Bayesian Inference and Decision Theory

Unit 1: A Brief Tour of Bayesian Inference and Decision Theory

v1.1

What this Course is About

- You will learn a way of thinking about problems of inference and decision-making under uncertainty
- You will learn to construct mathematical models for inference and decision problems
- You will learn how to apply these models to draw inferences from data and to make decisions
- These methods are based on Bayesian Decision Theory, a formal theory for rational inference and decision making

Logistics

- Web site
  - http://seor.vse.gmu.edu/~klaskey/SYST664/SYST664.html
  - Blackboard site: http://mymason.gmu.edu
- Textbook and Software
  - Hoff, A First Course in Bayesian Statistical Methods, Springer, 2009 (*Free softcopy from Mason library*)
  - Other recommended texts on course web site
  - We will use R, a free open-source statistical computing environment: http://www.r-project.org/. R code for many textbook examples is on author's web site
  - Late in the semester we will use JAGS, an open-source package for Markov Chain Monte Carlo simulation (interfaces with R): http://mcmc-jags.sourceforge.net/
- Requirements
  - Regular assignments (30%): can be handed in on paper or through Blackboard
  - Take-home midterm (35%) and final (35%)
- Office hours
  - Official office hours are 3:30-5:30 PM Wednesdays in my office and via Blackboard Collaborate
  - I respond to questions by email and am available by appointment
- Course delivery
  - 4:30-7:10 Mondays, ENGR 1107 or online via Blackboard Collaborate; all classes recorded
- Policies and Resources
  - Academic integrity policy
  - Read the policies and resources section of the syllabus

Course Outline

- Unit 1: A Brief Tour of Bayesian Inference and Decision Theory
- Unit 2: Random Variables, Parametric Models, and Inference from Observation
- Unit 3: Statistical Models with a Single Parameter
- Unit 4: Monte Carlo Approximation
- Unit 5: The Normal Model
- Unit 6: Gibbs Sampling
- Unit 7: Hierarchical Bayesian Models
- Unit 8: Bayesian Regression and Analysis of Variance
- Unit 9: Additional Monte Carlo Methods
- Unit 10: Hypothesis Tests, Bayes Factors, and Bayesian Model Averaging

(Later units are subject to change)

Learning Objectives for Unit 1

- Describe the elements of a decision model
- Refresh knowledge of probability
- Apply Bayes rule for simple inference problems and interpret the results
- Use a graph to express conditional independence among uncertain quantities
- Explain why Bayesians believe inference cannot be separated from decision making
- Compare Bayesian and frequentist philosophies of statistical inference
- Compute and interpret the expected value of information (VOI) for a decision problem with an option to collect information
- Download, install and use R statistical software

Bayesian Inference

- Bayesians use probability to quantify rational degrees of belief
- Bayesians view inference as *belief dynamics*
  - Use evidence to update prior beliefs to posterior beliefs
  - Posterior beliefs become prior beliefs for future evidence
- Inference problems are usually embedded in decision problems
- We will learn to build *models* of inference and decision problems

"All models are wrong but some are useful"
George Box

Decision Theory

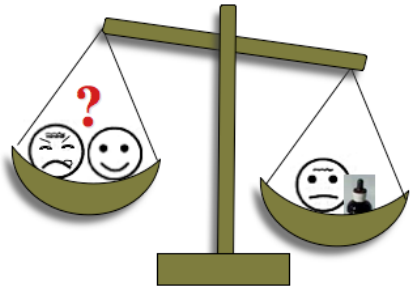
- Decision theory is a formal theory of decision making under uncertainty
- A ***decision problem*** consists of:
  - Possible actions: $\{a\}_{a \in A}$
  - States of the world (usually uncertain): $\{s\}_{s \in S}$
  - Possible consequences: $\{c\}_{c \in C}$ (of action in a state)
- Question: What is the best action?
- Answer (according to decision theory):
  - Measure "goodness" of consequences with a *utility function* u(c)
  - Measure likelihood of states with probability distribution p(s)
  - Best action *with respect to model* maximizes expected utility:

$$a^* = \underset{a}{\operatorname{argmax}}\{E[u(c)|a)\}$$

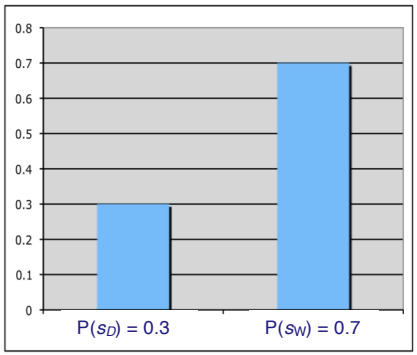
For brevity, we may write E[u(a)] for E[u(c) | a]

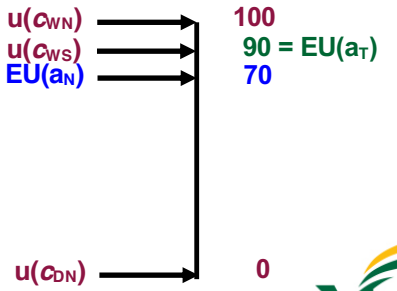
- Caveat emptor:
  - How good it is *for you* depends on fidelity of model to your beliefs and preferences

Illustrative Example:
Highly Oversimplified Decision Problem



- Decision problem: Should patient be treated for disease?
  - We suspect she may have disease but do not know
  - Without treatment the disease will lead to long illness
  - Treatment has unpleasant side effects
- Decision model:
  - Actions: a_T (treat) and a_N (don't treat)
  - States of world: s_D (disease now) and s_W (well now)
  - Consequences: c_{WN} (well shortly, no side effects), c_{WS} (well shortly, side effects), c_{DN} (disease for long time, no side effects)
  - Probabilities and Utilities:
    - $P(s_D) = 0.3$
    - $u(c_{WN}) = 100$, $u(c_{WS}) = 90$; $u(c_{DN}) = 0$
- Expected utility:
  - Treat: $.3\times90 + .7\times90 = 90$
  - Don't treat: $.3\times0 + .7\times100 = 70$
- Best action is a_T (treat patient)



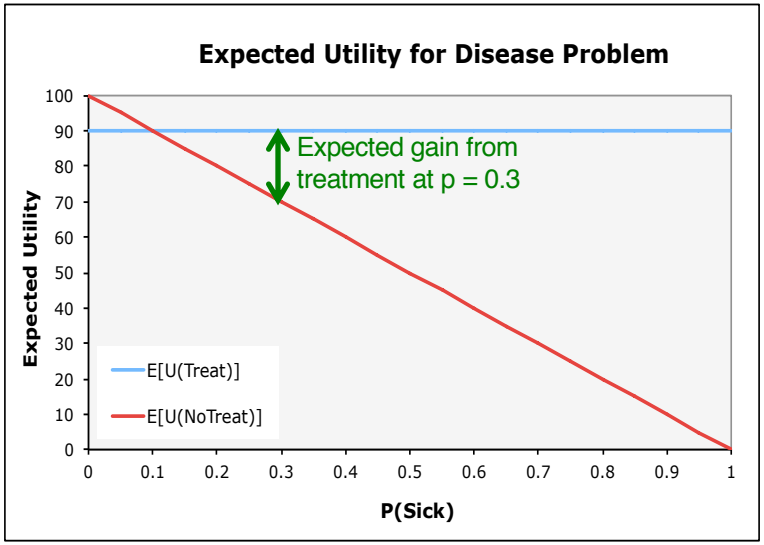


Sensitivity Analysis:
How Optimal Decision Varies with Sickness Probability

- Expected utility of not treating depends on the probability $p = P(s_D)$ of having the disease
  - $E[U|a_T] = 90$
  - $E[U|a_N] = 0p + 100(1-p) = 100(1-p)$
  - We should treat if $p > 0.1$, don't treat if $p < 0.1$
- When we are unsure about the value of p we may want to explore how the optimal decision changes as we vary p
- If our estimate is near the crossover point, we may want to gather information to refine our estimate of p
- *We will use Bayesian inference to refine our estimate of the probability*

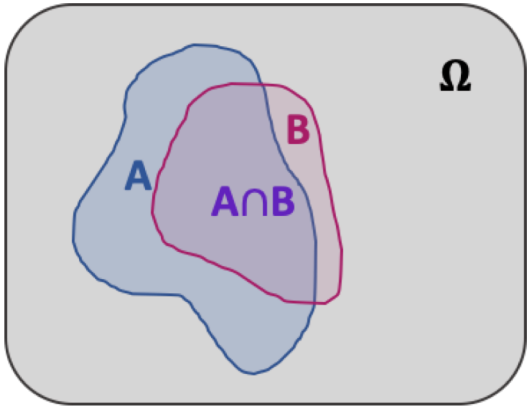
No dependence on p

Decreases as p increases

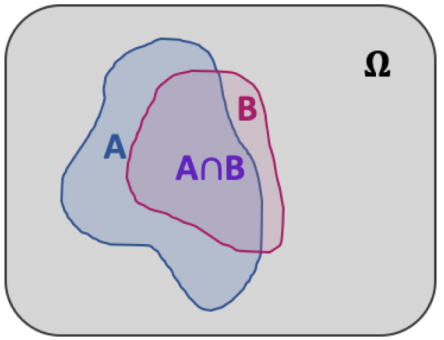


Interlude: Review of Probability Basics

- Probability is a mathematical representation for uncertainty
- We assign probability to **events**
  - An event A is a subset of the **sample space** Ω
- A **probability distribution** is a function on events that satisfies:
  - $P(A) \geq 0$ for all events A
  - $P(\Omega) = 1$
  - If $A_i \cap A_j = \emptyset$, then $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$
- From these properties we can derive others, e.g.:
  - $P(A) \leq 1$ for all events A
  - $P(\emptyset) = 0$
  - If $A \subset B$ then $P(A) \leq P(B)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for events A and B

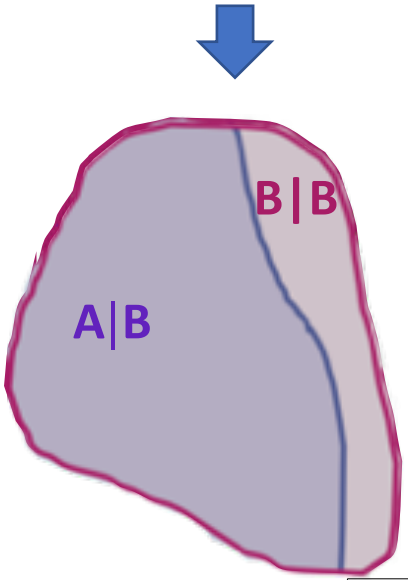


Conditional Probability



- The **conditional probability** $P(A|B)$ satisfies:
  - $P(A|B)P(B) = P(A \cap B)$
  - If $P(B) > 0$ then $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- A and B are **independent** if $P(A|B) = P(A)$
  - This implies $P(A \cap B) = P(A)P(B)$
- The **law of total probability** is:
  - If $B_i \cap B_j = \varnothing$ and $\Omega = B_1 \cup B_2 \cup \cdots$ then

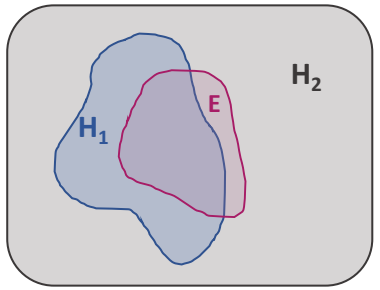
$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$



Bayes Rule: The Law of Belief Dynamics

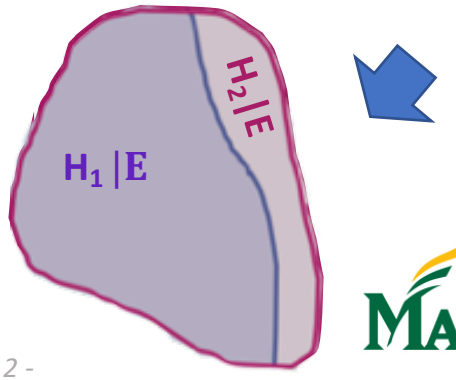
- Objective: use evidence to update beliefs
  - $H_1, \ldots H_n$: exclusive and exhaustive hypotheses $H_i \cap H_j = \varnothing$, $\Omega = H_1 \cup H_2 \cup \cdots$
  - E: evidence (with positive probability) $P(E)>0$
- Procedure: apply **Bayes Rule**:

$$P(H_i|E) = \frac{P(H_i \cap E)}{P(E)} = \frac{P(E|H_i)P(H_i)}{P(E)} = \frac{P(E|H_i)P(H_i)}{\sum_j P(E|H_j)P(H_j)}$$



- Bayes Rule (odds likelihood form):

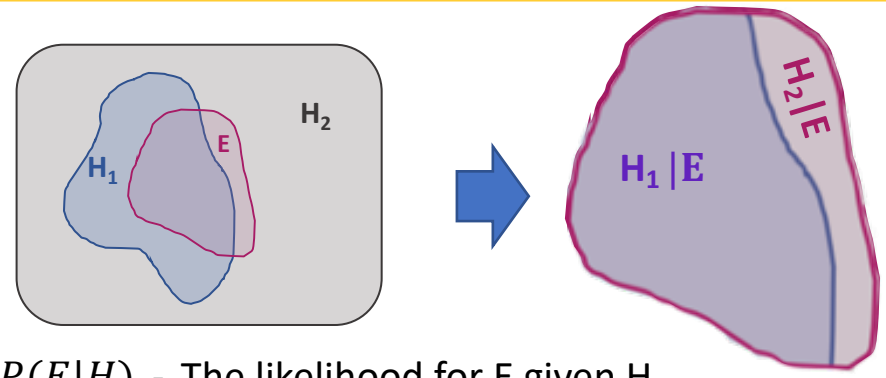
$$\frac{P(H_i|E)}{P(H_j|E)} = \frac{P(E|H_i)P(H_i)}{P(E|H_j)P(H_j)} \quad \left[P(E)>0,\ P(H_2)>0\right]$$



Interpreting Bayes Rule

- Bayes Rule (odds likelihood form):

$$\frac{P(H_i|E)}{P(H_j|E)} = \frac{P(E|H_i)P(H_i)}{P(E|H_j)P(H_j)}$$



- Terminology:

$P(H)$ - The prior probability of H

$P(E)$ - The predictive probability of E

$\frac{P(E|H_i)}{P(E|H_j)}$ - The likelihood ratio for H_i versus H_j

$P(E|H)$ - The likelihood for E given H

$P(H|E)$ - The posterior probability of H given E

$\frac{P(H_i)}{P(H_j)}$ - The prior odds ratio for H_i versus H_j

- *The posterior probability of H_i increases relative to H_j if the evidence is more likely given H_i than given H_j*

Probability Review: Summary

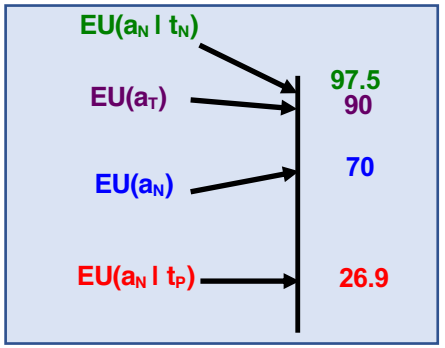
- Events: subsets of sample space
- Probability:
  - Maps event to number between 0 and 1
  - Measures how likely event is to occur
  - Satisfies basic rules
- Conditional probabilities measure how likely an event is given that another event has occurred
- Bayes rule tells us how probabilities change when we get new evidence

Extending the Disease Example: Gathering Information

- We can perform a test before deciding whether to treat the patient
  - Test has two outcomes: t_P (positive) and t_N (negative)
  - Quality of test is characterized by two numbers:
    - *Sensitivity*: Probability that test is positive if patient has disease
    - *Specificity*: Probability that test is negative if patient does not have disease
- Test characteristics:
  - Sensitivity: $P(t_P \mid s_D) = 0.95$
  - Specificity: $P(t_N \mid s_W) = 0.85$
- How does the model change if test results are available?
  - Take test, observe outcome t
  - Revise prior beliefs $P(s_D)$ to obtain posterior beliefs $P(s_D|t)$
  - Re-compute optimal decision using $P(s_D|t)$

Disease Example with Test

- Review of Problem Ingredients:
  - $P(s_D) = 0.3$ (*prior probability of disease*)
  - $P(t_P \mid s_D) = 0.95$; $P(t_N \mid s_W) = 0.85$ (*sensitivity & specificity of test*)
  - $u(c_{WN}) = 100$, $u(c_{WS}) = 90$; $u(c_{DN}) = 0$ (*utilities*)



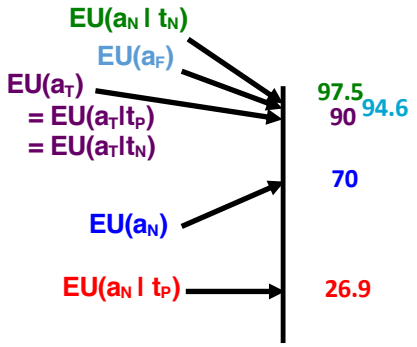
- If negative test: $P(s_D|t_N) = \frac{P(t_N|s_D)P(s_D)}{P(t_N|s_D)P(s_D) + P(t_N|s_W)P(s_W)}$
  - $P(s_D \mid t_N)$ = (0.3 x 0.05)/(0.3 x 0.05 + 0.7 x 0.85) = 0.025
  - $EU(a_N \mid t_N) = 0.025 \times 0 + 0.975 \times 100 = 97.5$
  - $EU(a_T \mid t_N) = 0.025 \times 90 + 0.975 \times 90 = 90$
  - Best action is not to treat
- If positive test: $P(s_D|t_P) = \frac{P(t_P|s_D)P(s_D)}{P(t_P|s_D)P(s_D) + P(t_P|s_W)P(s_W)}$
  - $P(s_D \mid t_P)$ = (0.3 x 0.95)/(0.3 x 0.95 + 0.7 x 0.15) = 0.731
  - $EU(a_N \mid t_P) = 0.731 \times 0 + 0.269 \times 100 = 26.9$
  - $EU(a_T \mid t_P) = 0.731 \times 90 + 0.269 \times 90 = 90$
  - Best action is to treat
- Optimal policy is to treat if positive; don't treat if negative
  - We will call this strategy a_F (FollowTest)

Decision Model: Summary

- To model a decision problem we specify:
  - Possible actions
  - Possible outcomes
  - Probabilities for outcomes given actions
  - Utilities for outcomes
- To find the best decision we calculate the expected utility for each outcome and choose the best
  - Sometimes we minimize loss (negative utility) instead of maximizing utility
- We can gather information to refine our estimate of the probabilities
  - Our optimal decision then depends on what we find out

Should We Gather Information?

- Reminder of problem ingredients:
  - $P(s_D) = 0.3$ (*prior probability of disease*)
  - $u(c_{WN}) = 100$, $u(c_{WS}) = 90$; $u(c_{DN}) = 0$ (*utilities*)
  - $P(t_P \mid s_D) = 0.95$; $P(t_N \mid s_W) = 0.85$ (*sensitivity & specificity of test*)
- Expected utility after doing test:
  - If test is positive we should treat, with $EU(a_T) = 90$
  - If test is negative we should not treat, with $EU(a_N \mid t_N) = 97.5$
- Probability test will be positive (use law of total probability):
  - $P(t_P) = P(t_P \mid s_D)\, P(s_D) + P(t_P \mid s_W)\, P(s_W) = 0.95x0.3 + 0.15x0.7 = 0.39$
- Expected utility of FollowTest strategy (treat if test is positive, otherwise not):
  - $EU(a_F) = P(t_P)\, EU(a_T \mid t_P) + P(t_N)\, EU(a_N \mid t_N)$
    $= 0.39 \times 90 + (1-0.39) \times 97.5 = 94.575$
- $EU(a_F)$ is larger than $EU(a_T) = 90$ so we should do the test



Expected Value of Information

- *Expected Value of Sample Information* (EVSI) is gain in expected utility from doing a test
  - EVSI for our medical example is 94.575 – 90 = 4.575
- Expected Value of Perfect Information (EVPI) is gain in expected utility from perfect knowledge of an uncertain variable
  - For medical example:
    - Suppose an oracle will tell us whether patient is sick
    - 30% chance we discover she is sick and treat - utility 90
    - 70% chance we discover she is well and don't treat - utility 100
    - Expected utility if we ask the oracle 0.3 x 90 + 0.7 x 100 = 97
    - Therefore EVPI = 97 - 90 = 7
- EVPI ≥ EVSI ≥ 0
  - EVSI = 0 if information won't change your decision

Should We Collect Information?

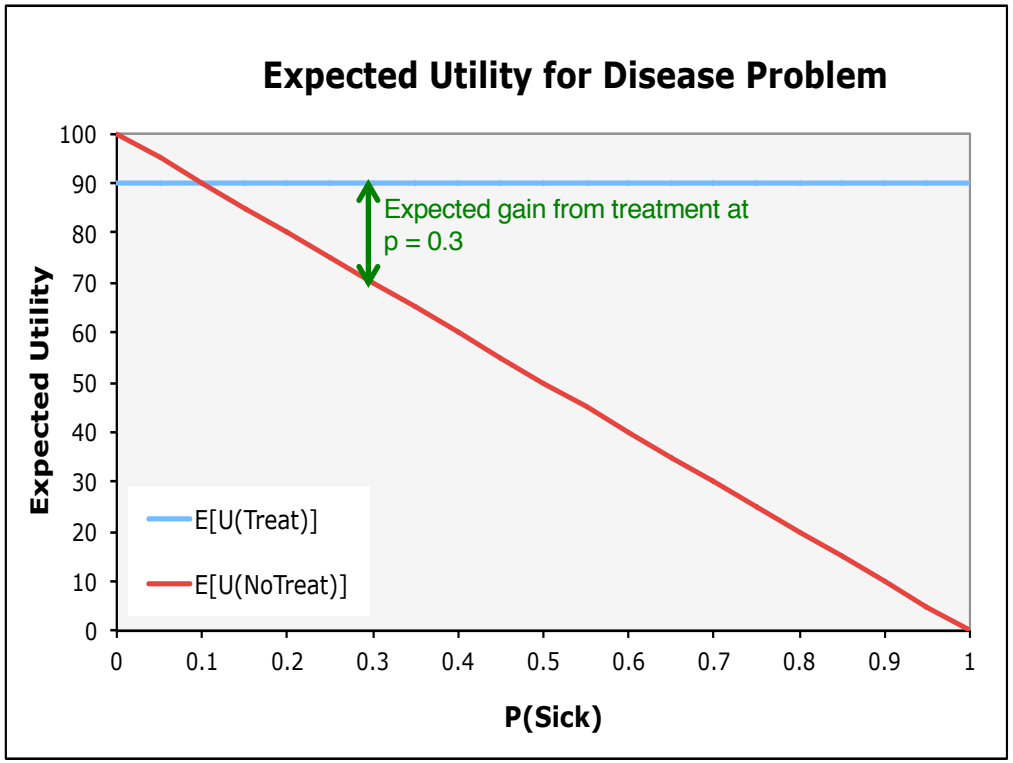
- General Principle: *Free information can never hurt*
- Whether we should do the test depends on whether utility gain EVSI=4.575 is greater than cost of information
- To analyze decision of whether to collect information:
  - Find maximum expected utility option if we don't collect information
  - Compute its expected utility U_0
  - Find EVPI
  - Compare EVPI with cost of information
  - If EVPI is too small in relation to cost then stop; otherwise, compute EVSI
  - Compare EVSI with cost of information
  - Collect information if expected utility gain is greater than cost of information

Strategy Regions for Medical Decision (Without Test)

Expected utility of not treating depends on the probability $p = P(s_D)$ of having the disease

- $E[U|a_T] = 90$
- $E[U|a_N] = 0p + 100(1-p) = 100(1-p)$

- The ***strategy regions*** for the decision (without test):
  - a_T if $p > 0.1$
  - a_N if $p < 0.1$
- What are the strategy regions if we do a test?



Expected Utility of FollowTest Policy as Function of Prior Probability $p = P(s_D)$

- FollowTest strategy treats if test is positive and otherwise not

World State	Probability P(t|s) P(s)	Action	Utility
Sick, Positive	.95*p*	Treat	90
Sick, Negative	.05*p*	NoTreat	0
Well, Positive	.15(1-*p*)	Treat	90
Well, Negative	.85(1-*p*)	NoTreat	100

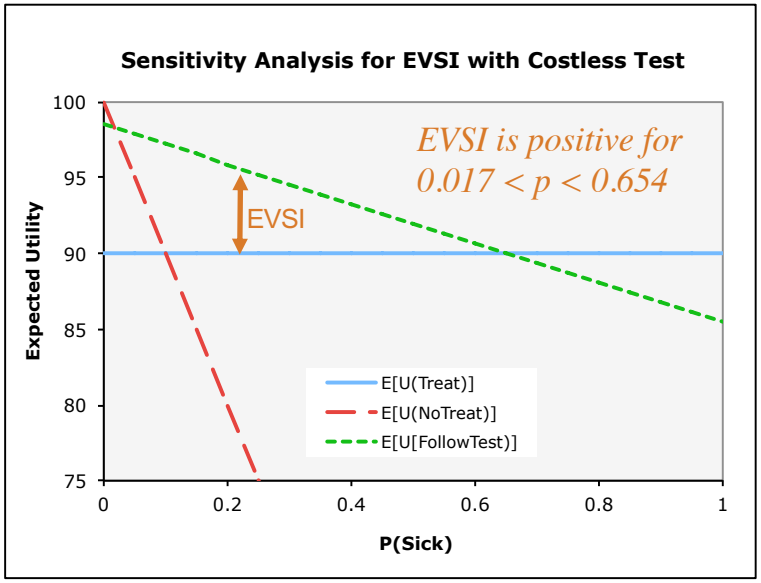
Before doing test, we think:

- *There are four possibilities for disease status and test results. Their probabilities are shown in the table*

  $P(s, t) = P(t|s)\,P(s) = P(s|t)\,P(t)$
- *We treat if test is positive and don't treat if test is negative, with utilities shown in last column.*
- *We multiply probability times utility for each world state and sum to get the expected utility of FollowTest*

$$E[U|a_F] = 0.95p\times 90 + .05p\times 0 + 0.15\,(1-p)\times 90 + 0.85\,(1-p)\times 100$$
$$= 98.5 - 13p$$

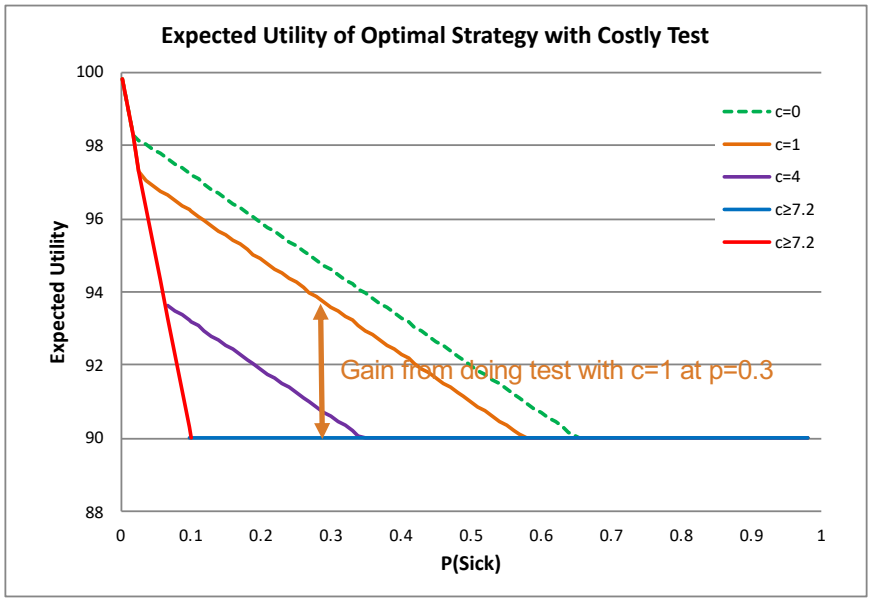
Strategy Regions for Medical Decision (With Test)



Region	Optimal Strategy
$p < 0.017$	NeverTreat
$0.017 < p < 0.654$	FollowTest
$p > 0.654$	AlwaysTreat

- **FollowTest**: $EU(a_F) = 98.5 - 13p$
- **AlwaysTreat**: $EU(a_T) = 90$
  - FollowTest is better when $98.5 - 13p > 90$ or $p < 8.5/13 = 0.654$
- **NeverTreat**: $EU(a_N) = 100(1 - p)$
  - FollowTest is better when $98.5 - 13p > 100(1-p)$ or $p > 1.5/87 = 0.017$

Strategy Regions for Costly Test:



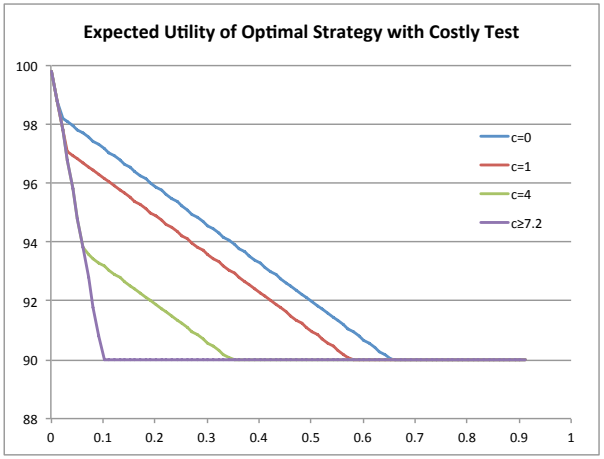
- *Test is worth doing if gain is larger than cost.*
- *Range of values for which test is worth doing:*
  $(1.5+c)/87 < p < (8.5-c)/13$

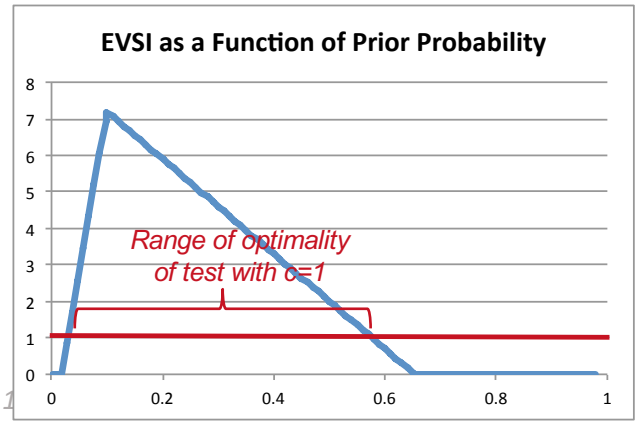
- **FollowTest**: $EU(a_F) = 98.5 - 13p - c$
- **AlwaysTreat**: $EU(a_T) = 90$
  - FollowTest is better when
    $98.5 - 13p - c > 90$ or $p < (8.5-c)/13$
- **NeverTreat**: $EU(a_N) = 100(1 - p)$
  - FollowTest is better when
    $98.5 - 13p - c > 100(1-p)$ or $p > (1.5+c)/87$

(c is cost of test)

EVSI and Costly Test

- Information collection is optimal when EVSI is greater than cost of test
- Probability range where testing is optimal depends on cost of test
- For a test with cost c:
  - Testing is optimal if $(1.5+c)/87 < p < (8.5-c)/13$





Summary : Value of Information and Strategy Regions

- Collecting information may have value if it might change your decision
  - Expected value of perfect information (EVPI) is utility gain from knowing true value of uncertain variable
  - Expected value of sample information (EVSI) is utility gain from available information
- In our example, EVSI is positive for $0.017 < p < 0.654$
  - If $0.017 \le p \le 0.1$ EVSI is $87p - 1.5$
  - If $0.1 \le p \le 0.654$ EVSI is $8.5 – 13p$
  - If $p = 0.3$ EVSI is $8.5 – 13p = 4.6$ (testing is optimal)
- Costly information has value when EVSI is greater than cost of information
- In our example:
  - If $0.017 \le p \le 0.1$ Test if $87p - 1.5 > c$ (where c is cost of test)
  - If $0.1 \le p \le 0.654$ Test if $8.5 – 13p > c$
  - If $p = 0.3$ Test if $4.6 > c$ (test if c is less than 4.6)

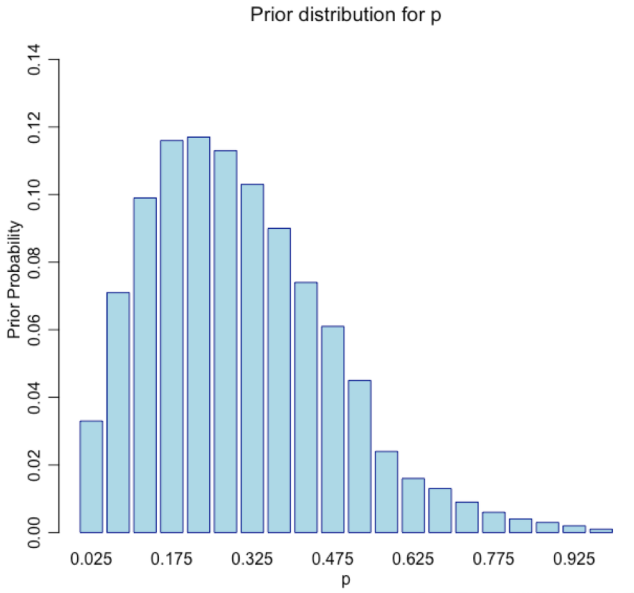
What if a Probability is Unknown?

- The model for our medical example depends on several parameters
  - Prior probability of disease
  - Sensitivity of test
  - Specificity of test
- Usually these probabilities are estimated from data and/or expert judgment
  - "Randomized clinical trials have established that Test T has sensitivity 0.95 and specificity 0.85 for Disease D"
  - "Given the presenting symptoms and my clinical judgment, I estimate a 30% probability that the patient has Disease D."
- How does a Bayesian combine data and expert judgment?
  - Use clinical judgment to quantify uncertainty about as a probability distribution
  - Gather data
  - Use Bayes rule to obtain posterior distribution for the unknown probability
  - If appropriate, use clinical judgment to adjust results of studies to apply to a particular patient

Example: Bayesian Inference about a Probability (with a very small sample)

- Assign prior distribution to possible values of disease probability *p*
  - Although *p* can be any real number between zero and 1, we pretend there are only 20 equally spaced possible values
  - Our prior distribution is consistent with our estimate p=0.3
- Observe 10 *independent and identically distributed* (iid) cases
  - $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 X_{10}) = (0, 1, 0, 0, 0, 0, 1, 0, 0, 1)$
  - Cases 2, 7, and 10 have disease; the rest do not
- How do we find the posterior distribution of the unknown probability?

The unknown probability actually has a continuous range of values. We will treat continuous distributions later. For now we approximate with a finite set of values.

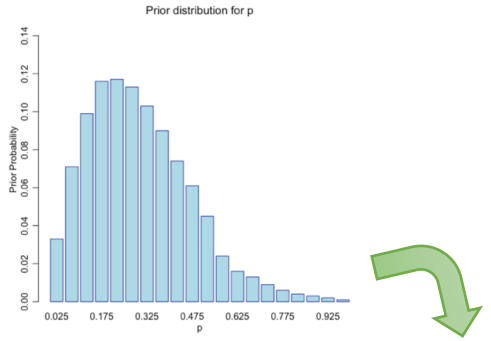


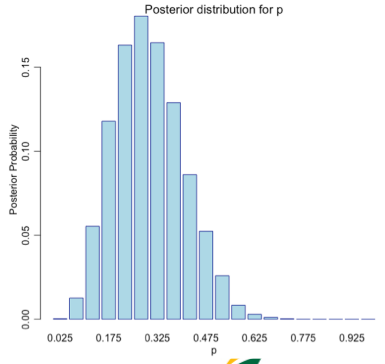
Posterior Distribution of Disease Parameter

- Applying Bayes Rule
  - We observed 3 cases of disease in 10 trials
  - Likelihood of data is $p^3(1-p)^7$
  - Multiply prior $g(p)$ times likelihood $p^3(1-p)^7$ and divide by sum:

$$g(p \mid \underline{x}) = \frac{g(p)p^3(1-p)^7}{\sum_{p'} g(p')p'^3(1-p')^7}$$

- Notice that the posterior distribution depends only on the number of cases with and without the disease
  - Cases with and without the disease are *sufficient* for inference about *p*

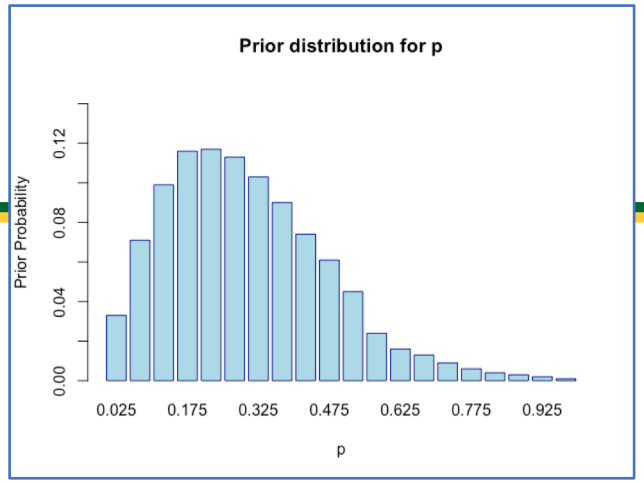


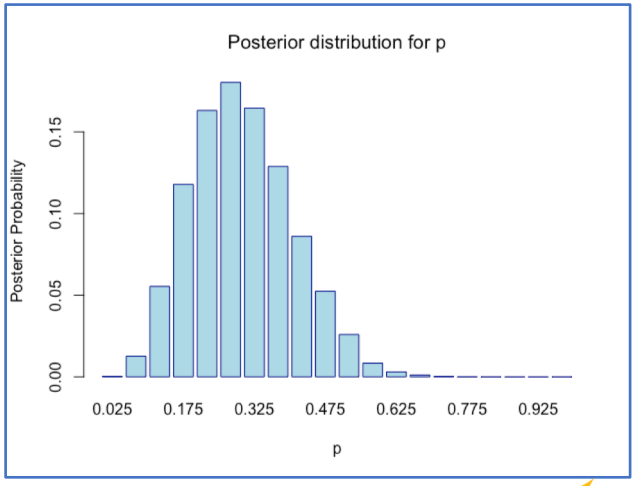


Underscore indicates a vector: $\underline{x}=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})$

Bayesian Inference Example: R Code

```r
# Unit 1 Bayesian Updating Example: Inference about an unknown probability
# Calculate and plot prior and posterior distribution
# for disease probability:  Discretized prior;
# sample of 10 cases, 3 having disease

# Assume the unknow probability p can take on one of 20 evenly spaced values
# between 0.025 and 0.975
pVals <- seq(length=20,from=0.025,to=0.975)

# Prior distribution: choose a prior distribution centered around 0.3
priorDist <- c(0.033, 0.071, 0.099, 0.116, 0.117, 0.113, 0.103,
               0.090, 0.074, 0.061, 0.045, 0.024, 0.016,
               0.013, 0.009, 0.006, 0.004, 0.003, 0.002, 0.001)

# Verify that the expected value of p is 0.3
sum(priorDist*pVals)

# Plot the prior distribution as a bar chart
barplot(priorDist,main="Prior distribution for p",
        xlab="p", ylab="Prior Probability",names.arg=pVals,
        border="darkblue", col="lightblue",ylim=c(0,0.15))

# Calculate the posterior distribution of p after observing sample of
# 10 cases, 3 having disease
numobs=10    # Number of observations
numd=3       # Number having the disease
lik = pVals^numd*(1-pVals)^(numobs-numd)  # Likelihood of data given p
pl <- priorDist * lik                     # prior times likelihood
postDist <- pl/sum(pl)                    # result of Bayes rule

# Plot the posterior distribution
barplot(postDist,main="Posterior distribution for p",
xlab="p", ylab="Posterior Probability",names.arg=pVals,
border="darkblue", col="lightblue")
```





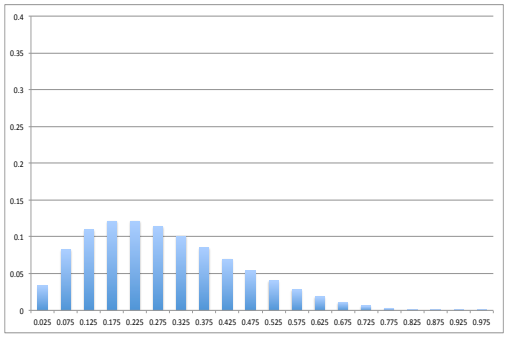
Horizontal axis is p = P(Sick);
height of bar is probability that P(Sick) = p

Bayesian Learning and Sample Size

- When the sample size is very large:
  - The posterior distribution will be concentrated around the maximum likelihood estimate and is relatively insensitive to the prior distribution
  - We won't go too far wrong if we act as if the parameter is equal to the maximum likelihood estimate
- When the sample size is very small:
  - The posterior distribution is highly dependent on the prior distribution
  - Reasonable people may disagree on the value of the parameter
- When the sample size is moderate, Bayesian learning can be a big improvement on either expert judgment alone or data alone
  - Achieving the benefit requires careful modeling
  - This course will teach methods for constructing Bayesian models
- A powerful characteristic of the Bayesian approach is the flexibility to tailor results to moderate-sized sub-populations
  - Bayesian estimate "shrinks" estimates of sub-population parameters toward population average
  - Amount of shrinkage depends on sample size and similarity of sub-population to overall population
  - Shrinkage improves estimates for small to moderate sized sub-populations

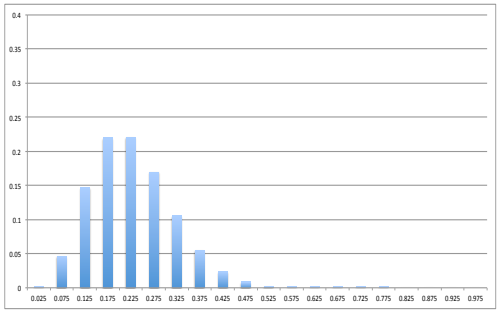
Effect of Sample Size on Posterior Distribution

Sample size 5: 1 with, 4 without

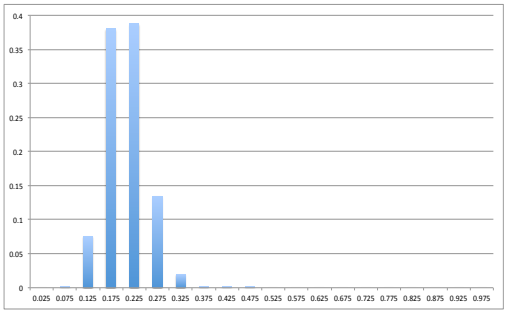


- These plots show the posterior distribution for Θ when:
  - Prior distribution is uniform
  - 20% of patients in sample have the disease
- Posterior distribution becomes more concentrated around 1/5 as sample size gets larger

Sample size 20: 4 with, 16 without



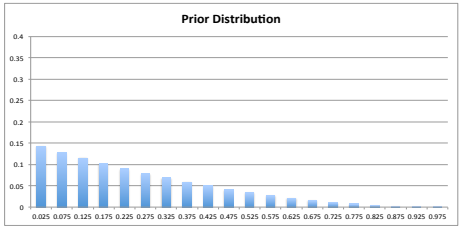
Sample size 80: 16 with, 64 without

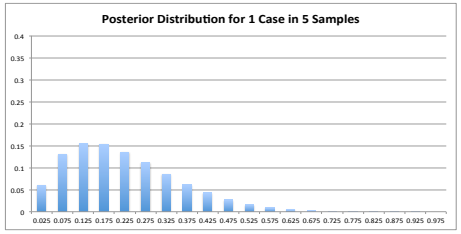


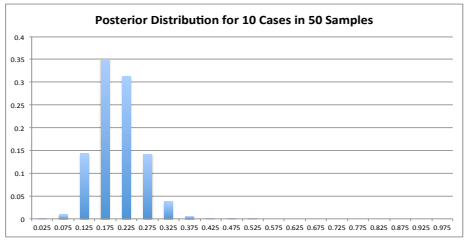
Horizontal axis is $\theta = P(s_D)$; height of bar is probability that $\Theta = \theta$

Sample Size and Impact of the Prior Distribution

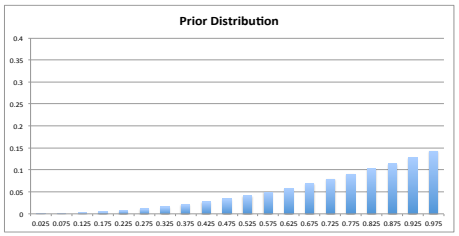
- Prior distribution favors low probabilities:

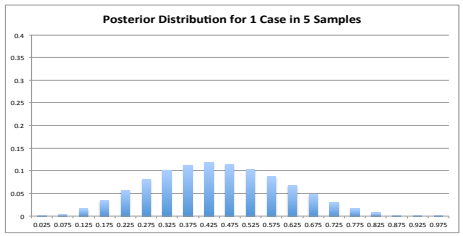


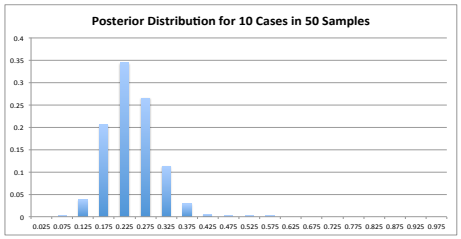




- Prior distribution favors high probabilities:







Horizontal axis is θ = $P(s_D)$; height of bar is probability that $\Theta = \theta$

- Bayesian inference "shrinks" posterior distribution toward prior expectations
  - Posterior distribution for smaller sample is more sensitive to prior distribution
  - Posterior distribution for larger sample is less sensitive to prior distribution

Some Concepts of Probability

- **Classical** - Probability is a ratio of favorable cases to total (equipossible) cases
- **Frequency** - Probability is the limiting value as the number of trials becomes infinite of the frequency of occurrence of a type of event
- **Logical** - Probability is a logical property of one's state of information about a phenomenon
- **Propensity** - Probability is a propensity for certain kinds of physical event to occur
- **Subjective** - Probability is an ideal rational agent's degree of belief about an uncertain event
- **Algorithmic** - The algorithmic probability of a finite sequence is the probability that a universal computer fed a random input will give the sequence as output (related to Kolmogorov complexity)
- **Game Theoretic** - Probability is an agent's optimal "announced certainty" for an event in a multi-agent game in which agents receive rewards that depend on both forecasts and outcomes

Probability *really is* none of these things.
Probability *can represent* all of these things.

The Frequentist

- A frequentist believes:
  - Probability can be legitimately applied only to repeatable problems
  - Probability is an objective property in the real world
  - Probability applies only to random processes
  - Probabilities are associated only with collectives not individual events
- Frequentist Inference
  - Data are drawn from a distribution of known form but with an unknown parameter (this includes "nonparametric" statistics in which the unknown parameter is the distribution itself)
  - Distribution may arise from explicit randomization or may be considered "close enough" to random
  - Inference treats data as random and parameter as fixed
  - For example: A sample $X_1,\ldots X_N$ is drawn from a normal distribution with mean Θ . A 95% confidence interval is constructed. The interpretation is:

    If an experiment like this were performed many times we would expect in 95% of the cases that an interval calculated by the procedure we applied would include the true value of Θ .
- A frequentist can say nothing about *any individual experiment* about θ!

The Subjectivist

- A subjectivist believes:
  - Probability as an expression of a rational agent's degrees of belief about uncertain propositions.
  - Rational agents may disagree. There is no "one correct probability."
  - If the agent receives feedback her assessed probabilities will in the limit converge to observed frequencies
- Subjectivist Inference:
  - Probability distributions are assigned to unknowns (parameters and observations).
  - Condition on knowns; use probability to express uncertainty about unknowns
  - For example: A sample $X_1,\ldots X_N$ is drawn from a normal distribution with mean θ having prior distribution $g(\theta)$. A 95% posterior credible interval is constructed, and the result is the interval (3.7, 4.9). The interpretation is:

    Given prior distribution for θ and observed data, the probability that θ lies between 3.7 and 4.9 is 95%.
- A subjectivist can draw conclusions about what we should believe about θ and about what we should expect on the next trial

The Bayesian Resurgence

- Bayesian inference is as old as probability
- Subjective view fell into disfavor in 19th and early 20th centuries
  - Positivism, empiricism, and quest for objectivity in science
  - "Paradoxes" and systematic deviation of human judgment from Bayesian "norm"
- There has been a recent resurgence
  - Computational advances make calculation possible for complex models
  - Bayesian models can coherently integrate many different kinds of information
    - Physical cause and effect
    - Logical implication
    - Informed expert judgment
    - Empirical observation
  - Unified theory and methods for data-rich and data-poor problems
  - Clear connection to decision making

Comparison: Understandability, Subjectivity and Honest Reporting

- Often the Bayesian answer is what the decision maker really wants to hear.
- Untrained people often interpret results in the Bayesian way.
- Frequentists are disturbed by dependence of the posterior interval on "subjective" prior distribution.

  It is more important that stochastics provides a means of communication among researchers whose personal beliefs about the phenomena under study may differ. If these beliefs are allowed to contaminate the reporting of results, … how are the results of different researchers to be compared?

  - H. Dinges

- Bayesians say the prior distribution is not the only subjective element in an analysis.
- Bayesian probability statements are always subjective, but statistical analyses are often done for public consumption. Whose probability distribution should be reported?
  - For large samples, a good Bayesian analysis and a good frequentist analysis will often be consistent
  - If results are sensitive to the prior distribution, a Bayesian analyst should report this sensitivity and present a range of results obtained from a range of prior distributions

Comparison: Generality

- Subjectivists can handle problems the frequentist approach cannot (in particular, problems with not enough data for sound frequentist inference).
- Frequentist statisticians say this comes at a price -- when there are not enough data the result will be highly dependent on the prior distribution.
- Subjectivists often apply frequentist techniques but with a Bayesian interpretation
- Frequentists often apply Bayesian methods if they have good frequency properties

Coherence and Rationality

- In the mid 20th century, several authors proposed systems of axioms intended to characterize rational behavior
  - Proofs that decision-makers satisfying these axioms must be expected utility maximizers
  - Proofs that decision-makers not satisfying these axioms are vulnerable to exploitation ("Dutch book")
  - Well-documented systematic departures of human decision-making from expected utility maximization
- A decision-maker is called *coherent* if she behaves as a maximizer of expected utility
- Should coherence be equated with rationality?

Axioms for Probability

De Groot, 1970

There is a qualitative relationship of relative likelihood $\prec$, that operates on pairs of events, that satisfies the following conditions:

SP1. For any two uncertain events A and B, one of the following relations holds: $A \prec B$, $A \succ B$ or $A \sim B$.

SP2. If A_1, A_2, B_1, and B_2 are four events such that $A_1 \cap A_2 = \varnothing$, $B_1 \cap B_2 = \varnothing$, and if $A_i \preceq B_i$, for $i = 1,2$, then $A_1 \cup A_2 \preceq B_1 \cup B_2$. If in addition $A_i \prec B_i$ for either $i=1$ or $i=2$, then $A_1 \cup A_2 \prec B_1 \cup B_2$.

SP3. If A is any event, then $\varnothing \preceq A$. Furthermore, there is *some* event A_0 for which $\varnothing \prec A_0$.

SP4. If $A_1 \supset A_2 \supset \ldots$ is a decreasing sequence of events, and B is some event such that $A_i \succeq B$ for $i=1, 2, \ldots$, then $\bigcap_{i=1}^{\infty} A_i \succeq B$.

SP5. There is an experiment, with a numerical outcome between the values of 0 and 1, such that if A_i is the event that the outcome x lies within the interval $a_i \leq x \leq b_i$, for $i=1,2$, then $A_1 \prec A_2$ if and only if $(b_1 - a_1) \leq (b_2 - a_2)$.

Axioms for Utility

Watson and Buede, 1987

A *reward* is a prize the decision maker cares about. A *lottery* is a situation in which the decision maker will receive one of the possible rewards, where the reward to be received is governed by a probability distribution. There is a qualitative relationship of relative preference $\prec^*$, that operates on lotteries, that satisfies the following conditions:

SU1. For any two lotteries L_1 and L_2, either $L_1 \prec^* L_2$, $L_1 \succ^* L_2$, or $L_1 \sim^* L_2$. Furthermore, if L_1, L_2, and L_3 are any lotteries such that $L_1 \prec^* L_2$ and $L_2 \prec^* L_3$, then $L_1 \prec^* L_3$.

SU2. If r_1, r_2 and r_3 are rewards such that $r_1 \prec^* r_2 \prec^* r_3$, then there exists a probability p such that $[r_1: p; r_3: (1\text{-}p)] \sim^* r_2$, where $[r_1:p; r_3:(1\text{-}p)]$ is a lottery that pays r_1 with probability p and r_3 with probability $(1\text{-}p)$.

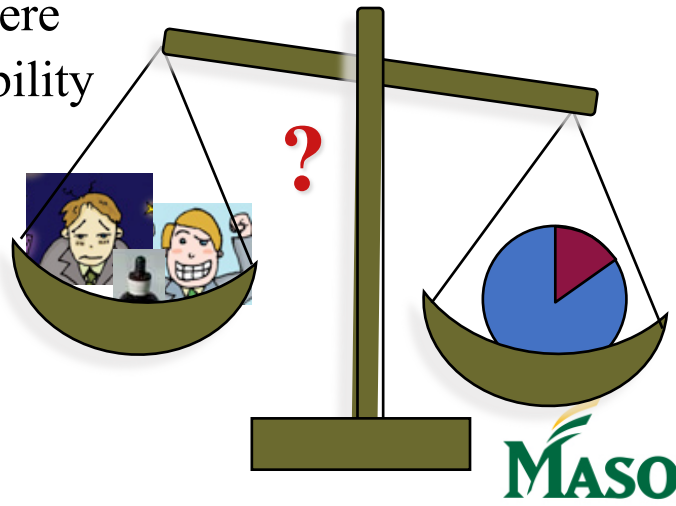
SU3. If $r_1 \sim^* r_2$ are rewards, then for any probability p and any reward r_3, $[r_1: p; r_3: (1\text{-}p)] \sim^* [r_2: p; r_3: (1\text{-}p)]$

SU4. If $r_1 \succ^* r_2$ are rewards, then $[r_1: p; r_2: (1\text{-}p)] \succ^* [r_1: q; r_2: (1\text{-}q)]$ if and only if $p > q$.

SU5. Consider three lotteries, $L_i = [r_1: p_i; r_2: (1\text{-}p_i)]$, $i = 1, 2, 3$, giving different probabilities of the two rewards r_1 and r_2. Suppose lottery M gives entry to lottery L_2 with probability q and L_3 with probability $1\text{-}q$. Then $L_1 \sim^* M$ if and only if $p_1 = qp_2 + (1\text{-}q)p_3$.

Probabilities and Utilities

- If your beliefs satisfy SP1-SP5, then there is a probability distribution Pr(·) over events such that for any two events A_1 and A_2, $\Pr(A_1) \geq \Pr(A_2)$ if and only if $A_1 \succeq A_2$.
- If your preferences satisfy SU1-SU5, then there is a utility function $u(\cdot)$ defined on rewards such that for any two lotteries L_1 and L_2, $L_1 \succeq^* L_2$ if and only if $\mathrm{E}[u(L_1)] \geq \mathrm{E}[u(L_2)]$, where $\mathrm{E}[\cdot]$ denotes the expected value with respect to the probability distribution Pr(·).



Why be a Bayesian?

- Arguments from theory
  - A *coherent* decision maker uses probability to represent uncertainty, uses utility to represent value, and maximizes expected utility
  - If you are not coherent then someone can make "Dutch book" on you (turn you into a "money pump")
- Pragmatic arguments
  - Useful and principled methodology for modeling inference, decision and learning
  - Analyze engineering tradeoffs between accuracy, complexity and cost
  - Represent and incorporate both empirical data and informed engineering judgment
  - Handle small, moderate and large sample sizes and parameter sets
  - Interpretability of results and understandability of model
- Arguments from experience
  - Successful applications attributed to decision theory

What do you think?

Unit 1: Summary and Synthesis

- Bayesian statistics is a theory of rational belief dynamics
- We took a broad-brush tour of Bayesian methodology
- We applied Bayesian thinking to a simplified medical example that illustrates many of the concepts we will be learning this semester
- Bayesian decision theory provides a methodology for rational choice under uncertainty
- The twentieth century has seen a resurgence of interest in subjective probability and an increased understanding of the appropriate role of subjectivity in science
  - Most statistics texts and courses take a frequentist approach but this is changing
  - The inventors of probability theory thought of it as a logic of enlightened rational reasoning. In the nineteenth century this was replaced by a view of probability as measuring "objective" propensities of "intrinsically random" phenomena
  - Bayesian methods often require more computational power than traditional frequentist methods
  - The computer revolution has enabled the Bayesian resurgence

References for Unit 1

- Bayes, Thomas. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370- 418, 1763.
- Bashir, S.A., *Getting Started in R*, http://www.luchsinger-mathematics.ch/Bashir.pdf
- Dawid, A.P. and Vovk, V.G. (1999), Prequential Probability: Principles and Properties, *Bernoulli*, 5: 125-162.
- de Finetti, Bruno. *Theory of Probability: A Critical Introductory Treatment*. New York: Wiley, 1974.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D., Bayesian Data Analysis (2nd edition), Chapman & Hall, 2004. Chapter 1
- Hájek, Alan, "Interpretations of Probability", *The Stanford Encyclopedia of Philosophy (Summer 2003 Edition)*, Edward N. Zalta/\(ed.), URL = <http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/>.
- Lee, P. *Bayesian Statistics: An Introduction, 4th ed.* Springer, 2012. Chapter 1
- Li, Ming and Vitanyi, Paul. *An Introduction to Kolmogorov Complexity and Its Applications.* (2nd ed) Springer-Verlag, 2005.
- Nau, Robert F. (1999), *Arbitrage, Incomplete Models, And Interactive Rationality*, working paper, Fuqua School of Business, Duke University.
- Neapolitan, R. *Learning Bayesian Networks,* Prentice Hall, 2003.
- Jaynes, E., *Probability Theory: The Logic of Science*, Cambridge University Press, 2003)
- Savage, L.J., *The Foundations of Statistics*. Dover, 1972.
- Shafer, G. *Probability and Finance: It's Only a Game*, Wiley, 2001.
- von Mises R., 1957, *Probability, Statistics and Truth*, revised English edition, New York: Macmillan