Model Based Statistics in Biology.

Part II. Quantifying Uncertainty and Evidence.

Chapter 7.1 Three Modes of Inference. Many Varieties.

ReCap Part I (Chapters 1,2,3,4)
Quantities and quantitative reasoning
ReCap (Ch5)
Data equations partition the observed value into the model component and the residual.
The sum of the squared residuals allows us to compare one model to another.
It allows us to quantify the improvement in fit.
ReCap (Ch 6)
Empirical frequency distributions express all of the information our data.
We use probability models to characterize our data with a few parameters
Empirical distributions and probability models have many uses.

Today: Three modes of inference.

Wrap-up
There are three modes of statistical inference, two based on measurement of uncertainty (frequentist and priorist) and one based on measurement of evidence.
Evidentialist inference addresses the question: What is the weight of evidence?
Priorist inferience addresses the question: What do I believe?
Frequentist inference address the question: What is my decision?
Hypothesis testing addresses two varieties of frequentist inference:
Yes/No rejection of the null hypothesis at a fixed error rate
Ranked rejection of the null hypothesis.

Three Modes of Inferential Statistics

Royall (1997) used the following example to illustrate the differences among frequentist, priorist, and evidentialist inference .

A physician prescribes a diagnostic test for a patient. The diagnostic test is known to give false positives 2% of the time. This is called Type I error. The test is also known to give false negatives 5% of the time. This is called a Type II error.

		Test result	
		Positive	Negative
Disease D	Present	0.95	0.05
	Absent	0.02	0.98

For the physician and the patient we have two hypotheses.

Hypothesis A. The patient does not have the disease

Hypothesis B. The patient does have the disease

A patient tests positive.

The physician asks:

1. What does the result tell me about Hypothesis A vs B?
2. What do I believe, now that I have the test result?
3. What do I do, now that I have the result?

The evidence in favor of disease presence B is less then convincing. For the test results we have (0.95/0.02)/0.98/0.05) = 2.4. If the physician starts with neutral belief (50:50 odds) then we have prior odds*test odds = (0.5/0.5)(0.95/0.05) = 19 :1 posterior odds, which seems convincing. However, if we use an infection rate of 10% (very high) we have (0.1/0.9)(0.95/0.05) = 2.1 :1 posterior odds. Which is not convincing.

For a disease known to be rapidly lethal the physician may decide to recommend treatment anyhow. That is, the physician may make a decision against criteria for false positives and false negatives, instead of going with just the evidence or with prior probabilities updated by evidence.

Statistical inference, as it has developed in the 20th and 21st century, addresses each of the three questions.

Likelihood inference addresses question 1.

Priorist inference addresses question 2.

Frequentist inference addresses question 3.

All three modes are based on the principle of likelihood. Statisticians as well as scientists differ on which mode is best. Statisticians do agree on the use of likelihood (in the technical sense) for statistical inference.

What is likelihood?

Definition: A model that makes the data more probable (best predicts the observed data) is said to be more likely to have generated the data.

Three Modes of Statistical Inference

Evidentialist inference (likelihood inference) uses a measure of evidence, the likelihood ratios. Here is an example.

The first published clinical trial was carried out by James Lind, who administered five different antiscorbutics to 12 sailors with scurvy on the British vessel Salisbury in May of 1747 (Lind 1753). Lind allocated two sailors to each of six different daily treatments for a period of fourteen days. The six treatments were:
1.1 litres of cider;
twenty-five millilitres of elixir vitriol (dilute sulphuric acid);
18 millilitres of vinegar three times throughout the day before meals;
half a pint of sea water;
two oranges and one lemon continued for six days only (when the supply was exhausted);
a medicinal paste made up of garlic, mustard seed, dried radish root and gum myrrh.

Lind reported the result as follows.

> *The most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them being at the end of six days fit for duty… The other was the best recovered of any in his condition; and being now deemed pretty well, was appointed nurse to the rest of the sick.*

Lind started with the idea that scurvy was a form of rot that could be halted by treatment with an acid. The expected result was 10 cures in 12 trials = 0.833
The observed result was 2 cures in 12 trials = 0.167

Here is a table of the observed result. Odds = p/(1-p)

	Proportion	Odds	OR	
Observed cure = 2/12	0.167	0.2	0.04	0.4 to 1 for theory
Acid theory,cure = 10/12	0.833	5	25	25 to 1 against theory

The evidence is against the acid cure theory.
Inferential statistics were unknown at the time. Looking back, we can say the evidentialist inference would have been appropriate. We can say that inference to an infinite number of repeats (frequentist inference) was not warranted because there was too little information on how the subjects were selected to repeat the study. Looking back Was inference from a prior probability (priorist inference) warranted? No, for lack of sufficient evidence for a sound prior probability. In 1601, James Lancaster served three teaspoons of lemon juice to men on one of four vessels sailing from England to India. Despite low rates of scurvy on the ship that received the lemon juice, and high rates of mortality (over 100 died) on the other three ships, little further research was conducted until James Lind's experiment in 1747. Lancaster's result (3:1 odds) was a flimsy prior for a trial even at the time, leaving aside modern ethics of consent with respect to sailors impressed into the British navy.
Bartholomew, M. (2002). James Lind and scurvy: A revaluation. *Journal for Maritime Research* 4: 1–14.

Three Modes of Statistical Inference.

Priorist (Bayesian) inference

The term Bayesian refers to Thomas Bayes (1702–1761), who used a deductive argument to prove that probabilistic limits could be placed on an unknown event (Bayes 1763). However it was Pierre-Simon Laplace (1749–1827) who introduced (as principle VI) what is now called Bayes' theorem, and applied it to celestial mechanics, population statistics, reliability, and jurisprudence (Laplace 1774, 1812). Laplace adopted the principle of indifference (uniform prior), then dropped it in favor of the law of large numbers (Stigler). In the 19th century priorist inference was rejected on logical grounds (Venn, Boule). Priorist inference re-emerged on stronger logical grounds in the 20th century (Keynes 1921, Jeffreys Feinberg 1971), based on an expanded version of Laplace's theorem. 20th century priorist inference called itself Bayesian, even though it was founded on Laplace's theorem, not the rule Bayes proved.
20th century priorist inference starts with a prior probability, to which evidence (a generalized likelihood ratio) is applied, to produce a posterior probability. Except in simple cases, this requires massive computation. Alan Turing used priorist inference, together with massive calculation, to break the code used by the German armed forces in World War II.

Fine (1973) listed 5 varieties of priorist inference:
Empirical. Uses facts and physical properties of the world. It is based solely on experiment, observation, and an appropriate probability model.
Logical. Based on logic, either inductive or deductive.
Interpersonal. Beliefs held in common, not based on personal belief
Pragmatic. Based on practical and useful beliefs, rather than logically true beliefs.
Personalistic. Based on individual opinions, judgements, and experience.

"Your priors are going to be a problem."
Simon Routh, Toronto, Ont.

"Who'd have thought they'd get you for tax evasion."
Nick Kanellis, Brooklyn, N.Y.

"I'm afraid they'll give you life."
Dan Crowe, Chicago, Ill.

In everyday life we tend to be priorists. We update personal belief with evidence. Unfortunately we also tend to be selective about observations, usully in favor of those that confirm rather than conflict with belief. In the lab, hunches and intuition are an important part of the priors we use. When science is presented in a public context, or used to develop standards (as in engineering) we need substantiated prior probabilities.

Proposed captions for a cartoon in the New Yorker magazine.

Bayes 1763
Boule
Feinberg 1971
Fine, T.L. 1973.Theories of Probability. Academic Press.
Jeffreys
Keynes 1921
Laplace 1774, 1812
Stigler
Venn

Priorist inference address Royal's second question: what do I believe, after applying evidence to a prior belief?
Priorist inference starts with a prior probability. This is multiplied by a generalized likelihood (a measure of the weight of evidence) to obtain a posterior (after the fact) probability.
If the physician starts with a 'neutral' (50:50) chance that the patient has the disease , then we have (0.5/0.5)(0.95/0.05) = 19 :1 posterior odds
But is 50% really the right prior probability? The physician might have ordered the test because of other information leading to a belief that the patient had a better than even chance of having the disease. At 2:1 odds for belief in presence of disease we have (2/1)(0.95/0.05) = 38:1 odds of having the disease.
But maybe the patient insisted on the test, even though the physician doubted that the patient had the disease. At a prior probability of 2:1 against the disease, we have (1/2)(0.95/0.05) = 9.5 to 1 chance of having the disease.
This illustrates the principle characteristic of 20th century priorist inference: It begins with a prior probability, not with evidence. Where the prior probability is based on strong consensus, inference is reliable. Where the prior probability lacks consensus, the posterior probability is open to doubt and dismissal.

Three Modes of Statistical Inference

Frequentist inference relies on the law of large numbers. Jacob Bernoulli published a rigorous mathematical proof of the law of large numbers in his Ars Conjectandi (The Art of Conjecturing) in 1713. Laplace (1812) used the law of large numbers to support classical (pre 20th century) priorist statistics. However, frequentist inference as it developed in the 20th century begins with a measure of evidence, then appeals to the law of large numbers to calculate a probability from the evidence.
Two varieties developed, the first by Fisher (1922), and the second by Neyman and Pearson (1933). Fisher defined likelihood, used it to make parameter estimates, then used the likelihood ratio to obtain a probability, to sort evidence into categories. Neyman and Pearson adopted Fisher's concept of likelihood, testing the null hypothesis and calculating a probability. They used the probability to declare a decision against the null hypotheses at an error rate fixed in advance. Fisher was the first to criticize this approach. These critiques have continued ever since but have failed to dislodge the Neyman Pearson variety of frequentist inference from standard practice.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc. London Ser. A 222 309–368;
Neyman, J.; Pearson, E. S. (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". Phil. Trans. R. Soc. Lond. A. 231 (694–706): 289–337.

Strictly speaking frequentist inference applies to indefinitely repeatable experiments. It also applies to repeatable protocols that produce consistent numbers.
An example is the age of mothers in B4605/B7220. The data were drawn each year from a new class of students. Since 1997, the mean value has converged to a predictable value of around 29 years of age.

Null hypothesis testing addresses Royal's third question – Does the physician act based on A (no disease) or on B (disease). It addresses this question by using the long-run error rate on choosing A. The error on rejecting A, on a positive test result, is 2% so we reject A. leaving B as surviving the test, but not confirmed.

We'll call Hypothesis A the "just luck" hypothesis, that the positive test was due to chance factors (the diagnostic test is not perfect). Here is another example.

Three Modes of Statistical Inference

Frequentist inference rests on rejection of the 'JUST LUCK' hypothesis.

Five people go fishing and catch 8 fish. Izaak Walton catches 7 of the 8 fish.
Izaak Walton says this is skill. The others say it is just luck. How to decide ?
The formal machinery for statistical decision making is to compare the observed outcome (7 of 8 fish caught by 1 in 5 fishers) to all possible outcomes, if success was just luck.

List outcomes on board.
How many arrangements of 8 fish, such that Izaak Walton has 7 ?

IW = 7	IW = 7	IW = 7	IW = 7
A = 1	B = 1	C = 1	D = 1

Given the question we consider even more extreme outcomes (IW = 8)

IW = 8	IW = 7	IW = 7	IW = 7	IW = 7
	A = 1	B = 1	C = 1	D = 1

Next we assign probability to each outcome, based on chance alone.
(which we will call the JUST LUCK hypothesis or H_o).

For each fish there is a 1 in 5 chance that Izaak Walton will catch it, under the JUST LUCK hypothesis.

IW = 8	IW = 7 A = 1	IW = 7 B = 1	IW = 7 C = 1	IW = 7 D = 1
$(1/5)^8$	$(1/5)^7$	$(1/5)^7$	$(1/5)^7$	$(1/5)^7$
0.00000256	0.0000128	0.0000128	0.0000128	0.0000128

There are five different outcomes--four ways for IW to catch 7 fish, one way to catch 8.
We assign a probability to each outcome, assuming the JUST LUCK hypothesis.
Then calculate the probability of this aggregate outcome (7 or more to IW) by chance alone. p = 0.00005376

IW catching 7 of 8 fish could be 'just luck,' but this would only happen 5 times in 10,000 if it were just luck.

An onlooker, with no stake in the matter, would have no reason to disagree with rejecting the decision that 7 of 8 fish was not "just luck." Frequentist inference rests on rejecting the JUST LUCK hypothesis. It does not, however, prove the alternative, greater skill.

Three Modes of Statistical Inference

Frequentist rejection of the JUST LUCK hypothesis.

Type I and II error - Definition

Neyman and Pearson defined Type I and II error according to the following table.

	Reject H_o	Cannot reject H_o
H_o False	OK	II
H_o True	I	OK

Type I "reject a true null" puts us at risk of a false positive finding or conclusion.
Type II "false null not rejected" puts us at risk of a false negative finding or conclusion.

The decision is about the null hypothesis, not about the alternative. With hypothesis testing all we can say is "We reject the null in favor of the alternative." We do not 'accept the alternative" when we reject the null. Conversely we do not accept the null just because we cannot reject it. We merely retain it due to uncertainty (Type II error).
Type I error is error on the side of credulity.
An example is risk is trusting a useless drug or believing that a trend based on 3 straight increases is 'more than just chance.'

Type II error is error on the side of skepticism. The risk is discarding a drug that is effective, or believing that the higher rate of cancer among smokers is 'just chance," or believing that three points falling close to a line indicate a reliable trend.

Type I and II errors,with their emphasis on hypothesis testing and confidence limits, have become central to the way statistics are currently practiced, Remembering which is which is not easy when it comes to applications. If in doubt, pull up the table. Type I error is about a false positive conclusion not about a "false positive." Keep in mind that when you reject the null, or take it as weakly supported, you are putting yourself at a known error rate of being wrong. That error rate is called Type I. The same thought sequence applies to Type II error. If you can't reject the null, you are at risk of being wrong. That error is Type II.

Type I and II error is all about making a decision. But surely there is more to the practice of science than declaring decisions.

The Null Hypothesis H_o

In science, the null is absence of knowledge or theory. In the absence of theory, such as a Mendelian ratio of purple to white flowers, the null is what we expect by "chance" -- the "Just Luck" hypothesis. The alternative hypotheis H_A is what we hope to see remain standing from a test against the null. The alternative hypothesis is what we hope to see emerge from experiment or observational study.

We are really interested in H_A not H_o so why do we use the null model?
We use the H_o because it is easier to disprove than prove a proposition. It is easier to put evidence against chance than to establish the research hypothesis H_A. It is easier to work out probabilities for H_o than it is with the research hypothesis H_A. To work out probabilities for the research hypothesis we have to say something about effect size.

This focus on the null, the JUST LUCK hypothesis, is not at all intuitive. It is a recent invention, as these things go. It was devised by RA Fisher in 1922. Compare this to the history of probability, which began with the writing of Blaise Pascal and Pierre de Fermat, then with calculation of betting odds in games of chance in the late 17th century. Application of probability to measurement began in the 18th century, with a Bayes' Rule for a probability interval (1763) and a fully developed theory with successful application by Laplace (1774, 1812).

The logic of the H_o is unfamiliar, not at all intuitive. It seems backwards, because it focuses on rejecting chance, rather than 'proving' anything.

Because it is unfamiliar and not at all intuitive, it often fails to work in a public setting. Picture a scientist carefully rejecting the null hypothesis (rejecting chance), when the jury is thinking about the effects on people: did the drug cause harm?

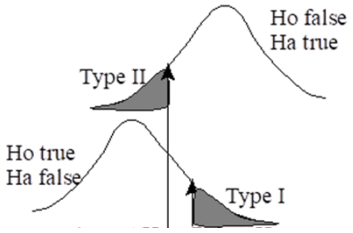
Graphical interpretation of Type I / Type II trade-off.

The upper curve shows distribution of outcomes in a population where there is an effect:

H_A true and H_o false

The lower curve shows the distribution of outcomes in a population where there is no effect:

H_A false and H_o true



The arrow on the left side shows the value of a statistical outcome of an experiment where Type II error is substantial, and so we cannot reject the null hypothesis. If, however, the null is false there is some Type II error (left of arrow on the upper curve). We are at risk of Type II error when we cannot reject the null hypothesis.

The arrow on the right shows the statistical value of the outcome of an experiment where Type I error is small, and so we reject the null. If we extend this arrow to the upper curve, we see that Type II error (to the left of the arrow) is now larger than in the first experiment.

This illustrates the trade-off between Type I and Type II error.

Choice of the Alternative (discovery) Hypothesis H_A

Contrary to Fisher, Neyman and Pearson argued for a decision theoretic approach. With this approach we state a pair of mutually exclusive hypotheses, the null H_o and the alternative H_A. We use a fixed Type I error to either reject or not reject the null.

We state the H_A/H_o pair with respect to the <u>population</u>, not the sample.
For example: H_a: $E(G_{control}) \neq E(G_{treatment})$

Where G is growth rate of control plants and plants treated with fertilizer.
E(G) is the expected (population) value of growth rate.

The H_A/H_o pair can be chosen in <u>several ways</u>

Often the research hypothesis is simply a matter of H_A = effect, H_o = no effect. But more sophisticated choice of pairs are possible, based on the current state of the science. These are more informative than simple yes/no decisions about evidence for an effect.

Here are three ways of stating the H_A/H_o pair, going from least informative to most informative.

1. H_a: $E(G_{control}) \neq E(G_{treatment})$ (stated on nominal scale) "two tail test"
   H_o: $E(G_{control}) = E(G_{treatment})$
2. H_a: $E(G_{control}) < E(G_{treatment})$ (stated on ordinal scale) "one tail test"
   H_o: $E(G_{control}) \geq E(G_{treatment})$
3. H_a: $E(G_{control}) = 1/5 * E(G_{treatment})$ (stated on ratio scale)
   H_o: $E(G_{control}) \neq 1/5 * E(G_{treatment})$

Statement of H_A on an ordinal scale (#2) is better than nominal scale (#1). More is learned by risking a prediction of "greater" than in making a prediction of "different"

Statement on ratio scale (#3) is better yet. More is learned than if prediction was on a less informative ordinal scale (#2). Statement on a ratio scale amounts to priorist inference. We state the H_A before obtaining data, take this as the prior, apply evidence, and arrive at a posterior probability.

Skillful choice of H_A improves analysis.

The H_A/H_o pair can be chosen any way we like, to make best use of what we already know about the question at hand.

All outcomes need to assigned either to H_o or H_A

Care is needed in defining the H_A/H_o pair, as this will determine how we calculate the Type I error (p-value). We compute Type I error differently for one-tailed than for two tailed tests.

Choice of H_A/H_o pair *MUST* be made before undertaking the analysis.
It is incorrect to use the difference of the observed means to choose a one tail test over a two tail test.

Standards for Type I error

	Reject H_o	Cannot reject H_o
H_o False	OK	20%
H_o True	5%	OK

Instead of Fisher's ranking of evidence according to Type I error. Neyman and Pearson decision theoretic inference requires that the criterion for Type I error be set before analysis. By tradition Type I error is set at $\alpha = 5\%$. This appears to be arbitrary. However, it has a rationale. It is a compromise between Type I and Type II error. If we reduce Type I error to 1%, we inevitably increase Type II error.

H_A $Y = \beta_0 + \Sigma\beta_x X +$ residual
H_A data = model + scatter

We reject the H_o at a Type I error of $\alpha = 5\%$

H_o $Y = \beta_0 +$ residual
H_o data = constant + scatter

We cannot reject the H_o if $\alpha > 5\%$
This puts us at risk of Type II error

We are at risk of Type II error when we cannot reject the null hypothesis. This risk decreases as sample size increases, decreases as variability of data decreases, decreases as the effect we wish to detect becomes larger, and decreases if we increase our tolerance for Type I error (note the trade-off described above).

Type II error can be calculated once we have the data, with means, standard deviations, and sample sizes. However, calculation of Type II error after the fact is of little value for the analysis we just did, except perhaps to tell us, too late, that our experiment was doomed to failure. Both high variance or small sample size can increase Type II error and so doom an experiment to failure.

Best practice is to estimate Type II error for the research we plan to undertake, from data with similar variability.

Why are the standards for Type I and II error different?
One answer (Royal 1997) is that a 5% Type II error criterion results in large costs to meet this standard. Another answer is that false positives (Type I error) can have a high cost if implemented.

Standards for Type II error

So how much Type II error should we tolerate? The conventional criterion is that we accept Type II error at 20%, while accepting Type I error at 5%. A competent statistician can compute the Type II error, given some information on the variability, the magnitude of the effect you wish to detect, your tolerance of Type I error (5% ?), and sample size.

Such calculations are called power analyses – the aim is to estimate the power of the research design to detect a difference, given variability, magnitude of effect, tolerance of Type I error, and sample size. Here, power to detect a difference is defined as $1-\beta$. The power of a design with 20% Type II error is 80%.

> Strangely, the motto chosen by the founders of the Statistical Society in 1834 was *Aliis exterendum*, which means "Let others thrash it out." William Cochran confessed that "it is a little embarrassing that statisticians started out by proclaiming what they will not do."
> E. A. Gehan and N. A. Lemak. 1995. *Statistics in Medical Research: Developments in Clinical Trials* (Plenum Press).

Null Hypothesis Testing in a Public Setting.

The logic of the null hypothesis assumes neutral onlookers. An example was catching fish, among friends willing to consider a statistical analysis. What about science in a public setting where the public does not consist of neutral onlookers? In a public setting we cannot assume neutral onlookers. People bring strong beliefs, different experiences, and divergent economic interests to any discussion of evidence in matters that affect us all.

For example, does vaccination increase the risk of autism ?
Parents of autistic children are not neutral onlookers. A vocal set of parents believe that vaccination causes autism and as a result the rate of vaccination against childhood diseases (mumps, measles, polio) has dropped in several locations in North America.
Is this a problem? Yes. It allows diseases to reappear in populations that are normally free of the disease. It puts at risk those children who would not survive the disease, even though most would. Having had measles (and chicken pox) myself I can easily look upon these diseases as just part of life. But what about the larger picture? What about the children, outside of my experience, that did not survive when I had the measles?

Null Hypothesis Testing in a Public Setting.

Another example. Are genetically modified organisms dangerous as food?
A majority of the American public believe that it is. Few biological scientists believe that it is. This issue is the largest gap between the beliefs of biologists and the beliefs of the public in North America. To an increasing degree research scientists function in a public setting that does not consist of neutral onlookers.

Scientists that use observational data face an additional problem. They cannot appeal to long run probabilities based on thousands of runs of an experimental protocol.

Finally there is the fact that null hypothesis testing is not intuitive and so does not do well in public settings.

Critique of Null Hypothesis Testing.

The case against Neyman-Pearson decision theoretic hypothesis testing began with Fisher and has been made repeatedly since. Why does it remain the focus of teaching statistics?

> In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an American Statistical Association discussion forum:
> Q:Why do so many colleges and grad schools teach p = 0.05?
> A: Because that's still what the scientific community and journal editors use.
> Q:Why do so many people still use p = 0.05?
> A: Because that's what they were taught in college or grad school.

The 'sociological inertia' described by Cobb has many components.

-Likelihood inference (Royal 1997) and priorist inference have yet to make their way into general purpose text books.

-Texts fail to provide clear reasons for using any one of the three modes—evidentialist, frequentist hypothesis testing, or 20th century priorist.

-Calculations of Type I and II error remain indispensible in the design of experiments where there are substantial monetary costs to increasing sample size, and real risks in allowing experimental intervention, as with medical research.

- p-values have become established as standards of evidence in the published literature, even though they are not measures of evidence. Manuscript and thesis reviewers often lack the background to evaluate hypothesis tests, and too often take p-values as evidence.

-Lack of computational machinery for priorist inference before the advent of programmable computers with adequate computational capacity.

Replacement of Fisher's Guidelines with Bright Lines.

Perhaps the most important reason for sociological inertia was the replacement of R.A. Fisher's guidelines for null hypothesis testing with Neyman-Pearson decision-theoretic brightlines. Fisher (1922, 1954 p 80) used categories.

> *If p is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05, and consider that higher values of [chisquare] indicate a real discrepancy.*

In *Design of Experiment* (Fisher 1966) refers to the exact value of Type I error in conjunction with strength of evidence.

> *"Convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1% we do not, in Inductive Inference, ever need to lose sight of the exact strength which the evidence has in fact reached."*

Conjunction of p-values with strengths of evidence in a single sentence appears to point at the p-value as a measure of evidence. Read more closely (Fisher 1956 p75), it argues in favor of likelihood inference over use of familiar levels of significance.

> *"For all purposes, and more particularly for the communication of the relevant evidence supplied by a body of data, the values of the Mathematical Likelihood are better fitted to analyse, summarize, and communicate statistical evidence of types too weak [i.e, confidence limits and p-values] to supply true probability statements; it is important that the likelihood always exists, and is directly calculable."*

Throughout his life Fisher argued vehemently against Neyman-Pearson decision theoretic hypothesis testing. Here is a quote from the middle of the last century.

> *I have been concerned for a number of years with the tendency of decision theory to attempt the conquest of all statistics. This concern has been founded, in large part, upon my belief that science does not live by decisions alone--that its main support is a different sort of inference.*
> John W. Tukey 1960. Conclusions vs Decisions. *Technometrics* 2: 423-433.

Critique of Bright Line Null Hypothesis Testing.

Critiques of null hypothesis testing continued throughout the 20th century, rising to a crescendo in this century, culminating in 2016 statement published by the American Statistical Association (Wasserstein and Lazar 2016.) The statement listed a series of "don'ts" mostly focused on the words "statistical significance." The 2016 statement was followed up by a 2019 statement (Wasserstein *et al*) that took the next step, recommending that declarations of "statistical significance" be abandoned. The 2016 statement was accompanied by a variety of publications that suggested alternatives to p-values. Curiously, the alternative proposed by Fisher (1956 p75) was absent.

The 2019 ASA statement was followed by a 2021 statement (Benjamini *et al*) iterating the necessity of publishing a measure of uncertainty (such as a p-value) in reporting statistical analyses.

Benjamini, Y. Et al. (2021). ASA President's Task Force Statement on Statistical Significance and Replicability. *Harvard Data Science Review*, *3*(3). https://doi.org/10.1162/99608f92.f0ad0287

Fisher 1924, 1954 p80, Statistical Methods for Research Workers. Hafner, New York.

Fisher 1935, 1966 8th Ed. Design of Experiments. Olver & Boyd London.

Fisher, R.A. 1956, 1973 3rd Ed. p 75. Statistical Methods and Scientific Inference. Hafner, London

Wasserstein,RL, and Lazar, NA 2016.)

Wasserstein, RL, Schirm, AL & Lazar,NA 2019. Moving to a world beyond "p < 0.05", The American Statistician, 73:sup1, 1-19.

Which Mode of Inference to Use?

The worked examples in this course will list reasons for choice among the 3 modes or inference. They will also list reasons for both varieties of frequentist inference --Fisher sorting versus a Neyman-Pearson fixed Type I error. An important criterion will be whether the results are repeatable (as with an experiment). A similarly important criterion is enough prior evidence to form a convincing prior probability. Another criterion will be a consideration of the costs of Type I error for subjects (as in the health sciences), for the experimenter, and for the public. And finally a last criterion: will the results be used in a public setting with non-neutral onlookers.

Report Evidence and Uncertainty, Not Just Uncertainty.

In this course we will use a measure of evidence, the likelihood ratio, along with a measure of uncertainty, such as a p-value

Certainty	Type I error					
very high	$p < 0.001$					
high	$p < 0.01$			High certainty with good or strong evidence		
some	$p < 0.05$		Some certainty Some evidence			
uncertainty	$p < 0.1$			Low certainty on good or strong evidence		
no certainty	$P > 0.1$	Uncertainty with Inadequate evidence				
		LR < 10	LR > 10	LR > 20	LR > 100	LR > 1000
Evidence		Inadequate	Some	Good	Strong	Very Strong

Calculating Likelihood Ratios.

Computer packages report statistics based on likelihood ratios, but do not automatically report the likelihood ratio itself. To obtain the likelihood ratio, we need functions that unpack a statistic (t, F, X^2, R^2) to reveal the likelihood ratio.

Box 7.7.1 lists functions to extract likelihood ratios from the statistics reported by computer packages. This listing is readily downloaded from the course website.

https://github.com/DavidCSchneider/StatisticalScience/blob/main/RefMaterial/CalculatingLikelihoodRatios.pdf

Your Turn

1. Find an ANOVA table of interest to you, with at least two explanatory variables.
   a) Using an appropriate formula from Box 7.1.1, calculate the LR for each term in the model (including any interactive effect terms).
   b) State a conclusion about each term, with respect to strength of evidence (LR) and one measure of uncertainty (p-value, standard deviation, standard error, confidence limit).
   When reporting evidence, use the LR to compare the alternative to null model.
   When reporting uncertainty with respect to the null model, restrict the statement to the null hypothesis. Do not extend it to the alternative hypothesis (either accept or reject).

Box 7.1.1 Calculating Likelihood Ratios

1. From the explained variance R^2

$$LR = (1 - R^2)^{-n/2}$$

2. From a goodness of fit statistic G

$$G = -2lnLR$$
$$LR = e^{G/2}$$

G is distributed as χ^2 with k degrees of freedom

$$G \sim \chi^2(k)$$

k is the change in number of degrees of freedom in the ANODEV table

4. From ΔDev the reduction in deviance (improvement in fit) in the ANODEV table

$$\Delta Dev = G$$
$$G = -2\ln(L_m \,/\, L_{full})$$

L_m is the likelihood of the reduced model

L_{full} is the likelihood of the full (unreduced) model

$$LR = (L_{full} \,/\, L_m)$$
$$LR = e^{G/2}$$

5. From the *t* statistic

$$LR = (1 + t^2/(n-2))^{n/2}$$

6. From the *F* statistic.

$$LR = \left(1 + F\left(\frac{df_{numerator}}{df_{denominator}}\right)\right)^{n/2}$$

7 From a model term displayed vertically in an ANOVA table.

$$LR = \left(\frac{SS_{term} + SS_{res}}{SS_{res}}\right)^{n/2}$$