in Babylon.

Read Daniel 1.5–15. What Jadad score do you give Daniel's study?

21. The following quotation is from an article by Jonah Lehrer which was published in the Dec 16, 2011 issue of *Wired*. Comment on his observations in light of what you have read in this chapter.

"When doctors began encountering a surge in patients with lower back pain in the mid-20th century, as I reported for my 2009 book *How We Decide*, they had few explanations. The lower back is an exquisitely complicated area of the body, full of small bones, ligaments, spinal discs, and minor muscles. Then there's the spinal cord itself, a thick cable of nerves that can be easily disturbed. There are so many moving parts in the back that doctors had difficulty figuring out what, exactly, was causing a person's pain. As a result, patients were typically sent home with a prescription for bed rest."

"This treatment plan, though simple, was still extremely effective. Even when nothing was done to the lower back, about 90 percent of people with back pain got better within six weeks. The body healed itself, the inflammation subsided, the nerve relaxed."

"Over the next few decades, this hands-off approach to back pain remained the standard medical treatment. That all changed, however, with the introduction of magnetic resonance imaging in the late 1970s. These diagnostic machines use powerful magnets to generate stunningly detailed images of the body's interior. Within a few years, the MRI machine became a crucial diagnostic tool."

"The view afforded by MRI led to a new causal story: Back pain was the result of abnormalities in the spinal discs, those supple buffers between the vertebrae. The MRIs certainly supplied bleak evidence: Back pain was strongly correlated with seriously degenerated discs, which were in turn thought to cause inflammation of the local nerves. Consequently, doctors began administering epidurals to quiet the pain, and if it persisted they would surgically remove the damaged disc tissue."

"But the vivid images were misleading. It turns out that disc abnormalities are typically not the cause of chronic back pain. The presence of such abnormalities is just as likely to be correlated with the absence of back problems, as a 1994 study published in The New England Journal of Medicine showed. The researchers imaged the spinal regions of 98 people with no back pain. The results were shocking: Two-thirds of normal patients exhibited 'serious problems' like bulging or protruding tissue. In 38 percent of these patients, the MRI revealed multiple damaged discs. Nevertheless, none of these people were in pain. The study concluded that, in most cases, 'the discovery of a bulge or protrusion on an MRI scan in a patient with low back pain may frequently be coincidental.'"

http://www.springer.com/978-3-319-21987-5

Science and Society
Understanding Scientific Methodology, Energy, Climate, and Sustainability
Swanson, E.S.
2016, XXIX, 276 p. 93 illus., 39 illus. in color.,
ISBN: 978-3-319-21987-5