Estimation

Parametric { Binomial, Poisson, Normal, Mixture Normal

MOM

MLE { Mathematical Solution, Newton-Raphson, EM algorithm

Nonparametric for the density $f(x)$

Kernel density estimation

Estimation is not good on the boundary.

cumulative distribution function $F(x) \to$ cdf

empirical cdf

$$F(x) = P(X \leq x)$$

$$F_n(x) = \frac{\#\text{number of } X_i \leq x}{n} = \frac{1}{n} \sum 1(X_i \leq x)$$

Let $Y_i = 1(X_i \leq x)$ $\qquad Y_i \overset{iid}{\sim} \text{Bernoulli}(F(x))$

$$F_n(x) = \frac{1}{n} \sum Y_i = \frac{\text{Bin}(n, F(x))}{n}$$

$$E F_n(x) = \frac{n F(x)}{n} = F(x)$$

Theorem DKW Inequality:

$$P\left(\sqrt{n} \cdot \sup_t \left| \hat{F}_n(t) - F(t) \right| > z \right) \leq 2\exp(-2z^2)$$

$z = \log n$

$$P\left(\sup_t \left| \hat{F}_n(t) - F(t) \right| > \frac{\log n}{\sqrt{n}} \right) \to 0$$

Application: ① Test the model assumption. Kolmogorov-Smirnov.

② Multiple testing: False Discovery Rate (FDR)

Hypothesis Testing: Yes / No question

H_0: Model assumption

Ronald A. Fisher

Elements of Hypothesis Testing

Hypothesis: H_0 null hypothesis (conventional knowledge)

H_a: Alternative hypothesis. (we want to demonstrate)

Example Clinical Trial $\tau = P($ Success of Surgical Procedure$)$

H_0: $\tau \leq 0.2$ H_a: $\tau > 0.2$ Hypothesis should never depends on the data

Composite Hypothesis

Example: τ = number of West Nile Virus

H_0: $\tau = 55$. H_a: $\tau \neq 55$

Single Hypothesis

Test Statistic: a function of the data

Rejection Region: sets of values that a researcher wants to

Rejection Region: sets of values that a researcher wants to reject H_0 $\quad R$

Types of Errors:

	Reject H_0	Fail-to-reject H_0
H_0	Type I error	✓
H_a	✓	Type II

False Discovery
False Positive = Type I

False Negative = Type II

Type I error, "α" significance level of α

Simple: $\alpha = P(R \mid H_0)$

$$\alpha = \sup_{P_0 \in H_0} P(R \mid P_0 \text{ is true})$$

Type II error: β

$$\beta = P(R^C \mid \theta \in H_a)$$

power: $= 1 - \beta = P(R \mid \theta \in H_a)$

Control Type I error α at 0.1, 0.05, 0.01,

minimize the type II error.

choose Rejection region $P(R \mid H_0) \leq \alpha$

distribution of $(T \mid H_0)$

Example 6.2.3:

Example 6.2.3 :

$H_0: \pi = 0.2$

$H_a: \pi = 0.5$

$n = 25$ subjects. Y be the number of success

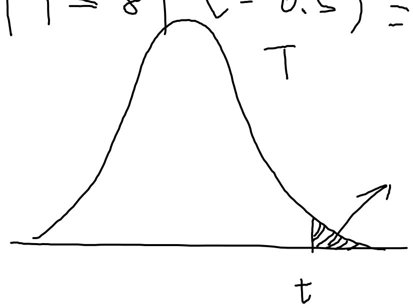
$Y \sim \mathrm{Bin}(25, \pi)$

$R = \{Y > 8\}$. Type I and II ?

$\text{Type I} = P(R \mid H_0) = P(Y > 8 \mid \pi = 0.2) = 0.047$

$\text{Type II} = P(R^c \mid H_a) = P(Y \leq 8 \mid \pi = 0.5) = 0.054$

P-value. $P(\quad \mid H_0)$



P-value is the smallest "α" that one can reject H_0

If $\alpha <$ P-value, Fail-to-reject H_0

$\alpha >$ P-value, reject H_0

If $n = 25$, $Y = 10$, $\alpha = 0.01$

$P(Y \geq 10 \mid \pi = 0.2) = 0.005 < \alpha$. $\rightarrow$ reject H_0

① Hypothesis ② "α"

③ data $\rightarrow$ Statistic

④ Find the distribution of T under H_0

⑤ Rejection region. {T, rejection} $\rightarrow$ make a decision

④' p-value, $\{p\text{-value}, \alpha\} \to$ make a decision

6.3 Generalized Likelihood Ratio Test (GLRT)

H_0: $\theta \in \Theta_0$

H_a, $\theta \in \Theta_1$

Test Statistic: $LR = \dfrac{\max_{\theta \in \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)}$

$$\Lambda = \frac{\max_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)}$$

Rejection region: $\{\Lambda > k_0^*\}$

Test for one population mean.

H_0: $\mu = 17.60 = \mu_0$

H_a: $\mu \neq 17.60$ $\qquad \sigma = 1.3$

$X_1\ X_2 \cdots X_n \sim N(\mu, \sigma^2)$

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}\right\}$$

H_0: $\max_{\theta \in \Theta_0} L(\mu) = L(17.60) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_i (x_i - 17.6)^2}{2\sigma^2}\right\}$

$H_0 \cup H_a$. $\max L(\mu) = L(\bar{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum (x_i - \bar{x})^2}{2\sigma^2}\right\}$

$$\Lambda = \frac{\exp\left\{-\frac{\sum (x_i - \bar{x})^2}{2\sigma^2}\right\}}{\exp\left\{-\frac{\sum (x_i - \mu_0)^2}{2\sigma^2}\right\}} = \exp\left\{\frac{\sum (x_i-\mu_0)^2 - \sum (x_i - \bar{x})^2}{2\sigma^2}\right\}$$

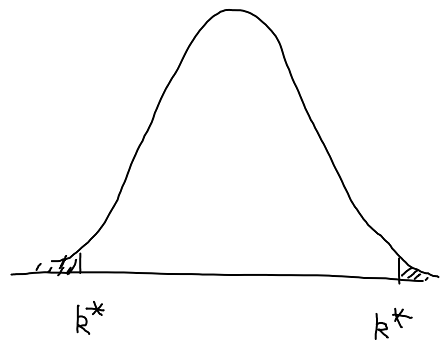
$$= c \cdot \exp\left\{ \frac{n(\bar{x}-\mu_0)^2}{2\sigma^2} \right\}$$

Reject Region $R = \{ \Lambda > k^* \} = \left\{ c \exp\left(\frac{n(\bar{x}-\mu_0)^2}{2\sigma^2} \right) > k^* \right\}$

$$= \left\{ \frac{n(\bar{x}-\mu_0)^2}{2\sigma^2} > k^* \right\} = \left\{ (\bar{x}-\mu_0)^2 > k^* \right\}$$

$$= \left\{ |\bar{x}-\mu_0| > k^* \right\} = \left\{ \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > k^* \right\}$$

Let $Z = \dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$



Under H_0: $Z \sim N(0,1)$

$$\alpha = P(|Z| > k^* \mid H_0)$$

$$k^* = Z_{\alpha/2}$$

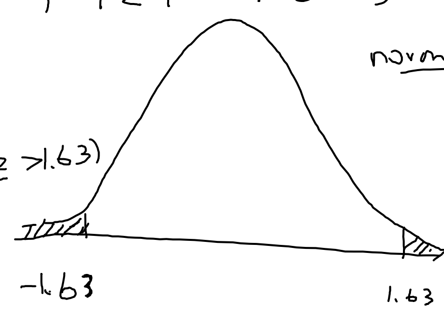
$$R = \left\{ \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > Z_{\frac{\alpha}{2}} \right\} = \left\{ |Z| > Z_{\frac{\alpha}{2}} \right\}$$

$$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{17.30-17.60}{1.3/\sqrt{50}} = -1.63$$

Reject Region $\alpha = 0.10$ $R = \{ |Z| > 1.645 \}$

normal

$\rightarrow$ Fail-to-reject H_0



p-value $= P(|Z| > 1.63 \mid H_0) = 2P(Z > 1.63)$

$= 0.103$

Some results

① H_0: $\tau = \tau_0$ vs H_a: $\tau \neq \tau_0$ Λ

$$2 \log \Lambda \,\big|\, H_0 \sim \chi^2_1 \quad \text{when } n \to \infty$$

(2) Nested null and alternative

$\underline{\underline{H_0}}$: $\theta \in \underline{\underline{\Theta_{p-r}}}$

H_a: $\theta \in \Theta_{\underline{\underline{p}}}$

$$2 \log \Lambda \,\big|\, H_0 \sim \chi^2_r \quad \text{when } n \to \infty.$$

$\underline{\underline{H_0}}$: rate is constant

H_a: rate is seasonal

$$\tilde{Y} = (Y_1, Y_2, \ldots, Y_{12})$$

<u>Multinomial</u> "c" category

$$\sum_{i=1}^{c} \pi_i = 1$$

$$(Y_1, Y_2, \ldots, Y_c) \sim \text{Multinomial}(n, \tilde{\pi})$$

$$P(Y_1 = y_1, \ldots, Y_c = y_c) = \frac{n!}{y_1! \cdots y_c!} \prod_{i=1}^{n} \pi_i^{y_i}$$

$$\hat{\pi}_i = \frac{y_i}{n}$$

Let $\theta_1, \theta_2, \cdots, \theta_{12}$ be the daily rate of suicide
$d_1, d_2, \ldots, d_{12}$ be the number of days in each month

$$\pi_i = \underline{\underline{d_i}}\, \theta_i$$

$$\tilde{Y} = \text{Multinomial}(n, \tilde{\pi})$$

$\underline{\underline{H_0}}$: $\theta_1 = \underline{\theta_2 = \cdots = \theta_{12}}$

$H_{\underline{a}}$. At least two $\underline{\theta\text{'s are not the}}$ same

H_a. At least two θ's are not the same

GLRT $\quad L(\tilde{\theta}) = \dfrac{n!}{y_1! \cdots y_{12}!} \prod_{i=1}^{n} (d_i \theta_i)^{y_i}$

$1 = \Sigma \pi_i = \Sigma d_i \cdot \theta$
$= \theta \Sigma d_i = \theta \cdot 365$

H_0: $\theta = \dfrac{1}{365}$

$$L(\hat{\theta}) = \frac{n!}{\prod y_i!} \prod \left(\frac{d_i}{365}\right)^{y_i}$$

$H_0 \cup H_a$: $\max L(\hat{\theta}) = L\left(\left(\frac{y_1}{n}, \frac{y_2}{n}, \cdots \frac{y_{12}}{n}\right)\right) = \dfrac{n!}{\prod y_i!} \prod \left(\frac{y_i}{n}\right)^{y_i}$

$$\Lambda = \prod \frac{\left(\frac{y_i}{n}\right)^{y_i}}{\left(\frac{d_i}{365}\right)^{y_i}} = \prod \left(y_i \cdot \frac{365}{n d_i}\right)^{y_i}$$

$$2 \log \Lambda = 2 \sum_i y_i \log\left(y_i \cdot \frac{365}{n \cdot d_i}\right)$$

Under H_0. $\quad 2 \log \Lambda = \chi^2_{11}$

P-value: $= P\left(\chi^2_{11} > \text{test statistic}\right)$

Reject the H_0