Boston Computational Immunology 2018 Summer School
Notes on foundations
T.B. Kepler 11 June 2018

What are we trying to accomplish

The successful scientist has expertise of two different types: *judgment* and *technique*. *What* should you and *how* do you do it. This true in the context of data analysis as well as in the laboratory.

Statistics is an aid to judgment, but not a replacement for it. Statistical testing does not prove or disprove hypotheses. Statistical methods provide you with an unbiased and objective means to form your own judgments.

My position in a nutshell


Figure 1.Malcolm Rorty

Malcolm Rorty graduated with an engineering degree from Cornelll in 1896 and began his long career with Bell Telephone as a telephone installer. After the war in 1918 he was made *chief statistician* for Bell[1]. That year, the first U.S. academic department of statistics was started at Johns Hopkins. Rorty became the President of the American Statistical Association in 1931. In his inaugural address to that body in 1930, he reflected on the challenge he faced putting together a staff of statisticians before statistics was an established academic discipline in the US. He wrote [1],

> …it was of small importance whether a candidate could, or could not, work out a coefficient of correlation, provided he knew instinctively when to distrust one. And in the end we found ourselves placing technically trained statisticians, mathematicians, and actuaries under the supervision of skilled logicians-of men who might have to learn the details of statistical technique, but who started with an instinctive knowledge of the scientific method as a whole.
>
> This, then, was our pragmatic answer to the question, "What is a statistician?"-the statistician must be instinctively and primarily a logician and a scientist in the broader sense, and only secondarily a user of the specialized statistical techniques. The statistician who knew only statistics was a danger and a blight. It was less important to know how to use statistics, than it was to know how and when not to use them.
>
> The statistical work in astronomy must be done primarily by astronomers, and in biology by biologists, and in each science by those who are trained primarily in that science and only secondarily in the statistical method.

Although a great deal has changed since that time, including the establishment and growth of statistics as an academic discipline, a great deal still remains the same. You are still best equipped to analyze your own data. You may recognize that there are aspects of the process that exceed your present technical skill, or for which you require additional consultation to inform your judgment. But you are the

[1] https://en.wikipedia.org/wiki/Malcolm_C._Rorty

investigator. This class is intended to allow you to acquire additional technical skills as required and to effectively weigh advice from professional data analysts in your decision making.

The question for us at this stage is: How does science work? How do we go from observations to propositions in science?

In daily life there is no problem. I can point to a swan and say, "that swan is white".
But how would I ever get to the "law" all swans are white.

Proof in Mathematics

A simple proof in mathematics takes the form: for all integers a, b, and c

$$a+b=a+c \Rightarrow b=c \tag{1.1}$$

This is one of the basic facts of arithmetic we learn early on. It's proof is not trivial (we can't use subtraction or negative numbers yet).

But that is not how scientific reasoning works. In fact, there is no "proof" at all in science.

Aristotle's syllogism

The logical form of the syllogism takes the form of

All swans are white
Zoe is a swan
Therefore: Zoe is white

This is a valid argument

How about

Zoe is a swan
Zoe is white
Therefore: all swans are white.

Is this a valid argument?

These 50 animals are all swans
All of them are white
Therefore: all swans are white

Is this any better?

How about this one:

We have studied these n individual bacteria
All of these n bacteria encode a polymerase
Therefore: All bacteria encode a polymerase

This is still not valid. Fortunately, that is not how science works.

Instead, the community of scientists constructs a shared *model* of the world. In the current model, bacteria cannot replicate their genome without a polymerase. Bacteria that cannot replicate cannot survive. Therefore, all extant bacteria must have a polymerase. If we found a microbe that did not have a polymerase, we would want to know if it really is a bacterium before we abandon our model.

Each experiment invites us to change our model of the world, sometimes filling in spots of uncertainty, and sometimes requiring large-scale reorganization.

150 years ago or so, Pasteur hypothesized that microbes cause furuncles (skin abcesses) (Pasteur L (1880). On the extension of the germ theory to the etiology of certain common diseases.)

Expectation given the hypothesis: If microbes cause furuncles, then if one extracts material from the furuncle, microbes will be found in it. Furthermore, if one extracts material and does not find microbes, the hypothesis will be weakened. (Why not disproven?)

Pasteur examined 5 furuncles from 3 people and cultured microorganisms from all of them. He thus concluded, “it appears certain that every furuncle contains an aerobic microscopic parasite, to which is due the local inflammation and the pus formation that follows.

What do you make of this claim?

A thought experiment.

Close your eyes and imagine coming home to your place of residence. Open the door and walk inside. Can you describe what you see? Walk to the refrigerator. Open the door. What is the first thing you see inside? Is there an egg in the refrigerator? If so, take one out and hold it in your hand. Now loosen your grip and let the egg fall to the ground. What happens? Can you see the result?

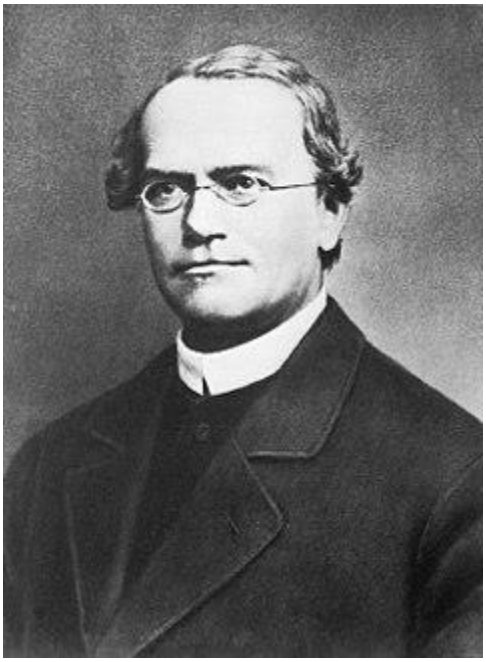
Were you able to carry out this exercise? Were you able to “see” the room, the refrigerator, the egg? Were you able to experience the sensation of holding the egg in your hand? Were you able to “predict” the result of dropping the egg? How were these things made possible? This happens because you have a cognitive model of the world that allows you to predict the results of your actions. You always experience the world in the context of your cognitive model. The shared scientific model is of just this type. We often think of scientific knowledge as comprising a large set of facts, including laws. But these propositions do not constitute scientific knowledge, they describe it. Scientific knowledge as kind of shared, or social cognitive model, and a cognitive model cannot be entirely represented by propositions.

A model is the means by which we predict what is going to happen. We feed a scenario into our model, and ask what will happen.

We test models by running experiments and judging whether they effectively predict the outcome. But no single test can establish the validity of a model in one fell swoop. There are no proofs in science. In the early days of quantum mechanics, Louis Debroglie conceived a mental model in which electrons behaved like waves. He devised an experiment to test his model. His model passed the test. But many people were still not convinced. Fast-forward 93 years and 95% of American adults (and 100% of Americans aged 18-29) own a mobile phone. These devices we designed and built under the assumption that electrons behave like waves. Now there is very little dissent.

We will move on now to examine the paper in which Gregor Mendel proposed and tested his model for inheritance of simple traits.

Mendel's Hybridization Experiments



Gregor Mendel was a Moravian monk and scientist who carried out a set of experiments on plant hybridization using the garden pea, *Pisum sativum*. He published the paper reporting this work in 1866 [2]. This paper is now regarded as establishing the foundation of modern genetics. We are going to examine the way Mendel used experimentation to establish new knowledge. We have chosen this paper because its conclusions have proven durable and provide the firm basis for enormous numbers of applications. The experiments are straightforward, and easy to understand.

In this paper, Mendel proposed a quantitative model for inheritance under plant hybridization that explained several known phenomena and he designed and carried out years-long experiments to test his model.

The main point of this exercise is to dissect the paper to reveal the key concepts that underlie the logic of experimentation. These will turn out to be the key concepts underlying experimental statistics, too. Understanding them in a more familiar, non-mathematical context will smooth the way to understanding the statistics that embodies them.

The primary concepts the understanding of which we want to reinforce are

- Phenomena to be explained
- Hypothesis
- Experiment
- Hypothesis Test
- Interpretation

The phenomena to be explained

The community of plant hybridists included the horticulturalists, whose interest was in commercial application, and the scientists who were interested in a variety of theoretical questions. Both groups were interested in understanding or controlling the *characters* that are passed from one generation to the next. Some characters, such as color or flavor, were of direct interest to horticulturalists. Scientists, on the other hand, took the term more abstractly to refer to any minimal observable characteristic by which to strains could be distinguished. Mendel used seed color, seed morphology, and five other characters.

Mendel encapsulates several phenomena, recognized by all people who have studied plant hybridization or engage in hybridization for practical purposes, by referring to

- "The striking regularity with which the same hybrid forms always reappeared whenever fertilization took place between the same species".

Preceding investigators cited, more specifically [3],

- Many characters that differ in the parents are not blended in the offspring, but retain one or the other of the two parental types.
- The distinguishing character in first generation of hybrids all resemble that of one of the two parents.
- It does not matter which parent, male or female, exhibits the dominant character. The results either way are indistinguishable.
- The second and subsequent generations produce plants exhibiting a mixture of parental characters.

Mendel does not list all of these phenomena explicitly, but assumes the state of the field is known to his readers. He goes on to remark on the proposition that "no generally applicable law governing the formation and development of hybrids has been successfully formulated." Thus he sets out his primary aim: to formulate a set of propositions through which the relevant phenomena can be understood and predicted.

Hypothesis

Mendel does not state his hypotheses explicitly before he reports his results. But we can construct the hypotheses that his model implies. His model is that there are characters
The experimental set-up

Offspring resulting from crosses between distinct strains were classified into three categories. The third category contained those crosses whose progeny resembled one or the other of the parents rather than being intermediate in form. It was known that the first generation reliably resembled one or the other of the parents, but in subsequent generations, both types could be found [Naudin, 1864 as cited in Roberts]

In this paper, Mendel advances a *model* of inheritance that he goes on to *test* using several strains of garden pea.

The notion of model is ubiquitous is science and appears in many places, under slightly different guises. Its basic meaning is that it's *a thing that represents another thing*. Usually, it is used to represent another thing because it is more convenient for its purpose than is the thing represented. In biology, we have model organisms, which stand in for whole classes of organisms. *D. melanogaster* is a model for organisms with genes, *C. elegans* for organisms that undergo development. In biomedicine, we have animal models for human disease, such as the EAE mouse for human multiple sclerosis. Here, it refers first to a mental construct that may be used to explain and predict the phenomenon under investigation. In some cases, the mental model can be made more precise by employing a mathematical model to express part of the idea. Mendel uses both. We will introduce a further meaning of model, that of a measurement model, later on.

The model he proposes is that for the dichotomous traits of interest, there are two latent determinants. He is distinguishing between the evident trait and the unobserved factor that gives rise to the trait. He is drawing a distinction between the *phenotype* and the *genotype*. Where there a given trait is dichotomous, that is, it exhibits two phenotypes, Mendel suggests that there are actually three

genotypes. So if the phenotypes are D and R, he proposes that the three underlying genotypes can be labeled dd, dr, and rr.

He treats these individual determinants as independent elements under sexual reproduction. In other words, any offspring gets one determinant from each parent. If the parental genotype is dr, either determinant is equally likely to be passed to any given offspring. Furthermore, the two parents are indistinguishable in terms of their contribution to the progeny genotype: "it is perfectly immaterial whether the dominant character belongs to the seed plant or to the pollen plant; the form of the hybrid remains identical in both cases." (p9)

According to this simple model, the genotypic unions produce genotypic progeny according to
dd x dd -> dd
dd x rd -> ½ rd + ½ dd
dd x rr -> rd
rd x rd -> ¼ dd + ½ rd + ¼ rr
rd x rr -> ½ rr + ½ rd
rr x rr -> rr

The second line indicates that the progeny of a dd and rd union is **equally likely** to be rd or dd.

Finally, he suggests that one of the two determinants dominates the other. That is, when both are present, the trait is identical to that of one of the two parental types. dd and dr both underlie D, and rr underlies R.

He then tests this model by carrying out several experiments.

First, Mendel tests several different "pure lines", finding those that consistently produce plants with a given trait. Within the context of his model, he says that these are homogeneous in their determinants. *We would say they are homozygous.* He is thus setting up the experiment by preparing plants to be homozygous in the loci of interest. According to his model, one line has genotype aa (and phenotype A), ond the other has genotype bb (and phenotype B).

Note the way he describes these results on page 5 (emphasis added):

> All the other varieties yielded perfectly constant and similar offspring; at any rate, no *essential* difference was observed during two trial years.

He is not saying that no differences were observed, but that those that were observed were not essential differences. It is not clear how one tells which differences are essential and which are not.

Next, he crosses these lines and observes that only one phenotype is produced. This allows him to determine which trait, A or B, is dominant. He next ("the first generation from the hybrids") self-fertilizes the first-generation hybrids and counts the numbers of progeny of each phenotype (Table 1).

Table 1. Summary of results from the first generation from the hybrids.

experiment	phenotype 1	phenotype 2	np1	np2
Expt 1	roundish seeds	wrinkled seeds	5474	1850
Expt 2	yellow seeds	green seeds	6022	2001
Expt 3	violet-red flowers	white flower	705	224
Expt 4	simply inflated pods	constricted pods	882	299
Expt 5	green pods	yellow pods	428	152
Expt 6	axial inflor.	radial inflor.	651	207
Expt 7	long stem	short stem	787	277

Mendel knows that his model predicts that in some sense, the proportion of progeny with the dominant phenotype will be ¾, and is satisfied that his data are consistent with that prediction. But we are going to go one step further. We are going to try to be more precise in our understanding of what his model predicts.

Experiments 1 and 2 involve counting seeds from individual plants. Mendel does not give an exhaustive tabulation of the more than 500 plants in the two experiments, but he does give results from ten plants from each.

Table 2. Excerpt from experiment 1.

Plant	round	wrinkled
1	45	12
2	27	8
3	24	7
4	19	10
5	32	11
6	26	6
7	88	24
8	22	10
9	28	6
10	25	7

Table 3. Excerpt from experiment 2.

Plant	yellow	green
1	25	11
2	32	7
3	14	5
4	70	27
5	24	13
6	20	6
7	32	13
8	44	9
9	50	14
10	44	18

In the second experiment he takes individuals from the second generation from the hybrids (the progeny counted in Table 1), and allows them to self-fertilize. He counts progeny and classifies the parent according to the progeny in produces. He finds, first of all, that all plants with the recessive phenotype give rise exclusively to recessive-phenotype progeny.

Among F2 plants with a dominant phenotype, he finds that some give rise to dominant progeny only, and others give rise to both dominant and recessive progeny, and do in a 3:1 ratio. He thus classifies each F2 dominant plant according to whether it gives rise to dominant only or mixed progeny (Table 4).

experiment	phenotype 1	dominant	mixed	no.plants
Expt 1	roundish	193	372	565
Expt 2	yellow	166	353	519
Expt 3	violet-red	36	64	100
Expt 4	simply inflated	29	71	100
Expt 5	green	40	60	100
Expt 6	axial	33	67	100
Expt 7	long	28	72	100
Expt 5 repeat.	green	35	65	100

Probability

We use probabilities to represent the degree of certainty in the occurrence of a particular experimental outcome. Oddly, perhaps, there is no universal agreement on just what that means.

Suppose we have an experimental protocol, and intend to carry out the experiment with our system of choice. Although we may not explicitly write one down, we have a conceptual model that represents our understanding of the system being studied and provides a means to predict the outcome of the experiment. The model typically comprises a model of the system as understood prior to the experiment tentatively enlarged to encompass an experimental hypothesis.

For Mendel, the base model represented a partial understanding of heredity and knowledge about the various phenomena observed in plant breeding. The model extension represented his conception of the mechanism of heredity, with abstract carriers of trait information combined between the two parents and transferred to the offspring. The base model played a crucial role in the design of the experiment, and the hypothetical extension of that model gave him predictions of what the experimental outcome would be if the model were true.

Mendel was able to enumerate all the possible outcomes of the experiment: no smooth peas out of all n peas, one smooth out of n, and so on. But his model did not enable his predicting a sharp outcome. He could only describe what the most likely outcomes would be. This is where the idea of probability arises, although he did not describe it in modern terms. He seems to have thought that his model predicted that the observed proportion of smooth peas would be close to ¾, but he couldn't say just how close, and certainly not how close how often or with what *probability*.

Nowadays, we like to proceed by taking a probabilistic model to represent the generation of the experimental outcome, assigning a probability to each of the possible outcomes. In the case of Mendel's pea hybridization experiments, we take the *binomial* model as representing our conceptual model of the generation of the experimental data. The binomial model itself is based on a submodel—the Bernoulli model—for each of n individual observations, each of which has just two possible mini-outcomes. For Mendel, these would be smooth or wrinkled, yellow or green, and so on depending on the specific experiment. Each of these binary outcomes is given a probability p by the Bernoulli model. If the individual observations all have the same probability and are independent of each other, the Binomial model holds and provides the probability of each of the possible n + 1 outcomes.

But we still don't know what this probability *is*. The dominant definition of probability among scientists and other professional users of statistics, is that probability is the relative frequency of an outcome as more and more *replicate* experiments are done. We don't actually do all those experiments, but in doing the one experiment we will do, we are obliged to consider what other outcomes we might have obtained and place ours within the context of all these others. This perspective gives rise to hypothesis testing and confidence interval estimation.

Interlude with R

The frequentist interpretation is intended to make probability an objective property of an experimental system. The Bayesian interpretation emphasizes probability as related

Suppose your are betting on the Belmont stakes, where Justify's odds are 4-5. These odds indicate that a bet of $5 pays a profit of $4, or a total of $9. This is the equivalent of giving Justify a 4/9 probability of winning.

If you bet the $4 and the horse wins, you get $9 back. If the horse loses, you get nothing back. So your expected winnings are -$4 + p*$9 = -$4 + 4/9 * $9 = 0. This is if you accept the probability given by the odds maker. If your own subjective probability is higher, you rate the expected winning as higher as well, and you will be willing to take the bet. This is the approach to probability that Bayesian statistics uses.

The Bayesian makes explicit use of conditional probability and the likelihood. The probability of getting x smooth peas out of n peas total depends (is contingent on) the probability that any given pea is smooth. This conditional probability is written $P\left(x \mid p,n\right)$. The Bayseian is perfectly happy flipping this around and asking, what is the probability that the Bernoulli parameter is p, given that we found x peas in our experiment. In order to do this, we have to have some subjective probability for every possible value of p. We call this the prior probability and denote it $\pi\left(p\right)$. Then Bayes' rule reads

$$P\left(p \mid x,n\right) = \frac{P\left(x \mid p,n\right)\pi\left(p\right)}{\int_0^1 dp' P\left(x \mid p',n\right)\pi\left(p'\right)} \tag{1.2}$$

Back to R

1. Rorty MC. Statistics and the Scientific Method. Journal of the American Statistical Association. 1931;26(173):1-10. doi: 10.2307/2278253.

2. Mendel G. Experiments in plant hybridization (1865). Verhandlungen des naturforschenden Vereins Brünn) Available online: www mendelweb org/Mendel html (accessed on 1 January 2013). 1996.
3. Roberts HF. Plant Hybridization before Mendel. Princeton, NJ: Princeton University Press; 1929.