Nonparametric Inference

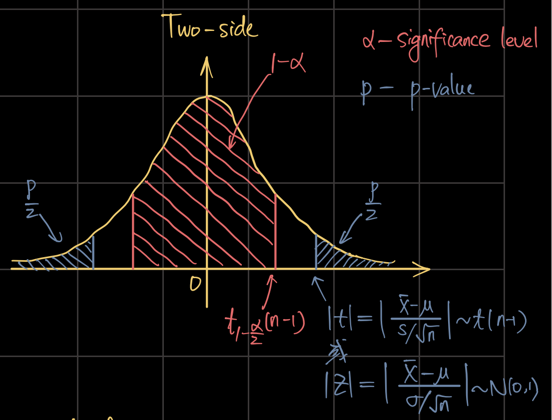
1. Hypothesis Testing

- H_0: Null hypothesis
- H_1: Alternative hypothesis (study hypothesis：即题目中假设啥，啥就是H_1).

— Notion. η: 总体中位数 (population median)

μ: 总体均值 (population mean)

θ: 总体比例/概率 (population probability / probability).



— Conclusion: 要么是 we reject H_0 in favor of H_1 要么是 we do not reject 绝不可能 prove H_0 from a sample.

$|t| < t_{1-\frac{\alpha}{2}}(n)$ 或 $p > \alpha$ $|t| > t_{1-\frac{\alpha}{2}}(n)$ 或 $p < \alpha$

2. Wilcoxon Signed Ranks Test 符号秩检验 (WSR Test)

- Assumption: 观察值来自对称分布，且随机并独立.
- Example:

Observed value	2491	2485	3433	2575	2521	2451	2550	2540
d_i	−9	−15	933	75	21	−49	50	40
sign	−	−	+	+	+	−	+	+
rank	1	2	8	7	3	5	6	4

Hypothesis Testing: H_1: $\eta \neq 2500$.

$\Rightarrow d_i = x_i - 2500$

取$|d_i|$排序.

排序并给出 sign (符号) 和 rank (位次)

计算正/负秩和: $W+ = 8+7+3+6+4 = 28$. $W- = 1+2+5 = 8$,

统计量 $W = \min(W+, W-) = \min(28, 8) = 8$

查 WSR Test 表.

	size α (one-sided)			
	5%	2.5%	1%	0.5%
	size α (two-sided)			
n	10%	5%	2%	1%
1	–	–	–	–
2	–	–	–	–
3	–	–	–	–
4	–	–	–	–
5	0	–	–	–
6	2	0	–	–
7	3	2	0	–
8	5	3	1	0
9	8	5	3	1

Critical value 临界值.

Rejection Region (RR) = {W− : W− ≤ 3}

The observed value of W− = 8 does not lie in RR,

so we do not have significant evidence to reject the Null Hypothesis (H_0).

In other words, there is insufficient evidence that the median is different from 2500.

3. Mann-Whitney U Test (双变量)

- Assumption: 观察值独立，且其分布在各自总体中相同.
- Example:

$\frac{1}{3} \times (2+3+4) = 3$ $\frac{1}{2} \times (5+6) = 5\frac{1}{2}$

Males	74	72	77	76	76	73	75	73	74	75
Rank	$5\frac{1}{2}$	1	13	$10\frac{1}{2}$	$10\frac{1}{2}$	3	8	3	$5\frac{1}{2}$	8
Group	A	A	A	A	A	A	A	A	A	A
Females	77	75	78	79	77	73	78	79	78	80
Rank	13	8	16	$18\frac{1}{2}$	13	3	16	$18\frac{1}{2}$	16	20
Group	B	B	B	B	B	B	B	B	B	B

Hypothesis Testing: H_1: $\eta_A \neq \eta_B$.

Test Statistic 检验统计量.

R_i = 第i组的秩和. $R_A = 5\frac{1}{2} + 1 + 13 + 10\frac{1}{2} + 10\frac{1}{2} + 3 + 8 + 3 + 5\frac{1}{2} + 8 = 68$

$R_B = 13 + 8 + 16 + 18\frac{1}{2} + 13 + 3 + 16 + 18\frac{1}{2} + 16 + 20 = 142$

R_i的最小值理论为 $\frac{1}{2}n(n+1) = 1+2+3+\cdots+n$. 通常写作 $U_A = R_A - \frac{1}{2}n_A(n_A+1)$, $U_B = R_B - \frac{1}{2}n_B(n_B+1)$. $\Rightarrow U = \min(U_A, U_B)$.

在这里 $n_A = n_B = 10$. 则 $R_A = 68 - \frac{1}{2}\times10\times11 = 13$, $R_B = 142 - \frac{1}{2}\times10\times11 = 87$. $\Rightarrow U = \min(U_A, U_B) = 13$.

查M-W U Test表. (5% two-sided)

n_1 \ n_2	2	3	4	5	6	7	8	9	10	11
2	–	–	–	–	–	–	0	0	0	0
3	–	–	–	0	1	1	2	2	3	3
4	–	–	0	1	2	3	4	4	5	6
5	–	0	1	2	3	5	6	7	8	9
6	–	1	2	3	5	6	8	10	11	13
7	–	1	3	5	6	8	10	12	14	16
8	0	2	4	6	8	10	13	15	17	19
9	0	2	4	7	10	12	15	17	20	23
10	0	3	5	8	11	14	17	20	23	26
11	0	3	6	9	13	16	19	23	26	30

Critical value

因此 $RR = \{U : U \le 23\}$

The observed value is $U = 13$, which lies in RR.

Hence we reject H_0 in favor of H_1.

In other words, there is statistically significant evidence that ... .

Parametric Inference

1. Point Estimators & Likelihood

(1) Point Estimation: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

- Unbiased: $E[t(X)] = \theta$.
- Consistent: $Var[t(X)] \to 0$ $(n \to \infty)$

(2) Maximum Likelihood Estimator (MLE)

- The likelihood function: $L(\theta; x_1, x_2, \cdots, x_n) = \underbrace{\prod_{i=1}^{n} Pr(X = x_i)}_{\text{discrete}}$ or $L(\theta) \propto \underbrace{\prod_{i=1}^{n} f(\theta; x_i)}_{\text{continuous}}$
- The log-likelihood function: $\ell(\theta) = \log(L(\theta; x)) = \log\left(\prod_{i=1}^{n} Pr(X=x_i)\right) = \sum_{i=1}^{n} \log(Pr(X=x_i))$.
- $\hat{\theta}_{MLE} = \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\theta; x)$ or $\underset{\hat{\theta}}{\operatorname{argmax}}\, \ell(\theta)$ $\longrightarrow$ $\frac{\partial \ell(\theta)}{\partial \theta} = \ell'(\theta) = 0$. 确保是局部最大值，需使 $\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \ell''(\theta) < 0$

(3) Relative Likelihood

- The relative likelihood function: $R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{MLE})}$
- The relative log-likelihood function: $r(\theta) = \log R(\theta) = \log L(\theta) - \log L(\hat{\theta}_{MLE}) = \ell(\theta) - \ell(\hat{\theta}_{MLE})$.

(4) The sample information: $k(x) = -\ell''(\hat{\theta}_{MLE})$

(5) Newton-Raphson method: 无法找出 $\ell'(\theta) = 0$ 的根时. 重复过程: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$. 直到 $x_{n+1} = x_n$.

$$\Rightarrow \theta^{(n+1)} = \theta^{(n)} - \frac{\ell'(\theta^{(n)})}{\ell''(\theta^{(n)})} \quad \text{直到 } \theta^{(n+1)} = \theta^{(n)}$$

2. Interval Estimation and Multiparameter Models

(1) Interval Estimation using Likelihood.

用相对似然函数 $R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{MLE})}$，令 $L(\theta) \geq p \times L(\hat{\theta}_{MLE})$ i.e. $R(\theta) \geq p \in [0,1]$

$\Rightarrow$ 对数相对似然函数 $r(\theta) = \ell(\theta) - \ell(\hat{\theta}_{MLE}) \geq \log(p)$.

$\Rightarrow$ 50%似然区间定义为 $r(\theta) \geq \log(0.5) = -0.693$. 根据Newton-Raphson方法. $g(\theta) = r(\theta) - \log(0.5) = 0$.

- Example: $X \sim Exp(\theta)$, $\theta > 0$. Hint: $f(x) = \theta e^{-\theta x}$.
  - Likelihood: $L(\theta) \propto \prod_{i=1}^{n} f(x_i) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$
  - Log-likelihood: $\ell(\theta) = n\log\theta - \theta\sum_{i=1}^{n} x_i \Rightarrow \ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i \Rightarrow \hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}} = 0.016$
  - Relative Likelihood: $R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{MLE})} = \frac{\theta^n e^{-\theta\sum x_i}}{\hat{\theta}_{MLE}^n e^{-\hat{\theta}_{MLE}\sum x_i}} = \frac{\theta^{24} e^{-1539\theta}}{0.016^{24} e^{-0.016\times 1539}}$
  - Log relative likelihood: $r(\theta) = \ell(\theta) - \ell(\hat{\theta}_{MLE}) = 24\log\theta - 1539\theta - 24\log(0.016) + 0.016 \times 1539 = 24\log(\theta) - 1539\theta + 123.9$.

- 50% likelihood interval : $r(\theta) \geq \log(0.5) = -0.693 \Rightarrow 24\log(\theta) - 1539\theta + 123.9 \geq -0.693$. 用软件画图大致得 $\hat{\theta} \in (0.011, 0.019)$.

- Newton-Raphson : $\theta_B^{(j+1)} = \theta_B^{(j)} - \dfrac{g(\theta_B^{(j)})}{g'(\theta_B^{(j)})}$ where $g(\theta_B) = 24\log(\theta_B) - 1539\theta_B + 124.593 \Rightarrow g'(\theta_B) = \dfrac{24}{\theta_B} - 1539$.

  θ_B: Bound $\Rightarrow$ θ_L – Lower bound, θ_U – Upper bound

Iteration	$\theta^{(j)}$	$g(\theta^{(j)})$	$g'(\theta^{(j)})$	$\frac{g(\theta^{(j)})}{g'(\theta^{(j)})}$
0	看图大约得 0.011	-0.573	642.82	-0.0009
1	0.012	-0.023	461	-0.0001
2	θ_L: 0.012	-0.023	461	-0.0001

Iteration	$\theta^{(j)}$	$g(\theta^{(j)})$	$g'(\theta^{(j)})$	$\frac{g(\theta^{(j)})}{g'(\theta^{(j)})}$
0	看图大约得 0.019	0.232	-275.84	-0.00084
1	0.020	-0.07555	-339	0.0002
2	θ_U: 0.020	-0.07555	-339	0.0002

$\Rightarrow$ A 50% likelihood interval for $\hat{\theta}_{MLE}$ is $(0.012, 0.020)$

(2) Confidence Interval

Assume that $X_1, \cdots, X_n$ are independent $N(\mu, \sigma^2)$. And $\hat{\mu}_{MLE} = \bar{X}$, 构建总体均值 μ_T 的区间估计.

⓪ Pivotal funtion 枢轴量方程，记为 $PIV(\theta_T, X)$.

① One-sample t-interval and t-test.

- σ is known. $PIV(\mu_T, X) = \dfrac{\bar{X} - \mu_T}{\sigma/\sqrt{n}} = Z \sim N(0,1)$
- σ_T is unknown. $PIV(\mu_T, X) = \dfrac{\bar{X} - \mu_T}{S/\sqrt{n}} = t \sim t(n-1)$. $\Rightarrow |PIV| \leq t_{1-\frac{\alpha}{2}}(n-1) \Rightarrow CI: \left(\underbrace{\bar{X}}_{\text{estimate}} \pm t_{1-\frac{\alpha}{2}}(n-1) \cdot \underbrace{\frac{S}{\sqrt{n}}}_{\text{estimate standard error}}\right)$.

② Two-sample t-interval and t-test. (parametric equivalent of a Mann-Whitney test).

Assuming that $\sigma_1 = \sigma_2 = \sigma_T$, a pivotal function for $\mu_1 - \mu_2$ is derived as follows.

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_T^2}{m}\right) \text{ independently of } \bar{Y} \sim N\left(\mu_2, \frac{\sigma_T^2}{n}\right). \Rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma_T^2\right).$$

- σ_T is known. $PIV(\mu_1 - \mu_2; X, Y) = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\sigma_T^2}} \sim N(0,1)$.
- σ_T is unknown. Pooled estimate of σ is usually denoted S_p, where

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{m+n-2} \Rightarrow PIV(\mu_1 - \mu_2; X, Y) \sim t(m+n-2)$$

(3) Intervals based on Approximate Confidence (large sample).

① Wilks Intervals

当 θ 设置为其真实值 θ_T 后. $PIV(\theta) = 2\left(\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)\right) \sim \chi_1^2$, approximately.

$\Rightarrow$ A confidence interval stems from the result that $Pr\left(2\left(\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)\right) \leq \chi_1^2(c)\right) \approx c$.

$\Rightarrow$ An approximate $100c\%$ confidence (Wilks) interval for θ : $\left\{\theta : 2\left(\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)\right) \leq \chi_1^2(c)\right\}$.

$\Rightarrow$ In terms of the relative log-likelihood function: $\{\theta : -2r(\theta) \le \chi_1^2(c)\} \Rightarrow \{\theta : r(\theta) \ge -\frac{1}{2}\chi_1^2(c)\}$ 在样本数量较大时有效.

※ Wilks Interval 具有不变性 (Invariance property)

若 $\beta = g(\theta)$, 且 $g(\cdot)$ is monotonic. If (θ_a, θ_b) is a Wilks interval for θ then, $(g(\theta_a), g(\theta_b))$ would be a Wilks interval for β.

② Wald Intervals

当 θ_T 是 θ 的真实值时, $\hat{\theta}_{MLE} \sim N(\theta_T, \frac{1}{k(x)})$, approximately, where sample information $= k(x) = -l''(\hat{\theta}_{MLE})$.

$\Rightarrow PIV(\theta) = \dfrac{\hat{\theta}_{MLE} - \theta_T}{\sqrt{1/k(x)}} \sim N(0,1)$. $\Rightarrow$ An approximate $100c\%$ confidence (Wald) interval for θ: $\left(\hat{\theta}_{MLE} \pm \Phi^{-1}(1-\frac{1-c}{2})\sqrt{\frac{1}{k(x)}}\right)$. $\alpha = 1-c$

※ Not invariant, and always symmetric.

※ MLE with 2 parameters

仍要检查是否有局部最大值. 引入 Hessian of $l(\cdot)$ w.r.t. (with respect to) $\hat{\mu}, \hat{\sigma}$. That is $H = \begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l}{\partial \sigma \partial \mu} & \frac{\partial^2 l}{\partial \sigma^2} \end{pmatrix}\Bigg|_{\hat{\mu}, \hat{\sigma}}$.

For the turning point $\hat{\mu}, \hat{\sigma}$ to be a maximum, H must be $\underbrace{\text{negative definite}}_{\forall a,\ a^T H a < 0}$.

$\Rightarrow$ Newton - Raphson method: $\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} - H^{-1} g$.

where $H = \begin{pmatrix} \frac{\partial^2 l}{\partial \theta_1^2}\Big|_{\hat{\theta}^{(j)}} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2}\Big|_{\hat{\theta}^{(j)}} & \cdots \\ \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1}\Big|_{\hat{\theta}^{(j)}} & \frac{\partial^2 l}{\partial \theta_2^2}\Big|_{\hat{\theta}^{(j)}} & \cdots \\ \vdots & \vdots & \end{pmatrix}$ and $g = \begin{pmatrix} \frac{\partial l}{\partial \theta_1}\Big|_{\hat{\theta}^{(j)}} \\ \frac{\partial l}{\partial \theta_2}\Big|_{\hat{\theta}^{(j)}} \\ \vdots \end{pmatrix}$

这么丧尽天良的傻逼公式
只需要在R中用optim()就行了

(4) Confidence Regions

去死吧

Lecture 12 & 13

3. Hypothesis Testing ($H_0: B(\theta)=0$ vs. $H_1: B(\theta)\neq 0$)

(1) Generalized Likelihood Ratio Test (GLRT) $\Lambda = \dfrac{L(\hat{\theta}_{H_1})}{L(\hat{\theta}_{H_0})}$.

value of θ satisfying $B(\theta)=0$ which maximizes the likelihood (i.e. the restricted MLE).

If H_0 is true then in the limit as $n\to\infty$, $2\lambda = 2\left(\ell(\hat{\theta}_{H_1}) - \ell(\hat{\theta}_{H_0})\right) \sim \chi^2_r$

MLE of θ (under H_1)

Reject H_0 if $2\lambda > \chi^2_r$, r = number of restrictions imposed by H_0.

- Example: $X \sim \text{Poisson}(\theta)$. $H_0: \theta=\theta_0$ vs. $H_1: \theta\neq\theta_0$

  - Likelihood: $L(\theta) = \prod_{i=1}^{n} \dfrac{e^{-\theta}\theta^{x_i}}{x_i!}$ $\Rightarrow$ Log-likelihood: $\ell(\theta) = \sum_{i=1}^{n} \log \dfrac{e^{-\theta}\theta^{x_i}}{x_i!} = \sum_{i=1}^{n}\left(-\theta + x_i\log\theta - \log(x_i!)\right)$

  - 在H_0下. 假设$\theta=\theta_0$, 则 $\ell(\hat{\theta}_{H_0}) = \sum_{i=1}^{n}\left(-\theta_0 + x_i\log\theta_0 - \log(x_i)\right) = -n\theta_0 + \left(\sum_{i=1}^{n}x_i\right)\log\theta_0 - \sum_{i=1}^{n}\log(x_i!)$

    在H_1下. $\ell'(\theta) = -n + \dfrac{1}{\theta_0}\sum_{i=1}^{n}x_i = 0$ $\Rightarrow$ $\hat{\theta}_{MLE} = \dfrac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$. $\Rightarrow$ $\ell(\hat{\theta}_{H_1}) = -n\bar{x} + \left(\sum_{i=1}^{n}x_i\right)\log\bar{x} - \sum_{i=1}^{n}\log(x_i!)$

  - GLRT: $2\lambda = 2\left(\ell(\hat{\theta}_{H_1}) - \ell(\hat{\theta}_{H_0})\right) = 2\left(-n\bar{x} + \left(\sum_{i=1}^{n}x_i\right)\log\bar{x} + n\theta_0 - \left(\sum_{i=1}^{n}x_i\right)\log\theta_0\right)$.

    $$= 2\left(\left(\sum_{i=1}^{n}x_i\right)\left(\log\frac{\bar{x}}{\theta} - 1\right) + n\theta_0\right).$$

补充已知: $\sum_{i=1}^{200}x_i = 1800$, $H_0: \theta=10$ vs. $H_1: \theta\neq 10$.

$\Rightarrow$ GLRT: $2\lambda = 2\left(1800\left(\log\dfrac{1800/200}{10} - 1\right) + 200\times 10\right) = 20.7021$

- Rejection Rejion: $RR = \{x: 2\lambda > \chi^2_r(1-\alpha)\}$ $\xrightarrow[\alpha=5\%]{r=1}$ $RR = \{x: 2\lambda > \chi^2_1(0.95) = 3.8415\}$.

- Conclusion: We reject H_0 because this hypothesis does not look plausible in the light of the observed data.

※ A Good Testing Procedure?

	Reject H_0	Do not reject H_0
H_0真	Type 1 error: α	$1-\alpha$
H_1真	$1-\beta$	Type 2 error: β

$\Rightarrow Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha$ significance level. (显著性水平)

$\Rightarrow Pr(\text{Reject } H_0 \mid H_1 \text{ is true}) = 1-\beta$ power (功效)

(2) GLRT for multinomials

① GLRT for a multinomial Distribution

- Data: $x_1, x_2, \cdots, x_k$. where $\sum_{i=1}^{n}x_i = n$

- Model: $X = \{X_1, X_2, \cdots, X_k\} \sim Mu(n,\theta)$, where $\sum_{i=1}^{k}\theta_i = 1$.

H_0: $\theta_1, \theta_2, \cdots, \theta_k$ 可以用 p 个未知参数 $\phi_1, \phi_2, \cdots, \phi_p$ $(p<k-1)$

$\Rightarrow$ Simple examples: $H_0: \theta_i = \frac{1}{k}$ $(p=0)$, H_0: the frequencies follow a Binomial(k,ϕ) distribution $(p=1)$.

$\Rightarrow$ Likelihood: $L(\theta) = K\prod_{i=1}^{k}\theta_i^{x_i}$ $\Rightarrow$ Log-Likelihood: $\ell(\theta) = \log(K) + \sum_{i=1}^{k} x_i \log(\theta_i)$

$\Rightarrow$ Computing $2\lambda = 2\left(\ell(\hat\theta_{H_1}) - \ell(\hat\theta_{H_0})\right)$ $\Rightarrow$ GLRT statistic: $2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right)$. where O_i - observed frequencies (H_0 is true).
E_i - expected frequencies

$\Rightarrow RR = \left\{x: 2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) > \chi^2_{k-1-p}(1-\alpha)\right\}$. where k - the number of levels (or categories)

p - the number of parameters being estimated.

- Example: (Sales)

Product	1	2	3	4	5
Frequency (Observed)	6	14	4	4	2
Expected	6	6	6	6	6

H_0: 顾客随机购买特定产品导致 $\theta_i = \frac{1}{5}$

$\Rightarrow$ Expected (under H_0): $30 \times \frac{1}{5} = 6$.

$\Rightarrow$ The test statistic: $2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right)$

$\Rightarrow$ The observed value of the test statistic: $2\left(6\log\left(\frac{6}{6}\right) + 14\log\left(\frac{14}{6}\right) + 4\log\left(\frac{4}{6}\right) + 4\log\left(\frac{4}{6}\right) + 2\log\left(\frac{2}{6}\right)\right) = 12.84$.

$\Rightarrow RR = \left\{x: 2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) > \chi^2_{5-1-0}(1-0.05) = \chi^2_4(0.95) = 9.49\right\}$. $\Rightarrow$ Reject H_0.

并没有估计参数.

- Alternative method: Pearson's χ^2 test.

② Goodness of Fit 拟合优度.

- Pearson's χ^2 test: $2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) \approx \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

  - Model: $f(x;\theta)$

  - Data: 假设 $x \in \mathbb{R}$ 遵循此模型. 数据被划分成 k 个不相交 (disjoint) 区间.

  - H_0: 数据是从模型 $f(x;\theta)$ 中抽取的 i.i.d. 样本.

    Under H_0 the statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1-p}$

    k = # of intervals / catagories

    p = # of parameters estimated under H_0

    基于 H_0 真的预期数据

  - $RR = \left\{x: \chi^2 > \chi^2_{k-1-p}(1-\alpha)\right\}$.

- Example: 根据 Hardy-Weinberg Equilibrium, 基因型 AA, Aa, aa 在人群中的频率分别为 $(1-\theta)^2$, $2\theta(1-\theta)$, θ^2.

Sample:

Blood Type	M	MN	N	Total
Observed Frequency	342	500	187	1029

H_0: $\theta_1 = (1-\theta)^2$, $\theta_2 = 2\theta(1-\theta)$, $\theta_3 = \theta^2$

H_1: the multinominal probabilities do not have the specified form.

- Likelihood: $L(\theta) = K\prod_{i=1}^{3}\theta_i^{x_i} = K\cdot\left((1-\theta)^2\right)^{342}\cdot\left(2\theta(1-\theta)\right)^{500}\cdot\left(\theta^2\right)^{187} = K\theta^{874}(1-\theta)^{1184}$.

$\Rightarrow L'(\theta) = K\cdot 874\theta^{873}(1-\theta)^{1184} - K\theta^{874}\cdot 1184(1-\theta)^{1183} = 0 \Rightarrow \hat\theta_{MLE} = \frac{x}{n} = \frac{874}{2058} = 0.4247$.

续表:

Expected Frequency	340.6	502.8	185.6

$1029(1-\hat\theta)^2 = 340.6$

$$\Rightarrow \chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{(342-340.6)^2}{340.6} + \frac{(500-502.8)^2}{502.8} + \frac{(187-185.6)^2}{185.6} = 0.03190758$$

只估计1个参数θ.

- $RR = \{x : \chi^2 > \chi^2_{3-1-1}(1-0.05) = \chi^2_1(0.95) = 3.84\}$ $\Rightarrow$ Do not reject H_0 in favor of H_1.

※ In comparison, the GLRT statistic is $2\sum_{i=1}^{3} O_i \log\left(\frac{O_i}{E_i}\right) = 2\left(342\log\left(\frac{342}{340.6}\right) + 500\log\left(\frac{500}{502.8}\right) + 187\log\left(\frac{187}{185.6}\right)\right) = 0.03190228$

(3) Properties of MLEs

① Invariance 不变性 : If $\beta = g(\theta)$, then $\hat{\beta} = g(\hat{\theta})$.

② Consistency 一致性: $n \to \infty$, $\hat{\theta} \to \theta_T$

※ Expected Log-Likelihood

- Result 1: $E_T\left[\frac{\partial l}{\partial \theta}\Big|_{\theta_T}\right] = 0$ $\Rightarrow$ log-likelihood 在 θ_T 转折.
- Result 2: $Var\left[\frac{\partial l}{\partial \theta}\Big|_{\theta_T}\right] = E_T\left[\left(\frac{\partial l}{\partial \theta}\Big|_{\theta_T}\right)^2\right]$, due to $\left(E_T\left[\frac{\partial l}{\partial \theta}\Big|_{\theta_T}\right]\right)^2 = 0$

  $\Rightarrow$ Fisher Information : $I_\theta \equiv E_T\left[\left(\frac{\partial l}{\partial \theta}\Big|_{\theta_T}\right)^2\right]$.
- Result 3 : $I_\theta \equiv E_T\left[\left(\frac{\partial l}{\partial \theta}\Big|_{\theta_T}\right)^2\right] = -E\left[\frac{\partial^2 l}{\partial \theta^2}\Big|_{\theta_T}\right]$ $\Rightarrow$ 非负, 因此转折点 θ_T 为最大值.

③ Asymptotic Normality 渐近正态性. (Large sample props).

$n \to \infty$: $(\hat{\theta} - \theta_T) \sim N(0, \theta_T^{-1})$, $\hat{\theta} \sim N(\theta_T, I_\theta^{-1})$

- Sample information matrix 样本信息矩阵 $K(x) = -H$ (MLE的对数似然函数 Hessian 的负值) 是 I_θ 的充分近似.

$\Rightarrow \hat{\theta}_{MLE} \sim N(\theta_T, K(x)^{-1})$.

Bayesian Inference

1. Introduction

(1) Non-Bayesian statistics is often called Frequentist statistics.

① Definitions for probability.

- Bayesian：一个人相信一个命题的程度
- Frequentist：在大量试验中，一个事件相对频率的"极限"。

② Difference

- Frequentist：概率只能分配给"random experiments"或"events"。
- Bayesian：概率可以分配给任何东西，无论是否有数据。

(2) Probability Statement

① Prior probability 先验概率：未观察到任何新数据前，根据已有知识、历史经验或假设，对某事件发生概率的初步估计。

eg. 从以往来看，吉林市的降水概率为30%，那么 Pr(下雨) = 30%（基于历史数据的先验概率）。

② Posterior probability 后验概率：观察到新数据之后对事件发生概率的更新估计。

基于 Bayes' theorem：$Pr(A|B) = \frac{Pr(A) \cdot Pr(B|A)}{Pr(B)}$

$Pr(A)$ — 先验概率
$Pr(B|A)$ — 似然函数
$Pr(B)$ — 边际概率

2. • Notion：$p(\theta)$ — prior beliefs

$p(X|\theta)$ — data

(1) 计算后验分布

- 对于连续的θ：$p(\theta|X) = \frac{p(\theta) \cdot p(X|\theta)}{p(X)} = \frac{p(\theta)\, p(X|\theta)}{\int p(\theta) p(X|\theta)\, d\theta} \propto p(\theta)\, p(X|\theta)$
- $p(x|\theta) = L(\theta|x) \Leftarrow L(\theta|x_1,\cdots,x_2) = \prod_{i=1}^{n} p(x_i|\theta)$ i.e. $p(\theta|X) = \frac{p(\theta) L(\theta|X)}{p(X)}$

⇒ For the case of one data point x & a parameter θ：

$$p(\theta|x) = \frac{p(\theta) \cdot L(\theta|x)}{p(x)}$$

- $p(\theta|x)$：the posterior distribution.
- $L(\theta|x)$：the likelihood for a parameter θ given the data x.
- $p(\theta)$：the prior distribution of the parameter θ.
- $p(x)$：the unconditional (marginal) distribution of the data.

(2) Normalising constant / Marginal distribution : $p(x)=\int p(\theta)p(x|\theta)\,d\theta=\int p(\theta)L(\theta|x)\,d\theta$.

$\Rightarrow p(\theta|x)\propto p(\theta)L(\theta|x)$ posterior $\propto$ prior $\times$ likelihood

(3) Conjugate prior 共轭先验

若先验分布 $p(\theta)\in$某分布族 且 后验分布 $p(\theta|x)\in$同一分布族，则 $p(\theta)$是 似然函数 $p(x|\theta)$的共轭先验.

3. Example: Tossing a coin (Heads or Tails) $\Rightarrow$ Binomial model: $p(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}$. 用Bayesian推断来估计成功比例 θ.

- Likelihood: $L(\theta|x)=p(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}\propto\theta^x(1-\theta)^{n-x}$
- Prior (Bata): $p(\theta)=\dfrac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$, $\theta\in(0,1)$ 其中 $B(\alpha,\beta)=\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- Posterior: $p(\theta|x)\propto p(\theta)\cdot p(x|\theta)$

$$\Rightarrow p(\theta|x)\propto\theta^{\alpha-1}(1-\theta)^{\beta-1}\cdot\theta^x(1-\theta)^{n-x}$$

$$\Rightarrow p(\theta|x)\propto\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

$$\Rightarrow \theta|x\sim Beta(\alpha+x,\beta+n-x)\quad\Rightarrow p(\theta|x)=\frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}{B(\alpha+x,\beta+n-x)}$$

$\Rightarrow$ 参数更新: $\alpha\to\alpha+x$, $\beta\to\beta+n-x$

有缘再续

※ 如何找到共轭先验分布.

似然函数 $L(\theta|x)=p(x|\theta)$	共轭先验分布 $p(\theta)$	参数意义
Binomial (θ)	Beta(α,β), $p(x)=\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	α,β: 成功/失败次数的先验假设
Poisson (λ)	Gamma(α,λ), $p(x)=\dfrac{x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}}{\Gamma(\alpha)}$	
Exp (λ)	Gamma(α,β), $p(x)=\dfrac{x^{\alpha-1}e^{-\frac{1}{\beta}x}}{\beta^{\alpha}\Gamma(\alpha)}$	α,β: 事件率的先验假设
Normal (μ,σ_0^2)	Normal (μ,σ^2)	μ,σ^2: 均值和方差的先验假设