# Introduction to Bayesian Inference

# Example: Medical Diagnosis (Gigerenzer, 2002)

- Consider women aged 40-50 with no family history of cancer and no symptoms of cancer
  - The proportion that have breast cancer is .008
- Conduct mammogram screening
  - If a woman has breast cancer, the probability of a positive mammogram is .90
  - If a woman does not have breast cancer, the probability of a positive mammogram is .07
- A woman undergoes the mammogram, the result is positive
- What should we infer?
  - What's the probability the woman has breast cancer?

# *Maximum Likelihood*

# Logic of Maximum Likelihood

- A general approach to parameter estimation
- The use of a model implies that the data may be sufficiently characterized by the features of the model, including the unknown parameters
- Parameters **govern** the data in the sense that the data **depend** on the parameters
  - Given values of the parameters we can calculate the (conditional) probability of the data
  - Mammogram (data) depends on breast cancer status (parameter)
  - When conventional statistical approaches discuss a "model" they usually refer to this dependence structure
- Maximum likelihood (ML) estimation asks: *"What are the values of the parameters that make the data most probable?"*

# Maximum Likelihood

- Specify the conditional probability of the data as $p(x \mid \theta)$
  - E.g., $\theta = (\mu, \sigma^2)$ $p(x \mid \mu, \sigma^2) = N(\mu, \sigma^2)$
  - This describes the structure as function of $x$

- When taken as a function of $\theta$, this is referred to as the ***likelihood***

$$L(\theta \mid x) = p(x \mid \theta)$$

- ML estimation then maximizes this function w.r.t. $\theta$, using the known values of $x$

# Medical Diagnosis
# Conditional Probability Distribution

- If a woman has breast cancer, the probability of a positive mammogram is .90

$$p(Mam = + \mid BC = Y) = .90 \quad p(Mam = - \mid BC = Y) = .10$$

- If a woman does not have breast cancer, the probability of a positive mammogram is .07

$$p(Mam = + \mid BC = N) = .07 \quad p(Mam = - \mid BC = N) = .93$$

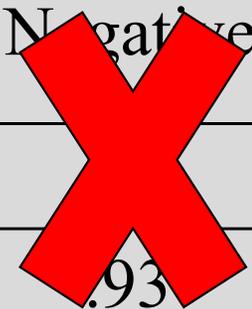|  | *Mammogram Result* | |
|---|---|---|
| *Breast Cancer* | Positive | Negative |
| Yes | .90 | .10 |
| No | .07 | .93 |

# Medical Diagnosis Maximum Likelihood

- A woman undergoes the mammogram, the result is positive

$$L(BC = Y \mid Mam = +) = p(Mam = + \mid BC = Y) = .90$$

$$L(BC = N \mid Mam = +) = p(Mam = + \mid BC = N) = .07$$

| | Mammogram Result | |
|---|---|---|
| *Breast Cancer* | Positive | Negative |
| Yes | .90 | |
| No | .07 | .93 |

- ML: *What is the value of BC that maximizes this? BC = Yes*

# *Bayesian Inference*

# Example: Medical Diagnosis (Gigerenzer, 2002)

- Consider women aged 40-50 with no family history of cancer and no symptoms of cancer
  - The proportion that have breast cancer is .008
- Conduct mammogram screening
  - If a woman has breast cancer, the probability of a positive mammogram is .90
  - If a woman does not have breast cancer, the probability of a positive mammogram is .07
- A woman undergoes the mammogram, the result is positive
- What should we infer?
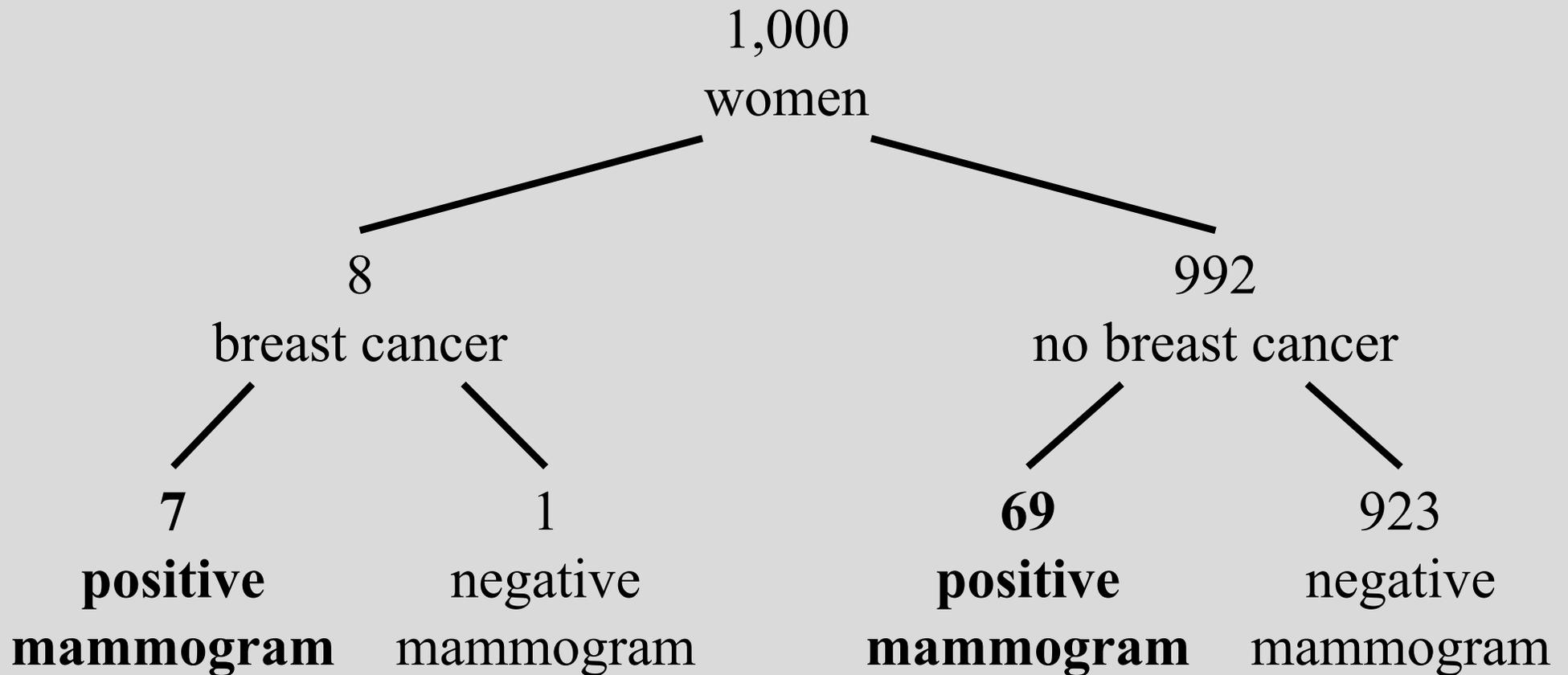  - What's the probability the woman has breast cancer?

# Example: Medical Diagnosis Natural Frequencies

- Consider 1,000 women aged 40-50 with no family history of cancer and no symptoms of cancer
  - 8 of the 1,000 women have breast cancer
- Conduct mammogram screening
  - Of the 8 women with breast cancer, 7 will have a positive mammogram
  - Of the remaining 992 women w/o breast cancer, 69 will have a positive mammogram
- A woman undergoes the mammogram, the result is positive
- What should we infer?
  - What's the probability the woman has breast cancer?

# Example: Medical Diagnosis
# Natural Frequencies

1,000
women

8
breast cancer

992
no breast cancer

**7**

**positive**
**mammogram**

1

negative
mammogram

**69**

**positive**
**mammogram**

923

negative
mammogram

$$p(Breast\ Cancer = \text{Yes} \mid Mammogram = \text{Positive}) = \frac{7}{7+69} = \frac{7}{76} \approx .09$$

# Bayes' Theorem In Action

- The setup
  - Two entities: $x$ (data) and $\theta$ (parameter)
  - We know the conditional probabilities $p(x \mid \theta)$, which tell us what to believe about $x$ if we knew the value of $\theta$
  - When we learn the value of $x$, what should we believe about $\theta$?
- We combine three things (two of which are important)
  - Relative conditional probabilities for $x$ given $\theta$, the likelihood
  - Previous probabilities about $\theta$
  - The marginal probability for $x$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

# Conditional Probability or Likelihood

Conditional probability of $x$ given $\theta$; likelihood for $\theta$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

Relative probability of $x$ given the value of $\theta$

Once value of $x$ is known, and this is viewed as a function of $\theta$ it is a ***likelihood***

# Medical Diagnosis
# Conditional Probability Distribution

- If a woman has breast cancer, the probability of a positive mammogram is .90

$$p(Mam = + \mid BC = Y) = .90 \quad p(Mam = - \mid BC = Y) = .10$$

- If a woman does not have breast cancer, the probability of a positive mammogram is .07

$$p(Mam = + \mid BC = N) = .07 \quad p(Mam = - \mid BC = N) = .93$$

| | Mammogram Result | |
|---|---|---|
| Breast Cancer | Positive | Negative |
| Yes | .90 | .10 |
| No | .07 | .93 |

# The Prior Distribution

Everything we know about $\theta$ **before** we observe value for $x$

Prior probability distribution for unknown $\theta$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

# Medical Diagnosis Prior Distribution for *Breast Cancer*

- The proportion of women aged 40-50 with no family history of cancer and no symptoms that have breast cancer is .008

- Assign a prior distribution

*p(Breast Cancer = Yes) = .008*

*p(Breast Cancer = No) = .992*

| Breast Cancer | |
|---|---|
| Yes | .008 |
| No | .992 |

# Marginal Probability

Conditional probability of $x$ given $\theta$; likelihood for $\theta$

Prior probability distribution for unknown $\theta$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

Marginal probability of $x$ (over $\theta$)

$$p(x) = \sum_{\theta} p(x \mid \theta) p(\theta)$$

Serves to normalize the distribution
Note that $p(x)$ does not vary with $\theta$
We will eventually discard it

# Medical Diagnosis
## Marginal Probability of *Mammogram* = +

$$p(x) = \sum_{\theta} p(x \mid \theta) p(\theta)$$

$p(x \mid \theta)$

|  | Mammogram | |
|---|---|---|
| BC | + | - |
| Yes | .90 | |
| No | .07 | .93 |

$\times$

$p(\theta)$

| BC | |
|---|---|
| Yes | .008 |
| No | .992 |

$=$

$p(x , \theta)$

| BC | |
|---|---|
| Yes | .00720 |
| No | .06944 |

$\Sigma_{\theta}$   .07664

# Posterior Distribution

Conditional probability of $x$ given $\theta$; likelihood for $\theta$

Prior probability distribution for unknown $\theta$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

Posterior probability distribution for unknown $\theta$ given $x$

Marginal probability of $x$ (over $\theta$)

Captures what we think about $\theta$ now that we have incorporated $x$

# Medical Diagnosis Posterior Distribution

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)} = \frac{p(Mam = + \mid BC)\, p(BC)}{p(Mam = +)}$$

| | Mammogram | |
|---|---|---|
| BC | + | - |
| Yes | .90 | |
| No | .07 | |

$\times$

| BC | |
|---|---|
| Yes | .008 |
| No | .992 |

$=$

| BC | |
|---|---|
| Yes | .00720 |
| No | .06944 |

.07664

.07664

The ***posterior probability*** that the woman has breast cancer

$=$

| BC | |
|---|---|
| Yes | .09 |
| No | .91 |

# Bayesian Inference as Updating for Probability-Based Reasoning

*"Prior to the mammogram, we didn't think the woman had breast cancer. After the positive result on the mammogram…we don't think the woman had breast cancer. So it seems the mammogram is irrelevant, or Bayes doesn't work"*

*"We are uncertain if the woman has breast cancer. And our language for uncertainty is probabilities. Before the mammogram, the probability of breast cancer is $p(BC = Y) = .008$. After the mammogram, the probability of breast cancer is $p(BC = Y \mid M = +) = .09$. These probabilities are expressions of our beliefs. We still think it's unlikely the patient has breast cancer, but our beliefs have shifted considerably from what they were."*

# Proportionality in the Posterior Distribution

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)} \propto p(x \mid \theta)\, p(\theta)$$

Marginal probability of $x$ (over $\theta$)
***Does not change for different values of $\theta$***

Discarding yields proportionality

# Medical Diagnosis
# Proportionality of Posterior Distribution

$$p(BC \mid Mam) = \frac{p(Mam = + \mid BC)\, p(BC)}{p(Mam = +)} \propto p(Mam = + \mid BC)\, p(BC)$$

| | Mammogram | |
|---|---|---|
| BC | + | - |
| Yes | .90 | |
| No | .07 | |

$\times$

| BC | |
|---|---|
| Yes | .008 |
| No | .992 |

$=$

| BC | |
|---|---|
| Yes | .00720 |
| No | .06944 |

.07664

.07664

Does not change for different values of $\theta$
Dividing simply serves to ***normalize*** the distribution

| BC | |
|---|---|
| Yes | .09 |
| No | .91 |

# *General Approach to Bayesian Modeling*

# General Approach to Bayesian Modeling

A ***Bayesianly justifiable*** analysis is one that

"treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes' theorem."

-- Rubin (1984, p. 1152)

# 3 Step General Approach to Bayesian Modeling

1.  Set up the full probability model: the joint distribution of all entities, including observables ($x$) and unobservables ($\theta$) in accordance with all that is known about the problem

2.  Condition on the observed data ($x$), calculate the conditional probability distribution for the unobservable entities ($\theta$) of interest given the observed data: the posterior distribution

3.  Examine fit, tenability/sensitivity of assumptions, reasonable conclusions?, respecify, summarize results, etc.

# Step 1

1.  Set up the full probability model: the joint distribution of all entities, including observables ($x$) and unobservables ($\theta$) in accordance with all that is known about the problem

$$p(x, \theta)$$

Difficult to do as a multivariate system

Joint probability of breast cancer and mammogram results

| | Mammogram | |
|---|---|---|
| *BC* | + | - |
| Yes | 0.0072 | 0.0008 |
| No | 0.06944 | 0.92256 |

# Step 1

1.  Set up the full probability model via a conditional distribution $p(x \mid \theta)$, and a prior $p(\theta)$

$$p(x, \theta) = p(x \mid \theta)\, p(\theta)$$

–   The prior probability of cancer

–   The conditional probability of the mammogram result, given cancer status

| BC  |      |
| --- | ---- |
| Yes | .008 |
| No  | .992 |

|     | Mammogram | |
| --- | --- | --- |
| BC  | +   | -   |
| Yes | .90 | .10 |
| No  | .07 | .93 |

# Step 2

2. Condition on the observed data ($x$), calculate the conditional probability distribution for the unobservable entities ($\theta$) of interest given the observed data: obtain the posterior distribution as

$$p(\theta \mid x) = \frac{p(x, \theta)}{p(x)}$$

$$= \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

$$\propto\ p(x \mid \theta)\, p(\theta)$$

# Bayes' Theorem for Discrete Variables

Conditional probability of $x$ given $\theta$; likelihood for $\theta$

Prior probability distribution for unknown $\theta$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)} = \frac{p(x \mid \theta)\, p(\theta)}{\sum_{\theta} p(x \mid \theta)\, p(\theta)} \propto p(x \mid \theta)\, p(\theta)$$

Posterior distribution for unknown $\theta$ given $x$

Marginal probability of $x$ (over $\theta$), does not depend on $\theta$, removal yields proportionality

# Bayes' Theorem for Continuous Variables

Conditional probability of $x$ given $\theta$; likelihood for $\theta$

Prior probability distribution for unknown $\theta$

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)} = \frac{p(x \mid \theta)\, p(\theta)}{\int_{\theta} p(x \mid \theta)\, p(\theta)\, d\theta} \propto p(x \mid \theta)\, p(\theta)$$

Posterior distribution for unknown $\theta$ given $x$

Marginal probability of $x$ (over $\theta$), does not depend on $\theta$, removal yields proportionality

# Bayes' Theorem

Effects a reversal of the conditional probability, $p(x \mid \theta) \rightarrow p(\theta \mid x)$

$$p(\theta \mid x) \propto p(x \mid \theta)\, p(\theta)$$

From $p(Mam \mid BC)$ to $p(BC \mid Mam)$

Confusion of these conditional probabilities is quite common

# *Contrasting Frequentist & Bayesian Inference (So Far)*

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | | |
| Prior | | |
| Solution | | |

- Data viewed as random, conditionally distributed given parameters
  - *Mammogram* is conditionally distributed given *BC*

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | Fixed (mostly) | |
| Prior | | |
| Solution | | |

- Frequentist inference: parameters are *fixed* and unknown
  - Frequentist perspective defines probabilities as long-run freq.
  - We don't know if the woman has breast cancer, but she either does or she doesn't, hardly something that invokes long-run frequencies

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | Fixed (mostly) | |
| Prior | No | |
| Solution | | |

- Random entities are assigned distributions
- Parameters aren't random → not assigned distributions

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | Fixed (mostly) | |
| Prior | No | |
| Solution | Point (e.g., ML) | |

- ML estimation attempts to find the value of the parameters that make the data seem most likely
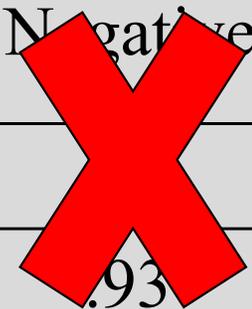
# Medical Diagnosis Maximum Likelihood

- A woman undergoes the mammogram, the result is positive

$$L(BC = Y \mid Mam = +) = p(Mam = + \mid BC = Y) = .90$$

$$L(BC = N \mid Mam = +) = p(Mam = + \mid BC = N) = .07$$

| | Mammogram Result | |
|---|---|---|
| *Breast Cancer* | Positive | Negative |
| Yes | .90 | |
| No | .07 | .93 |

- ML: *What is the value of BC that maximizes this? BC = Yes*

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | Fixed (mostly) | Random (mostly) |
| Prior | No | (even if not |
| Solution | Point (e.g., ML) | philosophically) |

- In a Bayesian analysis all unknown entities treated as random and assigned a distribution
    - Initially the data, but once observed, they are conditioned on
    - The parameters as well

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | Fixed (mostly) | Random |
| Prior | No | Yes |
| Solution | Point (e.g., ML) | |

- All unknown entities are assigned distributions
  - Don't know if the patient has breast cancer? Assign a distribution!

# Frequentist vs. Bayesian Inference

| Characteristic | Frequentist | Bayesian |
|---|---|---|
| Status of Data | Random | Random until obs. |
| Contribution of Data | Likelihood | Likelihood |
| Status of Parameters | Fixed (mostly) | Random |
| Prior | No | Yes |
| Solution | Point (e.g., ML) | Distribution |

- The "solution/answer" in a Bayesian analysis is a *distribution*

- Not trying to find a point, trying to find a distribution

  – Not trying to find the *estimate* of the parameter

  – Trying to find the *distribution* of a parameter

  – Can summarize if desired in the usual ways

# ML vs. Bayesian Estimation

- ML seeks to *maximize*     $L(\theta \mid x) = p(x \mid \theta)$
- Bayes seeks to *obtain*     $p(\theta \mid x) \propto p(x \mid \theta)\, p(\theta)$


- Finding a point vs. finding a distribution
- Other difference is the presence of the prior $p(\theta)$

*Reasoning As A Bayesian?*

# Do We Reason As A Bayesian?

- Active debate in psychology, neuroscience
- Some examples of natural occurrences in the research base and everyday life…but no
- Breast cancer example drawn from work of Gigerenzer; 24 physicians were explicitly asked for the probability that the woman has breast cancer:
  - 8 said < 10%, 8 said = 90% 8 said between 50% and 80%
  - Most not thinking about the prior, some not thinking of the data
  - Only 2 gave correct reasoning, others seemed to try to do so, but came up with the wrong answer
  - With natural frequencies, majority close to correct
- Research program by Kahneman & Tversky

# *Should* We Reason As A Bayesian?

- Characterization as "optimal" by psychology, neuroscience
- Should we include prior probability, condition on data, obtain posterior probabilities?
  - *Should we include the fact that only the proportion in the population that have breast cancer is .008?*
- If so, why?
  - Better descriptions of problems, models, and inference
  - Probability-based reasoning
  - Managing uncertainty
  - Synthesizing information over time
  - Has been shown to be effective in many disciplines
  - The answer to all your questions…
    http://en.wikipedia.org/wiki/Monty_Hall_problem

# *Concluding Thoughts & Summary*

# General Approach to Bayesian Modeling

A ***Bayesianly justifiable*** analysis is one that

"treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes' theorem."

-- Rubin (1984, p. 1152)

# 3 Step General Approach to Bayesian Modeling

1. Set up the full probability model: the joint distribution of all entities, including knowns ($x$) and unknowns ($\theta$) in accordance with all that is known about the problem

$$p(x, \theta) = p(x \mid \theta)\, p(\theta)$$

Common (shortcut) approach to specifying models for folks new to Bayesian analysis, but familiar with other approaches

A. Set up the model for the conditional probability of the data as you are used to: $p(x \mid \theta)$

B. List out all the unknown parameter(s) ($\theta$)

C. Specify prior distribution(s) for the parameter(s) $p(\theta)$, reflecting what is believed about the situation

# 3 Step General Approach to Bayesian Modeling

1. Set up the full probability model: the joint distribution of all entities, including knowns ($x$) and unknowns ($\theta$) in accordance with all that is known about the problem

$$p(x, \theta) = p(x \mid \theta)\, p(\theta)$$

2. Condition on the observed data ($x$), calculate the conditional probability distribution for the unknown entities ($\theta$) of interest given the observed data: the posterior distribution

$$p(\theta \mid x) = \frac{p(x, \theta)}{p(x)} = \frac{p(x \mid \theta)\, p(\theta)}{p(x)} \propto p(x \mid \theta)\, p(\theta)$$

3. Examine fit, tenability/sensitivity of assumptions, reasonable conclusions?, respecify, summarize results, etc.

# Wise Words on Probability

Probability is not really about numbers;
it is about the structure of reasoning

-- Glenn Shafer, quoted in Pearl, 1988, p. 77

# Summary

- Frequentist inference via ML
  - Overview of conceptions
  - Likelihood = conditional probability
- Bayesian inference
  - Prior, likelihood, marginal, posterior
  - Proportionality of posterior
  - Bayes as updater in probability-based reasoning
- Contrasting frequentist and Bayesian inference
- It's all about reasoning