Model Based Statistics in Biology.

Part II. Quantifying Uncertainty and Evidence.

Chapter 7 Statistical Inference

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops, which combined
models (what is the relation of scallop density to substrate?)
with statistics (how certain can we be?)

ReCap (Ch5)
Data equations partition variability into a model component and a random component.
Data equations apply to regression lines and to comparison of groups.
The sum of the squared residuals allows us to compare one model to another.
It allows us to quantify the improvement in fit, a key concept in statistics.

ReCap (Ch 6)
Frequency distributions are a key concept in statistics.
They are used to quantify uncertainty.
Empirical distributions are constructed from data
Theoretical distributions are models of data.

Today: Statistical inference.

Background

The word Probability derives from the latin term *probabilis* meaning "plausible" or generally approved." Mathematical treatment of probability begins with correspondence (1654) between Blaise Pascal (1623 - 1662) and Pierre de Fermat (1607 - 1665) concerning how to divide winnings in a game of chance. Jacob Bernoulli (*Ars Conjectandi* 1713) and Abraham de Moivre (*The Doctrine of Chance: or, a Method for Calculating the Probabilities of Events in Play* 1718) put probability on a sound mathematical foundation. For example de Moivre (p10) asked, concerning dice, what is the "Probability of drawing an Ace in three throws?" de Moivre (p10) calculates it at 91/216.

The word Statistics ultimately derives from the latin term *statisticum collegium* ("council of state"). The German Statistik, first introduced by Gottfried Achenwall (1749), originally designated the analysis of data about the state. It was introduced into English by Sir John Sinclair (1754 - 1835) who supervised the compilation of the Statistical Account of Scotland (21 vols., 1791-1799). Thus, statistics began as data to be used by governmental and (often centralized) administrative bodies. The use of the word broadened in the late 19th and early 20th century with the increasing use of inferential statistics founded on probability theory.

Descriptive statistics describe or summarize the observed measurements of a system.

Examples: Location: mean, median, mode
Dispersion: range, quartiles, standard deviations
coefficient of dispersion, coefficient of variation
Symmetry: quartiles, Pearson's moment coefficient

Inferential statistics are used to infer, predict, or forecast future outcomes, tendencies, and behaviors of a system, based on evidence from samples.

Inferential statistics began with Bayes (1763) and Laplace (1774) who independently discovered the use of conditional probability to make inferences from data. Bayes produced a rule for putting probabilistic limits on a single event. Laplace showed how to put a probability on an event given data and equally probable causes. Laplace (1786, 1812) then used the central limit theorem to produce a principle for inferring a probability from data and the sum of all causes (Stigler 1986, p 136). Laplace's Principle VI is now called Bayes' theorem, even though it does not occur in Bayes publication (Dale 1982).

Defining and drawing the sample. The sampling process (Yule and Kendall 1950 14th edition) consists of defining the unit, drawing samples, and applying a measurement protocol to the samples.

Defining the unit. A sampling unit has spatial and temporal limits. These are explicit when necessary. They are often implicit, however. For example a 1 ml sample from a bottle cast in the ocean has a stated volume from within the volume of water, taken from an implicitly defined volume in the water column.

Drawing the sample. This includes the temporal component of sampling. The water sample from a bottle cast has an implicit duration on the order of seconds, then on the order of a second for the time taken to trip the trigger and close the bottle at a specified depth in the water column. Batches of data consist of values drawn in a consistent way according to a protocol, but with no claim go be representative of some larger population. Batches of data as defined by Tukey (1977) are useful in exploratory data analysis. Samples are also drawn in a consistent way according to a protocol, with some implication that the values are representative of some larger population of potential samples.

In random sampling, also known as probability sampling, every item in the population has a known probability of being sampled. These probabilities are not necessarily equal. In simple random sampling, each item has an equal probability of occurring. Example of sampling with unequal probability include stratified sampling, multistage sampling, and cluster sampling.

In systematic sampling items are selected according to a schedule. An example is counting trees at sites every 100 m along a transect. These are simple to implement, but can lead to biased estimates unless the items in the frame have been randomized or the method is shown to produce the same result as a random sample.

Mechanical sampling occurs typically in sampling solids, liquids and gases, using devices such as probes, grabs, and scoops. Care is needed in ensuring that the sample is representative of the population.

In convenience sampling items are chosen haphazardly and in an unstructured manner. This is a commonly employed method.

Where random sampling is not possible it is important to evaluate the degree to which the sample represents the population. Inference from sample to a population results in an estimate. The goal is an accurate estimate that is as precise as possible.

Finite and infinite populations.

Finite populations. In an finite population we can identify whether a particular unit belongs to the population, but we can list all units. An example would be the list of people living in a city or a list of streams in a watershed. The list is called a frame.

Infinite populations. In an infinite population we can identify whether a particular unit belongs to the population, but we cannot list all units. The most common example we will encounter is notional: all possible measurements that could be made with a given measurement protocol.

Sample layout.

In a simple random survey, all units are placed randomly, and have the same sampling probability, which is 1/N, where N is the number of units in the frame. Here is a diagram.

Figure here

In a stratified random survey, the sampling probability is the number of units in each stratum of the frame. We can define a stratum any way we like, as long as each sampling unit belongs to one and only one stratum. The sampling probability is $1/N_i$ where N_i is the number of units in stratum i. Here is a diagram.

Figure here

We can use even more sophisticated and efficient designs, such as cluster sampling (Cochran and Cox), as long as we can calculate the sampling probability of any unit.

Precision is not the same as accuracy.

The precision of an estimate (such as a mean) is high if the uncertainty is low.
The accuracy of an estimate is high if the sample is representative of the population.

Example of sports and games

Soccer. Precision is how closely the shots at goal cluster. Accuracy is whether shots a are symmetrical around the target, such as the edge of the goal. Shots can be precise (cluster tightly) yet inaccurate if they are consistently off target. Scoring points depends on both precision (tight clustering) and accuracy (symmetry around the target).

Many games depend on both precision and accuracy.

Fixed target: Darts, hockey, basketball, curling

Random target (Boules family): Bocce, petanque, lawn bowling

Are inferential statistics necessary in science?

Some people will claim that a result is not to be trusted unless it is so clear that "you can drive a truck through." Platt (1964 Strong inference. *Science* **146**: 347-353) advocates this--obtain a clear result. A clear result produces agreement, without recourse to inferential statistics.

However, many fields of research involve substantial uncertainty that cannot be eliminated by manipulative control. A good example is epidemiology, in which potential sources of disease are isolated through sophisticated statistical methods. Inferential statistics are the central tool of epidemiology. They allow conclusions based on quantitative criteria open to examination.

Statistical methods can be used to estimate and remove sources of uncertainty. Statistical methods are used to increase the efficiency of experimental design, by removing extraneous effects through statistical calculation. This is called statistical control.

Statistical control

Q: But aren't manipulatively controlled experiments superior to statistically controlled observational studies ?

A: Yes they are, because we can reduce uncertainty and hence isolate cause.

Q: Shouldn't we then insist on manipulative experiments?

A: Yes, but manipulative experiments are not always feasible or ethical.

Q: When does statistical control become necessary?

A: It becomes necessary where manipulative control is unethical (*e.g.* human health, and to an increasing degree experiments on vertebrate animals)
It becomes increasingly necessary as uncontrollable variation increases.

Statistical control becomes increasingly important as the cost of manipulative control rises.

Statistical control becomes necessary when experimental control is impossible (*e.g.* weather effects on agricultural production, or on aquaculture, and much of the earth and ocean sciences.

Predictive Probability

With games of chance we can calculate probabilities from the mechanical properties of the instruments – a die with 6 sides, a roulette wheel with 38 slots, a deck of 52 cards. In the experiments we can do similar calculations. For example Hartry et al (1984) report the results from experiments on memory transfer in planarian worms through cannibalism. With a maze having 4 exit points, what is the chance of a planarian worm of arriving at the only exit with food?

The probability of arriving at the exit with food is $(0.5)^4 = 0.0625$, the odds are (1-0.0625)/0.0625 = 15 to 1 against arriving at the exit with food.
Once a worm 'learns' the maze (consistently arrives at the exit with food) the worm is fed to a cannibalistic conspecific to test for memory transfer, defined as the cannibal doing better than chance in the maze, without practice.

In experimental design we work with fixed criteria against chance, typically better then 5% chance or 0.95/0.5 = 19 to 1 odds. How many exits do we need to meet this criterion?
The probability of arriving at the only exit with food in a maze with 5 exit points is $(0.5)^5 = 0.03125$, the odds are (1-0.03125)/0.03125 = 31 to 1 against arriving at the exit with food. Based on predictive probability, we would choose the 5 exit maze to meet our criterion for "better than chance."

Likelihood relative to theory

Once we have drawn our data, we can calculate the likelihood of the experimental result and then compare this to theory. Here is an example-- the expected number of purple flowers in a dihybrid cross between pure strains of purple flowered pea plants (dominant trait) and white flowered plants (recessive trait).

Mendel (1865) reported 224 white and 705 purple flowing plants in a dihybrid cross. We can calculate the odds of this observed outcome, relative to the theoretical 3:1 ratio.

Observed Odds:	Purple/White = (705/929) / 224/929) = 705/224 = 3.147
Odds from theory:	Purple/White = 3:1
Odds ratio:	3.147 / 3 = 1.049

There is no evidence (even at a threshold odds ratio of 2:1) against theory.

Likelihood relative to chance.

Only rarely we do we have theory to use as a reference. In the absence of theory we use "just chance" as the reference likelihood. An example is the heterozygosity gradient in fruit flies.

$H = \hat{\beta}_o + \varepsilon$ $\quad \Sigma res^2 = \underline{0.1171}$

$H = \hat{\beta}_E \cdot E + Offset + \varepsilon$ $\quad \Sigma res^2 = \underline{0.0204}$

The reduction in squared deviation was: $\quad \Sigma res^2 = \underline{0.0966}$

The sum of the squared residuals was reduced from 0.1171 to 0.0204 by adding a parameter, the altitudinal gradient in heterozygosity, which we estimated at -0.1273 %/km, using least squares to obtain the estimate. The research model (with the slope) was LR = $(0.0204/0.1171)^{-7/2}$ = 453 times more likely than the reference model, that heterozygosity was unrelated to altitude. This likelihood ratio is "good evidence" against chance at 1% and odds of 99 to 1, but not "strong evidence" at 0.1% and odds of 999 to 1.

We conclude that the model with the gradient in heterozygosity β_E is more likely than the model without the gradient.

For the oat yield data, we had no prior estimate of the increase in yield. So we use "just chance" as the reference model. The model that includes the difference in means β_x improved the fit from 493 to 301, an improvement of 192.08/493.14 = 32%. The strength of the evidence depends on the number of observations. For the 8 observations in this example, the likelihood ratio for the research model (with the two means) was $LR = (493.14 / 301.16)^{-8/2}$ = 7.2 times more likely than the reference model, no relation between yield and treatment with Panogen. The evidence is less than adequate, if we take LR = 20 as a convenient criterion for adequate evidence. For a probability of 5%, more likely than not is 0.95/0.05 = 19:1 or nearly 20.

In both cases—pea genetics, and oat yields—we used likelihood inference. We calculated whether the research model was more likely to have generated the observed values than the "just chance" reference model. The "just chance" reference model is called the null model. Hypothesis testing against a null hypothesis is founded on likelihood inference.

Comparison of probability and likelihood

In this table the vertical bar | is read “given”

		Probability	Likelihood
Definition		Pr(Result | Parameter)	L(Parameter | Data)
	Example	Pr(3 successive heads | fair coin)	L(Fair coin | Data)
Inference		Prospective	Retrospective
Medical analogy		Prognosis	Diagnosis
	Example	Prognosis before a biopsy	Diagnosis after a biopsy
Games of chance		Betting before play	Betting during play
	Example	Roulette, dice	Bid whist, Bridge, Poker
Science		Experimental design	Analysis of experimental data
	Example	Prospective power analysis	Regression, ANOVA
Relation to data		Before taking data	After taking data
Calculations		From known parameters	From known data

References

Dale, A.I. 1982. Bayes or Laplace? An examination of the origin and early applications of Bayes' theorem. *Archive for History of Exact Sciences* 27: 23–47.

Hartry, A.L. ; P. Keith-Lee; W.D. Morton. 1964. Planaria: Memory transfer through cannibalism reexamined. *Science* 146: 274-275

Laplace, P.S. 1774.
Laplace, P.S. 1786
Laplace, P.S. 1812

Mendel, G. 1865. Experiments in plant hybridization (pp.1-39). Brünn: *Proceedings of the Natural History Society of Brünn*

Platt1964 Strong inference. *Science* 146: 347-353.

Stigler, S.M. 1986. *The History of Statistics. The Measurement Of Uncertainty. Before 1900*. Belknap Harvard University Press

Tukey, J.W. 1977. *Exploratory Data Analysis.* Addison-Wesley Publishing Company Reading, Mass.