Comparison of frequentist and Bayesian inference.
Class 20, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to explain the difference between the p-value and a posterior probability to a doctor.

2 Introduction

We have now learned about two schools of statistical inference: Bayesian and frequentist. Both approaches allow one to evaluate evidence about competing hypotheses. In these notes we will review and compare the two approaches, starting from Bayes' formula.

3 Bayes' formula as touchstone

In our first unit (probability) we learned Bayes' formula, a perfectly abstract statement about conditional probabilities of events:

$$P(A \,|\, B) = \frac{P(B \,|\, A)P(A)}{P(B)}.$$

We began our second unit (Bayesian inference) by reinterpreting the events in Bayes' formula:

$$P(\mathcal{H} \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}.$$

Now $\mathcal{H}$ is a hypothesis and $\mathcal{D}$ is data which may give evidence for or against $\mathcal{H}$. Each term in Bayes' formula has a name and a role.

- The prior $P(\mathcal{H})$ is the probability that $\mathcal{H}$ is true before the data is considered.

- The posterior $P(\mathcal{H} \,|\, \mathcal{D})$ is the probability that $\mathcal{H}$ is true after the data is considered.

- The likelihood $P(\mathcal{D} \,|\, \mathcal{H})$ is the evidence about $\mathcal{H}$ provided by the data $\mathcal{D}$.

- $P(\mathcal{D})$ is the total probability of the data taking into account all possible hypotheses.

If the prior and likelihood are known for all hypotheses, then Bayes' formula computes the posterior exactly. Such was the case when we rolled a die randomly selected from a cup whose contents you knew. We call this the deductive logic of probability theory, and it gives a direct way to compare hypotheses, draw conclusions, and make decisions.

In most experiments, the prior probabilities on hypotheses are not known. In this case, our recourse is the art of statistical inference: we either make up a prior (Bayesian) or do our best using only the likelihood (frequentist).

The Bayesian school models uncertainty by a probability distribution over hypotheses. One's ability to make inferences depends on one's degree of confidence in the chosen prior, and the robustness of the findings to alternate prior distributions may be relevant and important.

The frequentist school only uses conditional distributions of data given specific hypotheses. The presumption is that some hypothesis (parameter specifying the conditional distribution of the data) is true and that the observed data is sampled from that distribution. In particular, the frequentist approach does not depend on a subjective prior that may vary from one investigator to another.

These two schools may be further contrasted as follows:

Bayesian inference

- uses probabilities for both hypotheses and data.

- depends on the prior and likelihood of observed data.

- requires one to know or construct a 'subjective prior'.

- dominated statistical practice before the 20th century.

- may be computationally intensive due to integration over many parameters.

Frequentist inference (NHST)

- never uses or gives the probability of a hypothesis (no prior or posterior).

- depends on the likelihood $P(\mathcal{D} \,|\, \mathcal{H}))$ for both observed and unobserved data.

- does not require a prior.

- dominated statistical practice during the 20th century.

- tends to be less computationally intensive.

Frequentist measures like p-values and confidence intervals continue to dominate research, especially in the life sciences. However, in the current era of powerful computers and big data, Bayesian methods have undergone an enormous renaissance in fields like machine learning and genetics. There are now a number of large, ongoing clinical trials using Bayesian protocols, something that would have been hard to imagine a generation ago. While professional divisions remain, the consensus forming among top statisticians is that the most effective approaches to complex problems often draw on the best insights from both schools working in concert.

4 Critiques and defenses

4.1 Critique of Bayesian inference

1. The main critique of Bayesian inference is that a subjective prior is, well, subjective. There is no single method for choosing a prior, so different people will produce different priors and may therefore arrive at different posteriors and conclusions.

2. Furthermore, there are philosophical objections to assigning probabilities to hypotheses, as hypotheses do not constitute outcomes of repeatable experiments in which one can measure long-term frequency. Rather, a hypothesis is either true or false, regardless of whether one knows which is the case. A coin is either fair or unfair; treatment 1 is either better or worse than treatment 2; the sun will or will not come up tomorrow.

4.2 Defense of Bayesian inference

1. The probability of hypotheses is exactly what we need to make decisions. When the doctor tells me a screening test came back positive I want to know what is the probability this means I'm sick. That is, I want to know the probability of the hypothesis "I'm sick".

2. Using Bayes' theorem is logically rigorous. Once we have a prior all our calculations have the certainty of deductive logic.

3. By trying different priors we can see how sensitive our results are to the choice of prior.

4. It is easy to communicate a result framed in terms of probabilities of hypotheses.

5. Even though the prior may be subjective, one can specify the assumptions used to arrive at it, which allows other people to challenge it or try other priors.

6. The evidence derived from the data is independent of notions about 'data more extreme' that depend on the exact experimental setup (see the "Stopping rules" section below).

7. Data can be used as it comes in. There is no requirement that every contingency be planned for ahead of time.

4.3 Critique of frequentist inference

1. It is ad-hoc and does not carry the force of deductive logic. Notions like 'data more extreme' are not well defined. The p-value depends on the exact experimental setup (see the "Stopping rules" section below).

2. Experiments must be fully specified ahead of time. This can lead to paradoxical seeming results. See the 'voltmeter story' in:
http://en.wikipedia.org/wiki/Likelihood_principle

3. The p-value and significance level are notoriously prone to misinterpretation. Careful statisticians know that a significance level of 0.05 means the probability of a type I error is 5%. That is, if the null hypothesis is true then 5% of the time it will be rejected due to randomness. Many (most) other people erroneously think a p-value of 0.05 means that the probability of the null hypothesis is 5%.

Strictly speaking you could argue that this is not a critique of frequentist inference but, rather, a critique of popular ignorance. Still, the subtlety of the ideas certainly contributes to the problem. (see "Mind your p's" below).

4.4 Defense of frequentist inference

1. It is objective: all statisticians will agree on the p-value. Any individual can then decide if the p-value warrants rejecting the null hypothesis.

2. Hypothesis testing using frequentist significance testing is applied in the statistical analysis of scientific investigations, evaluating the strength of evidence against a null hypothesis with data. The interpretation of the results is left to the user of the tests. Different users may apply different significance levels for determining statistical significance. Frequentist statistics does not pretend to provide a way to choose the significance level; rather it explicitly describes the trade-off between type I and type II errors.

3. Frequentist experimental design demands a careful description of the experiment and methods of analysis before starting. This helps control for experimenter bias.

4. The frequentist approach has been used for over 100 years and we have seen tremendous scientific progress. Although the frequentist herself would not put a probability on the belief that frequentist methods are valuable, shouldn't this history give the Bayesian a strong prior belief in the utility of frequentist methods?

5 Mind your p's.

We run a two-sample t-test for equal means, with $\alpha = 0.05$, and obtain a p-value of 0.04. What are the odds that the two samples are drawn from distributions with the same mean?

(a) 19/1   (b) 1/19   (c) 1/20   (d) 1/24   (e) unknown

answer: (e) unknown. Frequentist methods only give probabilities of statistics conditioned on hypotheses. They do not give probabilities of hypotheses.

6 Stopping rules

When running a series of trials we need a rule on when to stop. Two common rules are:
1. Run exactly n trials and stop.
2. Run trials until you see a certain result and then stop.

In this example we'll consider two coin tossing experiments.
Experiment 1: Toss the coin exactly 6 times and report the number of heads.
Experiment 2: Toss the coin until the first tails and report the number of heads.

Jon is worried that his coin is biased towards heads, so before using it in class he tests it for fairness. He runs an experiment and reports to Jerry that his sequence of tosses was $HHHHHT$. But Jerry is only half-listening, and he forgets which experiment Jon ran to produce the data.

Frequentist approach.
Since he's forgotten which experiment Jon ran, Jerry the frequentist decides to compute the p-values for both experiments given Jon's data.

Let θ be the probability of heads. We have the null and one-sided alternative hypotheses

$$H_0 : \theta = 0.5, \qquad H_A : \theta > 0.5.$$

Experiment 1: The null distribution is binomial(6, 0.5) so, the one sided p-value is the probability of 5 or 6 heads in 6 tosses. Using R we get

$$p = 1 - \texttt{pbinom(4, 6, 0.5)} = 0.1094.$$

Experiment 2: The null distribution is geometric(0.5) so, the one sided p-value is the probability of 5 or more heads before the first tails. Using R we get

$$p = 1 - \texttt{pgeom(4, 0.5)} = 0.0313.$$

Using the typical significance level of 0.05, the same data leads to opposite conclusions! We would reject H_0 in experiment 2, but not in experiment 1.

The frequentist is fine with this. The set of possible outcomes is different for the different experiments so the notion of extreme data, and therefore p-value, is different. For example, in experiment 1 we would consider $THHHHH$ to be as extreme as $HHHHHT$. In experiment 2 we would never see $THHHHH$ since the experiment would end after the first tails.

Bayesian approach.
Jerry the Bayesian knows it doesn't matter which of the two experiments Jon ran, since the binomial and geometric likelihood functions (columns) for the data $HHHHHT$ are proportional. In either case, he must make up a prior, and he chooses Beta(3,3). This is a relatively flat prior concentrated over the interval $0.25 \le \theta \le 0.75$.

See `http://mathlets.org/mathlets/beta-distribution/`

Since the beta and binomial (or geometric) distributions form a conjugate pair the Bayesian update is simple. Data of 5 heads and 1 tails gives a posterior distribution Beta(8,4). Here is a graph of the prior and the posterior. The blue lines at the bottom are 50% and 90% probability intervals for the posterior.
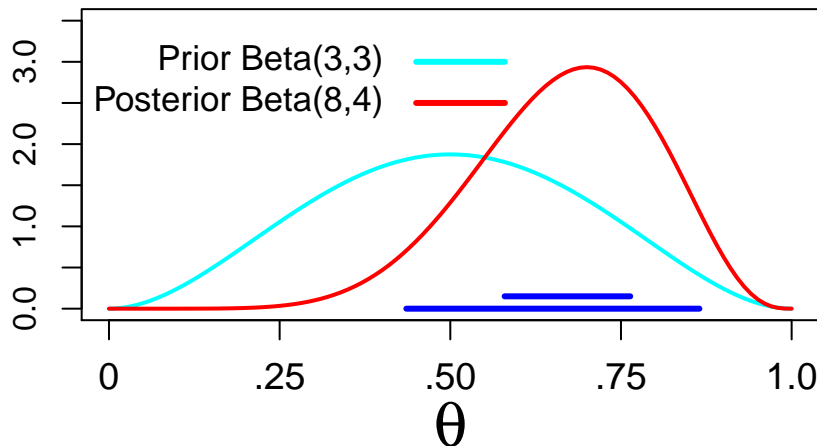
Prior and posterior distributions with 0.5 and 0.9 probability intervals

Here are the relevant computations in R:

Posterior 50% probability interval: `qbeta(c(0.25, 0.75), 8, 4)` $= [0.58\ 0.76]$
Posterior 90% probability interval: `qbeta(c(0.05, 0.95), 8, 4)` $= [0.44\ 0.86]$
$P(\theta > 0.50\,|\,\text{data}) = 1-$ `pbeta(0.5, posterior.a, posterior.b)` $= 0.89$

Starting from the prior Beta(3,3), the posterior probability that the coin is biased toward heads is 0.89.

7 Making decisions

Quite often the goal of statistical inference is to help with making a decision, e.g. whether or not to undergo surgery, how much to invest in a stock, whether or not to go to graduate school, etc.

In statistical decision theory, consequences of taking actions are measured by a utility function. The utility function assigns a weight to each possible outcome; in the language of probability, it is simply a random variable.

For example, in my investments I could assign a utility of d to the outcome of a gain of d dollars per share of a stock (if $d < 0$ my utility is negative). On the other hand, if my tolerance for risk is low, I will assign a more negative utility to losses than to gains (say, $-d^2$ if $d < 0$ and d if $d \geq 0$).

A decision rule combines the expected utility with evidence for each hypothesis given by the data (e.g., p-values or posterior distributions) into a formal statistical framework for making decisions.

In this setting, the frequentist will consider the expected utility given a hypothesis

$$E(U \,|\, \mathcal{H})$$

where U is the random variable representing utility. There are frequentist methods for combining the expected utility with p-values of hypotheses to guide decisions.

The Bayesian can combine $E(U \,|\, \mathcal{H})$ with the posterior (or prior if it's before data is collected) to create a Bayesian decision rule.

In either framework, two people considering the same investment may have different utility functions and make different decisions. For example, a riskier stock (with higher potential upside and downside) will be more appealing with respect to the first utility function above than with respect to the second (loss-averse) one.

A significant theoretical result is that for any decision rule there is a Bayesian decision rule which is, in a precise sense, at least as good a rule.