

STAT 325 – Handout 6

Bayes' Theorem (1.5)

Bayes' Theorem is my favorite theorem. It enables you to start with a prior probability about a hypothesis and produced an updated (sometimes called posterior) probability, conditional on new evidence/data. Bayes' Theorem applies to legal and medical cases, as well as many other areas of application. It even applies to philosophy, for some philosophers of science argue that Bayes' Theorem provides a framework for analyzing how science proceeds. There's even a book called *The Probability of God* that applies Bayes' Theorem to the question of whether God exists. Moreover, an entire school of thought about how to conduct statistical inference is based on Bayes' Theorem.

Example 6-1: Obama Votes

The following information pertains to voters interviewed by CNN following the 2008 Presidential election:

- 45% of whites voted for President Obama
- 95% of African-Americans voted for Obama
- 66% of voters of other races voted for Obama

a) The average of these three percentages is $(45+95+66)/3 \approx 68.67$. So, did 68.67% of the voters interviewed by CNN vote for Obama? Explain.

The following additional information about CNN's voters might be useful:

- 73% were white
- 13% were African-American
- 14% were of other races

Suppose that we choose one of the CNN voters at random. Define these events: $W = \{\text{voter selected is white}\}$, $A = \{\text{voter selected is African-American}\}$, $O = \{\text{voter selected is of other races}\}$, $V = \{\text{voter selected voted for President Obama}\}$.

b) Translate the above percentages (all six of them) into probability statements involving these events.

c) Represent these probabilities in a probability tree.

d) Use the Law of Total Probability to find the probability that a randomly selected CNN voter voted for President Obama.

e) Explain how the calculation in d) is different from the calculation in a).

Now suppose that one of the CNN voters is selected at random, and you learn that he/she voted for Obama. We will investigate how this information changes the probability that he/she belongs to a particular race.

f) Use the probability tree, and the definition of conditional probability, to determine $\Pr(W|V)$, $\Pr(A|V)$, and $\Pr(O|V)$.

g) How do these updated probabilities (after learning that the person voted for Obama) compare to the prior probabilities?

- **Bayes' Theorem:** If A is any event and B_1, B_2, \dots, B_k form a partition of the sample

space S , then $\Pr(B_* | A) = \frac{\Pr(A | B_*)\Pr(B_*)}{\sum_{i=1}^k \Pr(A | B_i)\Pr(B_i)}$.

- Notice that Bayes' Theorem applies when you know, or can easily find, one conditional probability but what you want is the *reverse* conditional probability.
- If the partition consists of just two events, then Bayes Theorem can be expressed

as: $\Pr(H | E) = \frac{\Pr(E | H)\Pr(H)}{\Pr(E | H)\Pr(H) + \Pr(E | H^c)\Pr(H^c)}$.

h) Confirm that applying Bayes Theorem rather than a probability tree provides the same answers as f).

Example 6-2: AIDS Testing

The ELISA test for AIDS was widely used in the mid-1990's for screening blood donations. As with most medical diagnostic tests, the ELISA test is not infallible. If a person actually carries the AIDS virus, experts estimate that this test gives a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry the AIDS virus, ELISA gives a negative result 92.6% of the time (the *specificity* of the test). Estimates at the time were that 0.5% of the American public carried the AIDS virus (the *base rate* with the disease).

- a) Suppose that a randomly selected person tests positive. Without doing any calculations, make a guess for the conditional probability, given this positive test result, that the person actually carries the AIDS virus.
- b) Translate the given percentages into well-defined probabilities.
- c) Express the probability of interest as a well-defined conditional probability.
- d) Use Bayes' Theorem to calculate the conditional probability of interest.
- e) Is this conditional probability smaller than you guessed?

To gain insight into what's happening here, imagine a hypothetical population of 1,000,000 people for whom these percentages hold exactly. Use the given percentages to fill in the following table, starting with the totals who do and do not carry the virus:

	Positive test	Negative test	Total
Carries AIDS virus			
Does not carry AIDS			
Total			1,000,000

- f) From the filled-in table, determine the conditional probability that a randomly selected person carries the AIDS virus given that he/she tests positive. Does this agree with your answer to d)?
- g) Use the table to explain why this probability turns out to be fairly small, compared to the sensitivity and specificity of the ELISA test.

Example 6-5: Forensic Evidence

Bayes' Theorem was applied by expert witnesses testifying in a rape trial in Pittsburgh in the mid 1980's. The defendant was accused of raping seven women in the Shadyside district of the city over a period from April 18, 1985, to January 30, 1986. By analyzing body secretion evidence taken from the scenes of the crimes, a forensic expert concluded that the assailant had the blood characteristics and genetic markers of type B, secretor, PGM 2+1-. She further testified that only 0.32% of the male population of Allegheny County had these blood characteristics and that the defendant himself was a type B, secretor, PGM 2+1-. The natural question to ask is how a juror should update his/her probability of the defendant's guilt in light of this forensic evidence.

a) Let G represent the event that the defendant is guilty, and let E represent the forensic evidence that the criminal's blood type was type B, secretor, PGM 2+1-. Let $P(G)$ represent the prior probability that a juror assigns to the defendant's guilt before hearing the forensic evidence. Rewrite Bayes' Theorem in terms of G and E for expressing how to find the updated probability of guilt, conditional on the forensic evidence, from the prior probability of guilt.

b) What is $P(E|G)$ in this situation?

c) What is $P(E|G^c)$ in this situation? [*Hint*: Assume that if the defendant did not commit the crimes, then some other "random" male in Allegheny County did.]

d) Use your answers to the preceding questions to express the updated probability of guilt $P(G|E)$ as a function of the prior probability of guilt $P(G)$.

e) Construct a graph of $P(G|E)$ as a function of $P(G)$, for values of $P(G)$ ranging from 0 to .5.

f) Calculate the updated probabilities of guilt $P(G|E)$ for the following prior probabilities $P(G)$: .5, .1, .01, .001, and .00000278.

The last entry in this list deserves special mention. The defense in this case argued that the prior probability of guilt should be 1 in 360,000, the estimated number of males in the appropriate age group in Allegheny County. The updated probability of guilt then becomes just 1 in 1150, the number of males with the same blood characteristics in the appropriate age group in Allegheny County.

Example 6-6: Estimating Goldfish

Suppose that you want to estimate the number of fish in a lake; call this number N . In a technique known as capture-recapture, you first capture a random sample of fish, let's say 10 fish. You mark these 10 fish so they can be identified later and then release them back into the lake. You allow them ample time to mix together with the other fish, and then you capture another random sample of fish, let's say 8 fish. Then you count how many of the fish captured in the second sample are marked as having already been caught in the first sample. Let's say you observe that 2 of these 8 fish had already been caught the first time.

Let E denote the evidence/data that 2 of the 8 fish caught in the second sample had also been caught in the first sample.

a) Determine $\Pr(E)$, as a function of N .

b) For what values of N is this probability positive?

c) Use R to graph this function, for values of $N \leq 100$ with positive probability.

```
N = (10:100)
probEgivenN = choose(10, 2) * choose(N-10, 6) / choose(N, 8)
plot(N, probEgivenN)
```

d) What value of N gives the largest probability of resulting in the observed evidence that 2 of the 8 fish captured in the second sample were marked?

e) Explain how this answer to d) makes intuitive sense.

A Bayesian approach to estimating the value of N would begin with prior probabilities that N equals a particular value. Then a Bayesian would update the probabilities of those values of N , based on the evidence/data E , using Bayes' Theorem.

Suppose for now that before taking the second sample we consider only 4 possible values of N : 10, 20, 40, 80, and we consider these to be equally likely. Let the event A mean that $N = 10$, B mean that $N = 20$, C mean that $N = 40$, and D mean that $N = 80$.

f) Determine the updated probabilities of these events A, B, C, D, conditional on the evidence/data E.

g) Which possible value of N is most likely, given the evidence/data E? Which is least likely? Which have become more likely than they were originally? Which have become less likely?

h) Now consider all 91 integer values of N from 10 to 100 as equally likely, prior to observing the second sample. Use Bayes' Theorem (with R) to determine the updated probabilities, given the evidence/data E.

```
N = (10:100)
priorprob = 1/length(N)
prior = rep(priorprob, times=length(N))
probEgivenN = choose(10,2)*choose(N-10,6)/choose(N,8)
numer = prior*probEgivenN
denom = sum(numer)
updated = numer/denom
plot(N,updated)
```

i) Which value of N has the largest updated probability? What is that probability?

j) Determine the smallest possible interval of values for N that has an updated probability of at least .9.