- Example

$X_1, X_2, \ldots X_6 \overset{iid}{\sim} Bern(\theta)$

$\vec{x} = \langle x_1, x_2, \ldots x_6 \rangle$.

If $\vec{x} := \langle 0,0,1,0,1,0 \rangle$,

$p(\vec{x} = \langle 0,0,1,0,1,0 \rangle; \theta) = \prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^2(1-\theta)^4$.

+ Statistical inferencing

- When θ is unknown but we want to know it, we'd use Statistical inferencing
- There are 3 goals:
  1. We want the best guess for θ; We'll call this $\hat{\theta}$.
  2. Confidence set - range of likely values of θ.
  3. Test theories about θ.
- Example: Consider $X_1, \ldots, X_6 \sim Bern(0.5)$;

$p(\vec{x} = \langle 0,0,1,0,1,0 \rangle; \theta = 0.5) = 0.5^6 = 0.0156$.

- How about $\theta = 0.25$? $\Rightarrow p(\cdots) = (0.25^2)(0.75)^4 = 0.0196$.
- Note that $P(\vec{x}; 0.25) > P(\vec{x}; 0.5)$ so $\theta = .25$ is a better estimate than $\hat{\theta} = 0.5$.
- Pick $\vec{\theta}$, a class of parametric models then use inferencing to get the best $\hat{\theta}$.

+ Likelyhood function $L(\theta, \vec{x})$

- The likelyhood function $L(\theta, x) = p(x, \theta)$ is a function of θ; which is just the (joint) pmf in which x is fixed. It gives the "likelyhood" (as a probability) of some θ occurring.
- So, it would make sense to take $\hat{\theta}$ to be the one with the biggest likelyhood — ie, the one that maximizes $L(\theta)$.
- $L(\theta)$ is not really necessarily a pdf — e.g. $\sum_{\theta \in \Theta} L(\theta, x) \neq 1$ always.

+ Maximum Likelyhood Estimator (MLE)

- $\hat{\theta}_{MLE} := \text{argmax}\{\mathcal{L}(\theta; x)\}, \theta \in \Theta$ is called the **MLE** & is the maximum point of $\mathcal{L}(\theta; x)$, if it exists.
- Note that for any fn g, $\text{argmax}(g) = \text{argmax}(7g) = \text{argmax}(\ln g)$, for e.g (fact from calc.).
  - This fact will be useful when looking for the MLE, since as we know, the maximum point for a function is found by taking the derivative, and it is often easier to take the derivative of $\ln g$ than just g.
- So if $\ell(\theta, x) := \ln(\mathcal{L}(\theta; x))$ (called the log likelyhood), then $\hat{\theta}_{MLE} = \text{argmax}\{\ell(\theta; x)\}$.

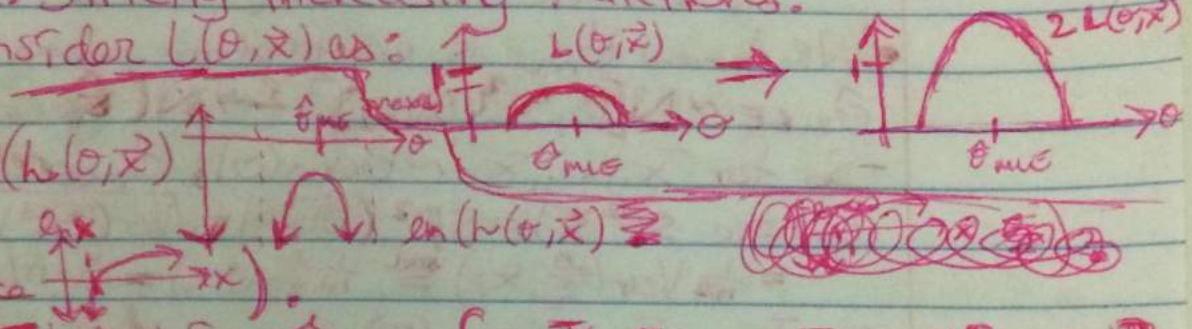
+ Side Note

With likelyhood, we really want to know how probable the value of θ is when given data is fixed — this is the "inverse question"!

1/31/2018

+ More on Maximum Likelyhood Estimator (MLE)

- $L(\theta; x_1, x_2, \ldots, x_n) = P(X_1, \ldots, X_n; \theta)$ ← joint density/mass function
- Goal of inferencing: Looking for most likely θ.
- $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \{L(\theta, \vec{x})\} = \operatorname{argmax}\{g(L)\} = \operatorname{argmax}\{\ln(\theta, \vec{x})\}$
- Note: A "positive monotonic transformation" g is one that preserves: $\forall x, y, x > y \Rightarrow g(x) > g(y)$; i.e. Strictly increasing functions.
- So, consider $L(\theta, \vec{x})$ as:



(since $\ln x$ ↑ with x).

- Ex: What is $\hat{\theta}_{MLE}$ for $X_1, X_2, \ldots, X_6 \sim Bern(\theta)$?

$L(\theta, \vec{x}) = P(\vec{x}, \theta) = \prod_{i=1}^{6} P(X_i; \theta) = \prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{1-x_i}$;

To find MLE, derive w.r.t. θ & set equal to 0:

$\ell(\theta; \vec{x}) = \ln\left(\prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{1-x_i}\right) = \sum_{i=1}^{6} \ln\left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$

$= \sum_{i=1}^{6}\left[x_i \ln(\theta) + (1-x_i)\ln(1-\theta)\right] = 6\bar{x}\ln\theta + \ln(1-\theta)\cdot(6-6\bar{x})$.

$\ell'(\theta; \vec{x}) = \frac{6\bar{x}}{\theta} - \frac{6-6\bar{x}}{1-\theta} \Rightarrow 0 \overset{set}{=} \ell'(\theta; \vec{x}) \Rightarrow \frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta} = 0 \Rightarrow$

$\frac{\bar{x}}{\theta} = \frac{1-\bar{x}}{1-\theta} \Rightarrow (1-\theta)\bar{x} = \theta(1-\bar{x}) \Rightarrow \theta = \bar{x} \Rightarrow \hat{\theta}_{MLE} = \bar{x} = \hat{p}$.

- So here, if $\vec{x} = \langle 0, 0, 1, 0, 1, 0 \rangle$, $\hat{\theta}_{MLE} = \bar{x} = 2/6 = 1/3$.
- $\hat{\theta}_{MLE} = \bar{X}$ is an estimator (a r.v.) & $\hat{\theta}_{MLE} = \bar{x}$ is an estimate.

$\bar{X}$ is r.v. of average of $\vec{x}$, & $\bar{x}$ is a realization of $\bar{X}$.

- Why do we care about MLE's?

① $\hat{\theta}_{MLE} \xrightarrow{p} \theta$ (converges in probability)

As n gets large, $\hat{\theta}_{MLE}$ looks more & more like θ.

- i.e. $\forall \varepsilon > 0$, $\lim_{n \to \infty} P(|\hat{\theta}_{MLE} - \theta| \geq \varepsilon) = 0$.

* ② Asymptotic Normality $\hat{\theta}_{MLE} \xrightarrow{d} N(\theta, SE[\hat{\theta}_{MLE}])$.

③ Hypothesis testing via:

~~Ret Reg~~ $= \left[\theta_0 \pm \frac{z_{\alpha}}{2} SE[\hat{\theta}_{MLE}]\Big|_{\theta_0}\right]$.

③ Efficiency: $SE(\hat{\theta}_{MLE})$ is theoretically the lowest SE for all consistent estimators (proposition 1).

• **Ex:** What is $\hat{\theta}_{MLE}$ for $X_1, \dots, X_n \overset{iid}{\sim} Geom(\theta)$, $\theta = (0,1)$
where $X \sim Geom(p) \Rightarrow P(X=x) = (1-p)^x p$?

$L_n(\theta; \vec{x}) = \prod_{i=1}^{n} P(X_i, \theta) = \prod_{i=1}^{n} (1-\theta)^{x_i} \theta \Rightarrow$

$\ell(\theta; \vec{x}) = \ln\left(\prod_{i=1}^{n} (1-\theta)^{x_i} \theta\right) = \sum_{i=1}^{n} \ln\left[(1-\theta)^{x_i} \theta\right] =$

$\sum_{i=1}^{n} \left[\ln(1-\theta)^{x_i} + \ln\theta\right] = \sum_{i=1}^{n} \left[x_i \ln(1-\theta) + \ln\theta\right] =$

$\ln(1-\theta) \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \ln\theta = n\bar{x}\ln(1-\theta) + n\ln\theta \Rightarrow$

$\ell'(\theta; x) = \frac{n}{\theta} - \frac{n\bar{x}}{1-\theta} \Rightarrow 0 \overset{set}{=} n\left(\frac{1}{\theta} - \frac{\bar{x}}{1-\theta}\right) \Rightarrow \frac{1}{\theta} = \frac{\bar{x}}{1-\theta} \Rightarrow$

$1 - \theta = \theta\bar{x} \Rightarrow 1 = \theta(\bar{x}+1) \Rightarrow \theta = \frac{1}{\bar{x}+1} \Rightarrow \hat{\theta}_{MLE} = \frac{1}{\bar{x}+1}$.

• We know $\hat{\theta}_{MLE}$ is asymptotic normal; ie.

$\hat{\theta}_{MLE} \sim N[\theta, SE^2(\hat{\theta}_{MLE})] \xrightarrow{p} N(\hat{\theta}_{MLE}, SE^2(\hat{\theta}_{MLE}))$.

- So for $X_1, \dots, X_n \sim Bern(\theta)$, $\hat{\theta}_{MLE} \sim N(\bar{x}, SE^2(\bar{x}))$,

where $SE^2(\bar{x}) = SE^2\left(\frac{\sum x_i}{n}\right) = \sqrt{Var\left(\frac{\sum x_i}{n}\right)}^2 = Var\left(\frac{\sum x_i}{n}\right)$

$= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) \overset{ind.}{=} \frac{1}{n^2} \sum_{i=1}^{n} (x_i) = \frac{1}{n^2} \cdot n(\theta)(1-\theta) = \frac{\theta(1-\theta)}{n}$,

- for $X_1, \dots, X_n \sim Geom(\theta)$, $\hat{\theta}_{MLE} \sim N\left(\frac{1}{\bar{x}+1}, SE^2\left(\frac{1}{\bar{x}+1}\right)\right)$, where SE is not so easy to calculate.

Goals of Inference Based on MLE (Frequentist World)

1. Point of estimation for θ.
2. Confidence sets for θ; (region of likely values).
3. Testing theories about θ.

Inference via MLE's

1. $\hat{\theta}_{MLE}$ is a number.
2. Confidence Interval.

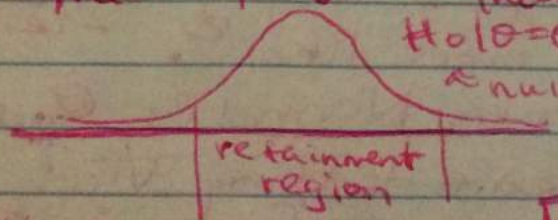
$CI_{\theta, 1-\alpha} := \left[\hat{\theta}_{MLE} \pm z_{\alpha/2} SE[\hat{\theta}_{MLE}]\right]$.

3. Hypothesis Testing

Two-Sided hypothesis test for θ:

$H_0 : \theta = \theta_0$ — default theory

$H_a : \theta \neq \theta_0$ — theory you wish to prove.



For $Bern(\theta)$, Ret. Region$_\alpha := \left[p_0 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}\right]$.

(level of confidence; p_0: best guess; $z_{\frac{\alpha}{2}}$: justified by CLT)

Recipe for Inference

Observe data → Pick $\mathcal{P}$ (Parametric model) → Do Inference via MLE.

- What's wrong with this?