考试，StatBI/E，2024/25

[请在此处填写您的姓名]

在周四，1月16日下午2点UTC+1之前提交（逾期提交将不予考虑）

目录

– 参考考试编织测试（参见 Absalon）以获取常见编织问题的可能解决方案。（作为备选方案，您可以编织到 HTML / DOC，打开生成的文档，并通过 “ 打印到 PDF ” 将此文档转换为 PDF；请特别注意仔细检查生成的 PDF 是否完整且外观符合预期。）

- 清楚地指出你正在回答的问题，并给出带有理由的完整答案，也就是说，仅仅在PDF中陈述一个代码块及其输出是不够的，也不会得分。例如，如果在问题12中，你被给定了随机变量X的n个独立同分布样本，并被要求提供一个其期望值的点估计，那么以下

```
> mean(data)
[1] 39.183
```

  是不够的，一个更好的替代方案可能是

  Answer to Problem 12
  经验平均值是期望值的点估计量。我们使用R来计算给定样本的经验平均值，并在下方打印结果

```
> mean(data)
[1] 39.183
```

  上述示例中的代码可以说是自解释的，因为它只是一行代码调用了一个具有自解释名称的单一命令，`mean`。较长的代码块通常不是自解释的。不要依赖报告的读者去猜测你在表格中试图实现/展示什么/用图表说明什么/实现什么，而应该提供包含文本+代码+输出的完整自包含答案。

- 始终解释并证明你的所有答案，并解释和记录代码。
- 如果您发现任何问题陈述中有问题，请在您的解答中明确指出并描述，并阐明您此后是如何进行的。

- 如果你无法完全解决一个问题，请写下你的想法并简要概述你的方法——正确的思路可能会为你赢得部分分数。
- 如果你使用既不属于R基础发行版，也未在课程中介绍的R包（用于解决每周练习的方案），你需要对这些包的正确运行负责。

- 始终确保在提交PDF之前进行双重检查。

数据

考试中需要使用的数据可以从Absalon下载。文件名为`exam202425.RData`，您可以通过在RStudio中打开上述文件或使用加载函数`load("exam202425.RData")`来加载数据。

问题 0：展示 [3]

最多可以获得三分的展示得分，例如，对于格式良好的PDF文件，对于整体自洽、结构良好的报告（包含图例和注释），对于良好记录的代码，以及对于统计声明和假设中的技术精确性。

问题 1：癌症的分子分类 [10]

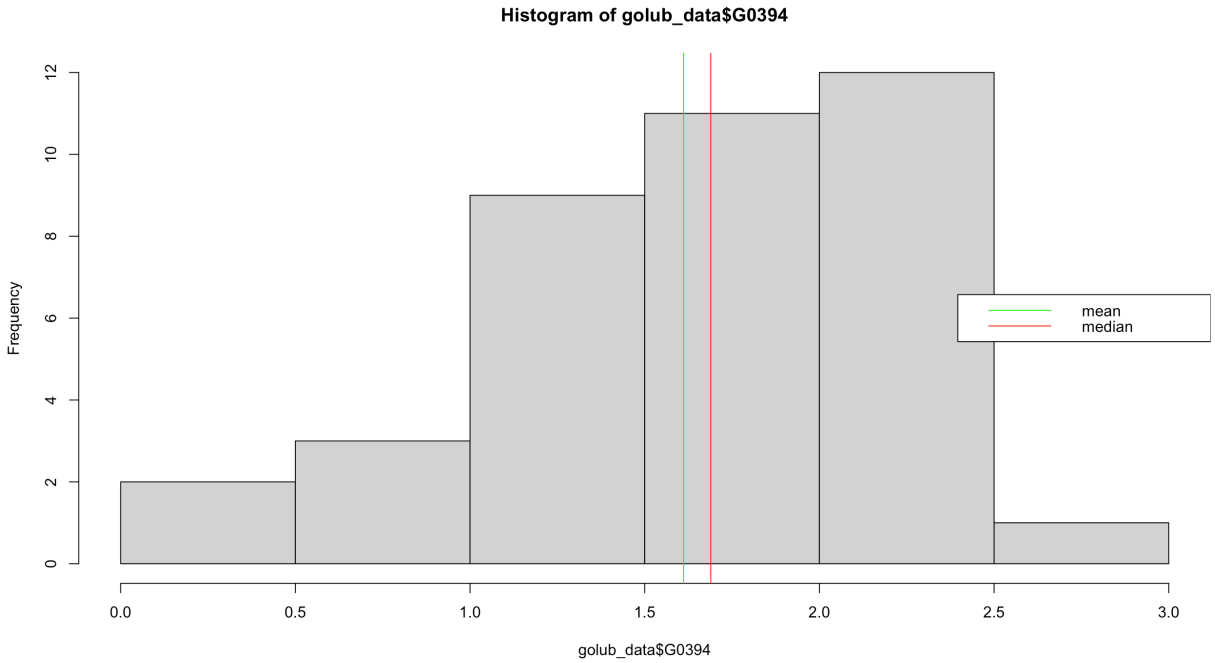
此练习的数据归功于Golub等人（1999年），他们调查了来自38个骨髓样本的基因表达水平，其中27个患有急性淋巴细胞白血病（ALL），11个患有急性髓系白血病（AML）。
此数据集记录了以下变量：

- `golub_data$cl`: 肿瘤类别，ALL 编码为 0，AML 编码为 1。
- `golub_data$G0394`: 基因394的表达水平
- `golub_data$G2037`: 基因2037的表达水平
- `golub_data$G2315`: 基因2315的表达水平
- `golub_data$G2316`: 基因2316的表达水平
- `golub_data$G2560`: 基因2560的表达水平

我们展示了数据框的头部以及下方的直方图（完整数据 `golub_data` 不需要用来回答问题，并且在考试 RData 文件中未提供）。

```
load("exam202425.RData")
# only the head is provided in the RData file
head(golub_head)
```

```
##   cl   G0394    G2037    G2315    G2316    G2560
## 1  0 1.88962  0.33593  0.16378  0.05593 -1.45769
## 2  0 1.58426 -0.46687 -0.69206 -0.38672 -1.39420
## 3  0 2.13685 -1.46227 -0.52619 -0.51714 -1.46227
## 4  0 1.72258  0.50311  0.19571  0.42125 -1.40715
## 5  0 1.43950 -1.42668  0.04565 -0.09057 -1.42668
## 6  0 2.17913  0.27125 -0.01277 -0.52405 -1.21719
```



问题 1.1 [3]

根据提供的信息，说明基因394的表达水平观测值是否超过60%位于1到2之间。通过明确引用你所依据的提供信息中的某些方面，详细阐述你的计算和论证过程来解释并证明你的答案。你是否缺少回答该问题的任何信息？请解释。

问题 1.2 [7]

现在考虑以下模型摘要：

```
load("exam202425.RData")
writeLines(golub_model_summary)
```

```
##
## Call:
## glm(formula = "cl ~ -1 + G2037 + G2560", family = binomial, data = golub_data)
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
## G2037   0.5916     0.6069   0.975 0.329640
## G2560   1.3351     0.3956   3.375 0.000739 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 52.679  on 38  degrees of freedom
## Residual deviance: 30.551  on 36  degrees of freedom
## AIC: 34.551
##
## Number of Fisher Scoring iterations: 5
```

1. 基于上述模型摘要，讨论你是否可以简化模型，并引用模型摘要中的信息来证明你的答案。对于你答案中涉及的任何统计检验，明确且正式地陈述零假设。2. 为简化起见，将系数四舍五入到1位小数（即，-2.288575将被四舍五入为-2.3）以回答此部分问题。使用上述提供的模型摘要，预测以下合成测试案例的肿瘤类别：

```
test_cases <- data.frame(G2037=c(-1, -1, 1, 1), G2560=c(-1, 1, -1, 1), predicted_class=c(NA, NA, NA, NA)
test_cases
```

```
##   G2037 G2560 predicted_class
## 1    -1    -1              NA
## 2    -1     1              NA
## 3     1    -1              NA
## 4     1     1              NA
```

讨论你是如何计算并得出预测结果的，并相应地在列 predicted_class 中填写两个肿瘤类别标签 AML 或 ALL 之一。打印包含已更新列 predicted_class 的 test_cases 数据框，该列现在包含预测的类别标签（ AML 或 ALL），而不是 NA 值。

3. 定义一个数据框 zerodf，使得在 R 中执行 predict(model, newdata = zerodf) 时，打印出以下 exact 输出

```
1
0
```

4. 是否可以根据提供的信息为系数估计提供置信区间？如果是，请提供置信区间，解释你的推理，并证明你的选择。如果不是，解释为什么提供的信息不足以做出判断。

问题 2：概率论 [20]

在这个问题中，你将处理一个样本空间 $\Omega = \{1, ..., 5000\}$ 以及定义在该样本空间之上的事件和随机变量。

在问题2.1中，你将定义概率分布P，在Ω上，你将在整个2.2 – 2.6中一直使用。

（作为备选方案，如果你尚未完成2.1以定义你的概率分布，但在后续部分需要用到，你可以考虑使用在$\Omega = \{1, ..., 5000\}$上的$P$概率分布，该分布通过对每个$\omega \in \Omega$通过`function(w)(w %% 250) / 622500`来定义，以此获得基于该分布的部分分数。）

问题 2.1：设置 [4]

将变量`mynumber`设置为您的KU学生邮箱地址中的3位数严格正整数，即，如果您的KU邮箱地址是`abc123@alumni.ku.dk` (或`abc123@ku.dk`)，则设置`mynumber <- 123`。（作为备选方案，并且只有在您没有KU邮箱地址的情况下，将`mynumber`设置为您的名字中字母数量与姓氏中字母数量的乘积，如果得到的数字不足三位，则在末尾添加7，即Quinn Sky将设置`mynumber <- 157`。）

为了定义在整个练习中将要使用的Ω上的概率分布P，我们将为每个基本事件$\omega \in \Omega$定义$P(\omega)$。为此，我们提供了作为RData文件)一部分的函数`setup_omega` (，该函数将生成一个数据框，指定样本空间$\{1, ..., 5000\}$上的概率分布。

示例说明 `setup_omega` 返回的数据框如何在 Ω 上定义概率分布。您的数字和概率分布将有所不同，这只是为了说明数据框。

函数 `setup_omega` 返回一个数据框，包含列 “ element ”、“ nnp ”（表示未归一化概率）、“ X ” 和 “ Z ” 。您为 `mynumber` 获得的表格可能看起来像以下这样（但数字不同）：

```
load("exam202425.RData")
dummy <- setup_omega(594)
head(dummy)
```

```
##   element nnp  X  Z
## 1       1  62 16 25
## 2       2 124 14 22
## 3       3 350  8 14
## 4       4 412 24 11
## 5       5 171  9  3
## 6       6 233  7  0
```

对于列 “ 元素 ” 中所述的每个元素$\omega \in \Omega = \{1, ..., 5000\}$，数据框指定了列 “ nnp ” 中所述的非标准化概率，即，在上面的例子中，事件$\{1\}$的概率将与62成比例，并且事件$\{2\}$的概率除以事件$\{1\}$的概率将等于$\frac{124}{62} = 2$。

要获取您在Ω上的概率分布P的定义，请调用上述数字`setup_omega(mynumber)`的函数`setup_omega`，并添加一列“ p ”，其中包含归一化概率$P(\omega)$。

计算 $P(A)$、$P(B)$、$P(A \cup B)$、$P(A, B)$ 和 $P(A \mid B)$ 以及 $A = \{1, 10, 100, 1000\}$ 和 $B = \{2, 20, 200, 2000\}$ 对于通过 `setup_omega(mynumber)$nnp` 定义的你的概率分布 P。
以分数形式准确表示结果，例如使用 $\frac{1}{3}$ 而不是小数 0.3333333 (。小数通常不是精确的，而是经过四舍五入或截断的，因此在中间计算过程中应避免使用小数，例如，$\frac{5}{43} \cdot \frac{(2+3)}{43} = \frac{25}{43^2} = \frac{25}{1849}$ 是精确的，而 `(5/43)*((2+3)/43)` 会得出 `0.01352082`，这与精确答案) 接近但不相等。

提示：我们提供了“ nnp ”列。显式处理归一化常数，而不是使用归一化概率。

问题 2.2：有效的概率分布 [2]

列出概率分布必须满足的4个性质。对于每个性质，提供一个数值示例，说明你在2.1中定义的概率分布P满足该性质，即定义一些事件并计算概率，以说明有效概率分布的性质。

问题 2.3：独立事件 [6]

定义两个事件 $D \subset \Omega$ 和 $E \subset \Omega$，在你在2.1中定义的概率分布 P 下它们是独立的。你必须选择 D 和 E，使得这两个集合都不是空集或不等于 Ω，也就是说，D 和 E 必须至少包含一个元素，但不能包含所有元素。

如果你得出结论认为此类事件不存在，请提供严谨的论证和演示，说明为什么此类事件不存在。

如果你得出存在此类事件的结论，请解释你是如何发现这些事件的推理过程：你是如何找到这两个事件的，搜索过程中考虑了哪些因素，你是否使用了某些R代码来帮助筛选事件，……？通过比较精确的计算结果（使用分数，即例如使用$\frac{1}{3}$而不是小数0.3333333，因为小数通常不是精确的，而是四舍五入或截断的），证明你提出的D和E确实是独立的。

提示：我们提供了“ nnp ”列。选择事件时，可能有助于确保$D \cap E$仅包含一个元素，然后计算出D或E中应包含哪些额外元素。使用“ p ”列中的归一化概率可能会由于数值问题而导致某些概率相等的错误答案。

问题 2.4：贝叶斯定理 [2]

用你选择的两个相依事件说明贝叶斯定理，使用你在2.1中定义在$\Omega = \{1, ..., 5000\}$上的概率分布$P$。证明你提出的$D$和$E$确实是独立的。

问题 2.5：随机变量 [4]

“ X ”列显示了一个随机变量的值，即$X : \Omega \to \mathbb{R}$是为每个元素$\omega \in \Omega = \{1, ..., 5000\}$定义的函数，并且“ X ”列显示的值$X(\omega)$ (同理适用于“ Z ”列和随机变量Z)。

使用在2.1中定义在$\Omega = \{1, ..., 5000\}$上的概率分布$P$实现离散随机变量$X$的概率质量函数。

使用您的R函数实现X的概率质量函数，计算$P(X \in \{0, 10, 20, 30, 40, 50\})$、$\mathbb{E}(X)$以及$X$的方差。

考虑以下 20 个独立同分布的随机变量 X 的观测值，如下 R 向量所示

```
c(3, 4, 8, 9, 4, 6, 5, 3, 4, 3, 5, 5, 3, 3, 8, 1, 4, 4, 6, 5)
```

```
##  [1] 3 4 8 9 4 6 5 3 4 3 5 5 3 3 8 1 4 4 6 5
```

并计算数据的对数似然。

问题 2.6：定义一个随机变量 [2]

在 Ω 上定义另一个随机变量 Y，使得该随机变量服从伯努利分布，并指出该伯努利分布的确切参数 $p \in (0, 1)$（使用分数，即例如用 $\frac{1}{3}$ 而不是小数 0.3333333，因为小数通常不是精确的，而是四舍五入或截断的）。
即，说明定义 Y 的函数以及确切的 p，使得 $Y \sim$ 满足伯努利 (p)。

提示：Y的分布特定于您在2.1中定义的Ω上的概率分布P。

问题 3：Q-Q 和拟合优度 [7]

设 R 为一个在 $\{1, ..., 7\}$ 中取值的随机变量，其概率质量函数为

$$r(x) = \begin{cases} a & \text{if } x = 1, \\ b & \text{if } x = 2, \\ c & \text{if } x = 3, \\ d & \text{if } x = 4, \\ e & \text{if } x = 5, \\ f & \text{if } x = 6, \\ g & \text{if } x = 7, \end{cases}$$

对于某些满足 $a + b + c + d + e + f + g = 1$ 的 $a, b, c, d, e, f, g \in (0, 1)$，也就是说，所有参数都严格大于 0 且严格小于 1，并且它们的总和为 1。

考虑以下 60 个独立同分布的随机变量 F 的观测值，由下面的 R 向量给出

```
c(5, 3, 6, 7, 5, 6, 5, 5, 4, 3, 6, 7, 5, 5, 6, 5, 5, 5, 1, 7, 3, 5, 7, 7, 7, 6, 7, 2, 5, 5, 2, 6, 5, 2,
```

```
##  [1] 5 3 6 7 5 6 5 5 4 3 6 7 5 5 6 5 5 5 1 7 3 5 7 7 7 6 7 2 5 5 2 6 5 2 5 6 5 6
## [39] 6 3 6 5 6 6 6 5 6 2 5 5 7 5 7 5 3 4 1 2 5 5
```

1. 对数据绘制与二项分布的Q-Q图，其中成功概率参数设置为$\frac{9}{14}$，试验次数设置为7。讨论并解释Q-Q图。

2. 对参数 a, b, c, d, e, f, g 的可能值进行有根据的猜测。（在这里，你可以为这些参数推导出具有某些特性的点估计量，但也可以简单地猜测参数。）

3. 提供一个统计上合格的答案，以判断上述具有概率质量函数 r 和你的参数猜测是否很好地描述了观测数据。基于适当的统计程序回答。证明你选择的程序的合理性，并评论其假设和结果。

4. 获取 R 期望值估计的 95% 自助置信区间。