About this class

The next two lectures are really coming from a statistics perspective, but we're going to discover how useful it is for the problems we are interested in!

Chapter 7 of Casella and Berger is a good reference for this material (most of this lecture is based on that chapter).

Statistics thinks largely about *samples*, particularly random samples.

Random variables (X_i): Functions from sample space to $\mathbb{R}$

Realized values of random variables: x_i

Random sample of size n from population $f(x)$: $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables with pdf or pmf $f(x)$

Point Estimators

Let's say we have a stream of values all coming from the same population (no changing with time): $x_1, \ldots, x_n$

Suppose the population is described by a pdf $f(x|\theta)$

We want to estimate θ

An *estimator* is a function of the sample: $X_1, \ldots, X_n$.

An *estimate* is a number, which is a function of the realized values $x_1, \ldots, x_n$

Think of an estimator as an algorithm that produces estimates when given its inputs

Can you think of a good estimator for the population mean?

Maximum Likelihood

Method for deriving estimators.

Let $\mathbf{x}$ denote a realized random sample

Likelihood function:

$$L(\theta|\mathbf{x}) = L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

If $\mathbf{X}$ is discrete, $L(\theta|\mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x})$

Intuitively, if $L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x})$ then θ_1 is in some ways a more plausible value for θ than is θ_2

Can be generalized to multiple parameters $\theta_1, \ldots, \theta_n$

Maximum Likelihood

For a sample $\mathbf{x} = x_1, \ldots, x_n$ let $\widehat{\theta}(\mathbf{x})$ be the parameter value at which $L(\theta|\mathbf{x})$ attains its maximum (as a function of θ, with $\mathbf{x}$ held fixed).

Then $\widehat{\theta}(\mathbf{x})$ is the maximum likelihood estimate of θ based on the realized sample $\mathbf{x}$. $\widehat{\theta}(\mathbf{X})$ is the maximum likelihood estimator based on the sample $\mathbf{X}$.

Note that the MLE has the same range as the parameter, by definition

Potential problems

- How to find and verify the maximum of the function?

- Numerical sensitivity

Differentiable Likelihood Functions

Possible candidates are the values of $\theta_1, \ldots \theta_k$ that solve:

$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0, (i = 1, \ldots, k)$$

Must check whether any such value of θ is in fact a global maximum (could be a minimum, an inflection point, a local maximum, and the boundary needs to be checked).

Normal MLE

Suppose $X_1, \ldots, X_n$ are iid $N(\theta, 1)$

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2}$$

Standard trick: work with the log likelihood

$$\log L(\theta|\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} -\frac{1}{2}(x_i - \theta)^2$$

Take the derivative, etc...

$$\frac{d}{d\theta} \log L(\theta|\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} (x_i - \theta)$$

$$\frac{d}{d\theta} \log L(\theta|\mathbf{x}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i - \theta) = 0$$

The only zero of this is for $\widehat{\theta} = \overline{\mathbf{x}}$

To show that this is, in fact, the maximum likelihood estimate:

1. Show it is a maximum:

$$\frac{d^2}{d\theta^2} \log L(\theta|\mathbf{x}) = \frac{1}{\sqrt{2\pi}}(-n) < 0$$

2. Unique interior extrememum, and a maximum – therefore a global maximum

Bernoulli MLE

Let $X_1, \ldots, X_n$ be iid Bernoulli(p)

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^y(1-p)^{n-y}$$

where $y = \sum x_i$

$$\log L(p|x) = y \log p + (n-y)\log(1-p)$$

If $0 < y < n$

$$\frac{d}{dp} \log L(p|x) = y\frac{1}{p} - (n-y)\frac{1}{1-p}$$

$$\frac{d}{dp} \log L(p|x) = 0 \Rightarrow \frac{1-p}{p} = \frac{n-y}{y}$$

Then $\widehat{p} = \frac{y}{n}$

Verify the maximum, and consider separately the cases where $y = 0$ (log likelihood is $n \log(1 - p)$) and $y = n$ (log likelihood is $n \log p$)

Binomial MLE, Unknown Number of Trials

Population is binomial (k, p) with known p and unknown k

$$L(k|\mathbf{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}$$

Maximizing by the differentiation approach is tricky

$$k \geq \max_i x_i$$

$$L(k|\mathbf{x}, p) > L(k-1|\mathbf{x}, p)$$

$$L(k|\mathbf{x}, p) > L(k+1|\mathbf{x}, p)$$

$$\frac{L(k|\mathbf{x},p)}{L(k-1|\mathbf{x},p)} = \frac{(k(1-p))^n}{\prod_{i=1}^n (k-x_i)}$$

Conditions for a maximum are:

$$(k(1-p))^n \geq \prod_{i=1}^n (k-x_i)$$

and

$$((k+1)(1-p))^n < \prod_{i=1}^n (k+1-x_i)$$

Solution: Solve the equation:

$$(1-p)^n = \prod_{i=1}^n (1-x_i z)$$

for $0 \leq z \leq \max_i x_i$. Call this $\widehat{z}$

$\widehat{k}$ is the largest integer equal to or less than $1/\widehat{z}$

MLE Instability

Olkin, Petkau and Zidek [JASA 1981] give the following example.

Suppose you are estimating the parameters for a binomial (k, p) distribution (both k and p unknown) and have the following data:

$$16, 18, 22, 25, 27$$

Turns out the ML estimate of k is 99.

Question – what do you think the ML estimate of p is?

But what if the data were slightly noisy, and the 27 should have been a 28?

The ML estimate of k is now 190!

What's going on here? Most likely the likelihood function is very flat in the neighborhood of the maximum

Bayesian Estimators

Classical vs. Bayesian approach to statistics

Classical: θ is an unknown but fixed parameter

Bayesian: θ is a quantity described by a distribution

Prior distribution describes ones beliefs about θ before any data is seen

A sample is taken and the prior is ten updated to take the data into account, leading to a *posterior distribution*

Let the prior be $\pi(\theta)$ and the sampling distribution be $f(x|\theta)$. Then the posterior is given by

$$\pi(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)/m(\mathbf{x})$$

Where $m(\mathbf{x})$ is the marginal distribution of $\mathbf{x}$, $\int f(\mathbf{x}|\theta)\pi(\theta)d\theta$

The posterior distribution can be used to make statements about θ, but it's still a distribution! For example, could use the mean of this distribution as a point estimate of θ.

Binomial Bayes Estimation

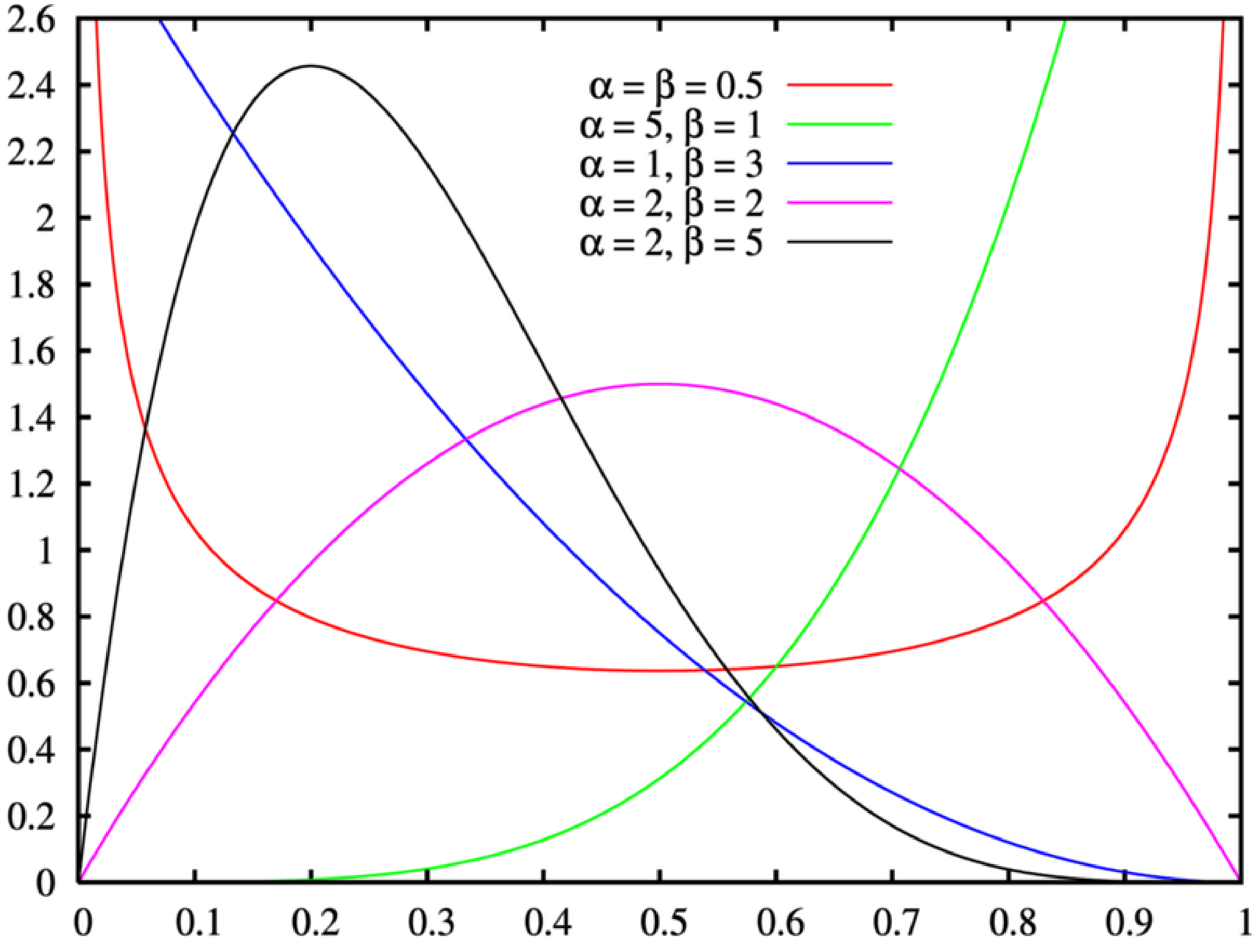
Let $X_1, \ldots, X_n$ be iid Bernoulli(p)

Let $Y = \sum X_i$

Suppose the prior distribution on p is beta(α, β) (really, I should subscript these, but for notational convenience I won't...)

Brief recap on the beta distribution – family of continuous distributions defined on $[0, 1]$ and governed by the two shape parameters.

A picture from wikipedia...



Probability density function

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

Nice fact: Mean is $\frac{\alpha}{\alpha+\beta}$

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$f(y) = \int_0^1 f(y|p) f(p) dp$$

$$= \int_0^1 \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp$$

$$= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

Then the posterior distribution is given by

$$\frac{f(y|p)\pi(p)}{f(y)}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

which is Beta$(y+\alpha, n-y+\beta)$!

Bayes estimate combines prior information with the data.

If we want to use a single number, we could use the mean of the posterior distribution, given by $\frac{y+\alpha}{n+\alpha+\beta}$

Normal MLE when μ and σ Are Both Unknown

$$\log L(\theta, \sigma^2|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}$$

Partial derivatives:

$$\frac{\partial}{\partial\theta}\log L(\theta, \sigma^2|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)$$

$$\frac{\partial}{\partial\sigma^2}\log L(\theta, \sigma^2|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\theta)^2$$

Setting to 0 and solving gives us:

$$\widehat{\theta} = \overline{x}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$