Study
- empirical studies
  - sample survey 抽样
  - observational 观察
  - experimental 实验 (有干预)
  - 1.1–1.2
- numerical summaries & graphical summaries — 1.3–1.5
- model fit → (plot) — 2.6

PPDAC — 3.
- target population : 目标群体
- study population : 实验可能 cover 到 ⊆ 所有 — study error
- sample protocol : sample size 个 sample randomly chosen from study population — sample error

estimation
- point estimate $\hat{\theta} = g(y_1, \cdots, y_n)$ — y: data — 具体 ⊆ 数
  - m.l.e : $L(\theta)$ 去 constant → $l(\theta)$ → $\frac{d}{d\theta} l(\theta) = 0$
- point estimator $\tilde{\theta} = g(Y_1, \cdots, Y_n)$ — Y: variable — function (有 distribution)
  - sample distribution — 4.1–4.2

$Y \sim G(\mu, \sigma)$	$\bar{Y} \sim G(\mu, \frac{\sigma}{\sqrt{n}})$	→ pivotal quantity $\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0,1)$
$Y \sim Po(\theta)$	$\bar{Y} \sim G(\theta, \sqrt{\frac{\theta}{n}})$	
$Y \sim Bi(n, \theta)$	$\bar{Y} \sim G(\theta, \sqrt{\frac{\theta(1-\theta)}{n}})$	
$Y \sim Exp(\theta)$	$\bar{Y} \sim G(\theta, \frac{\theta}{\sqrt{n}})$	

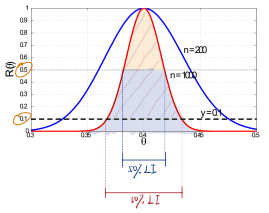
function : — 2.2–2.5

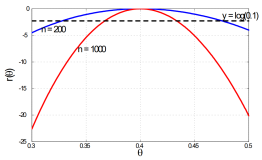
	likelihood	relative
\	$L(\theta) = \prod f(y_i; \theta)$	$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$
log	$l(\theta) = \log L(\theta)$	$r(\theta) = l(\theta) - l(\hat{\theta})$

interval :

likelihood interval LI 4.3

$\{\theta : R(\theta) \geq q\}$





n ↑ → narrower

> uniroot(function(x) (R(x) -q), lower, upper)$root

公式

confidence interval CI 4.4

$\{\theta : P(L(Y) \leq \theta \leq U(Y)) = p\}$

CL 100p%	CI
90%	$\bar{x} \pm 1.6449 \frac{\sigma}{\sqrt{n}}$
95%	$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
99%	$\bar{x} \pm 2.5758 \frac{\sigma}{\sqrt{n}}$

> pnorm(1.96) - pnorm(-1.96)
0.95

> qnorm((1+0.95)/2)
1.96

likelihood level (q) ⟷ 4.6 ⟷ confidence level (p)

找 c $p = P(W \leq c) = 2P(Z \leq \sqrt{c}) - 1$ $W \sim \chi^2_1$

$q = e^{-\frac{c}{2}}$

$p = P(\Lambda(\theta) \leq -2\log q) = 2P(Z \leq \sqrt{-2\log q}) - 1$

$\Lambda(\theta) \sim \chi^2_1$

100p% CI for $G(\mu, \sigma)$ 4.7 — μ unknown, σ unknown → CI (★)

4.5

- Chi-squared χ^2_k

$W_1, W_2, \cdots, W_n$ indep random variable, $W_i \sim \chi^2_{k_i}$ $\quad S = \sum_{i=1}^n W_i \sim \chi^2_{\sum k_i}$

$Z \sim G(0,1) \Rightarrow Z^2 = W \sim \chi^2_1$ $\quad W \sim \chi^2_1 \Rightarrow P(W \geq w) = 2(1 - P(Z \leq \sqrt{w}))$ $Z \sim G(0,1)$

$Z_1, Z_2, \cdots, Z_n \sim G(0,1)$ $\quad S = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$

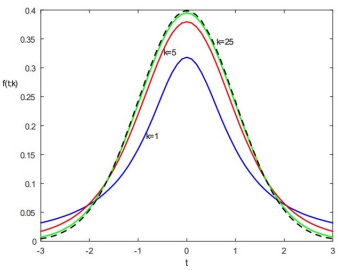
$W \sim \chi^2_2 \Rightarrow W \sim Exp(2)$ $\quad P(W \geq w) = e^{-\frac{w}{2}}$

- t distri

$Z \sim G(0,1)$ $\quad U \sim \chi^2_k \Rightarrow T = \frac{Z}{\sqrt{\frac{U}{k}}} \sim t_k$

$T \sim t(k)$ (d.f ↑)

k 越大, t(k) 图像越接近 $G(0,1)$ →



p-value = $P(D \geq d \,;\, H_0)$

$p-value$	Interpretation
$p-value > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p-value \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p-value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p-value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p-value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

$H_0: \theta = \theta_0$.
p-val $\leq 0.05 \Leftrightarrow$ 95% CI for θ 不包含 θ_0

5.1 $Bi(n,\theta)$ $D = |Y - n\theta_0|$ n large $\overset{CLT}{\Rightarrow} \frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} \sim G(0,1)$

5.2 $G(\mu,\sigma)$ $H_0: \mu = \mu_0$ $p\text{-val} = 2[1 - P(T \leq \frac{|\bar{y} - \mu_0|}{\frac{s}{\sqrt{n}}})]$ $T \sim t_{n-1}$

$H_0: \sigma^2 = \sigma_0^2$ $u = \frac{(n-1)s^2}{\sigma_0^2}$ $U \sim \chi^2_{n-1}$ $\begin{cases} P(U \leq u) > 0.5 & p\text{-val} = 2P(U \geq u) \\ P(U \leq u) < 0.5 & p\text{-val} = 2P(U \leq u) \end{cases}$

5.3 likelihood ratio test statistics $\lambda(\theta_0) = -2\log(R(\theta_0))$ $\lambda(\theta_0) \sim \chi^2_1$

6. Gaussian Response Model 比较 2 个 Variate X 与 Y in 关联性

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)$

$corr(X,Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_Y}$

- linear regression model : $Y_i \sim G(\alpha + \beta x_i)$ $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$
- check linear reg model { scatterplot & fitted line; residual plot $(x_i, \hat{r}_i)$; standard residual plot $(x_i, \hat{r}_i^*)$; standard residual plot $(\hat{\mu}_i, \hat{r}_i^*)$
- 2 population 比较 6.4
- General $G(\mu,\sigma)$ Response Model 6.5 $Y_i \sim G(\mu(x_i), \sigma)$ $\mu(x_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$
  $H_0: \beta_j = 0$ $T \sim t(n-k-1)$

7. Multinomial Model 比较 2 个 model

$\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log(\frac{y_j}{e_j})$ y_j = observed, e_j = expected $p\text{-value} = P(W \geq \lambda(\theta_0))$ $W \sim \chi^2(k-1-p)$

Pearson goodness of fit test 当有 observed 与 expected value 时直接套公式

$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j} \sim \chi^2_{k-1-p}$ $d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$ $p\text{-value} = P(D \geq d)$ $D \sim \chi^2(k-1-p)$

Two Way Table $d.f = (row-1)(col-1)$ ← 先跟据 distribution 求 e_j

8. Cause & Effect

causation

√associate, × casually related 原因