- Rebecka Jörnsten, jornsten@chalmers
  email best mode of communication

- Class MW 13:15 - 15 - perhaps change later if need be

- Course plan - preliminary only

  Masters, PhDs, PhDs other fields

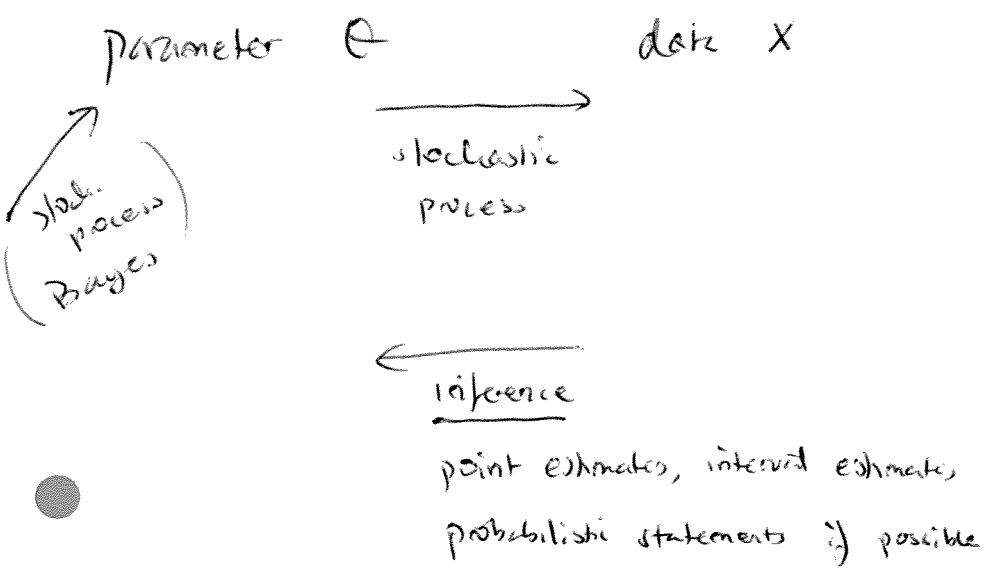
  - can adapt to some requests, spec. re modeling section

  - General idea - theoretical foundations, emph. on approx & assumptions.
    w1 - w4

  - w5 - w7 - model selection/distances
    - likelihood in more complex data structures

- Today - likelihood, setup & notation, definitions
  - Basic likelihood inference

Intro



inference

point estimates, interval estimates, probabilistic statements if possible

Frequentists – think of observed data as an instance of something that can be repeated under identical conditions

– Statistics computed from observed X ⇒ long-run properties of

Example $CI = [\hat{\theta} \pm 1.96\, se(\hat{\theta})]$

covers true θ in 95% of repeated experiments but we have no idea whether our specific interval covers or not.

Bayesian – inherent uncertainty/variability in real world too ⇒ prior belief $\pi(\theta)$

– observed data ⇒ update of belief into posterior ~~$p(\theta|x) \propto \pi(\theta)$~~

Both prior and posterior - probabilistic statements about θ (3)

$\Rightarrow$ Credible intervals $[\hat{\theta} \pm w(\theta)]$

posterior probability e.g. 95% that θ is covered by this interval.

Fisher, likelihood based inference

The likelihood based on observed data x

$\rightarrow$ provides some idea about uncertainty of θ

$\Rightarrow$ if possible convert to an exact probabilistic statement (for some data outcomes we may be able to do so)

- O/w use large-sample approximations, or approximations to particular likelihood function (quadratic = normal more later)

Notation & setup

- Stoch. process for data generation $P_\theta(x)$ or $f_\theta(x)$ or $p(x;\theta)$
- $X_1^n = \underset{\sim}{X}$ observed — joint density $p_\theta(x_1, \ldots, x_n)$

(would have $\theta_i = \theta(z_i)$ for x_i)
→ GLM

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1)\, p_\theta(x_2|x_1)\, p_\theta(x_3|x_1, x_2) \cdots$$

marginal / conditional

$$= \prod_{i=1}^{n} p_\theta(x_i | x_0^{i-1})$$

Markov of 1st order: $\prod_{i=1}^{n} p_\theta(x_i | x_{i-1})$

note $p_\theta(x_1)$ different from the other $p_\theta(x_i|x_{i-1})$

⇒ Time series

Q X_i's are independent

Likelihood (joint density)

$$L(\theta|\underset{\sim}{x}) = \prod_{i=1}^{n} p_\theta(x_i)$$

Bayes posterior $p(\theta|\underset{\sim}{x}) \propto \pi(\theta) L(\theta|\underset{\sim}{x})$

likelihood

prior belief

direct math expression for uncertainty of θ

As $n \to \infty$, $L(\theta|\underset{\sim}{x})$ dominates and two schools coincide

For smaller n, $\pi(\theta)$ can dominate so prior belief contributes largely to posterior

$\Rightarrow$ for this reason, <u>Uninformative priors</u>

popular use

Example $X \sim Bin(M, \theta)$

$\pi(\theta) \sim$ Uniform on $[0,1]$

Other popular examples - regularization priors

regression model

$$\begin{cases} X \sim N(z'\beta, \sigma^2) & X_i = z_i'\beta + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2) \\ \pi(\beta) \sim N(0, c) \end{cases}$$

Small c $\to$ β shrunk toward 0

large c $\to$ more like standard regression

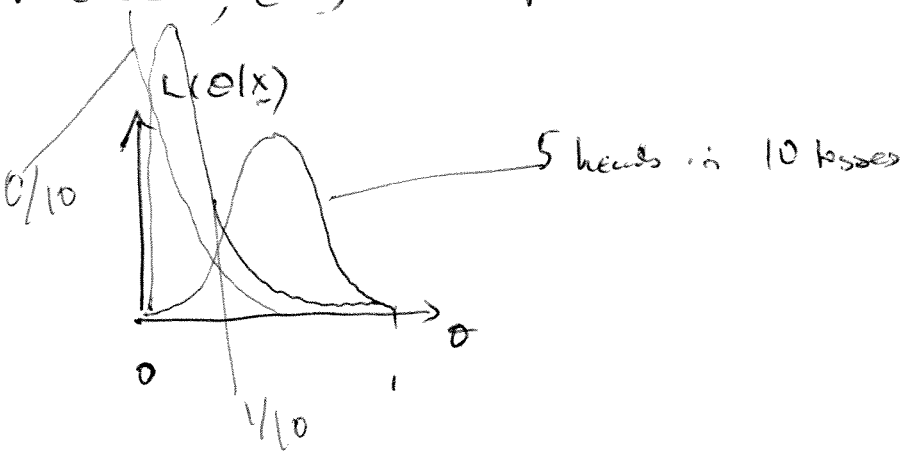
$c \to \infty$ uninformative

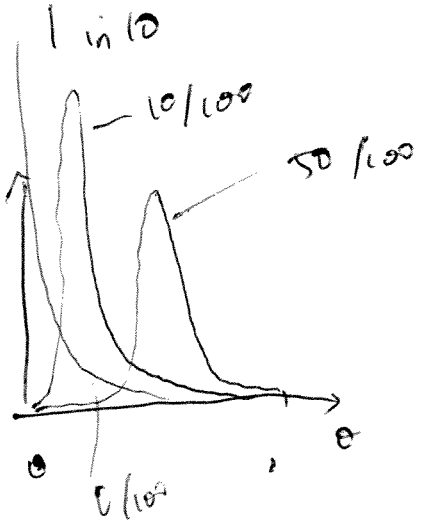
The likelihood function $L(\theta|\underset{\sim}{x})$

↑ fixed data

unknown parameter.

In simple case, 'easy' to plot $L(\theta|\underset{\sim}{x})$ as a function of θ





The likelihood is more 'flat' if data carries limited info about θ.

Thinking of $L(\theta|\underset{\sim}{x})$ as measuring likely data generating θ – logical choice for estimate is the maximum

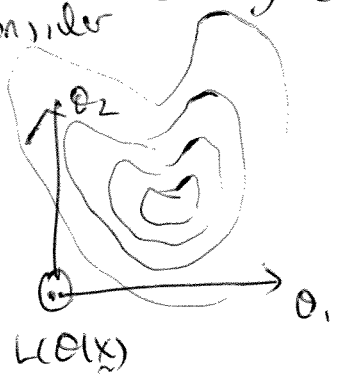
hint Shape of $L(\theta|\underset{\sim}{x})$ is what tells us about the uncertainty of this estimate

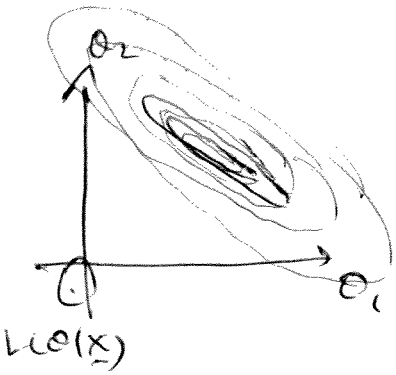
In non-scalar cases, not so easy to plot $L(\theta|\underset{\sim}{x})$

and shape can be even more important to

= by contours.

Consider





redundancy?

Working with the likelihood

Usually on a log-scale

$$\log L(\theta|\underset{\sim}{x}) = l(\theta|\underset{\sim}{x}) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

↳ can usually be converted to summary statistics like $\bar{x}$ and $\overline{x^2}$ etc

⇒ sufficiency later

So - to draw inference on θ we want to study $L(\theta|\underset{\sim}{x})$ or $l(\theta|\underset{\sim}{x})$

but we want to make sure our inference is not affected by the scale of the data we choose (perhaps arbitrarily) (e.g. inches vs cm)

If we transform $X \rightleftarrows y$, $x = x(y)$

the (change-of-variables) $\Rightarrow p_\theta(y) = p_\theta(x(y)) \left| \frac{\partial x}{\partial y} \right|$

$$\Rightarrow L(\theta|y) = L(\theta|x) \left| \frac{\partial x}{\partial y} \right|$$

To remove effect of variable transformation then $\Rightarrow$ look at <u>likelihood ratios</u>

$$\frac{L(\theta|y)}{L(\theta'|y)} = \frac{L(\theta|x)}{L(\theta'|x)}$$

Another very important question is the choice of parameterization $\eta = g(\theta)$

$\Rightarrow$ data should carry the same amount of info about η as it does about θ.

<u>Invariance</u>

((common used parameterization in $x \sim Bin(M, \theta)$ is the logit $\eta = \log \frac{\theta}{1-\theta}$)

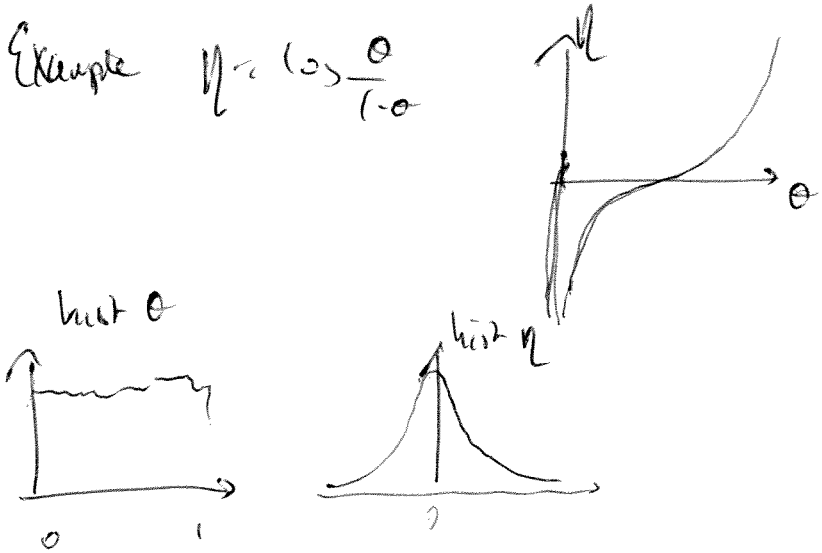
This is true for the likelihood ratio comparison.

Interesting to note though is if we're Bayesian — the parametrization does matter

$$p(\theta|\underset{\sim}{x}) \propto \pi(\theta) L(\theta|\underset{\sim}{x})$$

if $\pi(\theta)$ uninformative $\Rightarrow$ this may not be the case for $\eta = g(\theta)$!

Example $\eta = \log \frac{\theta}{1-\theta}$



Summarizing the likelihood function

→ Logical to use the location of the maximum for our estimate $\hat{\theta}_{MLE}$: point estimate of θ

→ but provides no measure of uncertainty

Idea – look at how sharp the likelihood function is around the maximum, i.e. the curvature of $L(\theta|x)$ near $\hat{\theta}$.

Indeed, if $L(\theta|x)$ is quadratic in θ this would be sufficient to describe the entire function.

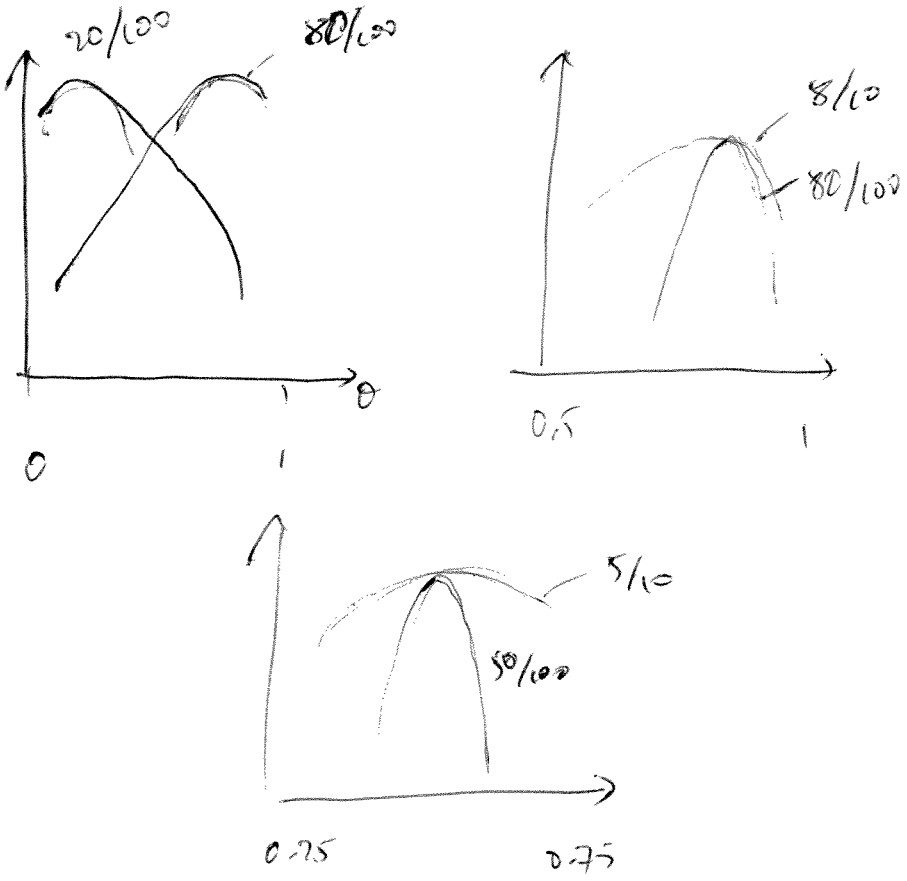
We call the (log) likelihood function $l(\theta|x)$ regular if it is well approximated by a quadratic function (in which case $\hat{\theta}$ + the curvature provides the complete information!)

Checking the quadratic approximation

Plot the normalized log likelihood

$$\log \frac{L(\theta|x)}{L(\hat{\theta}|x)}$$

(makes things comparable for different x as well [max @ 0])



As you can see, how well approximated the likelihood can be by a quadratic depends on sample size and outcome.

In general, for large n, approx works well.

Some concepts we will come back to later as well

Likelihood $L(\theta|\underset{\sim}{x})$ or $L(\theta)$

log likelihood $\ell(\theta|\underset{\sim}{x})$ or $\ell(\theta)$

The <u>score function</u> $S(\theta) = \frac{\partial}{\partial\theta} \log L(\theta)$

- The MLE $\{\theta : S(\theta) = 0\}$, denote by $\hat{\theta}$
- Curvature $I(\theta) = -\frac{\partial^2}{\partial\theta^2} \log L(\theta)$ (positive quantity)

Curvature @ MLE $I(\hat{\theta})$ – 'measures' the sharpness of the peak

– also called the <u>observed</u> Fisher information

(as opposed to the expected

$E_\theta\left[-\frac{\partial^2}{\partial\theta^2} \log L(\theta|\underset{\sim}{X})\right]$)

... more later

So is the quadratic approx OK?

$$\log L(\theta) \approx \log L(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

(Taylor expansion)

Since $S(\hat{\theta}) = 0$ by def

$$\Rightarrow \log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

(Special case - normal mean model

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\}$$

$$\log L(\mu) \propto -\frac{1}{2}(x-\mu)^2 \quad \text{Exactly quadratic}$$

Normal approx of $\hat{\theta}$ = Quad approx of $L(\theta)$ near $\hat{\theta}$.)

$$\frac{\partial}{\partial \theta} \log \frac{L(\theta)}{L(\hat{\theta})} = S(\theta) \approx I(\hat{\theta})(\hat{\theta} - \theta)$$

Now, for nice distribution families (more later)

One can show $E_\theta[S(\theta)] = E_\theta\left[\frac{\partial}{\partial \theta} \log L(\theta | X)\right] = 0$

and $V_\theta[S(\theta)] = E[I(\theta)]$

$$\Rightarrow I^{-1/2}(\hat{\theta}) S(\theta) \approx I^{1/2}(\hat{\theta})(\hat{\theta} - \theta)$$

$I^{-1/2}(\hat{\theta})$ — dimensionless

$$\Rightarrow \text{Var}(\hat{\theta}) \approx I^{-1}(\hat{\theta})$$

[Later, go through this in more rigour, CR lower bound etc]

Summary

If $l(\theta)$ approx quadratic near $\hat{\theta}$

$\Rightarrow$ can summarize $l(\theta)$ by the MLE $\hat{\theta}$ and the curvature $I(\hat{\theta})$

AND Have $\begin{cases} E(\hat{\theta}) = \theta \\ V(\hat{\theta}) \approx I(\hat{\theta})^{-1} \end{cases}$

Alternative - working with the likelihood function directly.

Values of θ that are supported by the data - large values for $L(\theta)$, or large values of LR $\frac{L(\theta)}{L(\hat{\theta})}$ (normalized L)

Suggests interval estimate $\left\{\theta : \frac{L(\theta)}{L(\hat{\theta})} > c\right\}$

But how do we choose c?

Use likelihood at least 5% of the maximum? or 10%?

This is tricky - we have no obvious cs here since $L(\theta)$ is data dependent not absolute.

$\Rightarrow$ try to convert this to a long-run freq or probabilistic statement.