PART 1. Clinical Trials 临床试验.

1. 盲法 (Blinding)

(1) 单向盲法 (Single): patient

(2) 双向盲法 (Double): patient & treatment team. (更科学. 紧急情况下应揭盲 (unblind)). 涉及手术和毒性不适用.

→ 影响: ① Patient: 知情的话会影响心理, 尤其精神治疗时.

② Treatment team: 更关注接受 new treatment 的患者, 影响患者态度, 从而影响治疗反应.

③ Evaluator: 更倾向于记录 new treatment 的有利反应, 尤其是主观评价的结果.

2. Hypothesis Test.

	Reject H_0	Fail to 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)

3. A - Placebo, B - Aicturan (new treatment), θ_i - Probability of a patient dying within 1 year having A or B.

The hypothesis we wish to test are $H_0: \theta_A = \theta_B$ vs. $H_1: \theta_A \neq \theta_B$ with significance level of α & power of $1-\beta$.

Suppose that the clinical difference $\delta = \theta_A - \theta_B \sim$ Normal. Therefore, a normal approximation the number of patiences is

$$N = \frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{(\theta_A - \theta_B)^2}\left[\Phi^{-1}(1-\tfrac{\alpha}{2}) + \Phi^{-1}(1-\beta)\right]^2 \quad \text{(向上取整)}$$

Proof.

假设有 $2N$ 个患者, 一半一半. 并设 (X_A, X_B) 表示死掉的患者数量. 所以有 $X_i \sim Binomial(N, \theta_i)$

若假设 N 足够大, 则可近似成正态分布. 即 $X_i \sim N(N\theta_i, N\theta_i(1-\theta_i))$ since $E[X_i] = N\theta_i$, $Var[X_i] = N\theta_i(1-\theta_i)$

记死亡率 $\hat{\theta}_i = \frac{X_i}{N}$. 那么 $E[\hat{\theta}_i] = \frac{1}{N}E[X_i]$. $Var[\hat{\theta}_i] = \frac{1}{N}Var[X_i]$. 则 $\hat{\theta}_i \sim N\left(\theta_i, \frac{\theta_i(1-\theta_i)}{N}\right)$.

因此 $\hat{\theta}_A - \hat{\theta}_B \sim N\left(\delta = \theta_A - \theta_B, \frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}\right)$

我们希望 $\delta > 0$ (用安慰剂的比用药的死的多), $T = \frac{(\hat{\theta}_A - \hat{\theta}_B) - \delta}{\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}} \xrightarrow{H_0: \delta = 0} \frac{\hat{\theta}_A - \hat{\theta}_B}{\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}} \sim N(0,1)$

由于 $Pr(\text{拒绝 } H_0 \mid \text{真实 } \delta) = 1-\beta$, 即 $Pr(|T| > \Phi^{-1}(1-\frac{\alpha}{2})) = 1-\beta$.

$$\Rightarrow Pr\left(\hat{\theta}_A - \hat{\theta}_B > \Phi^{-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}\right) = Pr\left(Z > \frac{\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}} - \delta}{\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}}\right) = 1-\beta$$

$$\Rightarrow Pr\left(Z \leq \frac{\delta - \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}}{\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}}\right) = 1-\beta \Rightarrow \frac{\delta - \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}}{\sqrt{\frac{\theta_A(1-\theta_A) + \theta_B(1-\theta_B)}{N}}} = \Phi^{-1}(1-\beta)$$

$$\Rightarrow \frac{\delta}{\sqrt{\frac{\theta_A(1-\theta_A)+\theta_B(1-\theta_B)}{N}}} = \Phi^{-1}(1-\tfrac{\alpha}{2}) + \Phi^{-1}(1-\beta)$$

$$\Rightarrow N = \frac{\theta_A(1-\theta_A)+\theta_B(1-\theta_B)}{(\theta_A-\theta_B)^2}\left[\Phi^{-1}(1-\tfrac{\alpha}{2}) + \Phi^{-1}(1-\beta)\right]^2, \quad \delta = \theta_A - \theta_B.$$

For countinue response: two treatment have mean responses (μ_A, μ_B), common std. σ. Using a similar argument as before:

$$N = \frac{2\sigma^2}{(\mu_A-\mu_B)^2}\left[\Phi^{-1}(1-\tfrac{\alpha}{2}) + \Phi^{-1}(1-\beta)\right]^2.$$

※ 实践中我们有限的患者进行试验.

① Reason: Patients are likely withdraw from the trial.

② Solution: Estimate the likely withdraw rate from similar existing studies and scale up the sample size appropriately to account for this.

4. Meta-Analysis: 将多个研究结果进行统计分析的组合.

(1) Main statistical objectives (Via. 2023 & 2022)

① Consistent and objective display of the data from different trials. ① 一致且客观地展示不同试验的数据.

② Testing of an overall hypothesis. ② 对总体的零假设进行检验.

③ Estimation of an average treatment effect. ③ 估计平均治疗效果.

④ Investigation of any statistical heterogeneity between trials. ④ 调查试验之间是否存在任何统计异质性.

(2) Binary response variables: Death (Failure) / Survival (Success).

	Failure	Success
Treatment	a	b
Control	c	d

$\Rightarrow$ 比值比: Odd ratio (OR) $= \frac{\text{odd of failure on treatment}}{\text{odd of failure on control}} = \frac{a/b}{c/d} = \frac{ad}{bc}$.

$\Rightarrow \log(OR) = \log\left(\frac{ad}{bc}\right) = \log a - \log b - \log c + \log d$.

$\Rightarrow Var[\log(OR)] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$. Proof: $Var[g(X)] = \left(\frac{dg}{dx}\right)^2 \cdot Var[X]$

$\Rightarrow Var[\log a] = \left(\frac{1}{a}\right)^2 \cdot \underbrace{a}_{Var[a]} = \frac{1}{a}$

$X \sim Bin(n,p) \xrightarrow{n\to\infty} Poi(\lambda = np)$.

(3) Meta分析法: 假设我们需要结合C个试验的结果

(Notion) · $\hat{Y}_c$ - the countinuous effect size in the c-th trial, eg. log(OR). $c = 1, 2, \cdots, C$

· W_c - the reciprocal (倒数) of the effect size variance, i.e. $W_c = \frac{1}{Var[\hat{Y}_c]}$. 作为权重意义出现, 方差越大 权重越小.

① Fixed effect approach 固定效应方法: 假设所有试验存在相同的潜在效应量 (underlying effect size): $\hat{Y}_c \sim N(\mu, \frac{1}{W_c})$.

The only unknown parameter is μ, as the variance of each individual trial is known.

(i) MLE yields the estimator $\hat{\mu}_{FE} = \frac{\sum_{c=1}^{C} W_c \hat{Y}_c}{\sum_{c=1}^{C} W_c}$

Proof. $\hat{Y}_c \sim N(\mu, \frac{1}{w_c})$, $f(\hat{Y}_c;\mu) = \frac{1}{\sqrt{2\pi/w_c}} \exp\left(-\frac{(\hat{Y}_c-\mu)^2}{2/w_c}\right)$

$\Rightarrow$ Likelihood function: $L(\mu) = \prod_{c=1}^{C} f(\hat{Y}_c-\mu) = \prod_{c=1}^{C} \frac{1}{\sqrt{2\pi/w_c}} \exp\left(-\frac{(\hat{Y}_c-\mu)^2}{2/w_c}\right)$

$\Rightarrow$ Log-likelihood function: $\log L(\mu) = \sum_{c=1}^{C} \log\left(\frac{1}{\sqrt{2\pi/w_c}} \exp\left(-\frac{(\hat{Y}_c-\mu)^2}{2/w_c}\right)\right) = \sum_{c=1}^{C}\left(-\frac{1}{2}\log\left(\frac{2\pi}{w_c}\right) - \frac{w_c}{2}(\hat{Y}_c-\mu)^2\right)$

$= -\frac{1}{2}\sum_{c=1}^{C}\log\left(\frac{2\pi}{w_c}\right) - \frac{1}{2}\sum_{c=1}^{C} w_c(\hat{Y}_c-\mu)^2$

$\Rightarrow$ Differentiating with respect to μ: $\frac{\partial}{\partial\mu}L(\mu) = -\frac{1}{2}\sum_{c=1}^{C} w_c \cdot 2(\hat{Y}_c-\mu)(-1) = \sum_{c=1}^{C} w_c(\hat{Y}_c-\mu) = 0$

$\Rightarrow$ Solving for μ: $\sum_{c=1}^{C} w_c\hat{Y}_c = \mu\sum_{c=1}^{C} w_c \Rightarrow \hat{\mu} = \frac{\sum_{c=1}^{C} w_c\hat{Y}_c}{\sum_{c=1}^{C} w_c}$ is unbiased, $E[\hat{\mu}_{FE}] = \frac{\sum_{c=1}^{C} w_c E[\hat{Y}_c]}{\sum_{c=1}^{C} w_c} = \mu$.

(ii) The variance of $\hat{\mu}_{FE}$ is $Var[\hat{\mu}_{FE}] = \frac{1}{\sum_{c=1}^{C} w_c}$.

Proof. $Var[\hat{\mu}_{FE}] = Var\left[\frac{\sum_{c=1}^{C} w_c\hat{Y}_c}{\sum_{c=1}^{C} w_c}\right] = \frac{\sum_{c=1}^{C} w_c^2 Var[\hat{Y}_c]}{\left(\sum_{c=1}^{C} w_c\right)^2} = \frac{\sum_{c=1}^{C} w_c^2 \cdot (1/w_c)}{\left(\sum_{c=1}^{C} w_c\right)^2} = \frac{1}{\sum_{c=1}^{C} w_c}$.

which yields the sampling distribution: $\hat{\mu}_{FE} = \frac{\sum_{c=1}^{C} w_c\hat{Y}_c}{\sum_{c=1}^{C} w_c} \sim N\left(\mu, \frac{1}{\sum_{c=1}^{C} w_c}\right)$.

(iii) A 95% confidence interval for the population effect μ is then given by $\hat{\mu}_{FE} \pm 1.96\sqrt{Var[\hat{\mu}_{FE}]}$.

② Random effect approach 随机效应法. 每个试验都有一个自己的真实效应量, i.e. $\hat{Y}_c \sim N(\mu_c, \frac{1}{w_c})$. $\mu_c \sim N(\mu, \tau^2)$.

The model can be simplify to $\hat{Y}_c \sim N(\mu, \frac{1}{w_c} + \tau_c^2)$. $\Rightarrow$ MLE: $\hat{\mu}_{RE} = \frac{\sum_{c=1}^{C} \frac{1}{\frac{1}{w_c}+\tau^2}\hat{Y}_c}{\sum_{c=1}^{C} \frac{1}{\frac{1}{w_c}+\tau^2}} \sim N\left(\mu, \frac{1}{\sum_{c=1}^{C} \frac{1}{\frac{1}{w_c}+\tau^2}}\right)$

Proof. $f(\hat{Y}_c;\mu) = \frac{1}{\sqrt{2\pi\left(\frac{1}{w_c}+\tau_c^2\right)}} \exp\left(-\frac{(\hat{Y}_c-\mu)^2}{2\left(\frac{1}{w_c}+\tau_c^2\right)}\right)$.

$\Rightarrow L(\mu) = \prod_{c=1}^{C} f(\hat{Y}_c;\mu)$ $\Rightarrow \log L(\mu) = \sum_{c=1}^{C} \log f(\hat{Y}_c-\mu) = -\frac{1}{2}\sum_{c=1}^{C}\log\left(2\pi\left(\frac{1}{w_c}+\tau^2\right)\right) - \frac{1}{2}\sum_{c=1}^{C}\frac{(\hat{Y}_c-\mu)^2}{\frac{1}{w_c}+\tau_c^2}$

$\Rightarrow \frac{\partial}{\partial\mu}\log L(\mu) = -\frac{1}{2}\sum_{c=1}^{C}\frac{2(\hat{Y}_c-\mu)(-1)}{\frac{1}{w_c}+\tau_c^2} = \sum_{c=1}^{C}\frac{\hat{Y}_c-\mu}{\frac{1}{w_c}+\tau_c^2} = 0$

$\Rightarrow \sum_{c=1}^{C}\frac{\hat{Y}_c}{\frac{1}{w_c}+\tau_c^2} - \mu\sum_{c=1}^{C}\frac{1}{\frac{1}{w_c}+\tau_c^2} = 0$ $\Rightarrow \hat{\mu}_{RE} = \frac{\sum_{c=1}^{C}\frac{\hat{Y}_c}{\frac{1}{w_c}+\tau^2}}{\sum_{c=1}^{C}\frac{1}{\frac{1}{w_c}+\tau^2}}$

在随机效用模型中，估计 $\tau_c^2 = \max\left(0, \frac{Q-(C-1)}{S_1 - \frac{S_1}{S_2}}\right)$

其中统计量Q用来衡量不同试验效果之间的差异 $Q = \sum_{c=1}^{C} w_c(\hat{Y}_c - \hat{\mu}_{FE})^2$，$(C-1)$为自由度.

S_1与S_2指权重和与权重平方和. 即 $S_1 = \sum_{c=1}^{C} w_c$，$S_2 = \sum_{c=1}^{C} w_c^2$. 用来调整Q值，以获得更稳健的$\tau_c^2$估计.

5. 概念题.

(1) 临床试验设计 Design of Clinical Trials (四个阶段 Phrase) (via. 2023)

① Clinical Pharmacology and Toxicity 临床药理学和毒性.

主要关注药物的安全性，并在健康的人类志愿者(非患者)身上进行. 同时建立剂量计划(dosing schedule)

These are primarily concerned with drug safety, and are performed on healthy human volunteers (not patients). Dosing schedules are also established.

② Initial Clinical Investigation for Treatment Effect 初始临床疗效研究.

在患者中进行小规模研究，观察药物的有效性和安全性。是否有证据表明该药物有效且无毒？如今，II期试验更多采用随机化设计。但在过去是没有对照组的。

These are small scale studies in patients, looking at the effectiveness and safety of the drug. Is there any evidence that the drug actually works and is non-toxic? It is becoming increasingly common for phase II trials to be randomized, but in the (fairly recent) past they were generally uncontrolled (i.e., no control group).

③ Full Scale Evaluation of Treatment 治疗的全面评估.

在药物被证明具有合理疗效并值得进行全面研究后。在随机对照试验中，将其与当前针对的相同疾病的治疗方法（现有标准治疗方法（或Placebo））比较

After a drug is shown to be reasonably effective and worthy of a full scale investigation, it is essential (必要) to compare it with current standard treatments for the same disease (and / or with a placebo) in a randomised controlled trial.

④ Post-marketing Surveillance 上市后监测.

药物在被批准使用后，应收集副作用的证据，以获取其长期有效性的信息。若产生严重副作用仍有可能被撤回。

If a drug is approved for use, its performance is still closely monitored for evidence of side effects (副作用) and to obtain more information about its long-term effectiveness. It is still possible for treatments to be withdrawn at this stage if they produce serious side effects which are too rare to be detected in a phase III trial, and only become apparent when data are available for large numbers of treated patients.

PART 2 Epidemiology 流行病学.

1. Sensitivity 敏感度 & Specificity 特异性

- Sensitivity = $\frac{\text{Number of diseases people who screen positive}}{\text{Total number of diseased people}}$ (有病的人显阳性)
- Specificity = $\frac{\text{Number of diseases people who screen negative}}{\text{Total number of healthy people}}$ (没病的人显阴性).

} 显示的都是准确性.

The negative of these statistics are:

- False negative rate (FNR) = 1 − Sensitivity (误诊阴性)
- False positive rate (FPR) = 1 − Specificity (误诊阳性)

2. Standardisation 标准化.

(Via. 2023)

Defination: ① Direct Standardisation: 用于两个具有不同特征的研究人群中公平地比较疾病率.

is used to fairly compare the rates of disease in 2 separate study populations with different demographics. (人口特征)

② Indirect Standardisation: 旨在计算研究人群中预期的死亡人数. 假设应用参考人群的特定年龄死亡率. 后比较=看疾病率

aims to compute the number of deaths expected in the study population if the age specific mortality rates from a reference population applied. Then one can compare the rates of disease in the study and reference populations.

(1) Direct Standardisation:

- Notion: (i) y_i — 研究人群中第 i 个年龄组的死亡人数. } Study population
  (2) n_i — 人数. } Study population
  (3) N_i — 参考人群中第 i 个年龄组的人数. — reference population
  } $i = 1, 2, \cdots, G$ (依据年龄分成G组).

① 粗死亡率: Crude mortality rate = $\frac{\sum_{i=1}^{G} n_i}{\sum_{i=1}^{G} y_i} = \frac{\text{总死亡人数}}{\text{总人数}}$ (研究人群).

② 特定年龄组死亡率: Age specific mortality rate for group i = $\frac{y_i}{n_i}$

③ 特定年龄预期死亡人数: Expected number of mortalities for group i = $\frac{y_i}{n_i} \times N_i$

④ 年龄标准化死亡率: Age standardised mortality rate (ASMR) = $\frac{\sum_{i=1}^{G} \frac{y_i}{n_i} \times N_i}{\sum_{i=1}^{G} N_i}$

⇒ 若 $y_i \sim \text{Poisson}(n_i \theta_i)$, 其中 θ_i 是死亡率. $\hat{\theta}_{i_{MLE}} = \frac{y_i}{n_i} \Rightarrow \text{Var}[\hat{\theta}_i] = \frac{\theta_i}{n_i} \Rightarrow \text{Var}[\text{ASMR}] = \text{Var}\left[\frac{\sum_{i=1}^{G} \hat{\theta}_i N_i}{\sum_{i=1}^{G} N_i}\right] = \frac{\sum_{i=1}^{G} N_i^2 \frac{\theta_i}{n_i}}{\left(\sum_{i=1}^{G} N_i\right)^2}$

(2) Indirect Standardisation

· Notion: r_i – 参考人群中第 i 个年龄组的死亡率. (reference population)

E – 预期研究人群的死亡数量 (study population) $\Rightarrow E=\sum_{i=1}^{G} n_i r_i$

Y – 研究人群的总死亡数 (study population) $\Rightarrow Y=\sum_{i=1}^{G} y_i$.

① 标准化死亡率比: Standardised Mortality Ratio (SMR) $=\frac{\text{Observed deaths}}{\text{Expected deaths}}=\frac{Y}{E}=\frac{\sum_{i=1}^{G} y_i}{\sum_{i=1}^{G} n_i r_i}$

$\Rightarrow$ $\begin{cases} SMR=1 \Rightarrow \text{死亡数与预期相同} \Rightarrow \text{研究和参考人群的死亡率风险相同.} \\ SMR\neq 1 \Rightarrow \text{研究人群中死亡数超过(低于)预期.} \end{cases}$

② 构建简单的泊松模型 $Y=\sum_{i=1}^{G} y_i \sim \text{Poisson}(ER)$, 其中 $\hat{R}_{MLE}=\frac{Y}{E}$

$\Rightarrow Var[SMR]=Var[\frac{Y}{E}]=\frac{Var[Y]}{E^2}=\frac{ER}{E^2}=\frac{R}{E}$. 若 R 已知 $(\hat{R}=\frac{Y}{E}) \Rightarrow 95\%CI: \left(SMR \pm 1.96\sqrt{\frac{Y}{E^2}}\right)$.

3. Measuring the association between a risk factor and disease 测量风险因素与疾病之间的关联.

	Disease	No disease	
Risk factor	a	b	$a+b$
No risk factor	c	d	$c+d$
	$a+c$	$b+d$	n

(1) 相对风险: Relative Risk (RRR) $=\frac{Pr(\text{disease in the group with the risk factor})}{Pr(\text{disease in the group without the risk factor})}=\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$ $\begin{cases} =1. \text{零相对风险(无差异)} \\ \neq 1. \text{有风险因素患病风险更高(低)} \end{cases}$

$\Rightarrow Var[\log(RR)]=\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}$ $\Rightarrow 95\%CI: \left(\log(RR) \pm 1.96\sqrt{Var[\log(RR)]}\right)$.

(2) 比值比: Odd Ratio (OR). Similarly to PART 1 - 4(2) $OR=\frac{\frac{a}{b}}{\frac{c}{d}}=\frac{ad}{bc}$. $Var[\log(OR)]=\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$.

(3) 归因(相差)风险度: Attributable Risk (AR) = Pr(disease with the risk factor) − Pr(disease without the risk factor).

$$=\frac{a}{a+b}-\frac{c}{c+d} \triangleq P_1-P_2.$$

(4) 人口归因风险: Population Attributable Risk (PAR): 从群体角度, 而非个体. (归因风险因素占总人口比例)

$$PAR=\frac{(a+c)-nP_2}{(a+c)}$$

nP_2 ← 减去总人数的预期无风险患病人数.

$(a+c)$ ↖ 患病总人数

$\Rightarrow$ 分层 Stratification (分成 G 组)

① 第 i 组的比值比 $OR_i=\frac{a_i/b_i}{c_i/d_i}=\frac{a_i d_i}{b_i c_i}$

$\Rightarrow$ 权重 $w_i=\frac{1}{Var[\log(OR_i)]}=\frac{1}{\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}} \approx \frac{b_i c_i}{n_i}$

② Mantel-Haenszel 方法: $OR_{MH}=\frac{\sum_{i=1}^{G} w_i OR_i}{\sum_{i=1}^{G} w_i}$ $\Rightarrow Var[\log(OR_{MH})]=\frac{\sum_{i=1}^{G} W_i^2 V_i}{(\sum_{i=1}^{G} w_i)^2}$, where $V_i=\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}$.

Survival Analysis 生存分析

1. Survival Analysis Fundamental.

- Denote: 生存时间 T (r.v.) $\Rightarrow$ cdf. $F(t)=\Pr(T\le t)$ $\longrightarrow$ pdf. $f(t)$
- The survival function: $S(t)=\Pr(T>t)=1-F(t)$, for any $t>0$. 表示在 t 时间后的存活概率
- The hazard function: $h(t)=\lim_{\Delta t\to 0}\frac{\Pr(T\le t+\Delta t\mid T\ge t)}{\Delta t}$ $\Rightarrow H(t)=\int_0^t h(u)\,du$
- Theorem: $h(t)=\frac{f(t)}{S(t)}$

Proof: $h(t)=\lim_{\Delta t\to 0}\frac{\Pr(T\le t+\Delta t\mid T\ge t)}{\Delta t}=\lim_{\Delta t\to 0}\frac{\frac{\Pr(T\le t+\Delta t\cap T\ge t)}{\Pr(T\ge t)}}{\Delta t}=\lim_{\Delta t\to 0}\frac{\Pr(t\le T\le t+\Delta t)}{\Pr(T\ge t)\Delta t}$

$=\frac{1}{\Pr(T\ge t)}\lim_{\Delta t\to 0}\frac{\Pr(t\le T\le t+\Delta t)}{\Delta t}$

$=\frac{1}{S(t)}F(t)=\frac{f(t)}{S(t)}$.

- Theorem: $S(t)=\exp\left(-\int_0^t h(u)\,du\right)$.

Proof: $\frac{dS(t)}{dt}=\frac{d(1-F(t))}{dt}=-f(t)$ $\Rightarrow h(t)=\frac{f(t)}{S(t)}=\frac{-\frac{dS(t)}{dt}}{S(t)}=-\frac{d\log S(t)}{dt}$

$\Rightarrow$ The exponential distribution: $T\sim Exp(\lambda)$ $\Rightarrow f(t)=\lambda e^{-\lambda t}$ $\Rightarrow F(t)=1-e^{-\lambda t}$ $\Rightarrow S(t)=e^{-\lambda t}$

$\Rightarrow h(t)=\frac{f(t)}{S(t)}=\lambda$ (hazard rate).

$\Rightarrow E[T]=\frac{1}{\lambda}$, $Var[T]=\frac{1}{\lambda^2}$.

- Hazard rate { Large → High hazard & Short survival; Small → Low hazard & Long survival
- Lack of memory property 无记忆性: 导致无论 t 如何变化 失败的可能性都一样. Really realistic!

$\Rightarrow$ The Weibull distribution: $T\sim Weibull(\lambda,\gamma)$ $\Rightarrow f(t)=\lambda^\gamma\gamma t^{\gamma-1}e^{-(\lambda t)^\gamma}$ $\Rightarrow F(t)=1-e^{-(\lambda t)^\gamma}$ $\Rightarrow S(t)=e^{-(\lambda t)^\gamma}$

$\Rightarrow h(t)=\lambda^\gamma\gamma t^{\gamma-1}$

$\Rightarrow E[T]=\frac{1}{\lambda}\Gamma\left(1+\frac{1}{\gamma}\right)$, $Var[T]=\frac{1}{\lambda^2}\left[\Gamma\left(1+\frac{2}{\gamma}\right)-\Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$.

where $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}\,dt$. or $\Gamma(n)=(n-1)!$, $n\in\mathbb{N}$.

2. Estimating the survial function.

(1) Empirical distribution function (EDF) 经验分布函数.

$\hat{S}_{EDF}(t)=\frac{\text{number surviving beyond } t}{\text{number in the sample}}=\frac{\sum_{i=1}^n I[t_i>t]}{n}$, where $I[t_i>t]=\begin{cases}1, & t_i>t\\ 0, & \text{otherwise}\end{cases}$

(2) Kaplan-Meier (KM) 比 EDF 更好.

- n 个观测时间 $(t_1,\dots,t_n)$, m 个独特时间 $(t_{(1)},t_{(2)},\dots,t_{(m)})$ 且最多 $(n-m)$ 个 censored 观测值. 加入 $t_{(0)}=0$. 且 $\Pr(T>t_{(0)})=1$

截尾/删失

$$S(t_{(i)}) = \Pr(T > t_{(i)}) = \Pr(T > t_{(i)} \cap T > t_{(i-1)})$$

$$= \Pr(T > t_{(i)} \mid T > t_{(i-1)}) \Pr(T > t_{(i-1)})$$

$$= \Pr(T > t_{(i)} \mid T > t_{(i-1)}) \Pr(T > t_{(i-1)} \mid T > t_{(i-2)}) \Pr(T > t_{(i-2)})$$

$$= \left[\prod_{j=1}^{i} \Pr(T > t_{(j)} \mid T > t_{(j-1)})\right] \Pr(T > t_{(0)})$$

$$= \prod_{j=1}^{i} \Pr(T > t_{(j)} \mid T > t_{(j-1)})$$

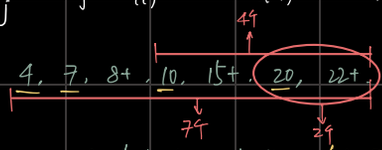
- Specifically, let

  - r_j: 在 $t_{(j)}$ 之前仍然存活的个体数量

  - $s_j = r_j - d_j$: 至少存活到 $t_{(j)}$ 的个体数量. d_j: 在 $t_{(j)}$ 发生死亡的数量. (不包括 censored 观测数据).

$\hat{S}_{KM}(t) = \prod_{j=1}^{i} \frac{s_j}{r_j}$, for $t_{(i)} \le t \le t_{(i+1)}$, $i = 1, 2, \cdots, m$. since the estimate of $\Pr(T > t_{(j)} \mid T > t_{(j-1)})$ will be $\frac{s_j}{r_j}$.

Oil: Survial data: 4, 7, 8+, 10, 15+, 20, 22+ (+ denotes a censored observation).



Here there are $m = 4$ distinct event times $(4, 7, 10, 20)$.

Event time $t_{(j)}$	Number at risk at time $t_{(j)}$ r_j	Number surviving beyond time $t_{(j)}$ s_j	Estimated survival function at $t_{(j)}$ $\hat{S}(t_{(j)})$
0	7	7	$7/7 = 1$
4	7	$7 - 1$ (只有1个 $t_{(j)} = 4$) $= 6$	$1 \times 6/7 = 6/7$
7	6	$6 - 1 = 5$	$6/7 \times 5/6 = 5/7$
10	4	$4 - 1 = 3$	$5/7 \times 3/4 = 15/28$
20	2	$2 - 1 = 1$	$15/28 \times 1/2 = 15/56$

⇒ Quantifying uncertainty in the survival 量化生存函数的不确定性

Recall: $\mathrm{Var}[\log(X)] \approx \frac{\mathrm{Var}[X]}{(E[X])^2}$. Proof: $\log(X) = \log(\mu) + \frac{1}{\mu}(X-\mu) - \frac{1}{2\mu^2}(X-\mu)^2 + \cdots$ ⇒ $\log(X) \approx \log(\mu) + \frac{1}{\mu}(X-\mu)$ ($\log(X)$ 在 μ 附近的 Taylor 展开)

- $\log(\hat{S}_{KM}(t)) = \log\left(\prod_{j=1}^{i} \frac{s_j}{r_j}\right) = \sum_{j=1}^{i} \log\left(\frac{s_j}{r_j}\right)$.

$E[\log(X)] \approx \log(\mu) + \frac{1}{\mu}(E[X-\mu]) = \log(\mu)$.

$E[\log^2(X)] \approx E\left[\left(\log(\mu) + \frac{1}{\mu}(X-\mu)\right)^2\right] = \log^2(\mu) + 2\log(\mu)\frac{1}{\mu}E[X-\mu] + \frac{1}{\mu^2}E[(X-\mu)^2]$.

$= \log^2(\mu) + \frac{\mathrm{Var}[X]}{\mu^2}$

Assumption: $s_j \sim \mathrm{Binomial}(r_j, \theta_j)$.

⇒ $\hat{\theta}_{j\,MLE} = \frac{s_j}{r_j}$

⇒ $\mathrm{Var}[\log(X)] \approx E[\log^2(X)] - (E[\log(X)])^2 = \frac{\mathrm{Var}[X]}{\mu^2}$

$\mathrm{Var}[s_j] = r_j \theta_j (1-\theta_j)$.

⇒ $\mathrm{Var}[\hat{\theta}_j] = \mathrm{Var}\left[\frac{s_j}{r_j}\right] = \frac{1}{r_j^2}\mathrm{Var}[s_j] = \frac{\theta_j(1-\theta_j)}{r_j} \approx \frac{\hat{\theta}_j(1-\hat{\theta}_j)}{r_j}$. $E[\hat{\theta}_j] = E\left[\frac{s_j}{r_j}\right] = \hat{\theta}_j$.

⇒ $\mathrm{Var}[\log(\hat{S}_{KM}(t))] = \sum_{j=1}^{i} \mathrm{Var}[\log(\hat{\theta}_j)] \simeq \sum_{j=1}^{i} \frac{\mathrm{Var}[\hat{\theta}_j]}{(E[\hat{\theta}_j])^2} = \sum_{j=1}^{i} \frac{1}{\hat{\theta}_j^2} \times \frac{\hat{\theta}_j(1-\hat{\theta}_j)}{r_j} = \sum_{j=1}^{i} \frac{1 - s_j/r_j}{s_j/r_j \cdot r_j} = \sum_{j=1}^{i} \frac{r_j - s_j}{r_j s_j}$.

⇒ $\mathrm{Var}[\hat{S}_{KM}(t)] = \left(E[\hat{S}_{KM}(t)]\right)^2 \mathrm{Var}[\log(\hat{S}_{KM}(t))] = \left(E[\hat{S}_{KM}(t)]^2\right) \sum_{j=1}^{i} \frac{r_j - s_j}{r_j s_j}$

$$\Rightarrow 95\%\ CI:\ \hat{S}_{KM}(t) \pm 1.96\sqrt{Var[\hat{S}_{KM}(t)]} = \hat{S}_{KM}(t) \pm 1.96\, E[\hat{S}_{KM}(t)]\sqrt{\sum_{j=1}^{i} \frac{r_j - s_j}{r_j s_j}}$$