Statistical models & inference:

estimate and predict uncertain quantities, events

Probability: Measurement of *uncertainty*

Interpretations?

- Counting: equally likely outcomes (card and dice games, sampling)
- Frequency: rain tomorrow? Relative frequency of rain on "similar" days
- Subjective: measure of belief, judgement
  - includes counting, frequency cases, BUT:
  - unique events – the 2001 Duke/UNC Game
  - conditioning: rain tomorrow conditional on – ?
    Rain today ? WeatherChannel forecast?

PROBABILITY AND INFERENCE: INTRODUCTION

"Simplest" statistical inference context: DIAGNOSIS example

- $x = 0, 1$
  - Data, to be observed, result of experiment/observation
  - (Binary) Sample Space $x \in \{0, 1\}$ – discrete, simple
- $\theta = 0, 1$
  - Parameter, "Truth," state of nature, uncertain/unknown
  - Primary objective: Inference on θ
  - (Binary) Parameter Space $\theta \in \{0, 1\}$ – discrete, simple
- Uncertainty about outcomes described by conditional probabilities $p(x|\theta)$

NOTATION: $p(x|\theta)$, $p(x = 1|\theta = 1)$, $Pr(x = 1|\theta = 1)$

- $p(x|\theta)$ defines the Sampling model
  Where do these probabilities come from?
  Assumptions about Population and Random Sampling
- Measures of likelihood of observed data given states of nature
  - $x = 1$ is 9.9 times as likely when $\theta = 1$ than when $\theta = 0$
  - Likelihood ratios $r(x) = p(x|\theta = 1)/p(x|\theta = 0)$
  - $r(1) = 9.9$, $r(0) = 0.0111...$
- Evidence, Information to compare $\theta = 1 : 0$ given data x

What is the probability of disease?

- Which probability?
- Are "you" diseased?
- Based on what evidence and information?

Initial probability

Based on all past data, assumptions $\rightarrow$ "information set" H

- Take this as the population disease incidence rate (/100)
- Initial probability = Prior probability $p(\theta = 1)$
- Formally, explicitly: $p(\theta = 1) \equiv p(\theta = 1|H)$
  - Where does this probability come from?
  - Assumptions about Population and Random Sampling

- Odds on disease (initial or prior odds), Odds = Prob/(1-Prob) or

$$o(\theta = 1) = \frac{p(\theta = 1)}{(1 - p(\theta = 1))} = \frac{p(\theta = 1)}{p(\theta = 0)}$$

- Prob = Odd/(1+Odds) or $p(\theta = 1) = o(\theta = 1)/(1 + o(\theta = 1))$
- e.g.: $p(\theta = 1) = 0.01$ and $o(\theta = 1) = 0.0101...$

JOINT PROBABILITY DISTRIBUTION: $p(x, \theta) = p(x|\theta)p(\theta)$

- Marginal distribution for x via *Law of Total Probability*:

$$p(x) = p(x|\theta = 0)p(\theta = 0) + p(x|\theta = 1)p(\theta = 1)$$

- Bayes' theorem:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}$$

- Bayes' theorem reverses conditioning
- Now x is fixed, conditioned upon
- Initial/prior probability distribution for θ is revised/updated
- Final probability = Posterior probability $p(\theta|x)$

In odds form,

$$o(\theta = 1|x) = o(\theta = 1)r(x)$$

based on observed likelihood ratio $r(x) = p(x|\theta = 1)/p(x|\theta = 0)$

Posterior odds = Prior odds × Likelihood ratio

- Scientific learning/Statistical Inference: $p(\theta) \to p(\theta|x)$
- Mapping via likelihood function $p(x|\theta)$ from sampling model
- Formally, $p(\theta|H) \to p(\theta|H, x)$ where H = prior information
- *Technically:* Simply apply Bayes' theorem

"Final" inferences depend on both data x and prior information H

Sensitivity to prior:

- *Any* prior $a = p(\theta = 1)$ implies posterior $p(\theta|x)$
- Variation with respect to a for given data, hence fixed $r(x)$ ... ?
- graph $p(\theta = 1|x)$ versus a

Statistics By Example

Your mischievous brother owns a coin which you know to be loaded so that it comes up heads 70% of the time. He comes to you with a coin and wants to make a bet with you. You don't know if this is the loaded coin or a fair one. He lets you flip it 5 times to check it out, and you get 2 heads and 3 tails. Which coin do you think it is?

1. What is our unknown parameter of interest?
2. What is our data?
3. What is our probability model for the data?
4. How do we estimate our unknown parameter?

The Unknown Parameter of Interest

We want to know which coin it is. We could denote the coin type by θ, where $\theta \in \{\text{fair}, \text{loaded}\}$.

The Data

Our data are the 5 flips of the coin. The observed values are 2 heads and 3 tails. We could write X for the number of heads observed in 5 flips.

The Probability Model for the Data

We use a binomial model for the data, where the probability of a single flip being heads depends on the value of θ:

$$
\begin{aligned}
P(X = x|\theta = \text{fair}) &= \binom{5}{x}\left(\frac{1}{2}\right)^5 \\
P(X = x|\theta = \text{loaded}) &= \binom{5}{x}(.7)^x(.3)^{5-x}
\end{aligned}
$$

Mathematically, this model is just a function of two things: x and θ. For a particular value of θ, the model gives us a probability model for X (the probability of observing different values of x). Once we have observed X (here $X = 2$), we can think of the probability model as a function of θ alone. This function is called the likelihood and is denoted $L(\theta)$. This gives the probability of the particular observed data as a function of θ.

Estimating the Unknown Parameter - I

One way to make statements about the value of θ is Maximum Likelihood. The idea is that our best guess of the true value of θ is the value that maximizes the likelihood.

$$f(\theta|X=2) = \begin{cases} 0.3125 & \text{if } \theta = \text{fair} \\ 0.1323 & \text{if } \theta = \text{loaded} \end{cases}$$

So in this example, we would guess that the coin is fair.

Some problems with this approach:

- It can be hard to make useful statements of our confidence/estimation error. (In particular, we can't give a probability of being right.)
- It can be hard to incorporate other information.

Estimating the Unknown Parameter - II

A different (and our preferred) approach to making statements about the value of θ is the Bayesian approach. In this case we think of θ as a random variable. We start with a prior distribution for θ, $P(\theta)$, (representing our beliefs before seeing the data). We then observe the data and compute a posterior distribution, $P(\theta|X)$, by using Bayes' Theorem.

$$
\begin{aligned}
P(\theta|X) &= \frac{P(\theta, X)}{P(X)} \\
&= \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta} \\
&\propto L(\theta)P(\theta)
\end{aligned}
$$

Example

Since you know your brother reasonably well, you might think that there is a 60% chance that this is the loaded coin, before you get to flip it 5 times. You then get 2 heads and 3 tails. What is your posterior probability that the coin is loaded?

$$
\begin{aligned}
Pr(\text{loaded}|X) &= \\
&= \frac{Pr(X|\text{loaded})Pr(\text{loaded})}{Pr(X|\text{loaded})Pr(\text{loaded}) + Pr(X|\text{fair})Pr(\text{fair})} \\
&= \frac{\binom{5}{2}(.7)^2(.3)^3(.6)}{\binom{5}{2}(.7)^2(.3)^3(.6) + \binom{5}{2}(.5)^2(.5)^3(.4)} \\
&= \frac{0.07938}{0.07938 + 0.125} = 0.388
\end{aligned}
$$

Prior Sensitivity

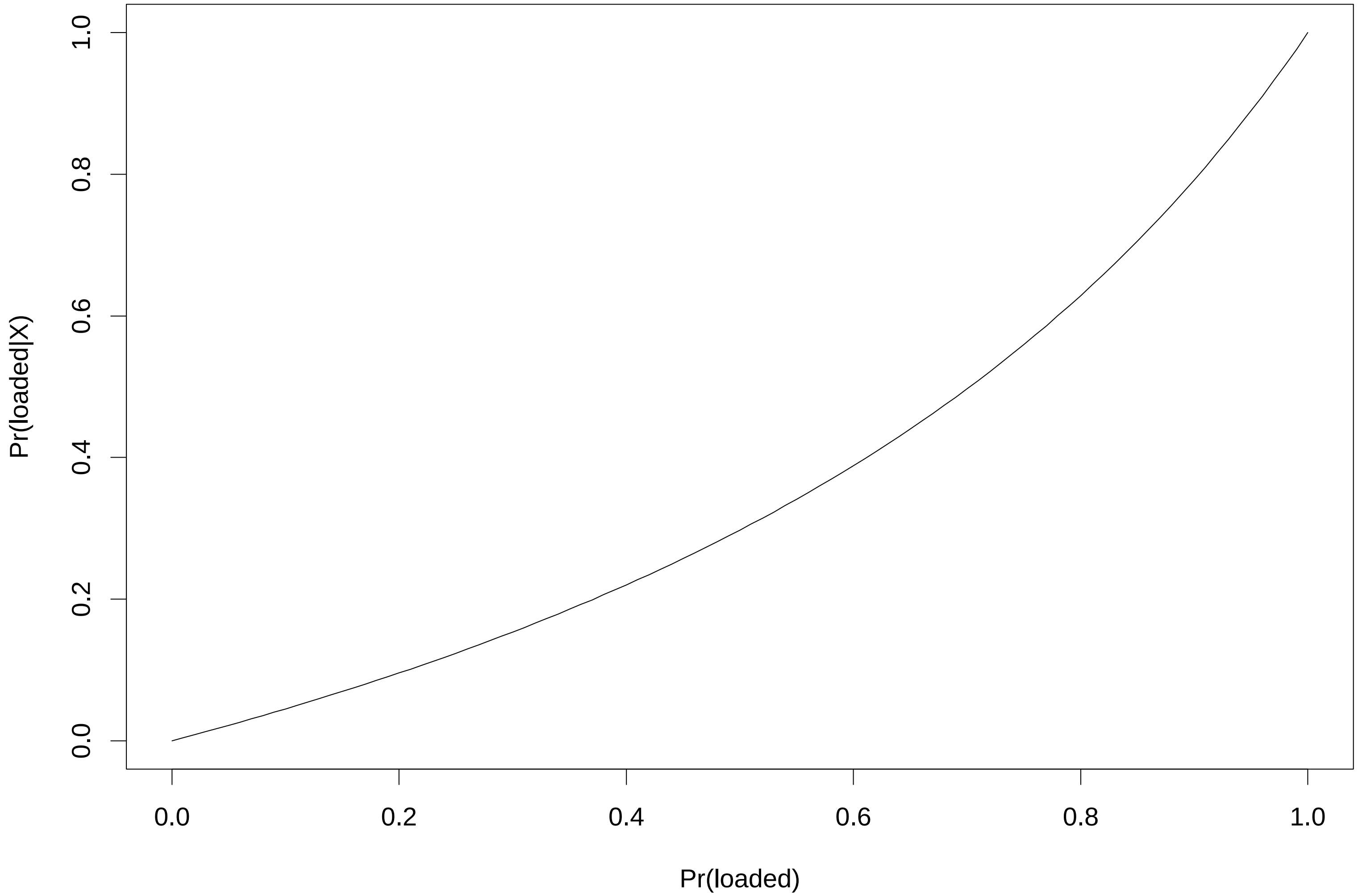
We could have made a different choice for a prior:

- $Pr(\text{loaded}) = .9 \Rightarrow Pr(\text{loaded}|X) = 0.792$
- $Pr(\text{loaded}) = .5 \Rightarrow Pr(\text{loaded}|X) = 0.297$

Note that the "non-informative" prior gives the "same" answer as the MLE. Also note that since the sample size is small, the result is highly sensitive to the choice of prior. As the sample size increases, the prior matters less.

Advantages of this approach:

- Our answer is a probability distribution for θ, so we can make probability statements and error estimates.
- We can incorporate other information into the prior.


Pr(loaded|X)
0.0
0.2
0.4
0.6
0.8
1.0
Pr(loaded)