## Hypothesis testing model

- One of $M \geq 2$ mutually exclusive hypotheses $H_0, H_1, \ldots, H_{M-1}$ is true
- **X** is a random variable whose value we can observe, and use, to decide which of the hypotheses is true
- If $H_i$ happens to be the true hypothesis, then the pmf of **X** is $P_i(u)$
- We can think of $P_i(u)$ as $p_{X|H_i}(u|H_i)$, the conditional pmf of **X** given that $H_i$ is true

  

## The decision rule

- We observe the value of **X** and announce our decision as to which hypothesis we believe to be true
- This decision may or may not coincide with reality — our decision may be $H_i$ when in fact $H_j$ is the true hypothesis
- The decision rule (which we are free to choose as we wish) assigns a hypothesis (the decision!) to each possible value of **X**

  

## The likelihood matrix L

- Likelihood matrix L has M rows (one for each hypothesis) and N columns (one for each value taken on by **X**)
- The entry in the i-th row and j-th column is $P_i(u_j)$, the (conditional) probability that $\mathbf{X} = u_j$ when $H_i$ is the true hypothesis
- The sum of the entries in each row is 1
- The decision rule is specified by shading one entry in each column of L

  

## (Conditional) error probabilities

- The (conditional) probability of a correct decision given that $H_i$ is true is the sum of the shaded squares on the i-th row
- The (conditional) probability of error given that $H_i$ is true is the sum of the unshaded squares on the i-th row
- Many statisticians are very uncomfortable with calling these parameters conditional probabilities

  

## Maximum-likelihood decision rule

- The hypothesis for which the probability of the observation is the maximum is called the maximum-likelihood (ML) decision when that observation is made
- For each observation (value of **X**), the ML decision rule maximizes the likelihood of the observation
- Operationally, the ML rule says: shade the largest entry in each column of L

  

## Probabilities of error for ML rule

- The ML decision rule is just one of many possible decision rules, and the various (conditional) probabilities of error can be computed as described before
- Exercise: The ML rule shades the largest entry in each column of L. But we can maximize the probability of a correct decision (when $H_i$ is true) by choosing the largest entry in the i-th row. Why does this better rule not work?

  

## Average error probability

- The (conditional) probability of an error (event E) given that the hypothesis $H_i$ is true is denoted $p_i(E) = P(E|H_i)$
- By the theorem of total probability,
  $$P(E) = \sum P(E|H_i) \cdot P(H_i) = \sum p_i(E) \cdot \pi_i$$
  where $\pi_i = P(H_i)$ is the probability that the hypothesis $H_i$ is true
- Many statisticians are very uncomfortable with the notion of probabilities being assigned to hypotheses

  

## Better to be right than to be wrong!

- Let $C = E^c$ be the event that a correct decision is made. Then,
  $$P(C) = \sum P(C|H_i) \cdot P(H_i) = \sum (1 - p_i(E)) \cdot \pi_i$$
- $P(C|H_i)$, the probability of a correct decision when hypothesis $H_i$ is true is just the sum of the shaded entries on the i-th row of the likelihood matrix L
- P(C) and P(E) are conveniently calculated via the joint probability matrix

  

## Likelihoods to joint probabilities

- The i-th row of the likelihood matrix L is just the conditional pmf of **X** given that $H_i$ is true (i.e. occurred)
- If we multiply the i-th row by $\pi_i = P(H_i)$, we get $P[\{\mathbf{X} = u_j\} \cap H_i]$, the probability that **X** has value $u_j$ and that $H_i$ is true
- Carrying out this operation on all the rows of L gives the joint probability matrix J whose (i,j)-th entry is $P[\{\mathbf{X} = u_j\} \cap H_i]$

  

## Properties of joint prob. matrix J

- J = KL where K is a diagonal matrix
- $K = \text{diag}[\pi_0, \pi_1, \ldots, \pi_{M-1}]$
- J is a Venn diagram whose (i,j)-th entry is the joint probability $P[\{\mathbf{X} = u_j\} \cap H_i] = P_i(u_j)\pi_i$
- The sum of all the entries in J is 1
- Since the row sums in L were 1, the row sums in J are just the $\pi_i$
- The column sums in J give us the unconditional pmf of **X**

  

## Error probabilities from J

- For an arbitrary decision rule specified by shaded entries in L, the probability of a correct decision is the sum of all the shaded entries in J
- The probability of error is the sum of all the unshaded entries in J
- The multiplications by the $\pi_i$ needed in the formula $P(E) = \sum p_i(E) \cdot \pi_i$ have all been taken care of while computing J from L

  

## Maximize P(C) by clever choice

- P(C), the average probability of a correct decision, is the sum of the shaded entries in J
- There is one shaded entry in each column
- Which entries are to be shaded is up to us!
- We can maximize P(C) (and thus minimize P(E)) by shading the largest entry in each column of J
- The corresponding decision rule is called a minimum-error-probability decision rule

  

## A Bayes is a Bayes is a Bayes …

- The minimum-error-probability decision rule chooses the largest entry in each column of J
- The minimum-error-probability decision rule is also called
  - a Bayes' or Bayesian decision rule
  - a maximum ***a posteriori*** probability (MAP or MAPP) decision rule
  - minimum risk decision rule
  - minimum cost decision rule

  

## The *a posteriori* probabilities

- Why is it called Bayes' rule?
- $p_{\mathbf{X}|H_i}(u_j|H_i) = P\{\mathbf{X} = u_j|H_i\}$ is the (i,j)-th entry in the likelihood matrix L
- Bayes' formula gives

  $P(H_i|\mathbf{X} = u_j) = P\{\mathbf{X} = u_j|H_i\} \bullet \pi_i / P\{\mathbf{X} = u_j\}$

  $= p_{\mathbf{X}|H_i}(u_j|H_i) \bullet \pi_i / P\{\mathbf{X} = u_j\}$
- $p_{\mathbf{X}|H_i}(u_j|H_i) \bullet \pi_i$ is the (i,j)-th entry in J
- $P\{\mathbf{X} = u_j\}$ is the j-th column sum in J

  

## More on *a posteriori* probabilities

- Bayes' formula:

  $P(H_i|\mathbf{X} = u_j) = P\{\mathbf{X} = u_j|H_i\} \bullet \pi_i / P\{\mathbf{X} = u_j\}$

  $= p_{\mathbf{X}|H_i}(u_j|H_i) \bullet \pi_i / P\{\mathbf{X} = u_j\}$
- $p_{\mathbf{X}|H_i}(u_j|H_i) \bullet \pi_i$ is the (i,j)-th entry in J
- $P\{\mathbf{X} = u_j\}$ is the j-th column sum in J
- Divide each entry in J by its column sum to convert the joint probability matrix J into the ***a posteriori*** probability matrix

  

## Example: Likelihood matrix

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $H_0$ | 0.2 | 0.3 | 0.2 | 0.3 |
| $H_1$ | 0.12 | 0.24 | 0.64 | 0.0 |
| $H_2$ | 0.16 | 0.1 | 0.1 | 0.64 |
| $H_3$ | 0.0 | 0.3 | 0.3 | 0.4 |

- Suppose that $P(H_0) = 0.4 = P(H_3)$ and $P(H_1) = 0.1 = P(H_2)$

  

## Example: Joint probability matrix

| | 0 | 1 | 2 | 3 | row sums |
|---|---|---|---|---|---|
| $H_0$ | 0.08 | 0.12 | 0.08 | 0.12 | 0.4 |
| $H_1$ | 0.012 | 0.024 | 0.064 | 0.00 | 0.1 |
| $H_2$ | 0.016 | 0.01 | 0.01 | 0.064 | 0.1 |
| $H_3$ | 0.00 | 0.12 | 0.12 | 0.16 | 0.4 |
| col sums | 0.108 | 0.274 | 0.274 | 0.344 | |

- $P(H_0) = 0.4 = P(H_3)$
- $P(H_1) = 0.1 = P(H_2)$

  

## The *a posteriori* probability matrix

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $H_0$ | 0.741 | 0.438 | 0.292 | 0.349 |
| $H_1$ | 0.111 | 0.088 | 0.234 | 0.000 |
| $H_2$ | 0.148 | 0.036 | 0.036 | 0.186 |
| $H_3$ | 0.000 | 0.438 | 0.438 | 0.465 |
| col sums | 1.000 | 1.000 | 1.000 | 1.000 |

  

## Rev. Thomas Bayes' contributions

- Formula for finding P(B|A) from P(A|B)
- The philosophical argument that in making decisions based on the observation that **X** had value $u_j$,
  - one must look at inverse probabilities $P(H_i|\mathbf{X} = u_j\}$ (i.e. ***a posteriori*** probabilities) instead of the likelihoods $P\{\mathbf{X} = u_j|H_i\}$
  - Decide in favor of the hypothesis with the largest ***a posteriori*** probability (MAP)



## The ECE 313 simplification

- The ***a posteriori*** probability matrix is found by dividing each entry in J, the joint probability matrix, by the corresponding column sum
- This does not change the relative values of the entries in the column — the largest entry in a column is still the same
- Don't bother computing the ***a posteriori*** probability matrix! Just work with J to find the Bayes' rule!



## Bayes' rule includes the ML rule…

- Bayes' decision rule depends on what the probabilities $\pi_i$ of the hypotheses are
- Consider the special case $\pi_i = M^{-1}$ for all i, that is, all M hypotheses are equally likely
- Since J is found by multiplying the i-th row of L by $\pi_i$ ( $= M^{-1}$ for all i) the largest entry in a column of J is in the same position as the largest entry in L
- Moral: ML decision rule = Bayes' decision rule for equally likely hypotheses



## How much is 2 + 2?

- Many statisticians and philosophers dislike Bayesian decision rules intensely
- "There are lies, there are damned lies, and there are statistics, …

  and then there are Bayes' decision rules"
- "The answer can be made to come out to whatever you want it to be!"
- "Your personal prejudices and beliefs get injected into the problem"



## Why is Bayes' decision rule hated?

- In many instances, it is difficult to assign probabilities to hypotheses
- Example: Almost all U.S. military radars have never ever "painted" an actual enemy target. What is $\pi_1$ in this case?
- In many instances where facts are hard to come by, the probabilities of hypotheses are based on beliefs
- "Your personal prejudices and beliefs get injected into the problem"



## Counterattack of the Bayesians

- Bayesian statisticians claim that the non-Bayesian people are using Bayesian statistics but refusing to admit the fact
- Non-Bayesians are implicitly assuming that the hypotheses are equally likely when they formulate the ML decision rule
- The non-Bayesians are also injecting their (unstated) belief that the hypotheses are equally likely when they make the implicit assumption

## 2 + 2 = 5 for large values of 2 …

- Consider the likelihood matrix and joint probability matrices shown below

L

| | 1 | 2 |
|---|---|---|
| $H_0$ | 0.4 | 0.6 |
| $H_1$ | 0.2 | 0.8 |

J

| | 1 | 2 |
|---|---|---|
| $H_0$ | $0.4\pi_0$ | $0.6\pi_0$ |
| $H_1$ | $0.2\pi_1$ | $0.8\pi_1$ |

- If $\pi_0/\pi_1 > 8/6$, that is, if $\pi_0 > 4/7$, $H_1$ will never be chosen by the MAP rule
- If $\pi_0/\pi_1 < 1/2$, that is, if $\pi_0 < 1/3$, $H_0$ will never be chosen by the MAP rule



## …and 2 + 2 = 5 for small values of 5

- Except when some entries in L are zero, it is always possible to find some probability assignment for the hypotheses such that the MAP rule chooses only one hypothesis
- In contrast, the ML rule cannot always choose only one hypothesis
- No row in L dominates another row (that is, every entry on the i-th row cannot be larger than the corresponding entry on the j-th row) Why ever not?



## Who's right and who's wrong?

- The Bayesian approach is very valuable if used carefully and thoughtfully
- The Bayesian approach can be abused very easily, and used to provide support for any theory that one wants to make up!
- The Bayesian approach should not be discarded out of hand as useless
- Kitchen knives are very useful tools but they can be very dangerous too!



## Statisticians do it with …

- The approach to decision theory taken in ECE 313 will puzzle most statisticians
- Statisticians use somewhat different terminology and somewhat different methods to derive the results that we obtained so easily via the L and J matrices
- False alarm = Type I error
  False dismissal = Type II error



## … the likelihood ratio

- We shall consider the simple case of two hypotheses $H_0$ and $H_1$ with corresponding pmfs $p_0(u)$ and $p_1(u)$ for the observation **X**
- If **X** is observed to have the value $u_i$, then the likelihood ratio $\Lambda$ has value

$$\Lambda = \frac{p_1(u_i)}{p_0(u_i)}$$

- As the name implies, $\Lambda$ is the ratio of the likelihoods $p_1(u_i)$ and $p_0(u_i)$ of the **X** value



## Thoughts about $\Lambda$

- The value of the likelihood ratio $\Lambda$ depends on the observation, that is, the observed value of the random variable **X**
- The likelihood ratio $\Lambda$ can be viewed as a random variable that happens to be a function of the random variable **X**
- Statisticians say that the likelihood ratio summarizes everything that we need to know about the decision problem

## The likelihood ratio test (LRT)

- Decision rules can be stated in terms of a "test" on the likelihood ratio
- Example: For a binary hypothesis testing problem, L has two rows. The i-th column has two entries $p_0(u_i)$ and $p_1(u_i)$, and the ML rule decides in favor of the larger
- Statisticians say that we are comparing the likelihood ratio $\Lambda$ to 1, and choosing $H_0$ or $H_1$ according as $\Lambda < 1$ or $\Lambda > 1$



## The ML rule as an LRT

- The ML decision rule is stated in terms of an LRT via the following picture
- If $u_i$ is the observed value of **X**, then compute the value of the likelihood ratio $\Lambda = p_1(u_i)/p_0(u_i)$

$$\Lambda \underset{H_0}{\overset{H_1}{\gtrless}} 1$$

- This says choose $H_1$ if $\Lambda > 1$ and choose $H_0$ if $\Lambda < 1$



## More on the LRT

- The LRT
  "choose $H_0$ if $\Lambda < 1$; choose $H_1$ if $\Lambda > 1$"
  is called a threshold test on the likelihood ratio $\Lambda$ because we are comparing $\Lambda$ to the threshold 1 and choosing $H_1$ if $\Lambda$ exceeds the threshold
- The MAP decision rule is also a threshold test on the likelihood ratio!



## The MAP rule as an LRT

- The MAP decision rule is stated in terms of an LRT via the following picture
- If $u_i$ is the observed value of **X**, then compute the value of the likelihood ratio $\Lambda = p_1(u_i)/p_0(u_i)$

$$\Lambda \underset{H_0}{\overset{H_1}{\gtrless}} \pi_0/\pi_1$$

- This says choose $H_1$ if $\Lambda > \pi_0/\pi_1$ and choose $H_0$ if $\Lambda < \pi_0/\pi_1$



## Does the MAP rule make sense?

- The ML rule is the MAP rule for $\pi_0 = \pi_1$
- The ML rule accepts $H_1$ if $\Lambda > 1$
- If $\pi_0 > \pi_1$, the MAP rule accepts $H_1$ only if $\Lambda > \pi_0/\pi_1$ where we know that $\pi_0/\pi_1 > 1$
- When $H_1$ is less likely than $H_0$, the MAP rule insists that it will accept $H_1$ only if there is strong evidence in favor of $H_1$
- A $\Lambda$ value barely larger than 1 is not enough: it better be "much larger"



## Decision regions

- Error probabilities are obained quite easily via L and J
- Purely in terms of symbols, let the set of all possible values taken on by **X** be partitioned into sets $\Gamma_0$ and $\Gamma_1$
- The decision regions $\Gamma_0$ and $\Gamma_1$ specify the decision rule
- If the observed value of **X** is in $\Gamma_i$, then decide that $H_i$ is true

## Error probabilities

- $p_0(E) = P(E|H_0) = P\{\mathbf{X} \quad \Gamma_1|H_0\}$
- $p_0(E) = P(E|H_0) = \sum_1 p_0(u)$ where the subscript on the summation sign means that the summation is over the set $\Gamma_1$
- Similarly, $p_1(E) = P(E|H_1) = \sum_0 p_1(u)$
- $P(E) = \pi_0 \bullet p_0(E) + \pi_1 \bullet p_1(E)$
  $= \pi_0 \bullet \sum_1 p_0(u) + \pi_1 \bullet \sum_0 p_1(u)$



## Average error probability

- $P(E) = \pi_0 \bullet p_0(E) + \pi_1 \bullet p_1(E)$
  $= \pi_0 \bullet \sum_1 p_0(u) + \pi_1 \bullet \sum_0 p_1(u)$
- Now add and subtract $\pi_0 \bullet \sum_0 p_0(u)$ to get
  $P(E) = \pi_0 \bullet \sum_1 p_0(u) + \pi_1 \bullet \sum_0 p_1(u)$
  $+ \pi_0 \bullet \sum_0 p_0(u) - \pi_0 \bullet \sum_0 p_0(u)$
  $= \pi_0 + \sum_0 [\pi_1 \bullet p_1(u) - \pi_0 \bullet p_0(u)]$
- In this sum, the first term is fixed, but we can affect the total by careful choice of $\Gamma_0$



## Minimum error probability

- $P(E) = \pi_0 + \sum_0 [\pi_1 \bullet p_1(u) - \pi_0 \bullet p_0(u)]$ where the sum is over the set $\Gamma_0$
- The choice of the set $\Gamma_0$ is up to us
- How can we choose $\Gamma_0$ to minimize $P(E)$?
- If we include in $\Gamma_0$ every value (and only those values) of $\mathbf{X}$ for which
  $$\pi_1 \bullet p_1(u) - \pi_0 \bullet p_0(u) < 0$$
  then the sum is as negative as possible



## Minimum error probability rule

- The set $\Gamma_0$ is chosen to include every value (and only those values) of $\mathbf{X}$ for which
  $$\pi_1 \bullet p_1(u) - \pi_0 \bullet p_0(u) < 0$$
- This is equivalent to saying that we decide in favor of $H_0$ exactly when the likelihood ratio $\Lambda = p_1(u)/p_0(u)$ is smaller than $\pi_0/\pi_1$
- This provides an alternative derivation of the minimum error probability decision rule



## Costs and risks

- $C_{i,j}$ is the cost if we decide that hypothesis $H_j$ is true when in fact hypothesis $H_i$ is true
- $C_{i,i}$ could be negative (indicating a reward) or it could be positive (representing the fixed costs associated with experiment)
- Assumption: $C_{i,j} > C_{i,i}$
- Costs are also called risks
- Can costs be determined?



## The average risk or cost

- Over a large number of trials, we incur different costs on different trials
- The average cost or the expected value of the cost depends on the choice of the decision regions $\Gamma_0$ and $\Gamma_1$
- Average cost
  $= [\pi_0 \bullet \sum_1 p_0(u)]C_{1,0} + [\pi_1 \bullet \sum_0 p_1(u)]C_{0,1}$
  $+ [\pi_0 \bullet \sum_0 p_0(u)]C_{0,0} + [\pi_1 \bullet \sum_1 p_1(u)]C_{1,1}$

## Minimum average risk or cost

- Average cost
  $= [\pi_0 \cdot \sum_1 p_0(u)]C_{1,0} + [\pi_1 \cdot \sum_0 p_1(u)]C_{0,1}$
  $+ [\pi_0 \cdot \sum_0 p_0(u)]C_{0,0} + [\pi_1 \cdot \sum_1 p_1(u)]C_{1,1}$
- Same trick of adding and subtracting gives that the average cost can be minimized by comparing the likelihood ratio $\Lambda$ to the threshold $(C_{1,0} - C_{0,0}) \cdot \pi_0 / (C_{0,1} - C_{1,1}) \cdot \pi_1$
- If $C_{1,0} = C_{0,1}$ and $C_{0,0} = C_{1,1}$, we get the MAP rule!



## Bayes' risk or cost

- The average cost or risk of the minimum cost rule is called the Bayes' risk
- We will not discuss costs any further (or require you to understand them on exams)
- For M ≥ 3 hypotheses, the calculations are messier — see the Lecture Notes from Fall 1997 on the class web page for some details if you are interested



## Summary

- We have discussed the minimum error probability decision rule (a.k.a. MAP or maximum a posteriori probability or Bayes' rule) at some length
- We introduced the notion of the likelihood ratio $\Lambda$ and the likelihood ratio test
- The ML decision rule and MAP decision rule are threshold tests on $\Lambda$
- The minimum-cost rule is also a threshold test on $\Lambda$