Data analysis

Ben Graham

MA930, University of Warwick

October 22, 2015

MLE recap

- $X_1, \ldots, X_n \sim f(x \mid \theta)$
- $L(X \mid \theta) = \prod_i f(x_i \mid \theta)$
- Ex: Poisson(λ)
- Ex: Uniform($0, \theta$)
- Ex: Exponential(θ)

Kullback–Leibler divergence

$$D_{KL}(P, Q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

- Non-negative
- Positive if the two distributions differ
- $\rightarrow$ *MLE* is consistant

10.4.1 Approximate MLE distribution

- iidrv random sample $X_1, \dots, X_n \sim f(x \mid \theta)$
- MLE $\hat{\theta}$
- Asymptotic distribution related to Fisher's Information matrix $I(\theta)$ (multivariate case $\theta \in \mathbb{R}^d$):

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{P} N(0, I^{-1}(\theta))$$

$$I_{ii}(\theta) = Var\left[\frac{\partial \log f(X \mid \theta)}{\partial \theta_i}\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i^2} \log f(X \mid \theta)\right]$$

$$I_{ij}(\theta) = Cov\left[\frac{\partial \log f(X \mid \theta)}{\partial \theta_i}, \frac{\partial \log f(X \mid \theta)}{\partial \theta_j}\right]$$

$$= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X \mid \theta)\right]$$

Ch 8 Hypothesis testing

- A hypothesis is a statement about a population parameter
- Null hypothesis, i.e. $H_0 : \theta = 0$
- Alternative hypothesis, i.e. $H_1 : \theta \neq 0$
- H_0 is often much simpler than H_1, i.e. a vector subspace.
- Question: Can you reject H_0?
  - Rejecting H_0 does not mean accepting H_1.
  - You never "accept H_0".
- Test statistic critical region: ie. reject H_0 if $|\bar{X}| > 3$

Rejecting does not mean rejecting

- Suppose there are 100 locations that might have oil at them
- Each location has a 4% chance of having oil
- Hypothesis test:
  - H_0 : oil vs H_1 : no oil
  - Reject H_0.
  - but ..
- Cost of exploratory digging $5M
- value of oil if found $250M
- $\mathbb{E}$[benefit of drilling]$=-5+0.04\times 250=5>0$

p-values

- A statistic $P = P(X)$ is a p-value if under H_0
  - $P \sim Uniform(0,1)$ or even if
  - $\forall t \in [0,1]$, $\mathbb{P}_{H_0}[P \leq t] \leq t$
- Rejecting H_0 corresponds to small p-values.
- i.e. Reject H_0 if $p \leq 0.05$
- https://xkcd.com/882/
- Simulate multi-arm experiment with Uniforms

Simple test

- $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$
- Neither value may be correct, but which is better?
- Type I error:
  - Reject H_0 under H_0
  - Size of the test $\alpha = \mathbb{P}_{H_0}(\text{Reject } H_0) \approx 0.05$
- Type II error:
  - Failing to reject H_0 under H_1 .
  - Power $\beta = \mathbb{P}_{H_1}(\text{Reject } H_0)$
    [prob of not type II error-ing under H_1]
- Possible critical region if $\theta_0 < \theta_1$:
  Reject H_0 in $C_\alpha = \{x : \hat{\theta}(x) > c_\alpha\}$
- Conflict between $\alpha$and β [experiment in R]
- Example: N coin tossed Bernoulli(θ). Biased?
- Example: Normal known mean.

Simple Likelihood ratio test

- $X_1, \ldots, X_n \sim f(x \mid \theta)$
- $L(X \mid \theta) = \prod_i f(x_i \mid \theta)$
- For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, the likelihood ratio test statistic is:
$$\lambda(x) = \frac{L(\theta_0 \mid x)}{L(\theta_1 \mid x)}$$
- Critical region $C_\alpha = \{x : \lambda(x) \leq c_\alpha\}$.
- Choose c_α such that $\mathbb{P}_{\theta_0}[\lambda(x) \leq c_\alpha] = \alpha$
- Power $\beta = \mathbb{P}_{\theta_1}[\lambda(x) < c_\alpha]$
- Example: Normal distribution $N(\mu, 1)$.
$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.
- Example: Exponential distribution plus constant.
$X \sim Exp(1) + \theta$.
$H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.
- Critical regions can be written in terms of sufficient statistics.

The Neyman-Pearson lemma

- $X_1, \ldots, X_n \sim f(x \mid \theta)$
- NP: The SLRT is as least as powerful as any other test of the same size.
- Proof:
  SLRT critical region C_α
  Another critical region C_1 with same size α
  $\mathbb{P}_{\theta_0}(C_\alpha) = \mathbb{P}_{\theta_0}(C_1)$
  $\mathbb{P}_{\theta_0}(C_\alpha \setminus C_1) = \mathbb{P}_{\theta_0}(C_1 \setminus C_\alpha)$
  $\mathbb{P}_{\theta_1}(C_\alpha \setminus C_1) \geq \mathbb{P}_{\theta_0}(C_1 \setminus C_\alpha)$
  $\mathbb{P}_{\theta_1}(C_\alpha) \geq \mathbb{P}_{\theta_0}(C_1)$

Simple null, composite alternative

- $H_0 : \theta = \theta_0$ vs
  $H_1 : \theta \neq \theta_0$ or $H_1 : \theta \geq \theta_0$ or $H_1 : \theta \leq \theta_0$ or $H_1 : \theta \in \mathbb{R}$
- Type I error:
  - Reject H_0 under H_0
  - Size of the test $\alpha = \mathbb{P}_{H_0}(\text{Reject } H_0) \approx 0.05$
- Uniformly most powerful test:
  - A test is UMP of size α if it has size α and
  - $\forall \theta_1 \in H_1$, the test's critical region is a most powerful test for testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \in \theta_1$.
  - Generally possible for one sided test, not for two sided tests.

Composite null, composite alternative LRT

- $H_0 : \theta \in \Theta_0$
- $H_1 : \theta \in \Theta_1$ or $H_1 : \theta \in \Theta \setminus \Theta_0$ or just $H_1 : \theta \in \Theta$
- $X_1, \ldots, X_n \sim f(x \mid \theta)$
- $L(X \mid \theta) = \prod_i f(x_i \mid \theta)$
- Def 8.2.1 For testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, the likelihood ratio test statistic is:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid x)}{\sup_{\theta \in \Theta_1} L(\theta \mid x)} = \frac{L(\hat{\theta}_0 \mid x)}{L(\hat{\theta}_1 \mid x)}$$

  where $\hat{\theta}_i$ is the MLE in Θ_i, $i = 1, 2$.
- Critical region $\{x : \lambda(x) \leq c\}$.

Wilk's theorem

- $H_0 : \theta \in \Theta_0$
- $H_1 : \theta \in \Theta$
- $d = dim(\Theta) - dim(\Theta_0)$
- $-2 \log \lambda(x) \to \chi^2_d$ under some regularity conditions on $f(x \mid \theta)$
- Ex: $X_1, \ldots, X_n \sim$N$(\theta, 1)$
  $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.
  $-2 \log \lambda(x) = n\bar{x} \sim \chi^2_1$ exactly
- Ex: multinomial distribution
  6-sided die
  $m \times n$ contingency table