Statistical Inference: Maximum Likelihood and Bayesian Approaches

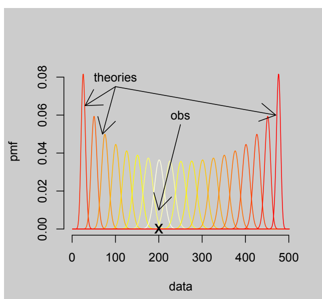
Surya Tokdar

From model to inference

- So a statistical analysis begins by setting up a model $\{f(x|\theta) : \theta \in \Theta\}$ for data X.
- Next we observe our actual data $X = x$.
- The pdfs/pmfs included in our model represent theories, the observations x is evidence.
- The goal of inference is to compare between these theories in light of the recorded evidence.

Example: Opinion poll

- n: # sampled = 500
- X: # in favor
- Model: $X \sim Bin(n, p)$, $p \in [0, 1]$.
- Obs: $X = 200$.



The likelihood function

- Clearly observed data will better match the prediction of some theories than others.
- In other words, some theories will better predict the particular observation than other theories.
- Better prediction means assigning a higher probability to observing $X = x$
- So we can assign scores to theories by this function of θ:

$$L_x(\theta) = f(x|\theta), \theta \in \Theta$$

- This is called the likelihood score/function.

Some words on the likelihood function

- $L_x(\theta)$ is a function of $\theta \in \Theta$.
  - it depends on the observed data x,
  - but for any single data analysis x is a fixed quantity.
- $\frac{L_x(\theta_1)}{L_x(\theta_2)} = 2$ implies the observed data is two times more likely to appear under theory θ_1 than under theory θ_2.
- For all technical purposes, one can work with $L_x(\theta)$ in the log-scale. That is, define the log-likelihood function

$$\ell_x(\theta) = \log L_x(\theta) = \log f(x|\theta), \theta \in \Theta.$$

- Log-scale comparisons are done by $\ell_x(\theta_1) - \ell_x(\theta_2)$.

Opinion poll likelihood

- Model $X \sim Bin(n, p)$, $p \in [0, 1]$. So

$$L_x(p) = \binom{n}{x} p^x (1-p)^{n-x}, \ p \in [0, 1]$$

and the log-likelihood function is

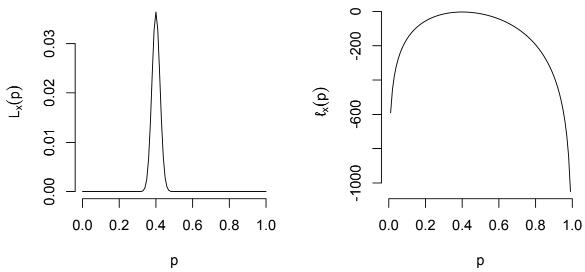
$$\ell_x(p) = \log \binom{n}{x} + x \log p + (n-x)\log(1-p), \quad p \in [0, 1]$$

- The first term on the r.h.s. does not involve p. So we write

$$\ell_x(p) = \text{const} + x \log p + (n-x)\log(1-p), \quad p \in [0, 1]$$

and don't care about the exact value of "const".
- Indeed, "const" disappears in differences $\ell_x(p_1) - \ell_x(p_2)$.

Graphs of likelihood and log-liklihood



Learning from the likelihood function

- Two goals
  1. To report a subset of attractive theories.
  2. To test a scientific hypothesis $\theta \in \Theta_0$, a subset of Θ.
- These may not be the only/most important goals
  - But capture the essence of "inference"
  - We'll get into other goals later
- Two aproaches to use $L_x(\theta)$ or $\ell_x(\theta)$ to come up with and interpret such a subset
  1. The maximum likelihood (ML) approach
  2. The Bayesian approach

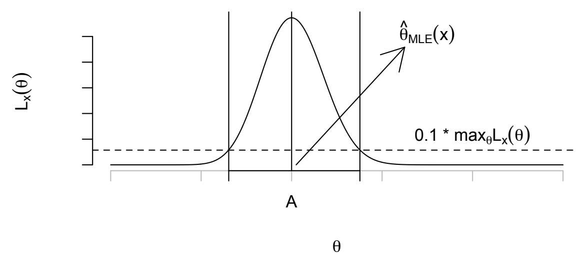
The ML approach

- Use $L_x(\theta)$ to split the parameter space into two subsets
  1. subset of "well supported" θ with high $L_x(\theta)$
  2. subset of "not-so-well-supported" θ with low $L_x(\theta)$.
- Can effect such a split by
  1. fixing a $k \in [0, 1]$
  2. setting the first set as

$$A_k(x) = \left\{ \theta \in \Theta : L_x(\theta) \geq k \max_{\tilde{\theta} \in \Theta} L_x(\tilde{\theta}) \right\}.$$

- Report support toward $\theta \in \Theta_0$ if $\Theta_0 \cap A_k \neq \emptyset$.

Graphical representation of ML approach



Choice of the threshold k

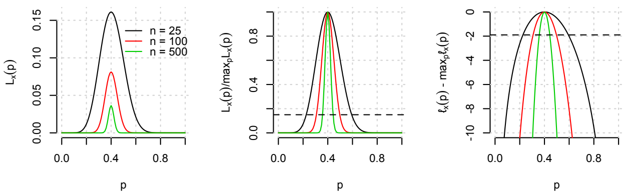
- With $k = 1$ we only report theories with the highest score,
  1. i.e., the set of maximum points of $L_x(\theta)$.
  2. Often there is one single point at which maximum is attained.
  3. When this happens, the maximum point is called the maximum likelihood estimate (MLE) and is denoted $\hat{\theta}_{MLE}(x)$.
- With $k = 0$ we report the whole set Θ – not making any use of the data.

Important questions

- How to choose k?
- Should the same k be chosen for all types of models?
- In the opinion poll example, should we use the same k when $n = 50$ as we do for $n = 500$?
- All these would boil down to:
  How to interpret the choice of k in a quantitative manner and how to communicate it to a reader?
- We will find an answer through the paradigm of classical statistics.

Back to Opinion poll

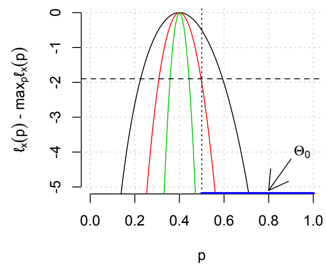
- Data X = number of students (out of n) in favor of a policy.
- Statistical model: $\{Bin(n, p) : p \in [0, 1]\}$.
- 3 cases: $n = 25, X = 10$; $n = 100, X = 40$; $n = 500, X = 200$



- Same $\hat{p}_{\text{MLE}}(x) = x/n = 0.4$, different $A_k(x)$ with $k = 0.15$.
  - $[0.23, 0.59]$; $[0.31, 0.49]$; $[0.36, 0.44]$.

More in favor than not?

- Any support toward $\theta \geq 0.5$?
- $A_{0.15}(x) \cap [0.5, 1] \neq \emptyset$ only for $(n = 25, X = 15)$.



The Bayesian approach

- Convert $L_x(\theta)$ into a plausibility score on Θ.
  - Uncertainty about any unknown quantity can be summarized by a pdf/pmf.
  - The parameter θ is one such quantity.
- Must have a pdf/pmf to describe θ before we observe data and one to describe it after we make the observation.

Prior and posterior

- Augment the model with a prior pdf $\pi(\theta)$ on Θ.
- $\pi(\theta)$ is the pre-data/*a priori* quantification of one's uncertainty about θ, with relative plausibility scores given by $\pi(\theta_1)/\pi(\theta_2)$.
- The post-data/*a posteriori* relative plausibility scores are

$$\frac{\pi(\theta_1|x)}{\pi(\theta_2|x)} = \frac{\pi(\theta_1)}{\pi(\theta_2)} \times \frac{L_x(\theta_1)}{L_x(\theta_2)}$$

  and correspond to the *posterior pdf*

$$\pi(\theta|x) = \frac{L_x(\theta)\pi(\theta)}{\int_\Theta L_x(\theta')\pi(\theta')d\theta'}, \quad \theta \in \Theta$$

From prior to posterior

- Posterior formula is not ad-hoc: driven by probability theory!
- A model $\{f(x|\theta) : \theta \in \Theta\}$, coupled with the prior $\pi(\theta)$ gives a joint quantification of (X, θ) as:

$$(X \mid \theta) \sim f(x|\theta), \quad \theta \sim \pi(\theta),$$
$$\text{i.e.,} \quad (X, \theta) \sim g(x, \theta) = f(x \mid \theta)\pi(\theta)$$

  where $g(x, \theta)$ is a pdf over $S \times \Theta$.
- By Bayes theorem, the conditional pdf of θ given $X = x$ is

$$\pi(\theta|x) = \frac{g(x, \theta)}{\int_\Theta g(x, \theta')d\theta'} = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta')\pi(\theta')d\theta'}$$

  which is just as before because $L_x(\theta) = f(x|\theta)$.

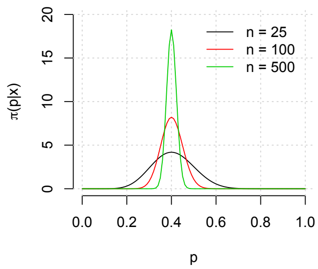
Reporting a Bayesian analysis

- $\pi(\theta|x)$ captures the entire post-data quantification of the uncertainty about θ.
- A report is essentially visual/numerical summaries of this pdf.
  - A plot of $\pi(\theta|x)$, if available, is most useful!
  - Numerical summaries include quantiles, mean, standard deviation, mode, high density regions, etc.
- 0.025-th & 0.975-th quantiles give a 95% posterior range of θ
- To evaluate evidence toward $\theta \in \Theta_0$, simply calculate

$$\Pr(\theta \in \Theta_0|x) = \int_{\Theta_0} \pi(\theta|x)d\theta.$$

Back to opinion poll

- For opinion poll example, take $\pi(\theta)$ to be the *Unif*$(0, 1)$ pdf.



- 95% posterior range: $[0.24, 0.59]$; $[0.31, 0.50]$; $[0.36, 0.44]$.
- $\Pr(\theta \geq 0.5|x)$: 0.163, 0.023, 0.00000369.

A two parameter problem

- Lactic acid concentrations $X_1, \cdots, X_n$ measured from cheese samples
- Model: $X_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$
- Model parameters $\mu \in (-\infty, \infty)$, $\sigma^2 > 0$.
- Care only about μ
  - Get a range for μ.
  - Is $\mu \leq 1$?

Why "problem"?

- Likelihood function $L_x(\mu, \sigma^2)$ compares (μ, σ^2) pairs.
- How to do it for μ alone?

ML approach: Profile likelihood

- ML reports a well supported set A_k of (μ, σ^2) values
- Look at all distinct values of μ that appear in A_k (paired with some σ^2). Report this set.
- Same as doing the following
  - Define profile likelihood: $L_x^*(\mu) = \max_{\sigma^2} L_x(\mu, \sigma^2)$.
  - Fix threshold $k \in [0, 1]$.
  - Report $A_k^*(x) = \{\mu : L_x^*(\mu) \geq k \max_{\mu'} L_x^*(\mu')\}$.

Bayes approach: marginal posterior pdf

- Prior pdf $\pi(\mu, \sigma^2)$ leads to posterior pdf $\pi(\mu, \sigma^2|x)$.
- But this describes a joint distribution of (μ, σ^2) given $X = x$.
- Interested only in μ? Integrate out σ^2

$$\pi^*(\mu|x) = \int \pi(\mu, \sigma^2|x) d\sigma^2$$

  and summarize μ based on the marginal pdf $\pi^*(\mu|x)$.

Integrated likelihood

- The marginal prior pdf is $\pi^*(\mu) = \int \pi(\mu, \sigma^2) d\sigma^2$.
- Conditional prior pdf of σ^2 given μ is $\tilde{\pi}(\sigma^2|\mu) = \frac{\pi(\mu, \sigma^2)}{\pi^*(\mu)}$.
- The marginal posterior pdf satisfies

$$\pi^*(\mu|x) = \frac{\tilde{L}_x(\mu)\pi^*(\mu)}{\int \tilde{L}_x(\mu')\pi^*(\mu')d\mu'}$$

  where

$$\tilde{L}_x(\mu) = \int L_x(\mu, \sigma^2)\tilde{\pi}(\sigma^2|\mu) d\sigma^2.$$

- Bayes: consider average support for μ over all σ^2.
- ML: consider maximum support for μ at the best σ^2.