精算建模：损失分布参考答案及批改评述（Chap 2）

庄源

日期：2025 年 9 月 21 日

目录

1 Question 4：χ^2 goodness-of-fit test on hospital claims

注. 本题制作了 EXCEL 解答，可通过查看表格中的公式了解详细步骤。[下载]

1.1 原题

A sample of 90 hospital claims of X is observed where $\bar{x} = 5010$ and $s^2 = 49,100,100$. Table 1 (of grouped data) was constructed in order to test the goodness-of-fit of:

1. an exponential model for X, and
2. a Pareto model for X (using the method of moments).

Complete the table and perform the appropriate χ^2 goodness-of-fit tests. Comment on the adequacy of fit.

表 1: Hospital claims data

	Interval	O_i(Obs)	E_i(Exp)	E_i (Pareto-MM)
1	$0 \sim 528$	14		
2	$528 \sim 1,118$	17		
3	$1,118 \sim 1,787$	9		
4	$1,787 \sim 2,559$	8		
5	$2,559 \sim 3,473$	7		
6	$3,473 \sim 4,591$	12		
7	$4,591 \sim 6,032$	7		
8	$6,032 \sim 8,063$	4		
9	$8,063 \sim 11,536$	5		
10	$11,536 \sim +\infty$	7		

1.2 参考答案

对于指数分布，有：

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{5010}$$

$$E_i = n\hat{\theta}_i = n\left[e^{-\hat{\lambda}c_i} - e^{-\hat{\lambda}c_{i+1}}\right]$$

对于 Pareto 分布，有：

$$\hat{\alpha} = \frac{2s^2}{s^2 - \bar{x}^2} = 4.0917,\ \hat{\lambda} = (\hat{\alpha} - 1)\bar{x} = 15489.29$$

$$E_i = n\hat{\theta}_i = n\left[\left(\frac{\hat{\lambda}}{\hat{\lambda} + c_i}\right)^{\hat{\alpha}} - \left(\frac{\hat{\lambda}}{\hat{\lambda} + c_{i+1}}\right)^{\hat{\alpha}}\right]$$

将上式代入各区间，可得到区间内预期索赔数 E_i，计算结果如下表所示：

表 2: 各区间内指数分布与 Pareto 分布预期索赔数与实际索赔数

Interval	O_i(Obs)	E_i(Exp)	E_i (Pareto-MM)
$0 \sim 528$	14	9.0023	11.5347
$528 \sim 1118$	17	8.9984	10.7949
$1118 \sim 1787$	9	9.0000	10.0973
$1787 \sim 2559$	8	8.9967	9.4297
$2559 \sim 3473$	7	9.0055	8.8109
$3473 \sim 4591$	12	8.9998	8.2186
$4591 \sim 6032$	7	8.9978	7.6823
$6032 \sim 8063$	4	8.9985	7.2302
$8063 \sim 11536$	5	9.0011	6.9730
$11536 \sim +\infty$	7	8.9999	9.2284

计算 χ^2 统计量:

$$\chi^2_{GF} = \sum_1^E (O_i - E_i)^2 / E_i$$

指数分布的 χ^2 统计量为 16.89（P 值为 0.0313，分布为 $\chi^2(8)$），Pareto 分布的 χ^2 统计量为 9.1418（P 值为 0.2426，分布为 $\chi^2(7)$）。从 P 值上来看，指数分布的 P 值非常小，仅为 3%，这意味着在常见的 5%、10% 显著性水平下，原假设都被拒绝，因此指数分布无法很好拟合原数据。对于 Pareto 分布来说，其 P 值很大，在常见的显著性水平下不会被拒绝，因此 Pareto 分布是索赔分布的较好拟合。

1.3 给分标准与批改评价

表 3: Question 4 给分标准（共 15 分）

采分点	分值
估计指数分布参数	2
估计 Pareto 分布参数	2
计算两种分布的 E_i	4
计算两种分布的 χ^2 统计量	4
论述分布对原始数据的拟合效果	3

本题中，由于四舍五入造成的数值精确度问题不扣分。批改中发现的问题有:

- 题目要求 *complete the table*，但是部分同学没有把表格附上，也没有说明 E_i 的值，因此无法得到“计算 E_i”的 4 分；
- 题目要求 *comment on the adequacy of fit*，部分同学计算完 χ^2 统计量后，没有评论各分布对数据的拟合效果，因此无法得到“论述分布对原始数据的拟合效果”的 3 分；

- 很多同学无法得到正确的 χ^2 分布自由度，请大家仔细阅读课本第 56 面，自由度为 $d = k-1-r$，其中 r 代表待估参数的个数。这里的自由度不是 $k-1$，因为分布的参数必须要通过估计才能得出。
- 部分同学的假设检验叙述不严谨，如果要做假设检验，必须要说明置信度水平 α。没说明置信度水平就直接拒绝，这部分同学视为论述错误。
- 部分同学认为：指数分布的 χ^2 统计量要大于 Pareto 分布的 χ^2 统计量，所以指数分布的拟合效果不好，这是不严谨的说法。因为这两个统计量所对应的 χ^2 分布自由度不一样，统计量的大小不能相提并论。更好的比较方式是去比较两个统计量对应的 P-value。还有部分同学在 5% 的置信度下做卡方检验，这部分同学可以得分。

2 Question 9: χ^2 goodness-of-fit test on home insurance policies

注. 本题制作了 EXCEL 解答，可通过查看表格中的公式了解详细步骤。[下载]

2.1 原题

The following claim data set of 40 values was collected from a portfolio of home insurance policies, where $\bar{x} = 272.675$ and $s = 461.1389$.

10	11	15	22	28	30	32	36	38	48	51
55	56	68	68	85	87	94	103	104	105	106
109	119	121	137	178	181	226	287	310	321	354
393	438	591	1045	1210	1212	2423				

It is decided to fit a Pareto distribution $X \sim \mathrm{Pareto}(\alpha, \lambda)$ to the data using the method of moments. Find these estimates, and use them to perform a χ^2 goodness-of-fit for this distribution by completing Table 4.

表 4: Interval data on 40 home insurance claims

Interval	Observed	Expected
$0, 42.594$	$*$	8
$42.594, 102.270$	$*$	8
$102.270, 196.444$	$*$	$*$
$196.444, 322.336$	$*$	$*$
$322.336, +\infty$	$*$	$*$

2.2 参考答案

对于 Pareto 分布，有：

$$\hat{\alpha} = \frac{2s^2}{s^2 - \bar{x}^2} = 3.0752,\ \hat{\lambda} = (\hat{\alpha} - 1)\bar{x} = 565.8668$$

$$E_i = n\hat{\theta}_i = n\left[\left(\frac{\hat{\lambda}}{\hat{\lambda} + c_i}\right)^{\hat{\alpha}} - \left(\frac{\hat{\lambda}}{\hat{\lambda} + c_{i+1}}\right)^{\hat{\alpha}}\right]$$

将上式代入各区间，可得到区间内预期索赔数 E_i，计算结果如下表所示：

表 5: 各区间内实际索赔数与实际索赔数

Interval	Observed	Expected
$0, 42.594$	9	8
$42.594, 102.270$	9	8
$102.270, 196.444$	10	8
$196.444, 322.336$	4	6
$322.336, +\infty$	8	10

计算 χ^2 统计量：

$$\chi^2_{GF} = \sum_{1}^{E} (O_i - E_i)^2 / E_i$$

该统计量为 1.8154（P 值为 0.4035，分布为 $\chi^2(2)$）。在 1%、5% 和 10% 的置信度下都不拒绝原假设，Pareto 分布对于原始数据的拟合较好。

2.3 给分标准与批改评价

表 6: Question 9 给分标准（共 15 分）

采分点	分值
估计 Pareto 分布参数	4
计算 E_i 并完成表格	4
计算 χ^2 统计量	4
进行假设检验或计算 P 值	3

本题中，由于四舍五入造成的数值精确度问题不扣分。批改中发现的问题有：

- 题目要求做假设检验，但是很多同学在算出 χ^2 统计量后就不继续写了，因此无法得到“进行假设检验或计算 P 值”的 3 分；
- 大家对于自由度的看法五花八门，很多同学无法得到正确的 χ^2 分布自由度，请大家仔细阅读课本第 56 面，自由度为 $d = k - 1 - r$，其中 r 代表待估参数的个数（Pareto 分布的待估参数有两个）。

3 Question 14 (Programming): Sampling from Weibull Distribution

注. 本题制作了示例代码（Rmarkdown），同学们可尝试运行。[下载]

3.1 原题

If $X \sim W(c,\gamma)$, then determine the form of F_X^{-1}. Use this to write R code for generating a random sample of 300 observations from a $W(0.04,2)$ distribution. Run the code and compare your sample mean and variance with the theoretical values.

3.2 参考答案与代码展示（如需运行，请下载示例代码）

3.2.1 F_X^{-1} 的推导

$x \sim W(c,\gamma)$，则 $F_X(x) = 1 - e^{-cx^\gamma}$，令 $u = F_X(x) \sim u(0,1)$，则 $x = F_X^{-1}(u) = \left[-\frac{\ln(1-u)}{c}\right]^{\frac{1}{\gamma}}$。

上述推导非常直白，简单来说，$F_X(x)$ 研究对应 x 下的累计分布函数，而 $F_X^{-1}(u)$ 则是探究给定累计分布函数的情况下，x 为多少。在了解 $F_X^{-1}(u)$ 的情况下，我们就有了一种非常好用的生成随机数的方法（教材第 40 面）：

1. 随机生成一个 u，u 满足 $[0,1]$ 上的均匀分布；
2. 利用 $F_X^{-1}(u)$ 找出 x，即可生成 Weibull 随机数。

3.2.2 均值和方差的理论值

由教材第 40 面，Weibull 变量的各阶矩为：

$$E(X^k) = \frac{1}{c^{k/\gamma}}\Gamma\left(1+\frac{k}{\gamma}\right) \tag{1}$$

代入 $c = 0.04, \gamma = 2$，得：

$$E(X) = \frac{1}{c^{1/2}}\Gamma\left(1+\frac{1}{2}\right) = 5\cdot\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 2.5\sqrt{\pi} \approx 4.431135$$

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{1}{c^{2/2}}\Gamma\left(1+\frac{2}{2}\right) - 4.431135^2 \approx 5.365046$$

也可以直接使用 R 语言计算：

```
# 使用 gamma 计算伽马函数
weibull_mean <-1/c^(1/gamma_weibull)*gamma(1+1/gamma_weibull)
weibull_second_moment <-1/c^(2/gamma_weibull)*gamma(1+2/gamma_weibull)
weibull_variance <- weibull_second_moment - weibull_mean^2
weibull_mean
```

```
## [1] 4.431135
```

```
weibull_variance
```

```
## [1] 5.365046
```

3.2.3 Weibull 随机数的生成

```
# 设置随机数种子，使代码可以复现
set.seed(19991201)
# 设定分布参数
c <- 0.04
gamma_weibull <- 2
# 使用均匀分布生成 Weibull 随机数
weibull_sample <- (-(log(1-runif(300))/c))^(1/gamma_weibull)
# 查看样本均值和样本方差
mean(weibull_sample)
```

```
## [1] 4.251381
```

```
var(weibull_sample)
```

```
## [1] 5.310349
```

生成的随机数均值和方差和理论值接近，但仍有一定差距（参考答案未必与同学们的结果一致，因为随机数种子有差异）。

3.3 给分标准与批改评价

表 7: Question 14 给分标准（共 15 分）

采分点	分值
正确推导 F_X^{-1}	5
正确计算理论均值与方差	5
使用 R 语言生成 Weibull 随机数	5

这道题大部分同学都无法拿到一半的分，主要原因有：

- 漏做题目，请大家务必看清楚原题一共有多少问，很多同学没有得出 F_X^{-1} 就开始写代码；
- 原书 45 面的 $\bar{F}_X(x)$ 不是累积分布函数，而是 $1-F_X(x)$；
- 有些同学在代码处出错。这些同学没有使用参考答案中的 Weibull 随机数生成方式，而是直接使用 `rweibull` 函数生成随机数。但大家有没有发现，R 里面的 Weibull 分布定义和教材是不一样的？所以很多同学会得出不符合预期的随机数。代码如下所示，如果想要更加深入地了解为什么要进行这样的变换，请参考本答案的6.2.1节。

```r
# 设置随机数种子，使代码可以复现
set.seed(2120223132)
# 设定分布参数
c <- 0.04
gamma_weibull <- 2
# 变为 R 语言中的参数
shape <- gamma_weibull # shape 为 2
scale <- c^(-1/gamma_weibull) # scale 为 5
weibull_sample <- rweibull(300, shape = shape, scale = scale)
# 查看样本均值和样本方差
mean(weibull_sample)
```

```
## [1] 4.417005
```

```r
var(weibull_sample)
```

```
## [1] 5.116821
```

4 Question 17: Parameter estimation and claim inflation in lognormal distribution

注. 本题列举了查找对数正态分布累积分布函数的几种 R 软件实现（Rmarkdown），同学们可尝试运行。[下载]

4.1 原题

Suppose that X has a lognormal distribution with parameters μ and σ^2.

(a) Show that the ML estimators of these parameters based on a random sample of size n take the form:

$$\hat{\mu} = \frac{\sum_1^n \log x_i}{n} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_1^n [\log x_i - \hat{\mu}]^2}{n}.$$

(b) A sample of 30 claims from a lognormal distribution gave

$$\sum_1^{30} \log x_i = 172.5 \text{ and } \sum_1^{30} (\log x_i)^2 = 996.675.$$

Using the method of maximum likelihood, estimate the mean size of a claim, and the proportion of claims which exceed 400 .

(c) Let $W = kX$ where $k > 0$. Show that W is also lognormal and determine its parameters.

4.2 参考答案

4.2.1 a：极大似然法估计 Lognormal 的参数

对于一个满足对数正态分布的随机变量，其概率密度函数满足：

$$f_X(x) = \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-(\log x-\mu)^2/2\sigma^2}\right]\frac{1}{x} \tag{2}$$

假设现在有 n 个样本，分别为 x_i $(i=1,\ 2,\ \cdots,\ n)$，则似然函数 $L(\mu,\sigma^2)$ 为：

$$\begin{aligned} L(x_i;\mu,\sigma^2) &= \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-(\log x_i-\mu)^2/2\sigma^2}\right]\frac{1}{x_i} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \sigma^{-n} \cdot e^{-\sum_{i=1}^{n}(\log x_i-\mu)^2/2\sigma^2} \cdot \prod_{i=1}^{n}\frac{1}{x_i} \end{aligned}$$

将似然函数取对数，得到 $l(\mu,\sigma^2)$：

$$l(x_i;\mu,\sigma^2) = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - n\log\sigma - \frac{\sum_{i=1}^{n}(\log x_i-\mu)^2}{2\sigma^2} - \sum_{i=1}^{n}\log x_i$$

对数似然函数 $l(x_i;\mu,\sigma^2)$ 对 μ 求偏导数得：

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(-1)\cdot 2(\log x_i-\mu) = 0$$

化简得：

$$\sum_{i=1}^{n}\log x_i - n\mu = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\log x_i}{n}$$

接下来我们求 $\hat{\sigma}^2$，请注意，这里是要对 σ^2 整体求导，而不是对 σ 求导。为了将过程表示得更清楚，不妨令 $t=\sigma^2$，这时的对数似然函数为：

$$l(x_i;\mu,t) = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - n\log\sqrt{t} - \frac{\sum_{i=1}^{n}(\log x_i-\mu)^2}{2t} - \sum_{i=1}^{n}\log x_i$$

对数似然函数 $l(\mu,t)$ 对 t 求偏导数得：

$$\frac{\partial l}{\partial t} = -\frac{n}{2t} + \frac{\sum_{i=1}^{n}(\log x_i-\mu)^2}{2t^2} = 0$$

化简得：

$$\hat{t} = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\log x_i-\hat{\mu})^2}{n}$$

4.2.2 b：极大似然法估计 Lognormal 参数的数值案例

只需简单的变换，就可以计算参数的估计值：

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\log x_i}{n} = \frac{172.5}{30} = 5.75$$

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{\sum_{i=1}^n (\log x_i - \hat{\mu})^2}{n} \\
&= \frac{\sum_{i=1}^n \left[(\log x_i)^2 - 2\hat{\mu}\log x_i + \hat{\mu}^2 \right]}{n} \\
&= \frac{\sum_{i=1}^n (\log x_i)^2}{n} - 2\hat{\mu}\frac{\sum_{i=1}^n \log x_i}{n} + \hat{\mu}^2 \\
&= \frac{\sum_{i=1}^n (\log x_i)^2}{n} - \hat{\mu}^2 \\
&= \frac{996.675}{30} - 5.75^2 = 0.16
\end{aligned}
$$

由教材第 47 面关于对数正态分布各阶矩的论述，赔款随机变量 X 的平均值为：

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2} = 340.3587$$

$X \sim Lognormal(\mu, \sigma^2)$，那么 $\log(X) \sim Normal(\mu, \sigma^2)$，这样就可以使用我们熟悉的正态分布计算概率了。赔款大于 400 的概率可由下式求得：

$$
\begin{aligned}
\Pr(X > 400) &= \Pr(\log X > \log 400) \\
&= 1 - \Phi\left(\frac{\log 400 - 5.75}{\sqrt{0.16}}\right) = 1 - \Phi(0.60366) = 0.273034
\end{aligned}
$$

其中 $\Phi(*)$ 是标准正态分布的累积分布函数。

上述数值可以通过几种方式计算，精确度各有不同：

1. 查正态函数表，常用的正态函数表是 SOA 考试 P 的正态函数表 [下载]，可查得 $\Phi(0.60) = 0.7257$；
2. EXCEL，在任意单元格内写下：=1-NORM.S.DIST((LN(400)-5.75)/SQRT(0.16),TRUE)，可得结果 0.273034；
3. R，下面列举了三种计算本题中数值的方式，有的直接使用对数正态分布的 R 函数，还有的使用对数正态与正态之间的关系得出答案，结果都为 0.2730344。[下载]

```r
plnorm(400, meanlog = 5.75, sdlog = 0.4, lower.tail = FALSE)
```

```r
1 - pnorm(log(400),mean = 5.75, sd = 0.4)
```

```r
1 - pnorm((log(400)-5.75)/0.4)
```

```
## [1] 0.2730344
```

4.2.3 c：通货膨胀下的对数正态分布

通过变量替换，便可以得出 kX 的分布：

$$F_W(w) = \Pr(W \leqslant w) = \Pr(kX \leqslant w) = \Pr\left(X \leqslant \frac{w}{k}\right) = F_X\left(\frac{w}{k}\right) \tag{*}$$

上式对 w 求导，得：

$$f_W(w) = \left[F_X\left(\frac{w}{k}\right)\right]' = \frac{1}{k}\left\{\frac{1}{\frac{w}{k}\cdot\sigma\sqrt{2\pi}}e^{-\frac{\left[\log\left(\frac{w}{k}\right)-\mu\right]^2}{2\sigma^2}}\right\}$$
$$= \frac{1}{w\sigma\sqrt{2\pi}}e^{-\frac{(\log w-\log k-\mu)^2}{2\sigma^2}}$$

所以，$W \sim Lognormal\left(\log k + \mu, \sigma^2\right)$。

事实上，可以不用这么大费周折。我们可以直接从(*)式中推断出 W 的分布类型。从上一小题中我们知道，对数正态分布随机变量的累积分布函数为：

$$F_X(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right) \tag{**}$$

由(*)和式(**)，

$$F_W(w) = F_X\left(\frac{w}{k}\right) = \Phi\left(\frac{\log\frac{w}{k} - \mu}{\sigma}\right) = \Phi\left[\frac{\log w - (\log k + \mu)}{\sigma}\right]$$

令 $\mu^* = \mu + \log k$，不难看出 $W \sim Lognormal\left(\mu^*, \sigma^2\right)$。

4.3 给分标准与批改评价

表 8: Question 17 给分标准（共 20 分）

小题	采分点	分值
a（共 10 分）	正确写出似然函数	2
	正确写出对数似然函数	2
	正确估计 μ	3
	正确估计 σ^2	3
b（共 6 分）	给出 μ 的估计值	1
	给出 σ^2 的估计值	2
	计算赔款随机变量的均值	1
	计算赔款大于 400 的概率	2
c（共 4 分）	计算 W 的累积分布函数	1
	计算 W 的概率密度函数或直接推导 W 的形式	2
	叙述 W 满足的分布参数	1

同学们能在这题拿到不错的成绩，但是出现了以下问题：

- 漏题。大部分同学漏掉了赔款随机变量的均值，请大家读题仔细一些。
- 跳步。第一小题叫大家"*Show*"，其实就是叫大家证明。直接给出参数估计的结果只能得到部分分数。
- 很多同学在极大似然估计时对 σ 求导而不是 σ^2 求导，可能算出来数字是一样的。但是 σ^2 整体是一个参数，不能够取其中的一部分来求导。

5 Question 18: Maximum likelihood estimation and interval estimation under reinsurance

5.1 原题

On a particular class of policy, claim amounts coming into Surco Ltd. follow an exponential distribution with unknown parameter λ. A reinsurance arrangement has been made by Surco so that a reinsurer will handle the excess of any claim above $\$10,000$. Over the past year, 80 claims have been made and 68 of these claims were for amounts below $\$10,000$; these 68 in aggregate value amounted to $\$220,000$. The other 12 claims exceeded $\$10,000$.

(a) Let X_i represent the amount of the i^{th} claim from the 68 claims beneath $\$10,000$. Show that the log-likelihood function is

$$\ell(\lambda) = 68 \log \lambda - \lambda \sum_{i=1}^{68} x_i - 120,000\lambda$$

Hence find $\hat{\lambda}$ and calculate an approximate 95% confidence interval for λ.

(b) Let Z denote the cost to the reinsurer of any claim X, and hence $X = Y + Z$. Determine an expression for $E(Z)$ in terms of λ. Estimate $E(Z)$ using maximum likelihood.

(c) Next year, claim amounts are expected to increase in size by an inflationary figure of 5%. Suppose that the excess of loss reinsurance level remains at $\$10,000$. Let Z^* represent the cost to the reinsurer of a typical claim next year. Estimate $E\left(Z^*\right)$. Using your answer in (18a) or otherwise, derive a 95% confidence interval for $E\left(Z^*\right)$.

5.2 参考答案

5.2.1 题目到底在讲什么？

这道题我非常喜欢，非常适合精算考试。这道题把极大似然估计考得很深很透，也把再保险和通货膨胀结合在一起，基本融合了教材 2.5 节的所有知识。很多同学不太清楚题目到底在讲什么，让我给大家讲细一点点：

有一家保险公司买了再保险，只要赔案损失超过 10000 美金，超过的那部分就由再保险承担。也就是说，保险公司对一个赔案最多只承担 10000 美金。对于这类损失超出 10000 美金的赔案，保险公司的赔付记录上只会有“赔出 10000 美金”的描述，损失到底是多少，我们并不太了解。现在一共有 80 个赔案，其中 68 个没超过 10000 美金，总额为 220000；剩下 12 个超过了 10000 美金。

在保险中，经常会出现这样的情况：我们知道损失数据大于某个值，但是并不知道它的确切值是多少，只知道它大于了一个确定的值（如本题中的 10000 美金）。保险公司购买再保险就会造成上述情况，这种数据并不是我们平常讨论的完整数据（Complete Data），而被称为“删失数据”（Censored Data）。删失数据的极大似然估计与完整数据略有不同，请大家参考教材第 63 面。

评论. 保险公司还会出现一种情况，他们可能完全不知道小于某个值的损失数据。这是保单中的哪个条款导致的呢？这种数据又被称为什么数据？请同学们查阅相关资料。

5.2.2 a（1）：删失数据的极大似然估计

已知 $f_X(x)=\lambda e^{-\lambda x}$，则似然函数为：

$$L(\lambda)=\left(\prod_{i=1}^{68} f_{X_i}(x_i)\right)\cdot\left(\bar{F}_X(10000)\right)^{12}$$

上述似然函数其实非常好理解。根据上一小节关于题目的讨论，有 68 个数据小于 10000，我们完全了解这些数据的确切值，在写似然函数的时候还是采取概率密度连乘的方式。剩下的 12 个数据大于 10000，我们也只知道它们大于 10000，不了解它们的真实值究竟是多少，这时就拿 $\bar{F}_X(10000)$ 代替原来的概率密度。

取对数，得：

$$\begin{aligned} l(x_i;\lambda) &= \sum_{i=1}^{68}\log f_{X_i}(x_i)+12\log\bar{F}_X(10000) \\ &= \sum_{i=1}^{68}\log\left(\lambda e^{-\lambda x_i}\right)+12\log e^{-10000\lambda} \\ &= 68\log\lambda-\lambda\sum_{i=1}^{68}x_i-120000\lambda \end{aligned}$$

对 λ 求导，得：

$$\frac{\partial l}{\partial\lambda}=\frac{68}{\lambda}-\sum_{i=1}^{68}x_i-120000=0$$

解得 $\hat{\lambda}=0.0002$。

5.2.3 a（2）：极大似然估计的渐进性质与区间估计

想要得出 λ 的区间估计，我们就必须得知道 $\hat{\lambda}$ 到底满足什么分布。大家在大二的时候学过数理统计，肯定学过 Fisher 信息矩阵。防止有些同学已经忘掉了，我在这里再简单介绍一下，你们也可以再参考教材 324 面附录 B.4，也有类似表述。在单参数分布中，令 $\hat{\theta}$ 为参数的估计值，θ 为参数真值，n 为样本量，则有[1]：

$$\hat{\theta}\dot{\sim}N(\theta,\frac{1}{I(\theta)}) \tag{3}$$

其中，“$\dot{\sim}$” 表示“近似为某某分布”，当大样本的时候[2]，这个近似才比较好[3]。$I(\theta)$ 是 Fisher 信息。

---

[1] 下面的表述和教材的表述稍微有些不一样，我这里统一使用了总体的 Fisher 信息。因为教材 324 面中假设独立同分布的完整数据，因此总体的 Fisher 信息即为单个观测点的 Fisher 信息乘以 n。两者都没问题，大家以我这里写的为准。

[2] 本题有 80 个样本，已经比较大了，所以可以放心使用渐进分布。

[3] 如果极大似然估计量是无偏的话，极大似然估计量的渐进方差是所有无偏估计中最小的，达到了克拉默-拉奥下界（Crámer–Rao lower bound）。这个下界的推导者之一 Calyampudi Radhakrishna Rao 于 2023 年 8 月 22 日去世，他的一生就是统计的 100 年。[Calyampudi Radhakrishna Rao 简介]

一般来说，Fisher 信息有两种方式表示：

$$I(\theta)=E\left[\frac{\partial}{\partial\theta}\log L(\mathbf{x};\theta)\right]^2=-E\left[\frac{\partial^2}{\partial\theta^2}\log L(\mathbf{x};\theta)\right] \tag{4}$$

其中，$\mathbf{x}$ 表示观测值形成的向量，$L(\mathbf{x};\theta)$ 是似然函数，取对数后变为对数自然函数。

仔细观察式(4)，我们可以比较两种方式下的 Fisher 信息的计算难度。一般来说，我们喜欢用 $-E\left[\frac{\partial^2}{\partial\theta^2}\log L(\mathbf{x};\theta)\right]$ 来计算 Fisher 信息，因为 $E\left[\frac{\partial}{\partial\theta}\log L(\mathbf{x};\theta)\right]^2$ 中的二阶矩一般比较难算。也就是说，现在我们要把对数似然函数关于参数求两次导数，求期望以后，加一个负号就是 $I(\theta)$。

之前已经知道了对数似然是多少，让我们求两阶导：

$$\begin{aligned}I(\lambda)&=-E\left[\frac{\partial^2}{\partial\lambda^2}\log L(\mathbf{x};\lambda)\right]\\&=-E\left[\frac{\partial^2}{\partial\lambda^2}\left(68\log\lambda-\lambda\sum_{i=1}^{68}x_i-120000\lambda\right)\right]\\&=-E\left[-\frac{68}{\lambda^2}\right]=\frac{68}{\lambda^2}\end{aligned}$$

使用 $\hat{\lambda}=0.0002$ 来估计 Fisher 信息，则有 $I(\lambda)=1.7\times10^9$，代入式(3)，得

$$\hat{\lambda}\dot{\sim}N(\lambda,\frac{1}{I(\lambda)})=N(\lambda,5.8824\times10^{-10})$$

所以，λ 的区间估计为：$\left(\hat{\lambda}-Z_{0.975}\times\sqrt{5.8824\times10^{-10}},\hat{\lambda}+Z_{0.975}\times\sqrt{5.8824\times10^{-10}}\right)$ 其中，$Z_{0.975}$ 是标准正态分布的 97.5% 分位数，数值上约为 1.96。

因此，λ 的 95% 区间估计为$(0.000152463, 0.000247537)$。

有同学还要问，刚才我们只讨论了单参数分布，要是现在有多参数分布，Fisher 信息会变成什么形式呢？答案是一个矩阵！设某个分布有两个参数要估计（如 Gamma 分布），分别为 θ_1 和 θ_2，Fisher 信息就是一个 2*2 的矩阵：

$$I(\theta_1,\theta_2)=\begin{pmatrix}-E\left[\frac{\partial^2}{\partial\theta_1^2}\log L(\mathbf{x};\theta_1,\theta_2)\right] & -E\left[\frac{\partial^2}{\partial\theta_1\partial\theta_2}L(\mathbf{x};\theta_1,\theta_2)\right]\\ -E\left[\frac{\partial^2}{\partial\theta_1\partial\theta_2}\log L(\mathbf{x};\theta_1,\theta_2)\right] & -E\left[\frac{\partial^2}{\partial\theta_2^2}\log L(\mathbf{x};\theta_1,\theta_2)\right]\end{pmatrix} \tag{5}$$

对 Fisher 信息矩阵求逆，得到的就是 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的协方差矩阵[4]。$\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的渐进联合分布是二维正态分布：

$$\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{pmatrix}\dot{\sim}N\left(\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix},[I(\theta_1,\theta_2)]^{-1}\right) \tag{6}$$

这个分布的图像长得有点像山峰：

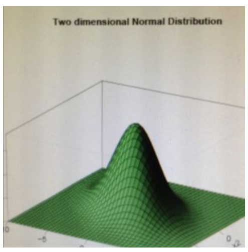

图 1: 一个二维正态分布的图像（张连增老师的微信头像）

[4]注意到了吗？在一维条件下，$\hat{\theta}$ 的方差即为 Fisher 信息的倒数，这就是一维情况下的“求逆”！

现在，你看得懂教材 44 面的推导了吗?

5.2.4 b：含再保险情况下对再保险责任的估计

因为 $X=Y+Z$，所以 $E(X)=E(Y)+E(Z)$。而教材第 62 面刚好讲过，在指数分布的情况下，有：

$$E(Y)=\frac{1}{\lambda}(1-e^{-\lambda M}) \tag{7}$$

所以：

$$E(Z)=E(X)-E(Y)=\frac{1}{\lambda}-\frac{1}{\lambda}(1-e^{-\lambda M})=\frac{1}{\lambda}e^{-\lambda M}$$

代入 $M=10000$，$\lambda=0.0002$，得$E(Z)=676.676$。

$E(Z)$ 也可以直接积分得出，在指数分布下这并不难。到了考试时，大家可能不记得(7)这个式子，所以硬着头皮推也很具实用价值。下面给出了推导过程：

$$\begin{aligned}E(Z)&=0\cdot F_X(M)+\int_M^{+\infty}(x-M)f_X(x)dx\\&=\int_M^{+\infty}(x-M)f_X(x)dx\\&=\int_0^{+\infty}yf_X(y+M)dy\ \cdots\cdots(\text{令}y=x-M)\end{aligned}$$

代入指数分布的概率密度函数做积分：

$$\begin{aligned}E(Z)&=\int_0^{+\infty}y\cdot\lambda e^{-\lambda(y+M)}dy\\&=e^{-\lambda M}\int_0^{+\infty}y\lambda e^{-\lambda y}dy\\&=-e^{-\lambda M}\int_0^{+\infty}y\,d\left(e^{-\lambda y}\right)\\&=-e^{-\lambda M}\left(ye^{-\lambda y}\Big|_0^{+\infty}-\int_0^{+\infty}e^{-\lambda y}dy\right)\ \cdots\cdots(\text{分部积分法})\\&=-e^{-\lambda M}\left[(0-0)+\frac{1}{\lambda}e^{-\lambda y}\Big|_0^{+\infty}\right]=\frac{1}{\lambda}e^{-\lambda M}\end{aligned}$$

5.2.5 c：通货膨胀下的再保险责任

教材第 65 面已经推导了指数分布下保险人的自留责任 $E(Y^*)$：

$$E\left(Y^*\right)=k\left[E(X)-\int_0^{\infty}y\lambda e^{-\lambda(y+M/k)}dy\right]=\frac{k}{\lambda}\left[1-e^{-\lambda M/k}\right]$$

其中 k 为通货膨胀率。

令 $X^*=kX$ 为通货膨胀后的总损失，利用 $E(X^*)=E(Y^*)+E(Z^*)$，则：

$$\begin{aligned}E(Z^*)&=E(X^*)-E(Y^*)\\&=kE(X)-\frac{k}{\lambda}\left[1-e^{-\lambda M/k}\right]\\&=\frac{k}{\lambda}e^{-\lambda M/k}\end{aligned}$$

上述结论也可以通过分部积分方式，使用与6.2.1节相似的方法进行推导。代入 $k=1.05$，$\lambda=0.0002$，可得$E(Z^*)=781.50$。

代入5.2.3节中得到的 λ 区间估计 $(0.000152463, 0.000247537)$，可得 $E(Z^*)$ 的 95% 置信度区间估计为(401.5152，1612.1923)。

5.3 给分标准与批改评价

表 9: Question 18 给分标准

小题	采分点	分值
a（共 10 分）	写出似然函数	2
	写出对数似然函数	2
	正确估计 λ	2
	给出 Fisher 信息	2
	正确进行区间估计	2
b（共 5 分）	正确计算 $E(Z)$	5
c（共 5 分）	正确计算 $E(Z^*)$	3
	正确给出 $E(Z^*)$ 的区间估计	2

这道题是全部作业题中情况最惨烈的题，平均分仍不足 10 分，大家的主要问题有：

- 漏题。（a）题叫大家"*Show*"，其实就是叫大家证明这个对数似然函数为什么长这个样子，而不是让大家直接拿来就用。只有从似然函数的构造开始向下得到对数似然，才能叫"*Show*"。此外，（a）和（c）要求大家得出 95% 置信区间，很多同学都漏掉了。
- 没有直接使用书中的二级结论，自行推导反而把题目做错了。教材第 62 面和 65 面的结论可以直接拿来解（b）和（c），有很多同学尝试自己推导，推错了。大家可以再看看书。
- 区间分布使用方法不当。在区间估计中，大多数同学使用的是中心极限定理，也就是下面这个式子：

$$\lim_{n\to\infty} \Pr\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < a\right) = \Phi(a) \tag{8}$$

  可最大的问题是，我们是使用极大似然法来估计 λ 的，这个估计量也不等于样本均值（在这道题中，也求不出样本均值的准确值），因此说 $\dfrac{\hat{\lambda} - \mu}{\sigma/\sqrt{n}}$ 也渐进服从正态分布是不对的。所以这个置信区间不能写成 $\left(\hat{\lambda} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\lambda} + 1.96\frac{\sigma}{\sqrt{n}}\right)$。极大似然估计量的方差应该使用 Fisher 信息矩阵求得，请大家参考5.2.3节。

6 Question 22 (Programming): K-S test on dataset `Theft`

注. 本题制作了数据集，可直接用于代码复现。[下载]

本题制作了示例代码（Rmarkdown），需配合数据集使用。[下载]

6.1 原题

Use Kolmogorov–Smirnoff tests to test fitness of the Weibull (ML and M%) and lognormal distributions to the Theft claim data.

6.2 参考答案与代码展示（如需运行，请下载数据集和示例代码）

6.2.1 从书本到 R 代码：Reparametrization

本题主要考察 K-S 检验的代码操作，书上已经有类似的代码，也有参数估计的标准答案，但细心的同学可能会发现：书上记录的 Weibull 分布和 R 软件里记录的 Weibull 分布的概率密度函数不一样。

以下是书上给到的概率密度函数：

$$f_X(x) = c\gamma x^{\gamma-1}e^{-cx^{\gamma}} \tag{9}$$

下面是 R 语言 `pweibull` 函数帮助中的概率密度函数，其中 a 是形状参数，σ 是尺度参数：

$$f(x) = (a/\sigma)(x/\sigma)^{a-1}\exp\left(-(x/\sigma)^{a}\right) \tag{10}$$

但实际上，上述两个公式是一回事！只不过用到的参数不太一样。令(9)式中的 γ 为 a，c 为 σ^{-a}，就能够得到(10)式。这种使用不同参数表达同一个分布的方式被称为 Reparametrization。在用极大似然法做估计的时候，某些形式的似然函数让极大似然估计的方程没有闭合解（Closed-form Solution）[5]，也有可能对算法的编写提出挑战，所以“换一种方式”表述分布的参数非常有必要。除了 Weibull 分布以外，Gamma 分布也经常要用到 Reparametrization（见书中第 43 面）。下面的很多与 Weibull 分布有关的代码虽然一开始使用的是书中 c 和 γ 的估计值，但最终都转换成了式(10)中的 a 和 σ，便于与 R 中的 `pweibull` 函数耦合，请同学们注意。

6.2.2 读入数据集

```r
library(readxl)
Theft <- read_excel("Chap_2_Dataset_Theft.xlsx")
# 将 Theft 从数据框变为向量，便于进行 K-S 检验
Theft <- Theft$Theft
```

[5] From Wikipedia: “In mathematics, a closed-form expression is a mathematical expression that uses a finite number of standard operations. It may contain constants, variables, certain well-known operations (e.g., + − × ÷), and functions (e.g., nth root, exponent, logarithm, trigonometric functions, and inverse hyperbolic functions), but usually no limit, differentiation, or integration. The set of operations and functions may vary with author and context.”

6.2.3 极大似然估计 Weibull 参数的 K-S 检验

书中第 46 面已经给出了 Weibull 分布参数的极大似然估计:

$\hat{c} = 0.00518,\ \hat{\gamma} = 0.71593$

下面的代码在 Reparametrization 后直接进行 K-S 检验:

```
c_mle_weibull <- 0.00518
gamma_mle_weibull <- 0.71593
ks.test(Theft,"pweibull",gamma_mle_weibull,
        c_mle_weibull^(-1/gamma_mle_weibull))
```

```
##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Theft
## D = 0.10079, p-value = 0.1746
## alternative hypothesis: two-sided
```

6.2.4 分位数估计 Weibull 参数的 K-S 检验

书中第 46 面已经给出了 Weibull 分布的分位数估计:

$\ddot{c} = 0.002494, \ddot{\gamma} = 0.847503$

仿照前面的代码，做 K-S 检验:

```
c_mpercent_weibull <- 0.002494
gamma_mpercent_weibull <- 0.847503
ks.test(Theft,"pweibull",
        gamma_mpercent_weibull,
        c_mpercent_weibull^(-1/gamma_mpercent_weibull))
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Theft
## D = 0.084857, p-value = 0.3532
## alternative hypothesis: two-sided
```

6.2.5 极大似然估计 Lognormal 参数的 K-S 检验

书中第 48 面已经给出了对数正态分布的分位数估计:

$\hat{\mu} = 6.62417,\ \hat{\sigma}^2 = 2.30306$

仿照前面的代码，做 K-S 检验：

```
mu_lognormal <- 6.62417
sd_lognormal <- sqrt(2.30306)
ks.test(Theft,"plnorm",mu_lognormal,sd_lognormal)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Theft
## D = 0.08673, p-value = 0.3274
## alternative hypothesis: two-sided
```

6.2.6 怎样能够得出 Weibull 分布参数的极大似然估计？(不作要求)

有些同学会疑问，书上对于 Weibull 分布 c 和 γ 的估计是怎么得出来的？使用计算机，这非常简单！极大化似然函数相当于做优化。因此可以直接调用 R 中的优化算法得出使得似然函数最大的参数；我们使用 R 语言中的 `optim` 函数最大化对数似然函数，并得出相应的参数估计。

```
weibull.fun<- function(parameter,x){
    shape_weibull <- parameter[1]
    scale_weibull <- parameter[2]
# 对数似然函数
    logL<- sum(log(dweibull(x,
                   shape=shape_weibull,
                   scale=scale_weibull)))
return(-logL)
}
# 因为 R 中只有最小化函数 optim()
# 我们只需要参数的值，最大化 logL 和最小化-logL 是一致的
# 初始化两个参数作为迭代初始值
theta0 <- c(0.5,100)
result <- optim(theta0,weibull.fun,x=Theft)
# 参数的值保存在 result$'par'中
# 有两个值，第一个是形状参数 第二个是尺度参数
# 换算成书上的参数方式
c_mle <- result[["par"]][2]^(-result[["par"]][1])
gamma_mle <- result[["par"]][1]
# 左边是我们估计的参数，右边是书上给出的答案
```

```
# 结果非常接近
c_mle; c_mle_weibull
```

```
## [1] 0.005194044
```

```
## [1] 0.00518
```

```
        gamma_mle; gamma_mle_weibull
```

```
## [1] 0.7156335
```

```
## [1] 0.71593
```

看来我们和书上的结果非常接近！在上面，我使用了 Reparametrization 略微简化了代码。

还有没有一步到位的方法呢？当然有！使用 `fitdistrplus` 这个包就可以一步到位：

```
library(fitdistrplus)
fitW <- fitdist(Theft,"weibull",method = "mle")
fitW[["estimate"]][["scale"]]^(-fitW[["estimate"]][["shape"]])
```

```
## [1] 0.005187123
```

```
fitW[["estimate"]][["shape"]]
```

```
## [1] 0.7158071
```

得到的结果也跟书上十分接近。

6.3 给分标准与批改评价

这道题对于同学们来说难度较大，大部分同学都空着。精算考试中，“看代码做题”的情况其实很常见，因此各位同学的代码不能荒废。除此以外，还有十几个同学的答案错得惊人的相似，答案到小数点后五位都是一样的，还请大家独立思考。

表 10: Question 22 给分标准（共 15 分）

采分点	分值
写了题号	4
题号后写了东西（不管对错）	4
使用任意一种编程语言进行 K-S 检验并附有代码（截图或抄写）	4
得出近似正确的 D 统计量与 P 值	3

极小部分同学不附代码，但也得出了近似正确的 D 统计量与 P 值，也给到满分或接近满分。

7 批改评述总结

本次作业共六题，各题分值总结如下：

表 11: 各题分值分布及班级卷面平均分

题号	分值	全班均分（不含迟交、漏交和严重抄袭）
4	15	10.78
9	15	11.84
14	15	6.73
17	20	15.51
18	20	8.92
22	15	7.81
总计	100	61.59

本次批改严格按照采分点给各位同学赋分，赋分方式与期末考试相似，大家可以看看自己的真实水平。做得最差的两道题是 14 和 18 题，大家可以多参考一下我给的答案。

本次卷面均分过低，漏答情况严重。几道编程题失分最多，也已尽量宽松批改，但情况仍未达预期。结合往年经验，降低同学们期末压力，张老师和我商议，**本次**按时提交的作业以下述方式给分。令 X 为卷面分，Y 最终作业分（登入给分系统），则有：

$$Y = \begin{cases} 80, & X < 80 \\ 10\sqrt{X}, & X \geqslant 80 \end{cases}$$

迟交的作业按照上述方式转换后，仍然按开学所讲，打 8 折处理。

助教学长的话：请大家别不过脑子就抄别人的答案。我认为，最重要的不是你交了一份看起来很漂亮的作业给我，而是你思考过了，期末考试哪怕没有你宿舍里那哥们儿或者你的好朋友，你自己都能考 60 分。你交份作业应付了事，糊弄糊弄助教，在我这儿当然没问题，在你哥们/姐妹面前吹吹牛虚荣一把也没问题，但别把自己给骗了。

关于这门课，我感触颇深。在你们前面三届的学长学姐们都告诉我一个经验：对于精算建模，得自己真正理解，否则期末就是给了你一模一样的题也不会。**平时到处抄，但期末连 20 分的题都没写满，在前几届学生中这样的人不在少数，请大家不要重蹈覆辙，别让期末考试卷子揭穿了你的谎言**。

我深知同学们对于不挂科和高 GPA 的渴望，因为在你们这个年纪，我也要为了保研和评奖焦虑；但千万不要舍本逐末失了德行。给分给得很宽容，请再记住：作业不是目的，而是过程。