Biostatistics

Guanxu Gleason WANG

University of Glasgow

版本：Past Paper of STAT5015 Biostatistics, UofG

更新：April 15, 2025

目录

1 Clinical Trials (临床试验)

1.1 Definitions

定义 1.1 (Randomisation 随机化) Patients are randomly allocated to the treatments.
患者被随机分配到治疗中.

定义 1.2 (Double-blind 双盲) None of the patient, the investigator or the medical team know which treatment is being given to the patient.
患者、调查员或医疗团队都不知道正在给患者进行哪种治疗.

定义 1.3 (Placebo-controlled 安慰剂对照) One of the groups is receiving a placebo – a “treatment” identical in all ways to the main treatment other than the active ingredient. A placebo has no active ingredient.
一组正在接受安慰剂 – 一种除了活性成分外, 在所有方面都与主要治疗相同的“治疗”. 安慰剂没有活性成分.

定义 1.4 (Four main stages to the design of clinical trials 临床试验设计的四个主要阶段)

Phase I – **Clinical Pharmacology and Toxicity**

– 临床药理学与毒理学

These are primarily concerned with drug safety, and are performed on healthy human volunteers (not patients). Dosing schedules are also established.
这些主要关注药物安全性，并在健康的人类志愿者（而非患者）身上进行。还建立了剂量安排。

Phase II – **Initial Clinical Investigation for Treatment Effect**

– 初始治疗疗效的临床研究

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 randomised, but in the (fairly recent) past they were generally uncontrolled (i.e. no control group).

这些是在患者中进行的小规模研究，观察药物的有效性和安全性。有证据表明该药物实际上有效且无毒吗？在 II 期临床试验中随机化变得越来越普遍，但在（相对较近的）过去，它们通常是未控制的（即没有对照组）。

Phase III – **Full Scale Evaluation of Treatment**

– 全面评估治疗方案

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

在一种药物被证明具有合理有效性和值得进行全面调查之后，将其与当前同种疾病的常规治疗方法（或安慰剂）进行随机对照试验比较是至关重要的。

Phase IV – **Post-marketing Surveillance**

– 上市后监测

If a drug is approved for marketing, its performance is still closely monitored for evidence of side effects 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withdrawn at this stage if they produce serious side effects which are too rare to be detected in a phase III trial, and only become apparent when data are available for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.
如果一种药物获得市场批准，其性能仍会密切监控以获取关于其副作用证据以及更多关于其长期有效性的信息。如果治疗在此阶段产生严重副作用，而这些副作用在 III 期临床试验中难以检测到，只有在大量治疗患者的数据可用时才会显现，那么此时仍有可能撤回治疗。

定义 1.5 (Meta-analysis and its main objectives 荟萃分析及其主要目标) Meta-analysis is the process of applying statistical methods to the problem of combining results from different clinical trial of the same treatment. Meta 分析是将统计方法应用于将同一治疗方法的不同临床试验结果相结合的问题的过程。

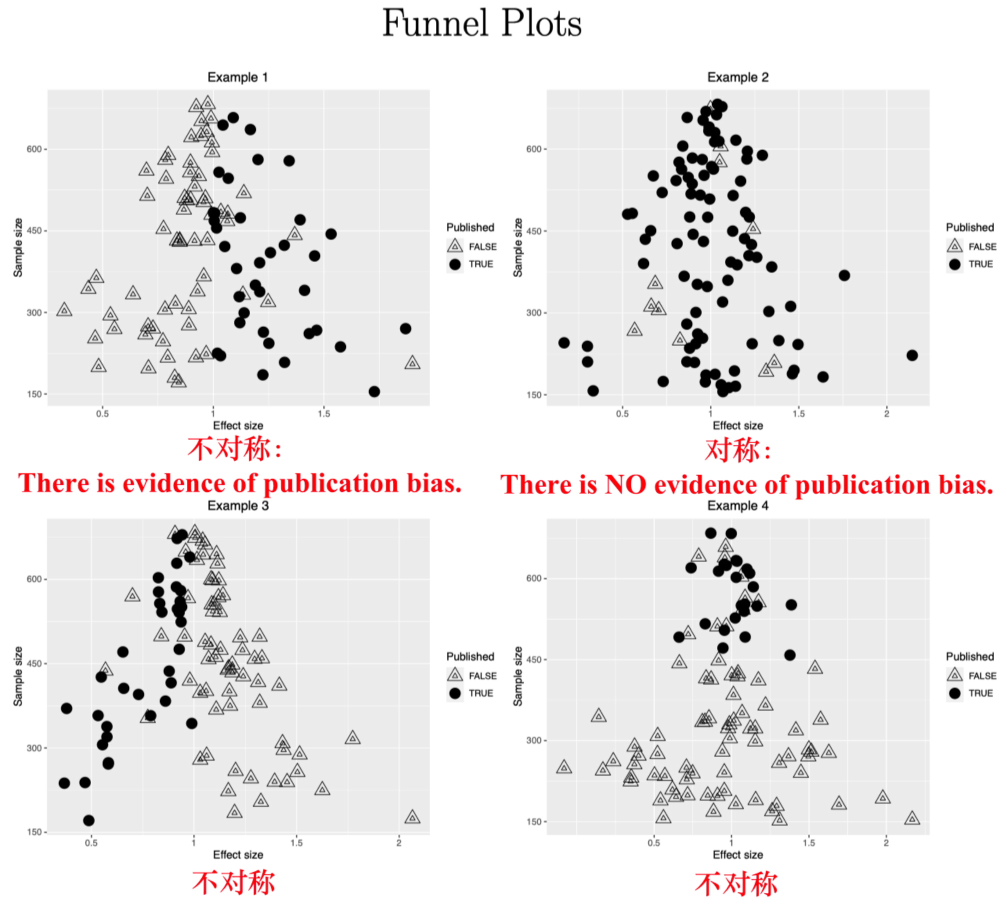
It has four main objectives:

- consistent, objective display of data from different clinical trials;
  一致、客观地展示来自不同临床试验的数据;
- testing an overall hypothesis; 测试一个总体假设;
- estimation of an average treatment effect; 平均治疗效果的估计;
- investigation of statistical heterogeneity between trials. 试验间统计异质性的研究.

定义 1.6 (Publication bias 发表偏差) It is widely recognised that clinical trials reporting large and significant treatment effects are more likely to be published in medical journals than trials with smaller, non-significant differences. 广泛认为，报告大型和显著治疗效果的临床试验比报告较小、非显著差异的试验更有可能发表在医学期刊上。Evidence of asymmetry in the funnel plots would suggest publication bias. 漏斗图中的不对称性证据表明存在发表偏倚。

例 1.1 (via. 2023.Q1(d)-iii, 2022.Q3(c)) When there is evidence of funnel plot asymmetry regarding the published studies, publication bias is only one possible explanation.

当存在关于已发表研究的漏斗图不对称的证据时，发表偏差仅是可能的解释之一。



Funnel Plots

不对称：
There is evidence of publication bias.

对称：
There is NO evidence of publication bias.

不对称

不对称

1.2 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

- Significance level (显著性水平): Type 1 error, 本来 H_0 是真的但是被拒绝, i.e.

$$\Pr\left(\text{Reject } H_0 \mid H_0 \text{ is true}\right) = \alpha.$$

- Power (功效): H_0 被正确拒绝, i.e.

$$\Pr\left(\text{Reject } H_0 \mid H_1 \text{ is true}\right) = 1-\beta.$$

1.3 Sample Sizes Calculations

定理 1.1 (临床试验所需患者数量计算) 假设有两种治疗方案: A – Placebo (安慰剂); B – 喂药.

记 θ_A 和 θ_B 分别为服用 Placebo 和 临床试验药品 的患者在 1 年内死亡的概率. 我们希望测试的假设 (hypothesis) 是

$$H_0 : \theta_A = \theta_B \quad \text{vs.} \quad H_1 : \theta_A \neq \theta_B$$

with significant level of α and power of $1-\beta$. 假设临床差异 (clinical difference) $\delta = \theta_A - \theta_B \sim \mathcal{N}$. 因此, 正态近似的患者数量为

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

证明. 假设有 $2N$ 个患者, 两个方案各有 N 个患者, 并设 (X_A, X_B) 为死亡患者数量, 因此有

$$X_i \sim \mathrm{Binomial}(N, \theta_i).$$

若 N 足够大, 则可近似为正态分布, 即

$$X_i \sim \mathcal{N}\Big(N\theta_i, N\theta_i(1-\theta_i)\Big).$$

记死亡率 $\widehat{\theta}_i = \frac{X_i}{N}$, 那么 $\widehat{\theta}_i \sim \mathcal{N}\left(\theta_i, \frac{\theta_i(1-\theta_i)}{N}\right)$, 因此

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

我们希望 $\delta > 0$, 即不吃药的死的多, 根据 t 检验, 得到统计量

$$T = \frac{(\widehat{\theta}_A - \widehat{\theta}_B) - \delta}{\sqrt{\frac{\theta_A(1-\theta_A)+\theta_B(1-\theta_B)}{N}}} \overset{H_0:\delta=0}{=\!=\!=} \frac{\widehat{\theta}_A - \widehat{\theta}_B}{\sqrt{\frac{\theta_A(1-\theta_A)+\theta_B(1-\theta_B)}{N}}} \sim \mathcal{N}(0,1)$$

由于

$$\Pr(\text{Reject } H_0 \text{ at significance level } \alpha \mid \text{true difference is } \delta) = 1 - \beta,$$

which corresponds to $\Pr\left(|T| > \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\beta$. 证明就到这里吧, 后面也没啥玩意了.

定理 1.2 (临床试验所需患者数量计算) For continue response: two treatment have mean responses (μ_A, μ_B), common std σ. Using a similar argument as before:

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

特别地: 实践中未必有足够患者进行试验.

- **Reason:** Patients are likely withdraw from the trial.
- **Solution:** Estimate the likely withdraw rate from similar existing studies and scale up the sample size appropriately to account for this.

1.4 Meta-Analysis

即, 将多个研究结果进行统计分析的组合.

1.4.1 Main statistical objectives

1. consistent and objective display of data from different clinical trials;
2. testing an overall hypothesis;
3. estimation of an average treatment effect;

4. investigation of statistical heterogeneity between trials.

1. 一致且客观地展示来自不同临床试验的数据；
2. 测试总体假设；
3. 估计平均治疗效果；
4. 调查试验之间的统计异质性。

1.4.2 Binary response variables: Death (Failure) / Survival (Success)

Odd Ratio (OR, 比值比)

	Failure	Success
Treatment	a	b
Control	c	d

$$\Rightarrow \quad \text{OR} = \frac{\text{odd of failure on treatment}}{\text{odd of failure on control}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

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

$$\Rightarrow \quad \text{Var}\left[\log\left(\text{OR}\right)\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

证明. Recall the Taylor expansion:

$$g(X) = g(\mu) + (X-\mu)g'(\mu) + \frac{(X-\mu)^2}{2}g''(\mu) + \cdots,$$

then

$$\text{Var}\left[g(X)\right] \approx \left[g'(\mu)\right]^2 \cdot \text{Var}\left[X\right] = \left(\frac{\mathrm{d}g(X)}{\mathrm{d}X}\right)^2 \cdot \text{Var}\left[X\right],$$

therefore

$$\operatorname{Var}[\log a] = \left(\frac{1}{a}\right)^2 \times a = \frac{1}{a}, \quad \left(X \sim \text{Binomial}(n,p) \xrightarrow{n\to\infty} \text{Poisson}(\lambda = np)\right).$$

. . . .

例 1.2 (via 2023.Q1(d)-iii) Table following was presented in a report of a meta-analysis of five trials of a new treatment for chronic kidney disease (CKD). Interpret the results in the table for each of the five studies as well as the result from all studies combined. What reservations, if any, may you attach to your interpretation? 请解释表中每个五项研究的成果以及所有研究合并的结果. 您对您的解释有何保留意见?

Study	Odd ratio (treatment/control)	95% Confidence Interval
1	0.42	0.15 to 1.08
2	0.33	0.07 to 1.98
3	0.59	0.29 to 1.01
4	0.41	0.18 to 1.23
5	0.25	0.03 to 0.88
All studies	0.44	0.41 to 0.82

Interpretation:

- In Table 1 the combined odds ratio is 0.44, so that the odds of death in the treated group is 0.44 times (around half) that of the control group.
- The confidence interval does not contain 1, so that the treatment is significantly better than the control.
- Note also that the confidence interval is still quite wide, so that the treatment effect is still not very precisely estimated. 请注意，置信区间仍然相当宽，因此治疗效果仍然没有非常精确地估计。

Reservations

- No information/detail is provided about the individual trials.
- Were they all randomised, double blind, adequately controlled?
- Were patient selection criteria similar? Were the study protocols (方案) in general similar?
- How were the five studies selected – is there a possibility of bias in the selection of studies (such as publication bias)?

. . . .

1.4.3 Meta-analytic models

假设有 C 次试验, 我们希望将这些试验的结果结合起来.

Notion

- $\widehat{Y}_c$ – the continuous effect size in the cth trial, 如: $\log(\mathrm{OR})$, 其中 $c = 1, 2, \cdots, C$.

- W_c – the reciprocal (倒数) of the effect size variance, i.e., $W_c = \dfrac{1}{\mathrm{Var}\left[\widehat{Y}_c\right]}$. 作为权重意义出现, 方差越大权重越小.

1. Fix effect effect approach 固定效应方法: 假设存在一个单一的真实效应量 (μ, true effect size, unknown), 每个 C 试验都估计了这个效应量, 从而得到简单模型

$$\widehat{Y}_c \sim \mathcal{N}\left(\mu, \frac{1}{W_c}\right), \text{ for } c = 1, 2, \cdots, C$$

(i) Then maximum likelihood estimation yields the estimator:

$$\widehat{\mu}_{\mathrm{FE}} = \frac{\sum_{c=1}^{C} W_c \widehat{Y}_c}{\sum_{c=1}^{C} W_c}.$$

证明. 有空再写, 反正就是 MLE. 然后还是 unbiased 的.

(ii) The variance of $\widehat{\mu}_{\mathrm{FE}}$ is $\mathrm{Var}\left[\widehat{\mu}_{\mathrm{FE}}\right] = \dfrac{1}{\sum_{c=1}^{C} W_c}$, which yields the sampling distribution:

$$\widehat{\mu}_{\mathrm{FE}} = \frac{\sum_{c=1}^{C} W_c \widehat{Y}_c}{\sum_{c=1}^{C} W_c} \sim \mathcal{N}\left(\mu, \frac{1}{\sum_{c=1}^{C} W_c}\right).$$

证明. $\mathrm{Var}\left[\widehat{\mu}_{\mathrm{FE}}\right] = \mathrm{Var}\left[\frac{\sum_{c=1}^{C} W_c \widehat{Y}_c}{\sum_{c=1}^{C} W_c}\right] = \frac{\sum_{c=1}^{C} W_c^2 \mathrm{Var}\left[\widehat{Y}_c\right]}{\left(\sum_{c=1}^{C} W_c\right)^2} = \frac{\sum_{c=1}^{C} W_c^2 \cdot (1/W_c)}{\left(\sum_{c=1}^{C} W_c\right)^2} = \frac{1}{\sum_{c=1}^{C} W_c}.$

(iii) A 95% confidence interval for the population effect μ is then given by

$$\widehat{\mu}_{\mathrm{FE}} \pm 1.96\sqrt{\mathrm{Var}\left[\widehat{\mu}_{\mathrm{FE}}\right]}.$$

(iv) Hypothesis test: $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$. The test statistic in this case under H_0 is given by

$$Z = \frac{\widehat{\mu}_{\mathrm{FE}}}{\sqrt{\mathrm{Var}\left[\widehat{\mu}_{\mathrm{FE}}\right]}} \sim \mathcal{N}(0, 1) \text{ under } H_0.$$

This test statistic can then be compared against the appropriate percentiles of a standard normal distribution, which would be $[-1.96, 1.96]$ for a 5% significance level.

2. Random effect approach 随机效应方法: 假设存在每个试验都有各自的真实效应量 (μ_c), 从而有

$$\widehat{Y}_c \sim \mathcal{N}\left(\mu_c, \frac{1}{W_c}\right), \text{ for } c = 1, 2, \cdots, C, \qquad \mu_c \sim \mathcal{N}(\mu, \tau^2).$$

This model thus partitions the variation in the data into two sources: (i) within trial uncertainty in the estimated effect captured by $\dfrac{1}{W_c}$; and (ii) between trial variation in the true effect sizes captured by τ^2. The model can be simplified to

$$\widehat{Y}_c \sim \mathcal{N}\left(\mu_c, \frac{1}{W_c} + \tau^2\right), \text{ for } c = 1, 2, \cdots, C.$$

so the two parameters to be estimated are the overall effect size μ and the between trial heterogeneity τ^2. Then the maximum likelihood estimator of μ conditional on τ^2 is given by

$$\widehat{\mu}_{\mathrm{RE}} = \frac{\sum_{c=1}^{C} \frac{1}{\frac{1}{W_c}+\tau^2}\widehat{Y}_c}{\sum_{c=1}^{C} \frac{1}{\frac{1}{W_c}+\tau^2}}.$$

在随机效应模型中, 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 $Q = \sum_{c=1}^{C} W_c \left(\widehat{Y}_c - \widehat{\mu}_{\mathrm{FE}}\right)^2$, $(C-1)$ 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, 已获得更稳健的 τ_c^2 估计.

2 Epidemiology (流行病学)

2.1 Definitions

2.2 Sensitivity (敏感度) & Specificity (特异性)

以下, 红色为患者实际有病 (阳性), 蓝色为患者实际没病 (阴性).

$$\left.\begin{aligned}
\text{Sensitivity} &= \frac{\text{Number of diseased people who screen positive}}{\text{Total number of diseased people}} \quad \text{(有病的人显阳性)} \\
\text{Specificity} &= \frac{\text{Number of diseased people who screen negative}}{\text{Total number of healthy people}} \quad \text{(没病的人显阴性)}
\end{aligned}\right\} \text{显示的都是}\underline{\text{准确性}}$$

这些统计量的反面是:

$$\text{False negative rate (FNR)} = \frac{\text{Number of diseased people who screen negative}}{\text{Total number of diseased people}} = 1 - \text{Sensitivity} \quad \text{(误诊为阴性)}$$

$$\text{False positive rate (FPR)} = \frac{\text{Number of diseased people who screen positive}}{\text{Total number of healthy people}} = 1 - \text{Specificity} \quad \text{(误诊为阳性)}$$

例 2.1 (via 2023.Q3(c)) A new screening test was developed for bronchitis, and a number of patients that potentially had the disease were given the test following a GP consultation. The results are depicted in the

		Disease		
		Yes	No	
Screening	Positive	32	29	61
Test	Negative	14	84	98
		46	113	159

contingency table below. **Q. Compute the sensitivity, specificity, false negative rate (FNR) and false positive rate (FPR) from the screening test. Interpret each statistic. A.**

$$\text{Sensitivity} = \frac{32}{46} = 0.70.$$

$$\text{Specificity} = \frac{84}{113} = 0.74.$$

$$\text{False negative rate (FNR)} = 1 - 0.70 = 0.30.$$

$$\text{False positive rate (FPR)} = 1 - 0.74 = 0.26.$$

A sensitivity of approximately 0.70means that 70% of diseased patients are correctly identified as having bronchitis. A false negative rate of approximately 0.30 means that 30% of diseased patients are incorrectly identified as not bronchitis. 约 0.70 的敏感性意味着 70% 的患病患者被正确识别为患有支气管炎. 约 0.30 的假阴性率意味着 30% 的患病患者被错误地识别为非支气管炎.

A specificity of approximately 0.74 means that 74% of healthy patients are correctly identified as not having bronchitis. A false positive rate of approximately 0.26 means that 26% of healthy patients are incorrectly identified as having bronchitis. 约 0.74 的特异性意味着 74% 的健康患者被正确识别为没有支气管炎. 约

0.26 的假阳性率意味着 26% 的健康患者被错误地识别为患有支气管炎.

2.3 Standardisation

2.3.1 Direct standardisation

Notion:

$$\left.\begin{array}{ll} \left.\begin{array}{l} y_i \text{ — 研究人群中第 } i \text{ 个年龄组的死亡人数} \\ n_i \text{ — 研究人群中第 } i \text{ 个年龄组的总人数} \end{array}\right\} & \text{study population} \\ N_i \text{ — 参考人群中第 } i \text{ 个年龄组的总人数} \quad \text{—} & \text{reference population} \end{array}\right\} i = 1, 2, \cdots, G \text{ (依据年龄分成 } G \text{ 组)}$$

1. 粗死亡率:

$$\text{crude mortality rate} = \frac{\sum_{i=1}^{G} n_i}{\sum_{i=1}^{G} y_i} = \frac{\text{总死亡人数}}{\text{总人数}} \quad \text{(study population)}$$

2. 特定年龄组死亡率:

$$\text{Age specific mortality rate for the } i\text{th group} = \frac{y_i}{n_i}$$

3. 特定年龄预期死亡人数:

$$\text{Expected number of mortalities for the } i\text{th group} = \frac{y_i}{n_i} \times N_i$$

4. 年龄标准化死亡率:

$$\text{Age standardised mortality rate (ASMR)} = \frac{\sum_{i=1}^{G} \frac{y_i}{n_i} \times N_i}{\sum_{i=1}^{G} N_i}$$

若 $y_i \sim \text{Poisson}(n_i\theta_i)$, 其中 θ_i 是死亡率, 则

$$\widehat{\theta}_{i_{\text{MLE}}} = \frac{y_i}{n_i} \quad \Rightarrow \quad \text{Var}\left[\widehat{\theta}_i\right] = \frac{\theta_i}{n_i} \quad \Rightarrow \quad \text{Var}\left[\text{ASMR}\right] = \text{Var}\left[\frac{\sum_{i=1}^{G} \widehat{\theta}_i N_i}{\sum_{i=1}^{G} N_i}\right] = \frac{\sum_{i=1}^{G} N_i^2 \frac{\widehat{\theta}_i}{n_i}}{\left(\sum_{i=1}^{G} N_i\right)^2}.$$

此公式可用于根据高斯假设计算基于 ASMR 的 95% 置信区间:

$$\text{ASMR} \pm 1.96 \times \text{Var}\left[\text{ASMR}\right].$$

2.3.2 Indirect standardisation

Notion:

$$\left.\begin{array}{ll} r_i \text{ — 参考人群中第 } i \text{ 个年龄组的死亡率} \quad \text{——} & \text{reference population} \\ \left.\begin{array}{l} E \text{ — 研究人群的预期死亡人数} \Rightarrow E = \sum_{i=1}^{G} n_i r_i \\ Y \text{ — 研究人群的总死亡人数} \Rightarrow Y = \sum_{i=1}^{G} y_i \end{array}\right\} & \text{study population} \end{array}\right\} i = 1, 2, \cdots, G$$

1. 标准化死亡率 (standardised mortality ratio, SMR):

$$\text{SMR} = \frac{\text{Observed deaths}}{\text{Expected deaths}} = \frac{Y}{E} = \frac{\sum_{i=1}^{G} y_i}{\sum_{i=1}^{G} n_i r_i}$$

$$\Rightarrow \begin{cases} \text{SMR} = 1 & \Rightarrow \quad \text{死亡人数与预期相同} \quad \Rightarrow \quad \text{研究和参考人群的死亡率风险相同} \\ \text{SMR} \neq 1 & \Rightarrow \quad \text{研究人群中死亡人数超过/低于预期} \end{cases}$$

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

$$\text{Var}\left[\text{SMR}\right] = \text{Var}\left[\frac{Y}{E}\right] = \frac{\text{Var}\left[Y\right]}{E^2} = \frac{ER}{E^2} = \frac{R}{E}.$$

实践中, R 是未知的, 并用 MLE 估计值代替. 因此，基于正态性，SMR 的大约 95% 置信区间为

$$\text{SMR} = \pm 1.96 \times \sqrt{\frac{Y}{E^2}}.$$

2.4 Measuring the Association between a Risk Factor and Disease (测量风险因素与疾病之前的关联)

1. 相对风险 (relative risk, RR):

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

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

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

2. 比值比 (odd ratio, OR) is similar to 1.4.2,

$$\mathrm{OR} = \frac{a/b}{c/d} = \frac{ad}{bc} \qquad \Rightarrow \mathrm{Var}\left[\log\left(\mathrm{OR}\right)\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

3. 归因 (相差) 风险度 (attributable risk, AR):

$$\begin{aligned}\mathrm{AR} &= \Pr(\text{disease with the risk factor}) - \Pr(\text{disease without the risk factor})\\ &= \frac{a}{a+b} - \frac{c}{c+d} \qquad \overset{d}{=\!=} P_1 - P_2\end{aligned}$$

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).

$$\mathrm{PAR} = \frac{(a+c) - nP_2}{(a+c)} \quad \begin{aligned}&\hookleftarrow -nP_2 : \text{减去总人数的预期无风险患病人数}\\ &\hookleftarrow (a+c) : \text{总患病人数}\end{aligned}$$

例 2.2 (via. 2022.Q2(b)) The table below presents the results of a case-control study, which was carried out to investigate the association between smoking and lung cancer.

Smoking (cigarettes smoked/day)	Lung Cancer (Cases)	No Lung Cancer (Controls)
0	48	172
1 – 10	162	272
11 – 20	321	134
21 – 35	111	15
35+	220	20
Total	862	613

The contingency table is:

	Lung Cancer (Cases)	No Lung Cancer (Controls)	Total
Smokers	$A = 814$	$B = 441$	1255
Non Smokers	$C = 48$	$D = 172$	220
Total	862	613	1475

The **attributable risk** (AR) is:

$$\mathrm{AR} = P_1 - P_2 = \frac{A}{A+B} - \frac{C}{C+D} = \frac{814}{1255} - \frac{48}{220} = 0.43.$$

The AR of 0.43 indicates that more than one-third of the lung cancer occurrences were due to the exposure of smoking. 超过三分之一的肺癌发生是由于吸烟暴露引起的。

The **population attributable risk** (PAR) is:

$$\mathrm{PAR} = \frac{(A+C) - nP_2}{(A+C)} = \frac{(862) - 1475 \times \frac{48}{220}}{862} = 0.62.$$

The PAR of 0.62 indicates the proportion of the lung cancer occurrences in the population which was attributable to the exposure of smoking. 表示在人群中由于吸烟暴露导致的肺癌发生比例。

. . . .

2.5 Stratification (分层)

与2.4类似, 只不过分成 G 组, 即 2×2 表格变为

1. 第 i 组的比值比: $\mathrm{OR}_i = \dfrac{a_i/b_i}{c_i/d_i} = \dfrac{a_i d_i}{b_i c_i}$, $\Rightarrow$ 权重: $w_i = \dfrac{1}{\mathrm{Var}\left[\log(\mathrm{OR}_i)\right]} = \dfrac{1}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \approx \dfrac{b_i c_i}{n_i}$.

	Disease	No disease	
Risk factor	a_i	b_i	$a_i + b_i$
No risk factor	c_i	d_i	$c_i + d_i$
	$a_i + c_i$	$b_i + d_i$	n_i

2. Mantel-Haenszel 方法:

$$\mathrm{OR}_{\mathrm{MH}} = \frac{\sum_{i=i}^{G} w_i \mathrm{OR}_i}{\sum_{i=i}^{G} w_i} = \frac{\sum_{i=i}^{G} \frac{b_i c_i}{n_i} \times \frac{a_i d_i}{b_i c_i}}{\sum_{i=i}^{G} \frac{b_i c_i}{n_i}} = \frac{\sum_{i=i}^{G} \frac{a_i d_i}{n_i}}{\sum_{i=i}^{G} \frac{b_i c_i}{n_i}}$$

$$\Rightarrow \mathrm{Var}\left[\log(\mathrm{OR}_{\mathrm{MH}})\right] = \frac{\sum_{i=i}^{G} w_i^2 v_i}{\left(\sum_{i=i}^{G} w_i\right)^2}, \quad \text{where } v_i = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

例 2.3 (via 2023.Q3(a)) The table of the lung cancer cases between smokers and non-smokers stratified by age is as following.

Q1. Calculate the odds ratio of developing lung cancer for the group who smoke tobacco compared to the group who do not smoke, unadjusted for age, and construct the corresponding 95% confidence interval. Comment on your results.

Smokes tobacco	Age < 50 years		Age $\geqslant 50$ years	
	with lung cancer	no lung cancer	with lung cancer	no lung cancer
Yes	68	22	43	47
No	23	13	45	66
Total	91	35	88	113

A. Ignoring the age stratification, the odds ratio of lung cancer is:

$$\mathrm{OR} = \frac{a \times d}{b \times c} = \frac{(68+43)\times(13+66)}{(22+47)\times(23+45)} = \frac{111 \times 79}{69 \times 68} = 1.87.$$

The variance of the $\log(\mathrm{OR})$ is:

$$\mathrm{Var}\left[\log(\mathrm{OR})\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} = \frac{1}{111} + \frac{1}{69} + \frac{1}{68} + \frac{1}{79} = 0.05.$$

The 95% CI for $\log(\mathrm{OR})$ is:

$$\begin{aligned} &\log(\mathrm{OR}) \pm 1.96 \times \sqrt{\mathrm{Var}\left[\log(\mathrm{OR})\right]} \\ &\log(1.87) \pm 1.96 \times \sqrt{0.05} \\ &(0.19, 1.07). \end{aligned}$$

Hence, the 95% CI for the OR is: $(\exp(0.19), \exp(1.07)) = (1.21, 2.92)$.

Interpretation: tobacco smokers have significantly higher odds of lung cancer (on average 87% higher odds) compared to non-smokers as the 95% confidence interval does not contain 1. 吸烟者患肺癌的几率显著高于非吸烟者 (平均高出 87%), 因为 95% 置信区间不包含 1.

Q2. Calculate the odds ratios of developing lung cancer for smokers against non-smokers for each group separately, i.e. age < 50 years and age $\geqslant 50$ years, and construct the corresponding 95% confidence intervals. Comment on your results and compare them to the results in Q1.

A(1). Among patients aged < 50 years, the odds ratio of lung cancer for smokers vs. non-smokers is:

$$\mathrm{OR}_{\text{age} < 50 \text{ years}} = \frac{a \times d}{b \times c} = \frac{68 \times 13}{22 \times 23} = 1.75.$$

The 95% CI for $\log(\mathrm{OR}_{\text{age} < 50 \text{ years}})$ is:

$$\begin{gathered}\log(\mathrm{OR}_{\text{age} < 50 \text{ years}}) \pm 1.96 \times \sqrt{\mathrm{Var}\left[\log(\mathrm{OR}_{\text{age} < 50 \text{ years}})\right]} \\ \log(1.75) \pm 1.96 \times \sqrt{\frac{1}{68} + \frac{1}{22} + \frac{1}{23} + \frac{1}{13}} \\ 0.56 \pm 1.96 \times \sqrt{0.18} \\ (-0.27, 1.39).\end{gathered}$$

The 95% CI for $\mathrm{OR}_{\text{age} < 50 \text{ years}}$ is $(\exp(-0.27), \exp(1.39)) = (0.76, 4.01)$.

Interpretation: Among patients of age < 50 years, smokers have on average 75% times higher odds of lung cancer compared to non-smokers. However, due to variability, the result does not appear significant since the 95% CI includes 1. 在年龄小于 50 岁的患者中, 吸烟者患肺癌的几率比非吸烟者平均高出 75%. 然而, 由于变异性, 结果并不显著, 因为 95% 置信区间包括 1.

A(2). Among patients aged $\geqslant 50$ years, the odds ratio of lung cancer for smokers vs. non-smokers is:

$$\text{OR}_{\text{age} \geqslant 50 \text{ years}} = \frac{a \times d}{b \times c} = \frac{43 \times 66}{47 \times 45} = 1.34.$$

The 95% CI for $\log(\text{OR}_{\text{age} \geqslant 50 \text{ years}})$ is:

$$\begin{aligned} &\log(\text{OR}_{\text{age} \geqslant 50 \text{ years}}) \pm 1.96 \times \sqrt{\text{Var}\left[\log(\text{OR}_{\text{age} \geqslant 50 \text{ years}})\right]} \\ &\log(1.34) \pm 1.96 \times \sqrt{\frac{1}{43} + \frac{1}{47} + \frac{1}{45} + \frac{1}{66}} \\ &0.29 \pm 1.96 \times \sqrt{0.08} \\ &(-0.26, 0.84). \end{aligned}$$

The 95% CI for $\text{OR}_{\text{age} \geqslant 50 \text{ years}}$ is $(\exp(-0.26), \exp(0.84)) = (0.77, 2.32)$.
Interpretation: Among patients of age $\geqslant 50$ years, smokers have on average 34% times higher odds of lung cancer compared to non-smokers. However, the result does not appear significant since the 95% CI includes 1. 在年龄不小于 50 岁的患者中, 吸烟者患肺癌的几率比非吸烟者平均高出 34%. 然而, 结果并不显著, 因为 95% 置信区间包括 1.

Comparison to Q1: We observe that: the odds ratios of lung cancer were not statistically significant (95% CIs include 1). The stratified odds ratios differ from the unadjusted (unstratified) odds ratio $\text{OR} = 1.87$ (both being lower). This is an indication that the effect of smoking on developing lung cancer changes when we stratify by age, hence age is a confounder. Therefore, we need to make use of the Mantel-Haenszel odds ratio for a more precise estimate, adjusting for the confounding effect of age. 我们观察到: 肺癌的比值比没有统

计学意义 (95% CI 包括 1). 分层比值比与未调整 (未分层) 的比值比 OR = 1.87 (两者都较低) 不同. 这表明, 当按年龄分层时, 吸烟对发展肺癌的影响会发生变化, 因此年龄是一个混杂因素. 因此, 我们需要利用 Mantel-Haenszel 比值比进行更精确的估计, 调整年龄的混杂效应.

Q3. Calculate the Mantel-Haenszel odds ratio and its corresponding 95% confidence interval. Comment on your results.

A.

$$\mathrm{OR_{MH}} = \frac{\sum_{i=i}^{G} w_i \mathrm{OR}_i}{\sum_{i=i}^{G} w_i} = \frac{\sum_{i=i}^{G} \frac{b_i c_i}{n_i} \times \frac{a_i d_i}{b_i c_i}}{\sum_{i=i}^{G} \frac{b_i c_i}{n_i}} = \frac{\sum_{i=i}^{G} \frac{a_i d_i}{n_i}}{\sum_{i=i}^{G} \frac{b_i c_i}{n_i}} = \frac{\frac{68\times 13}{91+35} + \frac{43\times 66}{88+113}}{\frac{22\times 23}{91+35} + \frac{47\times 45}{88+113}} = 1.45.$$

The variance of the $\mathrm{OR_{MH}}$ is:

$$\begin{aligned}
\mathrm{Var}\left[\log(\mathrm{OR_{MH}})\right] = \frac{\sum_{i=i}^{G} w_i^2 v_i}{\left(\sum_{i=i}^{G} w_i\right)^2} \quad &= \frac{\sum_{i=i}^{G} \left(\frac{b_i c_i}{n_i}\right)^2 \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}{\left(\sum_{i=i}^{G} \frac{b_i c_i}{n_i}\right)^2} \\
&= \frac{\left(\frac{22\times 23}{91+35}\right)^2 \left(\frac{1}{68} + \frac{1}{22} + \frac{1}{23} + \frac{1}{13}\right) + \left(\frac{43\times 66}{88+113}\right)^2 \left(\frac{1}{43} + \frac{1}{47} + \frac{1}{45} + \frac{1}{66}\right)}{\left(\frac{22\times 23}{91+35} + \frac{43\times 66}{88+113}\right)^2} \\
&= 0.06.
\end{aligned}$$

The 95% CI for $\log(\mathrm{OR}_{\mathrm{MH}})$ is:

$$\begin{aligned}
&\log(\mathrm{OR}_{\mathrm{MH}}) \pm 1.96 \times \sqrt{\mathrm{Var}\left[\log(\mathrm{OR}_{\mathrm{MH}})\right]}\\
&\log(1.45) \pm 1.96 \times \sqrt{0.06}\\
&(-0.11,\, 0.85).
\end{aligned}$$

The 95% CI for $\mathrm{OR}_{\mathrm{MH}}$ is $(\exp(-0.11), \exp(0.85)) = (0.90, 2.34)$.
After adjusting for the confounding effect of age, the OR of lung cancer did not differ significantly in smokers and non-smokers (95% CI includes 1) despite an OR of 1.45.

. . . .

3 Survival Analysis (生存分析)

3.1 Definitions

3.2 Fundamentals

Denote the time to **the event of interest** (e.g., death) as T, a positive random variable with c.d.f. $F(t) = \Pr(T < t)$ and p.d.f. $f(t)$.

- The **survival function** (生存函数):

$$S(t) = \Pr(T > t) = 1 - F(t), \quad \text{for any } t > 0,$$

  and is the probability of an individual surviving beyond time t (t 时刻后还活着).
  此外, 生存函数是一个单调递减函数, $S(0) = 1$ and decreasing to 0 as $t \to \infty$.

- The **hazard function** (危害函数) 是指在个体存活到该点的情况下, 个体在下一个小时间间隔内可能经历感兴趣事件 (e.g., death) 的概率:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \leqslant t + \Delta t \mid T \geqslant t)}{\Delta t}.$$

  There are many general shapes for the hazard function. The only restriction on $h(t)$ is that it is non-negative i.e., $h(t) \geqslant 0$.

- The **cumulative hazard function**:

$$H(t) = \int_0^t h(u)\,\mathrm{d}u.$$

定理 3.1 (Relationship between $h(t)$ and $S(t)$ (1))

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

证明.

$$\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(T \leqslant t + \Delta t \mid T \geqslant t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Pr(T \leqslant t + \Delta t \cap T \geqslant t) / \Pr(T \geqslant t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Pr(t \leqslant T \leqslant t + \Delta t)}{\Pr(T \geqslant t)\Delta t} \\
&= \frac{1}{\Pr(T \geqslant t)} \lim_{\Delta t \to 0} \frac{\Pr(t \leqslant T \leqslant t + \Delta t)}{\Delta t} \\
&= \frac{f(t)}{S(t)}.
\end{aligned}$$

定理 3.2 (Relationship between $h(t)$ and $S(t)$ (2))

$$S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right).$$

证明.

$$
\begin{aligned}
\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \frac{\mathrm{d}(1-F(t))}{\mathrm{d}t} &= -f(t) \\
\Rightarrow \quad h(t) = \frac{f(t)}{S(t)} = \frac{-\frac{\mathrm{d}S(t)}{\mathrm{d}t}}{S(t)} &= -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} \\
\Rightarrow \quad S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right)&.
\end{aligned}
$$

3.2.1 The exponential distribution

When the survival time $T \sim \mathrm{Exp}(\lambda)$ then

$$
F(t) = 1-\exp(-\lambda t) \;\; \left.\begin{array}{ll} \Rightarrow & f(t) = F'(t) = \lambda\exp(-\lambda t) \\ \Rightarrow & S(t) = 1-F(t) = \exp(-\lambda t) \end{array}\right\} \Rightarrow \; h(t) = \frac{f(t)}{S(t)} = \lambda.
$$

The mean and variance are

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

- Hazard rate (λ) $\begin{cases} \text{Large} & \rightarrowtail \quad \text{High hazard} \;\; \& \;\; \text{Short survival} \\ \text{Small} & \rightarrowtail \quad \text{Low hazard} \;\; \& \;\; \text{Long survival} \end{cases}$
- **Lack of memory property** (无记忆性): 导致无论 t 如何变化, 失败 (死) 的可能性都一样. NOT VERY REALISTIC!

3.2.2 The Weibull distribution

When the survival time $T \sim \text{Weibull}(\lambda, \gamma)$, where $\lambda > 0$ and $\gamma > 0$, the shape and scale parameters respectively, then:

$$\left.\begin{array}{ll} F(t) = 1 - \exp\{-(\lambda t)^\gamma\} & \Rightarrow \;\; f(t) = F'(t) = \lambda^\gamma \gamma t^{\gamma-1} \exp\{-(\lambda t)^\gamma\} \\ & \Rightarrow \;\; S(t) = 1 - F(t) = \exp\{-(\lambda t)^\gamma\} \end{array}\right\} \;\Rightarrow\;\; h(t) = \frac{f(t)}{S(t)} = \lambda^\gamma \gamma t^{\gamma-1}.$$

The mean and variance are

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

where $\Gamma(x) = \displaystyle\int_0^\infty t^{x-1} e^{-t}\,\mathrm{d}t,\; x \in \mathbb{R}$ or $\Gamma(n) = (n-1)!,\; n \in \mathbb{N}$

3.3 Exploratory analyses for survival data

3.3.1 Estimating the survival function

1. The **empirical distribution function (EDF)** estimator is defined as

$$\widehat{S}_{\text{EDF}}(t) = \frac{\text{number surviving beyond } t}{\text{number in the sample}} = \frac{\sum_{i=1}^n \mathbb{1}_{\{t_i > t\}}}{n}.$$

2. The **Kaplan-Meier (KM)** estimator (比 EDF 更好).

n 个观测时间 $(t_1, t_2, \cdots, t_n)$, m 个独特时间 $\left(t_{(1)}, t_{(2)}, \cdots, t_{(m)}\right)$ 且最多 $(n-m)$ 个 censomyred (截尾/删失) 观测值. 加入 $t_{(0)} = 0$, 且 $\Pr(T > t_{(0)}) = 1$.

$$\begin{aligned}
S\left(t_{(i)}\right) = \Pr\left(T > t_{(i)}\right) &= \Pr\left(T > t_{(i)} \cap T > t_{(i-1)}\right) \\
&= \Pr\left(T > t_{(i)} \mid T > t_{(i-1)}\right) \Pr\left(T > t_{(i-1)}\right) \\
&= \Pr\left(T > t_{(i)} \mid T > t_{(i-1)}\right) \Pr\left(T > t_{(i-1)} \mid T > t_{(i-2)}\right) \Pr\left(T > t_{(i-2)}\right) \\
&= \left[\prod_{j=1}^{i} \Pr\left(T > t_{(j)} \mid T > t_{(j-1)}\right)\right] \Pr\left(T > t_{(0)}\right) \\
&= \prod_{j=1}^{i} \Pr\left(T > t_{(j)} \mid T > t_{(j-1)}\right)
\end{aligned}$$

Notion

- 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, 其中 d_j 为在 $t_{(j)}$ 发生死亡的数量 (不包括 censomyred 观测数据)

Then the Kaplan-Meier estimator is given by

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{j=1}^{i} \frac{s_j}{r_j}, \quad \text{for } t_{(i)} \leqslant t < t_{(i+1)}, \ i = 1, 2, \cdots, m,$$

since the estimate of $\Pr\left(T > t_{(j)} \mid T > t_{(j-1)}\right)$ will be $\dfrac{s_j}{r_j}$.

例 3.1 (via 2023.Q2(b)) Suppose one has the following small sample of survival data (in days) where a $+$ denotes a censomyred observation:

$$2+, 3, 3, 5+, 6, 6, 9, 10+, 12, 13.$$

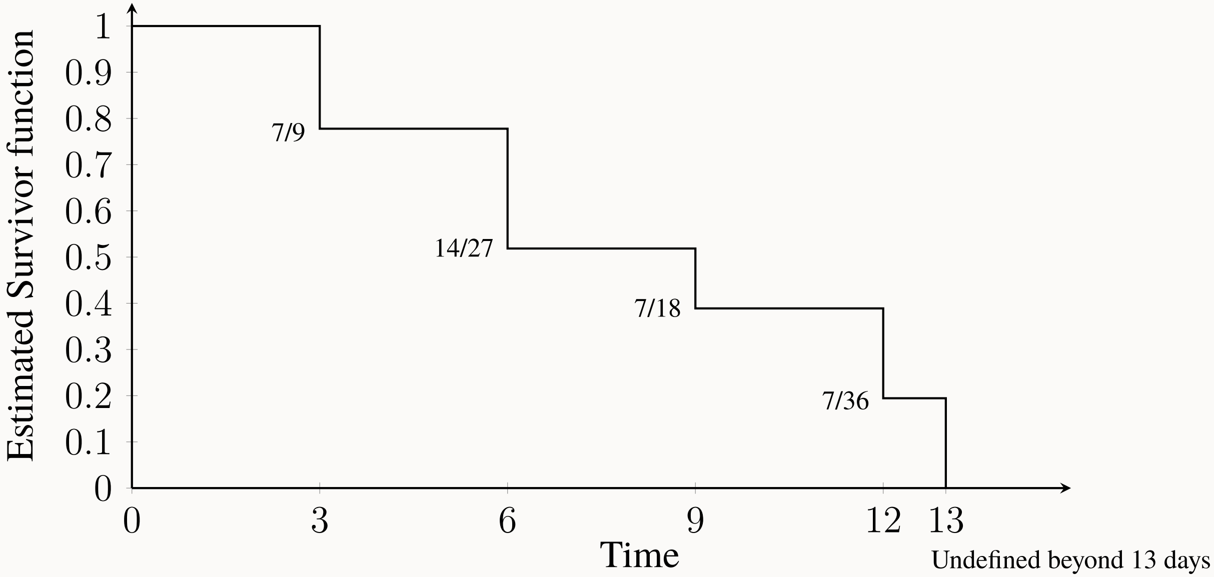
Here there are $m = 5$ distinct event times $(3, 6, 9, 12, 13)$, and hence one can construct the Kaplan-Meier estimate for these data in the form of a table as follows

Event time	Number at risk	Number surviving	Estimated survival
	at time $t_{(i)}$	beyond time $t_{(i)}$	at time $t_{(i)}$
t_i	r_i	$s_i = r_i - d_i$	$\widehat{S}\left(t_{(i)}\right)$
0	10 都活着	10	$10/10 = 1$
3	9 2+ 死了, 3 还没死	$9 - 2 = 7$	$1 \times 7/9 = 7/9 = 0.7778$
6	6 2+, 3, 3, 5+ 死了	$6 - 2 = 4$	$7/9 \times 4/6 = 14/27 = 0.5185$
9	4 9, 10+, 12, 13 活着	$4 - 1 = 3$	$14/27 \times 3/4 = 7/18 = 0.3889$
12	2 12, 13 还活着	$2 - 1 = 1$	$7/18 \times 1/2 = 7/36 = 0.1944$
13	1 只剩 13 活着	$1 - 1 = 0$	$7/36 \times 0/1 = 0$

It is now essential to plot such an estimate as in the figure below.

Q. What are the estimated survival probabilities at times 7 and 8?

A. *At times* 7 *and* 8 *the estimated survival probabilities are both the same and equal to* $14/27 = 0.5185$.



Q. Based on your answer to the question above, give one drawback to the Kaplan-Meier estimator of the survival function and explain briefly how this could be rectified.

A. *The previous answer illustrates that the Kaplan-Meier estimator is a step function, so the probabilities of surviving beyond times* 7 *and* 8 *are the same when one would expect the latter to be smaller. One could get around this discontinuous estimator by applying a smoother to it.*

前一个答案说明 *Kaplan-Meier* 估计量是一个阶梯函数，因此当人们预期后者更小的时候，超过 7 和 8 时间的生存概率是相同的。可以通过对其应用平滑器来绕过这个不连续的估计量。

. . . .

3.3.2 Quantifying uncertainty in the survival function

Recall $\operatorname{Var}[\log(X)] \approx \dfrac{\operatorname{Var}[X]}{(\mathbb{E}[X])^2} \quad \Leftrightarrow \quad \operatorname{Var}[X] \approx (\mathbb{E}[X])^2 \operatorname{Var}[\log(X)]$

证明. $\log(X)$ 在 μ 附近的 Taylor 展开为

$$\begin{aligned}
\log(X) &= \log(\mu) + \frac{1}{\mu}(X-\mu) - \frac{1}{2\mu^2}(X-\mu)^2 + \cdots \quad \approx \log(\mu) + \frac{1}{\mu}(X-\mu) \\
\Rightarrow \quad \mathbb{E}[\log(X)] &\approx \mathbb{E}[\log(\mu)] + \frac{1}{\mu}\mathbb{E}[X-\mu] \quad = \log(\mu) \\
\Rightarrow \quad \mathbb{E}\left[(\log(X))^2\right] &\approx \mathbb{E}\left[\left(\log(\mu) + \frac{1}{\mu}(X-\mu)\right)\right] \\
&= (\log(\mu))^2 + 2\log(\mu)\frac{1}{\mu}\mathbb{E}[X-\mu] + \frac{1}{\mu^2}\mathbb{E}\left[(X-\mu)^2\right] \\
&= (\log(\mu))^2 + \frac{\operatorname{Var}[X]}{\mu^2} \\
\Rightarrow \quad \operatorname{Var}[\log(X)] &\approx \mathbb{E}\left[(\log(X))^2\right] - (\mathbb{E}[\log(X)])^2 \quad = \frac{\operatorname{Var}[X]}{\mu^2} \quad = \frac{\operatorname{Var}[X]}{(\mathbb{E}[X])^2}
\end{aligned}$$

The Kaplan-Meier estimator is a product, and thus variances are difficult to compute. We overcome this by turning it into a sum by a natural log transformation as follows

$$\log\left(\widehat{S}_{\mathrm{KM}}(t)\right) = \log\left(\prod_{j=1}^{i}\frac{s_j}{r_j}\right) = \sum_{j=1}^{i}\log\left(\frac{s_j}{r_j}\right), \quad \text{for } t_{(i)} \leqslant t < t_{(i+1)}$$

Assumption $s_j \sim \mathrm{Binomial}\left(r_j, \theta_j\right) \quad \Rightarrow \quad \widehat{\theta}_{j_{\mathrm{MLE}}} = \frac{s_j}{r_j}$, so that

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

As a result, the variance of $\log\left(\widehat{S}_{\mathrm{KM}}(t)\right)$ for any $t_{(i)} \leqslant t < t_{(i+1)}$ can be computed as follows

$$\begin{aligned}
\mathrm{Var}\left[\log\left(\widehat{S}_{\mathrm{KM}}(t)\right)\right] = \sum_{j=1}^{i}\mathrm{Var}\left[\log\left(\frac{s_j}{r_j}\right)\right] &= \sum_{j=1}^{i}\mathrm{Var}\left[\log\left(\widehat{\theta}_j\right)\right] \\
&\simeq \sum_{j=1}^{i}\frac{\mathrm{Var}\left[\widehat{\theta}_j\right]}{\left(\mathbb{E}\left[\widehat{\theta}_j\right]\right)^2} \\
&= \sum_{j=1}^{i}\frac{1}{\widehat{\theta}_j^2} \times \frac{\widehat{\theta}_j\left(1-\widehat{\theta}_j\right)}{r_j} \\
&= \sum_{j=1}^{i}\frac{1-\frac{s_j}{r_j}}{\frac{s_j}{r_j}\cdot r_j} \quad = \sum_{j=1}^{i}\frac{r_j - s_j}{r_j s_j},
\end{aligned}$$

which assumes independence between successive time periods and uses the approximate variance result above. Using this result a second time the other way around (i.e. $\mathrm{Var}\left[X\right] \approx \left(\mathbb{E}\left[X\right]\right)^2 \mathrm{Var}\left[\log(X)\right]$) yields

$$\mathrm{Var}\left[\widehat{S}_{\mathrm{KM}}(t)\right] = \left(\widehat{S}_{\mathrm{KM}}(t)\right)^2 \sum_{j=1}^{i} \frac{r_j - s_j}{r_j s_j}, \quad \text{for } t_{(i)} \leqslant t < t_{(i+1)},\ i = 1, 2, \cdots, m.$$

Then assuming normality yields a pointwise 95% confidence interval for the survival function at any specified time point, $t > 0$, as

$$\widehat{S}_{\mathrm{KM}}(t) \pm 1.96\sqrt{\mathrm{Var}\left[\widehat{S}_{\mathrm{KM}}(t)\right]} \; = \; \widehat{S}_{\mathrm{KM}}(t) \pm 1.96\widehat{S}_{\mathrm{KM}}(t)\sqrt{\sum_{j=1}^{i} \frac{r_j - s_j}{r_j s_j}}.$$

3.3.3 Estimating the hazard function

A natural method of estimating the **hazard function** for a single sample of survival data is 计算给定事件时间发生事件的数量 与该时间点处于风险中的个体数量 之比, 即

$$\widehat{h}\left(t_{(j)}\right) = \frac{d_j}{I_j i_j},$$

where $d_j = r_j - s_j$ is the number of events at the jth event time $t_{(j)}, j = 1, 2, \cdots, m$ and $I_j = t_{(j+1)} - t_{(j)}$ is the length of the interval between successive (连续) real event times. 请注意，无法估计在始于最终死亡时间的区间内的 hazard function, 因为此区间是开放端点的.

Nelson (1972) and Aalen (1978) proposed estimating the **cumulative hazard function**, which is the total hazard up to time t, as:

$$\widehat{H}_{\text{NA}}(t) = \sum_{j=1}^{i} \frac{d_j}{r_j}, \quad \text{for } t_{(i)} \leqslant t < t_{(i+1)},$$

where again $d_j = r_j - s_j$ is the number of deaths in the jth time interval. This function is thus the sum of the estimated conditional probabilities of death from the first to the ith time interval for $i = 1, 2, \cdots, m$.

Since$S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right)$ and $H(t) = \int_0^t h(u)\,\mathrm{d}u$, then $S(t) = \exp(-H(t))$ and $H(t) = -\log S(t)$. As a result,

$$\widehat{H}_{\text{KM}}(t) = -\log\left(\widehat{S}_{\text{KM}}(t)\right),$$

is an alternative estimator of the cumulative hazard at time t.

3.3.4 Assessing the proportional hazards assumption

考虑两组随机分配接受标准治疗或新治疗的病人, 令 $h_S(t)$ 和 $h_N(t)$ 分别为标准治疗 (S, standard) 和新治疗 (N, new) 病人在时间 t 的死亡风险函数。比例风险假设 (**proportional hazards assumption**) 可以表示为:

$$h_N(t) = \Psi \times h_S(t) \quad \text{or} \quad \frac{h_N(t)}{h_S(t)} = \Psi,$$

for any non-negative time t where Ψ is a positive constant over time. 这里, 新治疗方案患者在时间 t 的风险 (hazard) 与标准治疗方案患者在同一时间的 hazard 成比例.

$$\Psi = \frac{h_N(t)}{h_S(t)} \begin{cases} < 1, \text{则 } h_N(t) \text{ 小于 } h_S(t), \text{ 因此新治疗方案的 } \textit{hazard} \text{ 比标准治疗方案小} \\ > 1, \text{则 } h_N(t) \text{ 大于 } h_S(t), \text{ 因此新治疗方案的 } \textit{hazard} \text{ 更大, 因此标准治疗方案更优越} \end{cases}$$

The estimate of the **cumulative hazard function** can be used to provide a simple graphical test of the proportional hazards assumption. Recall that two groups of patients have proportional hazards if $h_N(t) = \Psi \times h_S(t)$, and hence $H_N(t) = \Psi \times H_S(t)$. Moreover, as $H(t) = -\log(S(t))$ then

$$-\log\left(S_N(t)\right) = \Psi \times \left[-\log\left(H_S(t)\right)\right].$$

If we take logs again then

$$\log\left[-\log\left(S_N(t)\right)\right] = \log\left(\Psi\right) + \log\left[-\log\left(H_S(t)\right)\right].$$

因此，如果我们将对数累积风险函数（log cumulative hazard function）作图, 即 $\log H(t) = \log[-\log S(t)]$ 随时间作图, 并对比来自两个不同人群的样本, 那么如果这两条曲线呈现出平行的线条 (即曲线之间有一个固定的垂直差距), 也就是有一个常数差值 $\log(\Psi)$. 那就说明, 这两个群体之间满足比例风险假设（proportional hazards assumption）.

3.4 Hypothesis testing for survival functions

介绍名为 **log rank test** 的 χ^2 检验, 用于比较每个群体中观察到的事件数与在两个群体之间无差异的零假设下, 每个群体中对应预期事件数的比较. The hypotheses are:

- H_0: The survival functions for the two populations are the same.
- H_1: The survival functions for the two populations are different.

假设两个群体中有 m 个不同时间. Then suppose that at time $t_{(j)}$:

- d_{1j} 和 d_{2j}: 分别表示两个群体在时间 $t_{(j)}$ 发生的事件数. 不包括 censomyred observations. 因此 $d_j = d_{1j} + d_{2j}$.
- r_{1j} 和 r_{2j}: 分别表示两个群体在时间 $t_{(j)}$ 之前发生的事件数. 因此 $r_j = r_{1j} + r_{2j}$.

如果, 在时间区间 $[t_j, t_{j+1})$, 事件发生的个体占比估计为

$$\widehat{\theta}_j = \frac{d_j}{r_j} = \frac{d_{1j} + d_{2j}}{r_{1j} + r_{2j}}$$

那么, 在时间 t_j 的 **expected number of events** (预期事件数) 分别为

$$E_{1j} = r_{1j}\widehat{\theta}_j = \frac{r_{1j}d_j}{r_j}, \quad E_{2j} = r_{2j}\widehat{\theta}_j = \frac{r_{2j}d_j}{r_j}.$$

因此, 定义 **observed and expected numbers of events in each sample**:

- $E_1 = \sum_{j=1}^{m} E_{1j}$ and $E_2 = \sum_{j=1}^{m} E_{2j}$

- $O_1 = \sum_{j=1}^{m} d_{1j}$ and $O_2 = \sum_{j=1}^{m} d_{2j}$

Then the test statistic is given by:

$$\begin{aligned}\chi^2_{\text{LR}} &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \\ &= \frac{\left(\sum_{j=1}^{m} (E_{1j} - d_{1j})\right)^2}{\sum_{j=1}^{m} E_{1j}} + \frac{\left(\sum_{j=1}^{m} (E_{2j} - d_{2j})\right)^2}{\sum_{j=1}^{m} E_{2j}}\end{aligned}$$

在零假设下, 将近似服从 χ^2_1 分布.

. . . .

例 3.2 (via 2023.Q2(d)) Consider data on survival times from a clinical trial of 2 treatments, A and B as detailed below, where a + denotes a censomyred observation.

$$\begin{aligned}&A : 2,\quad 3+,\quad 5+,\quad 7,\quad 7,\quad 11 \\ &B : 4,\quad 5,\quad 6,\quad 8+,\quad 9+,\quad 11+\end{aligned}$$

Construct a table of the combined samples as follows.

Then the test statistic is given by:

$$\chi^2_{\text{LR}} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} = \frac{(4 - 3.27)^2}{3.27} + \frac{(3 - 3.73)^2}{3.73} = 0.31 < \chi^2_1(0.95) = 3.84.$$

$t_{(i)}$	d_{Ai}	d_{Bi}	d_i	r_{Ai}	r_{Bi}	r_i	E_{Ai}	E_{Bi}
2	1	0	1	6	6	12	$6 \times 1/12 = 1/2 = 0.5$	$6 \times 1/12 = 1/2 = 0.5$
4	0	1	1	4	6	10	$4 \times 1/10 = 2/5 = 0.4$	$6 \times 1/10 = 3/5 = 0.6$
5	0	1	1	4	5	9	$4 \times 1/9 = 4/9 = 0.44$	$5 \times 1/9 = 5/9 = 0.56$
6	0	1	1	3	4	7	$3 \times 1/7 = 3/7 = 0.43$	$4 \times 1/7 = 4/7 = 0.57$
7	2	0	2	3	3	6	$3 \times 2/6 = 1$	$3 \times 2/6 = 1$
11	1	0	1	1	1	2	$1 \times 1/2 = 1/2 = 0.5$	$1 \times 1/2 = 1/2 = 0.5$
Total	$O_A = 4$	$O_B = 3$					$E_A = 3.27$	$E_B = 3.73$

hence we fail to reject the null hypothesis and conclude that the survival functions for the two samples are the same.

. . . .

3.5 Modelling survival data – the Cox proportional hazards model

Let $h(t, \boldsymbol{z}_i)$ be the hazard function at time t for individual i with covariate vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{ip})$, and suppose we have data for $i = 1, 2, \cdots, n$ individuals. Then the **Cox proportional hazards model** (PHM) is defined as

$$h(t, \boldsymbol{z}_i) = h_0(t) \exp\left(\boldsymbol{z}_i^\top \boldsymbol{\beta}\right),$$

where, $h_0(t)$ is an arbitrary-shaped baseline hazard function and $\boldsymbol{\beta}$ is a set of unknown parameters in the regression part of the model that measure the effects of the potential explanatory variables on the hazard function for each individual i. The parameters act multiplicatively on the hazard function so that it can be written as the following product:

$$h(t, \boldsymbol{z}_i) = h_0(t) \exp(z_{i1}\beta_1) \exp(z_{i2}\beta_2) \cdots \exp(z_{ip}\beta_p).$$

The *proportional hazards model* (比例风险模型) 的使用源于以下事实: 对于任何两个具有协变量 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 的个体, 在任何时间 t 的 hazard function 之比完全不依赖于 t 或 $h_0(t)$, 即,

$$\frac{\boldsymbol{z}_i}{\boldsymbol{z}_j} = \frac{h_0(t) \exp\left(\boldsymbol{z}_i^\top \boldsymbol{\beta}\right)}{h_0(t) \exp\left(\boldsymbol{z}_j^\top \boldsymbol{\beta}\right)} = \exp\left((\boldsymbol{z}_i - \boldsymbol{z}_j)^\top \boldsymbol{\beta}\right).$$

Q. Describe two model assumptions made by the proportional hazards model (PHM).

A. The PHM assumes that the hazard functions for any two individuals are proportional over time, and the logarithm of the hazard rate is linear in each explanatory variable.

T. *PHM* 假设任何两个个体的风险函数随时间呈比例，且风险率的对数在每个解释变量中呈线性。

. . . .

例 3.3 (via 2023.Q2(c)-iii) The hazards for two subjects are proportional if:

$$h(t, \boldsymbol{z}_1) = C \times h(t, \boldsymbol{z}_2), \quad \forall t,$$

where C is a constant. The proportional hazards assumption for a set of patients can be assessed via plotting the log-cumulative hazard functions against survival time as seen in the lecture materials (一组患者的比例

风险假设可以通过将累积风险函数的对数与生存时间绘制成图来评估). Starting from the mathematical equation above, derive the log-cumulative hazard functions for $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, respectively. **Make sure to explain your reasoning behind every step**. Next, **explain what this plot should look like if the proportionality hazards assumption is satisfied**.

One may express this proportionality assumption using cumulative hazards instead by writing:

$$H(t, \boldsymbol{z}_1) = C \times H(t, \boldsymbol{z}_2).$$

The cumulative hazard function and survival function are linked by the relationship $H(t) = -\log(S(t))$. Hence

$$-\log(S(t, \boldsymbol{z}_1)) = C \times [-\log(S(t, \boldsymbol{z}_2))].$$

Taking the log again gives

$$\log[-\log(S(t, \boldsymbol{z}_1))] = \log(C) + \log[-\log(S(t, \boldsymbol{z}_1))].$$

Hence

$$\begin{aligned} X_1(t) &= \log[-\log(S(t, \boldsymbol{z}_1))] = \log[H(t, \boldsymbol{z}_1)] \\ X_2(t) &= \log[-\log(S(t, \boldsymbol{z}_2))] = \log[H(t, \boldsymbol{z}_2)]. \end{aligned}$$

To assess if the proportional hazards assumption is valid, one plots the quantities $\log[H(t, \boldsymbol{z}_1)]$ and $\log[H(t, \boldsymbol{z}_2)]$ against survival time and checks if the two curves are parallel.

The survival function corresponding to this model for individual i is given by:

$$\begin{aligned}
S(t|\boldsymbol{z}_i) &= \exp\left(-\int_0^t h(s,\boldsymbol{z}_i)\,\mathrm{d}s\right) \\
&= \exp\left(-\int_0^t h_0(s)\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)\,\mathrm{d}s\right) \\
&= \exp\left(-\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)\int_0^t h_0(s)\,\mathrm{d}s\right) \\
&= \left[\exp\left(-\int_0^t h_0(s)\,\mathrm{d}s\right)\right]^{\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)} \qquad = [S_0(t)]^{\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)}
\end{aligned}$$

where $S_0(t)$ is the baseline survival function and can be simply written in terms of the baseline hazard function as

$$S_0(t) = \exp\left(-\int_0^t h_0(s)\,\mathrm{d}s\right).$$

The probability density function from this model is obtained from

$$\begin{aligned}
f(t|\boldsymbol{z}_i) &= h(t|\boldsymbol{z}_i)S(t|\boldsymbol{z}_i) \\
&= h_0(t)\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)[S_0(t)]^{\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)}
\end{aligned}$$

Finally, given a sample of n observations of survival data $\{t_i, \delta_i, \boldsymbol{z}_i\}$, then from Section 3.2 the full data likelihood function is given by

$$\prod_{i=1}^{n} \left\{ (f(t_i|\boldsymbol{z}_i))^{\delta_i} (S(t_i|\boldsymbol{z}_i))^{1-\delta_i} \right\}.$$

where $\delta_i \in \{0, 1\}$, 即

- $\delta_i = 1$, 事件发生, 则我们观察到精确的死亡时间 t_i, 其概率为 $f(t_i|\boldsymbol{z}_i)$.
- $\delta_i = 0$, 数据删失, 则此人在 t_i 时还**活着**, 其概率为 $\Pr(T > t_i|\boldsymbol{z}_i) = S(t_i|\boldsymbol{z}_i)$.

. . . .

例 3.4 (via. 2022.Q2(a)) Represent the survival time by the random variable, T, and assume its cumulative distribution function is given below:

$$\begin{cases} \dfrac{41}{40} \dfrac{0.1t^2}{1 + 0.1t^2}, & 0 < t \leqslant 20, \\ 1, & t > 20. \end{cases}$$

The **survival function** is computed as:

$$S(t) = 1 - F(t) = 1 - \frac{41}{40} \frac{0.1t^2}{1 + 0.1t^2} = \frac{40(1 + 0.1t^2) - 4.1t^2}{40(1 + 0.1t^2)} = \frac{40 - 0.1t^2}{40(1 + 0.1t^2)}.$$

To compute the hazard function, we first need to compute the probability density function $f(t)$ by differentiating

$F(t)$ as follows:

$$\begin{aligned} f(t) &= \frac{\mathrm{d}}{\mathrm{d}x}F(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{41}{40}\frac{0.1t^2}{1+0.1t^2}\right) \\ &= \frac{41}{40}(0.1)\frac{2t(1+0.1t^2)-t^2(0.2t)}{(1+0.1t^2)^2} \\ &= \frac{41}{40}\frac{0.2t}{(1+0.1t^2)^2}, \end{aligned}$$

therefore:

$$f(t) = \begin{cases} \dfrac{41}{40}\dfrac{0.2t}{(1+0.1t^2)^2}, & 0 < t \leqslant 20, \\ 0, & t > 20. \end{cases}$$

The **hazard function** is derived as:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{41}{40}\frac{0.2t}{(1+0.1t^2)^2}}{\frac{40-0.1t^2}{40(1+0.1t^2)}} = \frac{8.2t}{(40-0.1t^2)(1+0.1t^2)}.$$

Consider the following sample of survival data:

$$2, 6, 8^+, 15^+, 19,$$

where a + denotes a censored observation. The **joint likelihood** for n observations is:

$$
\begin{aligned}
&L(t_1, \cdots, t_5, \delta_1, \cdots, \delta_5) \\
&= \prod_{i=1}^{5} \left\{ (f(t_i|\boldsymbol{z}_i))^{\delta_i} (S(t_i|\boldsymbol{z}_i))^{1-\delta_i} \right\} \quad = f(2) \times f(6) \times S(8) \times S(15) \times f(19) \\
&= \frac{41}{40} \frac{0.2(2)}{(1+0.1(2)^2)^2} \times \frac{41}{40} \frac{0.2(6)}{(1+0.1(6)^2)^2} \times \frac{40 - 0.1(8)^2}{40(1+0.1(8)^2)} \times \frac{40 - 0.1(15)^2}{40(1+0.1(15)^2)} \times \frac{41}{40} \frac{0.2(19)}{(1+0.1(19)^2)^2} \\
&= 7.27 \times 10^{-8}
\end{aligned}
$$

Therefore the **log-likelihood** is: $\log(L(t_1, \cdots, t_5, \delta_1, \cdots, \delta_5)) = -16.44$.

. . . .