Hypothesis Testing

- Hypothesis testing is one area where Bayesian methods and results differ sharply from frequentist ones
- Example: Suppose we have a coin and wish to test the hypothesis that the coin is fair
  - $H_0$: The coin is fair, $p(H) = p(T) = 0.5$
  - $H_1$: The coin is not fair (anything but the above)
- We divide the parameter space $\Theta$ into two pieces, $\Theta_0$ and $\Theta_1$, such that if the parameter $\theta$ is in $\Theta_0$ then the hypothesis is true, and if it is in $\Theta_1$ then the hypothesis is false.
- We observe a test statistic $x$, which is a function of the data $X$, and wish to decide, given $x$, whether or not to reject $H_0$

Classical Hypothesis Testing

- Classical statisticians recognize two kinds of errors (this is horrible terminology, but we are stuck with it)
  - A Type I error is made if we reject $H_0$ when it is true
  - A Type II error is made if we do not reject $H_0$ when it is false
- Choose a rejection region $R$:
  - $R = \{ x: \text{observing } x \text{ in } R \text{ leads to the rejection of } H_0 \}$
- Then the probability of making a Type I error is
  $p(x \in R | \theta \in \Theta_0)$
  and the probability of making a Type II error is
  $p(x \in \bar{R} | \theta \in \Theta_1)$

- Often we take $\Theta_0$ as a set containing a single point (simple hypothesis), so
  $\alpha = p(x \in R | \theta \in \Theta_0)$
  is well-defined. However, $\Theta_1$ is usually a collection of intervals (composite hypothesis) and
  $\beta = p(x \in \bar{R} | \theta \in \Theta_1)$
  has no definite value. The best one can do is to evaluate
  $p(x \in \bar{R} | \theta) = \beta(\theta)$
  as a function of $\theta$. $\beta(\theta')$ is the probability of making a Type II error if the true value of $\theta$ is $\theta'$. $\alpha(\theta) = 1 - \beta(\theta)$ is known as the power function of the test.

- We can construct different tests by
  - Choosing a different test statistic $x(X)$
    » But no real choice if $x$ is sufficient
  - Choosing a different rejection region $R$
  - Note that both of these are subjective choices!
Classical Hypothesis Testing

- A common practice is to choose things so that $\alpha$ is some fixed fraction like 0.05 or 0.01. This gives us an $\alpha$-level test, and the probability of making a Type I error is $\alpha$.
- Then, amongst the $\alpha$-level tests, the goal would be to choose that test such that the power function of the test dominates that of all other tests. Such tests are called “uniformly most powerful” (UMP) tests, and they have the smallest probability of committing a Type II error for any given value of $\theta$. Unfortunately, in general a UMP test does not exist.

Bayesian Hypothesis Testing

- So, we return to the Bayesian idea. We have to have a prior; we need the likelihood; then we can compute the posterior.
- Suppose we have two simple hypotheses. Thus, we have $H_0$ and $H_1$, and likelihood $P(X|H_0), P(X|H_1)$, and prior $p(H_0), p(H_1)$. Then the posterior odds are

$$\frac{p(H_0 | X)}{p(H_1 | X)} = \frac{p(X | H_0) \times p(H_0)}{p(X | H_1) \times p(H_1)}$$

Classical Hypothesis Testing

- However, from a Bayesian point of view these classical tests are all suspect. Since they depend on the probability of $x$ falling into some region $R$, they depend on the probability of data that might have been observed, but was not. Thus, they violate the Likelihood Principle.
- A Bayesian might well say that classical hypothesis tests commit a Type III error: Giving the right answer to the wrong question.
- A Bayesian test would have to depend on and be conditioned on just the data $X$ that was observed.

Bayesian Hypothesis Testing

- However, many realistic examples involve testing a compound hypothesis. For example, our coin tossing problem is to decide if the coin is fair, based on observations of the number of heads and tails. The expected proportion of heads is $\theta$. If the coin is fair, then $\theta=0.5$, and if it is not fair it is some other value. This means we will have to put a prior on $\theta$, for example, $U(0,1)$.
- Thus we are testing

$$H_0 : \theta = 0.5, \quad p(\theta | H_0) = \delta(\theta - 0.5)$$

against

$$H_1 : \theta \neq 0.5, \quad p(\theta | H_1) \sim U(0,1)$$
Bayesian Hypothesis Testing

- We calculate the posterior odds ratio

\[
\frac{p(H_0 \mid X)}{p(H_1 \mid X)} = \frac{\int p(H_0, \theta \mid X) d\theta}{\int p(H_1, \theta \mid X) d\theta} = \frac{\int_p(X \mid H_0, \theta) p(\theta \mid H_0) d\theta}{\int_p(X \mid H_1, \theta) p(\theta \mid H_1) d\theta} \times \frac{p(H_0)}{p(H_1)}
\]

- What we’ve done is simply to marginalize out the nuisance parameter \( \theta \) to get the posterior odds on the two hypotheses.

Bayesian Hypothesis Testing

- As with all Bayesian inference, the results of such a test depend on the prior. And in the case of a simple versus a compound hypothesis, the dependence is sensitive to the prior on \( \theta \), which means that one is less certain of the inference.

- One can look at a range of sensible priors to see how sensitive the results are.

- One can also look at particular classes of priors to see what the maximum evidence against the simple hypothesis is under that class of priors.

Bayesian Hypothesis Testing

- For the particular case of coin tossing, with \( h \) heads and \( t \) tails observed, this becomes

\[
\frac{p(H_0 \mid X)}{p(H_1 \mid X)} = \frac{\int_0^1 \theta^h (1 - \theta)^t \delta(\theta - 0.5) d\theta}{\int_0^1 \theta^h (1 - \theta)^t d\theta} \times \frac{p(H_0)}{p(H_1)}
\]

\[
= \frac{(h + t + 1)C_n^{h+t}(0.5)^{h+t}}{(n+1)C_n^n} \quad \text{with} \quad n = h + t
\]

- Example: Suppose we toss a coin 100 times and obtain 60 heads and 40 tails. What is the evidence against the hypothesis that the coin is fair?

- Assuming the priors we did for our calculation, we find on prior odds 1 that the posterior odds are

\[
101 \times \frac{C_{60}^{100}}{2^{100}} = 1.095
\]

- In other words, these data (slightly) favor the null hypothesis that the coin is fair!

  > Surprising to a frequentist since the two-sided p-value (tail area for obtaining 60 or more heads on 100 tosses given a fair coin) is 0.056 which would almost reject the null in an \( \alpha=0.05 \) level test.
**Bayesian Hypothesis Testing**

- Example: Suppose we toss a coin 100 times and obtain 60 heads and 40 tails. What is the evidence against the hypothesis that the coin is fair?
  - If we look at this example by comparing the simple hypotheses “fair” versus “biased coin with \( \theta=0.6 \)”, which is the most favorable prior \( \delta(\theta-0.6) \) on \( \theta \) for the alternative hypothesis, we still get
  
  \[
  \frac{p(H_0 | X)}{p(H_1 | X)} = \frac{0.5^{60}0.5^{40}}{0.6^{60}0.4^{40}} \times \frac{p(H_0)}{p(H_1)} = 0.134
  \]
  - We can consider this to be the maximum evidence against the null hypothesis. It is still over twice the classical two-sided \( p \)-value

**Bayesian Hypothesis Testing**

- Priors for binomial data
  - Jaynes suggests \( \text{beta}(\theta|0,0) \propto \theta^{1-1}(1-\theta)^{-1} \)
    - This has the advantage of agreeing with intuition if there is a good probability that either of the extremes \( \theta=0 \) or \( \theta=1 \) may be true (as with, for example, whether a random chemical taken off the shelf will or will not dissolve. Presumably if it dissolves the first time, it will each time and if it doesn’t dissolve the first time, it won’t dissolve any other time either)
    - However, if the number of heads or tails is 0 the posterior will not be normalizable and the test will give odds 0 or \( \infty \) which may or may not be desirable

---

**p-values**

- Recall that for a symmetric distribution a \( p \)-value is the tail area (one-sided) or twice the tail area (two-sided) under the probability distribution of \( x|\theta \) from where the data actually lie at \( x_0 \) to infinity:

\[
p\text{-value} = \int_{x_0}^{\infty} p(x|\theta_0)dx
\]

\[
p\text{-value} = \int_{-\infty}^{x_0} p(x|\theta_0)dx + \int_{x_0}^{\infty} p(x|\theta_0)dx
\]

---

**Bayesian Hypothesis Testing**

- Priors for binomial data
  - We used a flat prior for our analysis, for convenience
  - The Jeffreys prior is \( \text{beta}(\theta|1/2,1/2) \propto \theta^{1/2}(1-\theta)^{-1/2} \)
    - Show this! Hint: \( E(h|\theta)=n\theta \)
  - In practice the difference between flat and Jeffreys won’t make much difference since it’s just a difference of one extra head or tail
  - Informative conjugate priors are \( \text{beta} \) distributions \( \propto \theta^{x-1}(1-\theta)^{y-1} \). You may choose the parameters to match your prior knowledge
MCMC Simulation

- We can calculate our results using simulation (useful for when an exact solution is unavailable)
- We do this by simulating a random walk in both model space \( \{H_0, H_1\} \) and in parameter space \( \theta \). Thus we are simulating on both discrete and continuous parameters
- The key here is to allow ourselves to jump between our two models. This will in general be a M-H step
- Since the two models have differing numbers of parameters (in the coin tossing case, one model has zero parameters and the other has one) we will have to propose parameters and models simultaneously
- I will describe a technique known as reversible jump MCMC which is very effective

MCMC Simulation

- We see that \( \alpha \) is the ratio of two quantities of the form
  \[
  \beta_{mn} = \frac{p(X \mid H_m, \theta_m)p(H_m, \theta_m)}{q(H_m, \theta_m \mid H_n, \theta_n)}
  \]
  where \( m \) and \( n \) refer to the two states
- We have a great deal of latitude in picking \( q \). For example, we could choose it independent of state \( n \) (independence sampler):
  \[
  \beta_{mn} = \frac{p(X \mid H_m, \theta_m)p(H_m, \theta_m)}{q(H_m, \theta_m)}
  \]
And, we might factorize both $p$ and $q$:

$$
\beta_{\text{min}} = \frac{p(X \mid H_m, \theta_m) p(\theta_m \mid H_m) p(H_m)}{q(\theta_m \mid H_m) q(H_m)}
$$

- Example: Coin tosses. Choose prior on $H_m$, for example, $p(H_0) = p(H_1) = \frac{1}{2}$
- Choose a proposal, for example $q(H_0) = q(H_1) = \frac{1}{2}$
  
  » We’ll want to reconsider this
- If $m=0$, $\theta_0=\frac{1}{2}$, [strictly, $p(\theta_0 \mid H_0) = \delta(\theta_0-0.5)$], but if $m=1$ we need a prior on $\theta_1$. For simplicity we will take a uniform prior, as in our calculation
- We need also to consider the proposals $q(\theta_m \mid H_m)$ for $m=0,1$

$$
\beta_{\text{min}} = \frac{p(X \mid H_m, \theta_m) p(\theta_m \mid H_m) p(H_m)}{q(\theta_m \mid H_m) q(H_m)}
$$

- An excellent choice of $q(\theta_m \mid H_m)$ (if possible) would be to make it proportional to the posterior $p(\theta_m \mid X, H_m)$! For then we would get
  
  $$
  \beta_{\text{min}} \propto \frac{p(H_m)}{q(H_m)}
  $$
  
  which is a constant. Indeed, if we can also arrange things so that $\beta_{\text{min}} = 1$ we would get a Gibbs sampler!
- The latter can be done approximately by using a small training sample and picking $q(H_m)$ using the results

Amongst the many errors that people make interpreting frequentist results, one in particular is very common, and that is to quote a $p$-value as if it were a probability (e.g., that the null hypothesis is true, or that “the results occurred by chance”)!

- The approved use of $p$-values, on frequentist reasoning, is to report whether or not the $p$-value falls into the rejection region. This has the interpretation that if the null is true, then in no more than a fraction $\alpha$ of all cases will we commit a Type I error.
- The observed $p$-value is not a probability in any real sense! It is a statistic that happens to have a $U(0,1)$ distribution
- Despite their appeal, an observed $p$-value has no valid frequentist probability interpretation
• The observed p-value is \textit{not} a probability in any real sense! It is a statistic that happens to have a \textit{U}(0,1) distribution
  • If the observed p-values were real probabilities, we could combine them using the rules for probability to obtain p-values of combined experiments. Thus (on the null hypothesis of a fair coin), if we observed 60 heads and 40 tails and then independently observed 40 heads and 60 tails, the one-sided p-value for the combined experiment is evidently 0.5, whereas the one-sided p-values for two independent experiments are 0.028 and (1–0.028) respectively; the product is obviously not 0.5, contrary to the multiplication law
  • Similar results hold for two-sided p-values

• Furthermore, suppose you routinely reject two-sided at a fixed \( \alpha \)-level, say 0.05
  • Suppose in half the experiments the null was actually true
  • Finally, suppose that in those experiments for which the null is false, the probability of a given effect size \( x \) decreases monotonically as you go away from 0 (in either direction):

\[
p(e)
\]

\[
0 \rightarrow x
\]

• The absolute maximum evidence against the null hypothesis can be gotten by evaluating the likelihood ratio at the data. For example, if \( x \) is standard normal and we observe \( x = 1.96 \), which corresponds to an \( \alpha \) level of 0.05 (two tailed), we can calculate the likelihood ratio as

\[
p(x | H_0) = \frac{1}{2} \exp(-\frac{1}{2}1.96^2)
\]

\[
p(x | H_1) = \frac{1}{2} \exp(-\frac{1}{2}0^2)
\]

\[
p(H_0 | x) = 0.146\text{ on prior odds of 1}
\]

• Then amongst those experiments rejected with p-values in \((0.05-\varepsilon, 0.05)\) for small \( \varepsilon \), at least 30% will actually turn out to be true, and the true proportion can be much higher (depends upon the distribution of the actual parameter for the experiments where the null is false)
  • This says that under these circumstances, the Type I error rate (probability of rejecting a true null), \textit{conditioned} on our having observed \( p=0.05 \), is \textbf{at least} 30%!
  • Thus the numerical value of an observed p-value greatly overstates the evidence against the null hypothesis, which we already found for coin tosses.
### p-values

- Papers on this subject can be found on the web:
  - [http://makeashorterlink.com/?P3CB12232](http://makeashorterlink.com/?P3CB12232) (Paper by Berger and Delampady)
  - [http://makeashorterlink.com/?V2FB21232](http://makeashorterlink.com/?V2FB21232) (Paper by Berger and Sellke, with comments and rejoinder)
  - [http://www.stat.duke.edu/~berger/p-values.html](http://www.stat.duke.edu/~berger/p-values.html)
- Note that the papers by Berger and Delampady and by Berger and Sellke must be accessed from within the university network or by proxy server via the UVM VPN client.

### Bayesian Epistemology

- Bayesians measure the effect of new data $D$ on the relative plausibility of hypotheses by calculating the Bayes factor
  $$ F\left( \frac{H_0 \mid D}{H_1 \mid D} \right) = \frac{P(D \mid H_0)}{P(D \mid H_1)} $$
- Then we compute posterior odds from prior odds
  $$ O\left( \frac{H_0 \mid D}{H_1 \mid D} \right) = F\left( \frac{H_0 \mid D}{H_1 \mid D} \right) O\left( \frac{H_0}{H_1} \right) $$
- Bayes’ theorem allows us to calculate the effect of new data on various hypotheses and adjust posterior probabilities accordingly. It thus becomes a justification for the inductive method

### Falsification

- Popper proposed that a scientific hypothesis must be falsifiable by data. For example, the hypothesis that a coin has two heads can be falsified by observing one tail
- A hypothesis $H_0$ is falsifiable in Bayesian terms if, for some data $D$, its likelihood on $H_0$ is 0: $p(D \mid H_0) = 0$
- However, the requirement of falsifiability is too restrictive. In science, ideas are seldom, if ever, actually falsified. What usually happens is that old hypotheses are discarded in favor of new ones that new data have rendered more plausible, i.e., have higher posterior probability

### Ockham’s Razor

- “Pluralitas non est ponenda sine necessitate.”
  — William of Ockham
- Preferring the simpler of two hypotheses to the more complex, when both account for the data, is an old principle in science
  - Why do we consider
    $$ s = a + ut + \frac{1}{2}gt^2 $$
    to be simpler than
    $$ s = a + ut + \frac{1}{2}gt^2 + ct^3 $$
Ockham’s Razor

- One way to reflect the common scientific experience that simple hypotheses are preferable is to choose the prior probabilities so that the simpler hypotheses have greater prior probability (Wrinch and Jeffreys)
  - This is a “prior probabilities” interpretation of Ockham’s razor
  - Does it beg the question?
  - What principle should be used to assign the priors?

Simplicity

- We regard $H_0$ as simpler than $H_1$ if it makes sharper predictions about what data will be observed
- Hypotheses can be considered more complex if they have extra adjustable parameters (“knobs”) that allow them to be tweaked to accommodate a wider variety of data
- Complex hypotheses can accommodate a larger set of potential observations than can simple ones
  - “This coin has two heads” vs. “This coin is fair”
  - “This coin is fair” vs. “This coin has unknown bias $\theta$”
  - “The relationship is $s = a + ut + \frac{1}{2}gt^2$” vs. “The relationship is $s = a + ut + \frac{1}{2}gt^2 + ct^3$”

Simplicity

- Some years ago a possible planet was reported orbiting a pulsar. However,
  - The period of the planet was within 1% of half a year
  - The orbital phase of the planet is zero when the Earth-Sun-pulsar angle is $90^\circ \pm 1^\circ$
  - This suggests an alternative hypothesis that the orbital velocity of the Earth around the Sun, not a planet, may be responsible for the anomalous data
  - Assuming equal prior probabilities on $H_0$ and $H_1$, we see that $H_0$ predicts each piece of data with probability 0.02, whereas $H_1$ predicts each with probability 1; the posterior odds are 2500:1 in favor of the alternative

Plagiarism

- Compilers of mailing lists include bogus addresses to catch unauthorized repeat use of the list
- Mapmakers include small, innocuous mistakes to catch copyright violations
- Mathematical tables can be rounded up or down if the last digit ends in ‘5’ without compromising the accuracy of the table. The compiler can embed a secret “code” in the table to catch copyright violations
- In all cases, duplication of these errors provides prima facie evidence, useful in court, that copying took place
**Plagiarism**

- Example: a table of 1000 sines
  - Can expect to have a choice of rounding in 100 cases
  - Let $D = \text{“The rounding pattern is the same”}$
  - Let $C = \text{“The second table was copied from the first”}$
- Then
  \[
  P(D \mid C) = 1, \quad P(D \mid \overline{C}) \approx 10^{-30}
  \]
  \[
  F\left(\frac{C}{\overline{C}}\right) \approx 10^{30}
  \]

**Cheating**

- (McGill University: David Harpp & James Hogan)
  - If two students duplicate the same pattern of *wrong answers* on a multiple choice test, we may be suspicious that they collaborated or that one copied from the other
  - If we found that the two students had sat very near each other, our suspicions may be confirmed
    - (Or were tapping away on their cell phones!)
  - Repetition of the same errors yields high posterior probability that the errors were copied and not independently generated

**Evolutionary Biology**

- The principle of descent with modification underlies evolution
- *Pseudogenes* are genes that have lost essential codes, rendering them nonfunctional
- Nearly identical pseudogenes are observed in closely related organisms (e.g., chimpanzees and humans). By the same arguments as before, the posterior probability that this is due to actual copying from a common ancestor is vastly greater than the posterior probability that it is due to coincidence. This is powerful evidence in favor of evolution
- Similar evidence is provided by the fact that the genetic code is *redundant*. Several triplets of base pairs code for the same amino acids

**Mercury’s Perihelion Motion**

- Around 1900, Newtonian mechanics was in trouble because of the problem of Mercury’s perihelion motion
- Proposed solutions:
  - Rings of matter around the Sun, too faint to see
  - “Vulcan”, a small planet near the Sun, difficult to detect
  - Flattening of the Sun
  - Additional terms in the law of gravity (e.g., $yr^{-3}$ term where $y$ is an adjustable constant)
- Some solutions could be ruled out on observational grounds (Jeffreys-Poor debate, 1921)
- One could not rule out modifications to the law of gravity. The adjustable parameter $\gamma$ can be chosen to allow any motion $a$ of the perihelion
Mercury’s Perihelion Motion

- Along comes Einstein and the General Theory of Relativity, which predicts a very precise value for Mercury’s perihelion motion—no other value is possible
- Using contemporary figures
  - Poor gives a = 41.6" ± 2.0"
  - We have a_E = 42.98" for Einstein’s theory (E)
  - The conditional probability of data a on the hypothesis E that the true value is a_E is

\[
p(a \mid a_E) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{1}{2\sigma^2} (a - a_E)^2 \right]
= p(a \mid E)
\]

where \(\sigma = 2.0\"\) (error of observation)

Mercury’s Perihelion Motion

- The older theory F can be thought of as matching the observed value with a “fudge factor” \(a_F\)
- Observations of Mars, Earth and Venus can limit the fudge factor to \(|a_F| < 100"\)
- Assuming that Newton’s theory is approximately correct, we have a prior density of Mercury’s perihelion motion which we take for now to be \(N(0, \tau^2)\) with \(\tau = 50"\):

\[
p(a_F \mid F) = \frac{1}{\sqrt{2\pi} \tau} \exp \left[ -\frac{a_F^2}{2\tau^2} \right]
\]

The Bayes factor is

\[
\frac{p(a \mid E)}{p(a \mid F)} = \sqrt{1 + \frac{\tau^2}{\sigma^2}} \exp \left[ -\frac{D_E^2}{2} \right] \exp \left[ -\frac{D_F^2}{2(1 + \tau^2)} \right] = 26.0
\]

where

\[
D_E = \frac{a - a_E}{\sigma} = -0.69, \quad D_F = \frac{a}{\sigma} = 20.8, \quad \tau = \frac{\tau}{\sigma} = 25.0
\]

This is moderately strong evidence in favor of E.

The last two factors are \(O(1)\) and measure the “fit” of the two theories to the data. Nearly all of the Bayes factor is due to the first factor, which is known as the ‘Ockham factor’.
What’s Happening

• The Ockham factor arises from the fact that $F$ spreads its bets over a much larger portion of parameter space than does $E$. Essentially $E$ puts all its money on $a_f$, a precise value, spread out only by the error of observation $\sigma$. On the other hand, $F$ spreads its bets out over a range that is 25 times bigger, and hence most of its probability is “wasted” covering regions of the parameter space that were not observed.
• $E$ makes a sharp prediction, $F$ a fuzzy one
• When the data come out even moderately close to where $E$ predicts it will, $E$ is rewarded for the risk it took by getting a larger share of the posterior probability
• The factor of about 25 is just the dilution in probability that $F$ must sacrifice in order to fit a larger range of data

Objective Ockham’s Razor

• Berger invented an “objective” prior for these problems that sets lower bounds on the Bayes factor under certain assumptions
• The idea is to consider a class of priors and minimize the Bayes factor over that class
• We choose here the class of priors that is symmetric and nonincreasing relative to the origin 0 (which represents the pure Newtonian theory)
  • Idea is that large deviations are less likely than small ones, and we do not know a priori which direction the deviation from pure Newtonian physics will be

Objective Ockham’s Razor

• Restricted to this class of densities, the minimum value for the Bayes’ factor is found to be

$$F\left(\frac{E}{F}\right) \geq \frac{\pi}{2} \left[ \ln(|D_F| + 1.2) \right] \exp \left( - \frac{D^2}{2} \right) = 15.04$$

• Thus, 15.04 is an “objective” lower bound on the evidence that the data provide for $E$ and against $F$

Sharp Null Hypotheses

• One is entitled to ask, can a sharp point hypothesis ever be exactly true? In fact, such a hypothesis is almost never exactly true, since design defects in the experiment will almost certainly guarantee that the null is false. One does not need any data to know this! Thus, the real question is, when is a sharp null hypothesis a good approximation to the real situation?
  • The question becomes one of “practical” significance, because if you take lots and lots of data, as we’ll see, you’ll be almost certainly able to reject the null hypothesis at some point with a classical test, even though the effect or practical violation of the null hypothesis may be very small
Sharp Null Hypotheses

- So in real life we should test hypotheses of the form
  - $H_0: |\theta| \leq \varepsilon$, say, for some $\varepsilon$ that represents our understanding of when $\theta$ is insignificantly different from 0
- Popular orthodox tests are not designed to test such nulls
- Example: if a significance test on modern data were used to test Kepler’s laws, they would be rejected, even though Kepler’s laws are a very good zero-order approximation to planetary motion
- If a significance test were used to test whether atomic nuclei contained integer numbers of neutrons and protons, atomic weight data would reject (because of binding energy) even though the hypothesis is correct

Examples

- Case 1: Let the $p$-value be 0.05 so that $D_E = 1.95$. We can plot the Bayes factor versus $\bar{\tau}$
- How do $p$-values and posterior probabilities compare for sharp null hypotheses? Evidently, a small $p$-value is evidence against the null, but as we have seen, its numerical value overstates the evidence against the null
- Consider $a_E = a_F = 0$, so we center everything at 0. Then the Bayes factor is
  \[
  F_{E|F} = \frac{p(a|E)}{p(a|F)} = \sqrt{1 + \bar{\tau}^2} \exp \left\{ - \frac{D_E^2}{2} \left( 1 + \bar{\tau}^2 \right) \right\}
  \]
Examples

- From this we see that for a given $D_E$, the more vague the prior on the alternative (measured by $\tau$), the larger the Bayes factor in favor of the sharp prediction of the null.
- Theories with great predictive power (sharp predictions) are favored over those with vague predictions.

\[ F \left( \frac{E}{F} \right) = \frac{p(a \mid E)}{p(a \mid F)} = \tau \exp \left[ -\frac{D_E^2}{2} \right] \]

Ockham factor

Evidence factor

Examples

- For fixed $\tau$, and a standard deviation for a single observation of $\sigma$, with $n$ observations we will have:

\[ \sqrt{1 + \tau^2} = \sqrt{1 + \frac{n\tau^2}{\sigma^2}} = O(\sqrt{n}) \text{ for large } n \]
- Thus, asymptotically we have for large $n$:

\[ F \left( \frac{E}{F} \right) \sim \sqrt{n} \left( \frac{\tau}{\sigma} \right) \exp \left[ -\frac{z^2}{2} \right] \]

where $z=D_E$ is the z-score, or standardized variable $D_E$.

Examples

- Example 2: Give the alternative hypothesis the maximum advantage by choosing $a_F=a$ (i.e., adjust the fudge factor to agree exactly with the data) and let $\tau=0$ (alternative theory is infinitely sharp). Then the Bayes factor is:

\[ B = F \left( \frac{E}{F} \right) = \exp \left[ -\frac{D_E^2}{2} \right] \]

- For example, with $a/\alpha=1.95$ we get $B=0.15$ and posterior probability 0.13.

Examples

- This suggests a way to interpret $p$-values obtained on data with large $n$ (I. J. Good)
  - Let the $p$-value be $\alpha$
  - Compute the z-score $z_\alpha$ which gives this value of $\alpha$ for the $p$-value (use normal approximation and tables)
  - Take $\tau/\sigma = O(1)$ and compute

\[ B = F \left( \frac{E}{F} \right) = \sqrt{n} \exp \left[ -\frac{z^2}{2} \right] \]

- Then the posterior probability of the null is approximately

\[ p(H_0 \mid \text{data}) = \frac{B}{1+B} \]
Examples

• Here, where we have chosen a ridiculously favorable prior for the alternative, giving it every advantage, the posterior probability is nearly three times the p-value, which continues to overstate the evidence against the null
• It gets worse as $D_E$ gets larger:

<table>
<thead>
<tr>
<th>$D_E$</th>
<th>$B$</th>
<th>p-value</th>
<th>factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>0.011</td>
<td>0.0027</td>
<td>4.0</td>
</tr>
<tr>
<td>4.0</td>
<td>0.00034</td>
<td>0.000063</td>
<td>5.3</td>
</tr>
<tr>
<td>4.892</td>
<td>0.000063</td>
<td>0.000001</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Jeffreys-Lindley “Paradox”

• Choose a $p$-value $\alpha > 0$, however small. To this $p$-value there corresponds a $z$-score $z_\alpha$, and for large $n$ the Bayes factor against the alternative is

$$ B = \sqrt{n} \exp \left( -\frac{z_\alpha^2}{2} \right) $$

- Thus, for large enough $n$, a classical test can strongly reject the null at the same time that the Bayesian analysis strongly affirms it

Jeffreys-Lindley “Paradox”

• If $H$ is a simple hypothesis, $x$ the result of an experiment, the following two phenomena can exist simultaneously:
  - A significance test for $H$ reveals that $x$ is significant at the level $p < \alpha$ where the pre-chosen rejection level $\alpha > 0$ can be as small as we wish, and
  - The posterior probability of $H$, given $x$, is, for quite small prior probabilities of $H$, as high as $(1 - \alpha)$
• This means that the classical significance test can reject $H$ with an arbitrarily small $p$-value, while at the same time the evidence can convince us that $H$ is almost certainly true

• Example: A parapsychologist has a number of subjects attempt to “influence” the output of a hardware random number generator (which operates by radioactive decays). In approximately 104,490,000 events, 18,471 excess events are counted in one direction versus the other
• This is a binomial distribution. The standard deviation of the binomial distribution is $\sigma = \sqrt{n \alpha (1 - \alpha)}$ where $\alpha$ is the expected frequency of counts. Straightforwardly we find that $\sigma = 5111$ counts (using $\alpha = 0.5$)
• The classical significance test finds that the effect is significant at $18471/5111 = 3.61$ standard deviations, for a $p$-value of 0.0003 (two-tailed), using the approximation, excellent in this case, that the binomial distribution can be approximated by a normal distribution
Jeffreys-Lindley “Paradox”

- The Bayesian analysis is quite different. We have a genuine belief that the null hypothesis of no effect might be true.
  - To be sure, no point null is probably ever *exactly* true, because the random event generator might not be perfect. But tests of the generator are claimed to show that its bias is very small so a point null is a good approximation.

- On the alternative hypothesis, all we know is that the effect might be something, but we don’t know how much or even in what direction.
  - Parapsychologists call effects measured in the direction opposite to the intended one “psi-missing”, and it is considered evidence for a real effect—sort of “heads I win, tails you lose”.

- To reflect our ignorance we choose a uniform prior on the alternative.
- We’ve already seen the analysis of this problem in the coin-flipping problem. The result is
  \[
  \frac{p(H_0 \mid x)}{p(H_1 \mid x)} = 12 \frac{p(H_0)}{p(H_1)}
  \]
- In other words, although the classical test rejects $H_0$ with a very small $p$-value, the Bayesian analysis has made us twelve times more confident of the null hypothesis than we were!

- The $p$-value of 0.0003 corresponds to a $z$-score of $z = 3.61$ standard deviations on $n=104,490,000$. Our approximate formula for the Bayes factor yields
  \[
  B \approx \sqrt{104,490,000} \exp \left[ -\frac{(3.61)^2}{2} \right] = 15.1
  \]
- This is certainly in the right ballpark and confirms the approximate formula.
Connection with Ockham’s Razor

- This result, where the Bayesian answer is at great odds with the orthodox result, can be understood in terms of Ockham’s razor.
- The sharp null hypothesis $H_0$ is a special one that comes from our genuine belief (in our parapsychology example) that people cannot really influence the output of a random number generator by simply wishing it. The sharp hypothesis is inconsistent with nearly all possible data sets, since it is consistent only with the minuscule fraction of possible sequences for which the number of 0’s and 1’s are approximately equal. On the other hand, the alternative hypothesis is very open ended and would be consistent with any possible data.

Sampling to a Foregone Conclusion

- The phenomenon we have just discussed is closely related to the phenomenon of **sampling to a foregone conclusion**.
  - In classical significance testing, one is supposed to decide on everything in advance, and one is especially supposed to decide on exactly how much data to take in advance.
  - Failure to do this can lead to a situation where if you sample long enough, at some point with probability as close to 1 as you wish you will reject a true null hypothesis with as small a preset significance level as you wish.
  - This is **sampling to a foregone conclusion**.
- This phenomenon is peculiar to classical significance testing. It cannot occur in Bayesian testing.

Connection with Ockham’s Razor

- This means that the alternative hypothesis has an adjustable parameter, the effect size $\theta$, that the null hypothesis does not have. The null hypothesis makes a definite prediction $\theta=0.5$, whereas the alternative hypothesis can fit any value of $\theta$.
- Therefore the null hypothesis is **simpler** than the others in the sense we’ve been discussing. It has fewer parameters.
- Because of the Ockham factor, which naturally arises in the analysis, we can say that in some sense Ockham’s razor is a **consequence** of Bayesian ideas. The Ockham factor automatically penalizes complex hypotheses, forcing them to fit the data significantly better than the simple one before they will be accepted.

Connection with Ockham’s Razor

- To conclude: There are at least three Bayesian interpretations of Ockham’s razor
  - As a way of assigning prior probabilities to hypotheses, based on experience
  - As a consequence of the fact that complex hypotheses with more parameters, in their attempt to accommodate a larger set of possible data, are forced to waste prior probability on outcomes that are never observed. This automatic penalty factor favors the simpler theory
  - As an interpretation of the notion that when fitting data to empirical models, one should avoid overfitting the data
Stopping Rule Principle

- Sampling to a foregone conclusion is related to the stopping rule principle (SRP) according to which the stopping rule—how we decide to stop taking data—should have no effect on the final reported evidence about the parameter \( \theta \) obtained from the data.
- The SRP is a consequence of the Likelihood Principle.
  - Classical testing violates the stopping rule principle just as it violates the likelihood principle. Thus, “sample until \( n = 12 \)” and “sample until \( t = 3 \)” are different stopping rules that give different inferences in classical binomial testing, but the same inferences in Bayesian testing.
- The ability to ignore the stopping rule in Bayesian inference has profound implications for experimental design.
  - It is OK to stop if “the data look good” or “the data look horrible”. Breaking a prior decision to test \( n \) patients, for example, will not compromise the validity of the test.
    - Thus if the data look good and the new treatment looks very effective, it would be unethical not to break the protocol so that the patients on placebo can get the effective drug.
    - Likewise if the first 20 patients all died under the new treatment it would be unethical to continue.

Frequentist Work-arounds

- In frequentist hypothesis testing this problem can be avoided through the device of “\( \alpha \) spending”. That is, suppose we are in a drug clinical trial, and wish to stop the trial at some point to do a preliminary assessment of the results, to decide whether to continue the trial.
  - Terminate trial because of excess bad outcomes.
  - Terminate trial because drug is so effective it would be unethical not to give it to the placebo group.
  - A frequentist can “peek” at the data in advance if one is willing to “spend” some of the \( \alpha \) at that point; but if the test is continued, it will require a smaller \( \alpha \) at a later point to reach the preassigned \( \alpha \)-level for the overall trial.
  - However, this is much more complex and involved than the Bayesian approach.
Hypothesis Testing and Prior Belief

- Prior belief has to be a consideration in any kind of hypothesis testing. Thus, on the same data, different degrees of plausibility may accrue to a hypothesis.
- For example, consider shuffling a pack of alphabet cards. A subject is supposed to guess the letter on three cards. Suppose the subject names all three correctly. What do we think when
  - $H_1$: The subject is a child and is allowed to look at the cards
  - $H_2$: The subject is a magician, who only looks at the backs of the cards
  - $H_3$: The subject is a psychic, who only looks at the backs of the cards

Good’s Device

- The statistical evidence is the same, but the posterior beliefs are very different!
- This example illustrates a device due to I.J. Good for setting prior probabilities
  - It will take much more data to convince most people of the truth of $H_3$ than $H_2$, even though the evidence is identical. One can ask, “how much more data?” Then, by using Bayes’ theorem in reverse, we can estimate our prior on the hypotheses
    - How many times in a row would someone have to name a randomly-picked card before you would give 1:1 odds that he was a psychic?

Example

- Published report of an experiment in ESP (Soal and Bateman, 1954)
- Deck of 25 Zener cards (5 different designs) are shuffled, subject guesses order of the cards
  - Sampling without replacement, often misanalyzed by parapsychologists (P. Diaconis)
- Subject obtained $r=9410$ “hits” out of 37100 (vs. $7420 \pm 77$ expected) $f=0.2536$
- What should we think? $H_0 =$ no ESP: $p=0.20$, $f=r/n=0.2536$, which is $25.8 \sigma$ away from the expected value

Example

- Calculate probability given null hypothesis
  \[ p(\text{data} \mid H_0) = \binom{n}{r} p^r (1-p)^{n-r} \]
- Use Stirling approximation
  \[ n! \approx \sqrt{2\pi n}^{n+1/2} e^{-n} \]
  obtain
  \[ p(\text{data} \mid H_0) \approx \left[ \frac{n}{2\pi (n-r)} \right]^{n/2} \exp(nH(f,p)) \]
  where the cross-entropy is
  \[ H(f,p) = -f \ln f - (1-f) \ln \frac{1-f}{1-p} \]
Example

- The cross-entropy (Kullback-Leibler divergence)

\[ H(f, p) = -f \ln(f / p) - (1 - f)\ln\left(\frac{1 - f}{1 - p}\right) \]

measures the degree to which the observed distribution matches the expected distribution. Our *entropy* is just the cross-entropy relative to a uniform distribution.

- Plug in the data and find that

\[ p(data | H_0) = 0.00476 \exp(-313.6) \]

\[ = 3.15 \times 10^{-139} \]

This is very small! Does the subject have ESP?

- If we calculate the Bayes factor against the value of \( p \) that maximizes the cross-entropy (\( p=f \)) we still get huge odds against the null hypothesis.

\[ P = H(f, p) = \frac{p(data | P)p(P)}{\sum_{H_i} p(data | H_i)p(H_i)} \]

and if the sum in the denominator dominates the first term, the posterior probability of \( P \) will always be much less than 1.

Example

- The priors hardly matter. No matter what prior you choose on the alternative hypothesis, you’re going to get very strong evidence against the null hypothesis.

- So, does ESP exist?

Example

- The priors hardly matter. No matter what prior you choose on the alternative hypothesis, you’re going to get very strong evidence against the null hypothesis.

- So, does ESP exist?

Example

- The priors hardly matter. No matter what prior you choose on the alternative hypothesis, you’re going to get very strong evidence against the null hypothesis.

- So, does ESP exist?

Example

- Each of \( H_1, H_2, H_3, \ldots \) may have a prior probability \( \pi(H_i) \) that is much greater than that of the hypothesis \( P=H_f \) that the subject has genuine psychic powers, and each would adequately account for the data. As a result

\[ p(P | data) = \frac{p(data | P)p(P)}{p(data | P)p(P) + \sum_{H_i} p(data | H_i)p(H_i)} \]

and if the sum in the denominator dominates the first term, the posterior probability of \( P \) will always be much less than 1.

Example

- Each of \( H_1, H_2, H_3, \ldots \) may have a prior probability \( \pi(H_i) \) that is much greater than that of the hypothesis \( P=H_f \) that the subject has genuine psychic powers, and each would adequately account for the data. As a result

\[ p(P | data) = \frac{p(data | P)p(P)}{p(data | P)p(P) + \sum_{H_i} p(data | H_i)p(H_i)} \]

and if the sum in the denominator dominates the first term, the posterior probability of \( P \) will always be much less than 1.
Example

- Our information is *not* that someone actually performed the feat in question. It is that someone *reported to us* that the feat was performed.
- There are always hypotheses that haven’t been considered, and sometimes they may be raised to significance when data come along that support them.
- This is why careful consideration of all possible hypotheses is important.

Example

- Here’s another way to look at it: Let \( \psi \) be the hypothesis that psychic powers exist, and let \( D \) be the data from an experiment. Let \( C \) refer to an auxiliary hypothesis that cheating (or some other experimental defect) was involved. Then what we have really computed is not

\[
\frac{P(\psi | D)}{P(\bar{\psi} | D)} = \frac{P(D | \psi) P(\psi)}{P(D | \bar{\psi}) P(\bar{\psi})}
\]

but

\[
\frac{P(\psi | D, \bar{C})}{P(\bar{\psi} | D, \bar{C})} = \frac{P(D | \psi, \bar{C}) P(\bar{C}, \psi)}{P(D | \bar{C}, \bar{C}) P(\bar{C}, \bar{\psi})}
\]

Example

- We can compute

\[
P(D, \psi) = P(D, \psi, (C \lor \bar{C})) = P(D, \psi, C) + P(D, \psi, \bar{C})
\]

\[
P(D, \bar{\psi}) = P(D, \bar{\psi}, (C \lor \bar{C})) = P(D, \bar{\psi}, C) + P(D, \bar{\psi}, \bar{C})
\]

Example

- Of the four terms that make up the likelihood, i.e., \( P(D | \ldots) \), only one is very small, say \( h \), and the other three can be expected to be larger, say of order \( R \):

\[
P(D | \psi, C) = R_1 = R
\]

\[
P(D | \psi, \bar{C}) = R_2 = R
\]

\[
P(D | \bar{\psi}, C) = R_3 = R
\]

\[
P(D | \bar{\psi}, \bar{C}) = h << R
\]

[Comment: Although I have set the three \( R_j \)’s about equal to the same number \( R \), the important point is that they are large compared to \( h \). Probably \( R_1 \) and \( R_3 \) are about equal, whereas \( R_2 \) would be smaller, but much larger than \( h \).]
Example

- We need priors on $\psi$ and $C$. Assuming that $\psi$ and $C$ are independent [i.e., $P(\psi|C) = P(\psi)$ so $P(C|\psi) = P(C)$], we set

$$P(\psi|C) = P(\psi) = \varepsilon$$

$$P(C|\psi) = P(C) = \delta$$

so that

$$P(D, \psi, C) = R\varepsilon\delta$$
$$P(D, \psi, \overline{C}) = R\varepsilon(1 - \delta)$$
$$P(D, \overline{\psi}, C) = R\varepsilon(1 - \varepsilon)\delta$$
$$P(D, \overline{\psi}, \overline{C}) = h(1 - \varepsilon)(1 - \delta)$$

Example

- Then assuming that $\varepsilon, \delta << 1$, that the $R_i$ are of the same order and are large, and that $h$ is small compared to $\varepsilon, \delta$, we have

$$P(\psi|D) = \frac{R\varepsilon\delta + R\varepsilon(1 - \delta)}{R\delta + h(1 - \varepsilon)(1 - \delta)}$$

$$P(\overline{\psi}|D) = \frac{R\varepsilon}{R\delta + h} = O(\varepsilon/\delta)$$

$$P(C|D) = \frac{R\varepsilon\delta + R\varepsilon(1 - \varepsilon)\delta}{R\varepsilon(1 - \delta) + h(1 - \varepsilon)(1 - \delta)}$$

$$P(\overline{C}|D) = \frac{R\delta}{R\varepsilon + h} = O(\delta/\varepsilon)$$

Example

- But suppose that our prior on $\psi$ is << our prior on $C$. This means that we believe it a priori more likely that there’s an experimental defect than that psychic powers exist.

Then

$$P(\psi|D) = O(\varepsilon/\delta) << 1$$

$$P(\overline{\psi}|D) = O(\delta/\varepsilon) >> 1$$

so that the effect of doing the experiment has made us believe that it is highly likely that there is an experimental defect, while simultaneously maintaining our lack of belief in psychic powers, although with a lesser degree of skepticism (under our assumptions).

Example

- Laplace noted that those who make recitals of miracles “decrease rather than augment the belief which they wish to inspire; for then those recitals render very probable the error or falsehood of their authors. But that which diminishes the belief of educated men often increases that of the uneducated, always avid for the marvelous.”

- “Extraordinary claims demand extraordinary evidence.”

- “Educated” people can be wrong. Once the idea that stones might fall out of the sky was considered ludicrous. But to establish an extraordinary hypothesis requires extraordinary evidence, and it must be evidence that rules out more plausible (e.g., non-miraculous) alternatives.
Example

- In fact, the Soal/Bateman result was shown to have been
due to experimenter fraud—tampering with the data
records. This was shown by Betty Markwick, who
provided convincing evidence that Soal had altered the
record sheets in a systematic way so as to achieve an
excess of “hits”.

Making Decisions

- If one is testing hypotheses, it is for a reason. One does
not just sit around saying “Oh, well, I guess that
hypothesis isn’t true!”
  - There must be some action implied by making that
decision (an action can include “doing nothing”)
    » We decide to approve that drug
    » We decide to invest in that stock
    » We decide to publish our paper
    » Etc., etc.

Making Decisions

- Example: Testing a drug or treatment
  - Usually when testing a new treatment we will compare
it to an old treatment or a placebo
  - We won’t approve the new treatment if it isn’t better
than the old one
  - We might not approve the new treatment if it is
significantly more costly than the old one, unless it is
significantly better than the old one
    » But that judgement might depend upon whether you
were the patient, the drug company, the
government, or the insurance company!
  - We probably won’t approve the new treatment if it has
significant and adverse side-effects

Making Decisions

- This shows that not only is the probability of the states of
nature important, but we must also consider the
consequences of each state of nature (cost, side-effects,
desirable effects, and so on) given each of the possible
decisions that we might make. Thus we must decide
  - What are the states of nature?
  - What are the probabilities of each state of nature
  (Bayes)?
  - What actions are available to us?
  - What are the costs or benefits given each possible state
  of nature and each action (loss function)?
  - What are the expected costs or benefits of each action?
  - Which action is the best under the circumstances?
Making Decisions

- In drug testing, for example, the actions \( a \) we might take are
  - Approve the drug
  - Do not approve the drug
- As a result of our testing we will end up with posterior probabilities on the states of nature \( \theta \) which will, for example, include the cure rate of the new drug relative to the old one, information on side effects, etc.
- We will have to summarize the consequences of making various decisions about the drug as a utility or loss (=−utility). Call the loss function \( L(\theta, a) \); it is a function of the states of nature \( \theta \) and the actions \( a \)

\[ 
E_\theta(L(\theta, a)) = \int L(\theta, a)p(\theta \mid \text{data})d\theta 
\]

- Evidently, we would want to choose the action (approve, disapprove) depending upon which of these two actions gives the smallest loss
- If we were using utilities, we would maximize the expected utility

Making Decisions

- From this discussion we can see that losses/utilities play an important role in making decisions. Some aspects are objective (e.g., monetary costs); however many of them are subjective, just as priors are subjective.
  - The person who is affected by the decision is the one that must determine the loss/utility for this calculation
    - The insurance company, the drug manufacturer, the FDA, and the patient each will have a different loss/utility when choosing to use or approve, market, or use the drug. Each must use his own loss/utility when making the decision
    - The role of the statistician is to assist the user, but it is not to set the loss/utility (the same goes for the patient’s doctor)

Making Decisions

- Our course is not a course in decision theory. Useful books on decision theory are
  - *Smart Choices*, by Hammond, Keeney and Raiffa. Good introduction for lay people. Stresses the process of eliciting utilities. Discusses basics of probability
  - *Decision Analysis*, by Howard Raiffa, and *Making Decisions*, by Dennis Lindley. Good introductions