Biostatistics (BIOL0031132104) - Assignment #3
(released on Nov. 24th, due on Dec. 8th)

1. 卡特里娜飓风和丽塔飓风造成美国新奥尔良州大面积被洪水淹没，新形成了大量的沉积物。已知在飓风来临之前，新奥尔良的土壤中已存在高浓度的铅，这一危险的环境毒素。在飓风来临前，研究人员对 46 个地点的土壤铅含量进行了监测（毫克/千克），然后在飓风过后对每个地点的土壤再次进行了测量（Zahran 等人，2010）。研究人员计算了"土壤铅含量变化"（LogRatio），即飓风后的土壤铅含量与飓风前的土壤铅含量的比值的对数值（见数据文件“Q1_SoilLeadAndHurricanes.csv”）：其中小于零的数字表示飓风后土壤铅含量减少，大于零表示土壤铅含量增加，这个对数值近似呈正态分布。部分表格内容及每列说明见下。[20 marks]

Site	Soil_lead_Before_Katrina	Soil_lead_After_Katrina	Change	Ratio	LogRatio
1	711	310	-401	0.4360056259	-0.83
10	37	31	-6	0.8378378378	-0.18
11	46	53	7	1.152173913	0.14
12	90	21	-69	0.2333333333	-1.46
13	1049	648	-401	0.6177311725	-0.48
14	247	87	-160	0.3522267206	-1.04

Site: 地点编号
Soil_lead_Before_Katrina: 飓风前的土壤铅含量;
Soil_lead_After_Katrina: 飓风后的土壤铅含量;
Change 飓风后的土壤铅含量与飓风前的土壤铅含量的变化（原始值）;
Ratio: 飓风后的土壤铅含量与飓风前的土壤铅含量的比值;
LogRatio: 飓风后的土壤铅含量与飓风前的土壤铅含量的比值的对数值;

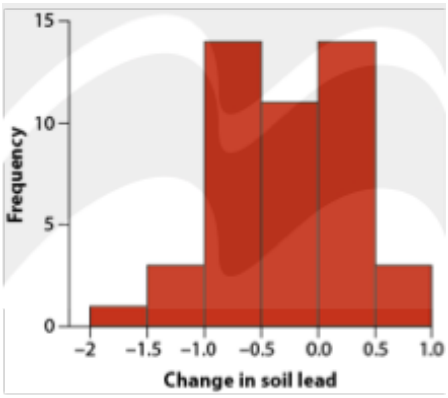
a. 选择合适的图类型并遵从制图原则来绘制土壤铅含量的变化（即主要针对表中的比值对数值“LogRatio”这一列的数据），并说明 46 个地点的土壤铅含量变化有什么特征，大部分地点是增加还是减少？[5 marks]

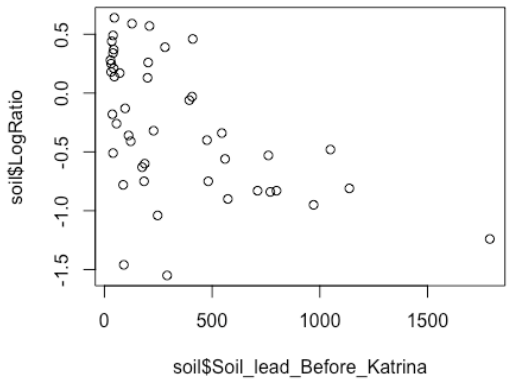
b. 确定土壤铅含量（比值对数值）平均变化的 95% 置信区间（计算公式参考：$\bar{Y} - t_{0.05(2),\ df}\ \mathrm{SE}_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2),\ df}\ \mathrm{SE}_{\bar{Y}}$，其中$t_{0.05(2),\ df=45} = 2.01$，或通过 R 语言 t.test()进行估计）。用语言描述该变化的性质：数据与“土壤铅含量增加”一致吗？还是与“土壤中铅含量降低”更一致？[5 marks]

c. 通过假设检验的 4 个步骤来检验飓风过后土壤铅的平均值是否发生了变化（提示：使用 R 语言的t检验来得出结果并给出结论）。[10 marks]

【参考答案】

(a) 答案可能各不相同（直方图或散点图等）[4 marks]。直方图如下左图。结果为负变化（表示减少）比正变化（增加）的地点更多[1 marks]；散点图如下右图，大多值落在 y 轴的 0 以下。





(b) 平均值的 95% 置信区间为 -0.274 ± 0.0851 $t_{0.05(2),\ df=45}$，其中 $t_{0.05(2),\ df=45}$ = 2.01，得出 -0.445 < μ < -0.103 [4 marks]。最合理的范围是负数，即负值，即平均土壤铅含量下降了约 0.10 至 0.45，数据与“土壤铅含量降低”一致；数据与铅升高不符。[1 marks]

(c) H_0：飓风过后，平均土壤铅含量没有变化（$\mu = 0$）。H_A：飓风过后土壤铅平均值发生了变化（$\mu \neq 0$）。[4 marks]

$\bar{Y}$ = -0.274，t=-3.22，df=45，$t_{0.05(2),\ df=45}$ =2.01。由于 t < -2.01，因此拒绝 H_0（P = 0.0024）。或 R 语言中 t.test(soil$LogRatio, mu = 0)直接得出 P = 0.002361。[4 marks]

结论：飓风过后，土壤中铅的平均浓度有所下降。[2 marks]

2. 下表中列出了 4 个不同正态分布的均值（mean）和标准差（SD）。对于每个分布（即每一行），计算单个值 *Y* 大于给定阈值（threshold）的概率，以及小于该阈值的值的概率（提示：可使用 R 语言中的“pnorm()”命令），保留小数点后 4 位即可。同时，请计算对该分布进行标准正态分布转换后的 *Z* 值，保留小数点后 2 位即可。[20 marks]

Mean	SD	threshold	Pr[*Y* > *threshold*]	Pr[*Y* < *threshold*]	*Z*
14	5	9			
15	3	18.5			
-23	4	-16			
14000	5000	9000			

【参考答案】

Mean	SD	threshold	Pr[*Y* > *threshold*]	Pr[*Y* < *threshold*]	*Z*
14	5	9	0.8413	0.1587	-1.00
15	3	18.5	0.1217	0.8783	1.17
-23	4	-16	0.0401	0.9599	1.75

14000	5000	9000	0.8413	0.1587	-1.00

（Pr[]两列的 8 个数值各 2 marks，*Z* 列 4 个数值各 1 mark，共计 20 分）

3. 螃蟹蜘蛛（*Thomisus spectabilis*）常栖息在花朵上以捕食前来访花的蜜蜂，正如 Chapter10 开头的照片所示（见下）。为了测试蜜蜂是否能区分有螃蟹蜘蛛和没有螃蟹蜘蛛的花朵，科研人员进行了试验，分别让 33 只蜜蜂在两朵花之间做出选择：一朵有螃蟹蜘蛛，另一朵没有。在这 33 次试验中，有 24 次试验中的蜜蜂选择了有螃蟹蜘蛛的花朵。在剩下的 9 次试验中，蜜蜂选择了没有螃蟹蜘蛛的花朵。基于这些数据，假定“蜜蜂选择有螃蟹蜘蛛的花朵”为“成功事件”，请进行适当的假设检验，采用二项检验计算 P 值，计算成功事件的比例及其置信区间（使用 Agresti-Coull 方法），并针对蜜蜂是否能区分两种花朵给出结论。同时请给出 R 代码。[20 marks]

© Ed Nieuwenhuys

【参考答案】

H_0: 蜜蜂对有无螃蟹蜘蛛的花无偏好（$p_1=p_2=0.5$）。[2 marks]

H_A: 蜜蜂对有无螃蟹蜘蛛的花有偏好（$p_1 \neq p_2$）。[2 marks]

蜜蜂选择有螃蟹蜘蛛的花的比例为 $p_1 = 24/33 = 0.727$ [2 marks]，$0.556 < p < 0.851$；[2 marks]

二项检验计算的 p-value = 0.0135 [2 marks]，因此拒绝 H_0 [2 marks]；

结论：蜜蜂更加偏好有螃蟹蜘蛛的花朵。[2 marks]

R 代码:

二项检验：binom.test(x = 24, n = 33, p = 0.5) [3 marks]

置信区间：binom.confint(x = 24, n = 33, method = "ac") [3 marks]

4. 库鲁病是新几内亚高原森林人患有的一种朊病毒疾病。过去，这种疾病曾通过食用已故亲属的尸体而传播，但这一仪式在大约 1960 年左右被终止。研究人员使用存档的组织样本研究了可能的抵抗库鲁病的基因变异。下表中的数据是所有患有该疾病的年轻和老年个体的朊蛋白基因密码子 129 的基因型。由于老年人长期暴露在库鲁病毒中，该群体中异常常见的基因型可能预示着具有抵抗力的基因型。[28 marks]

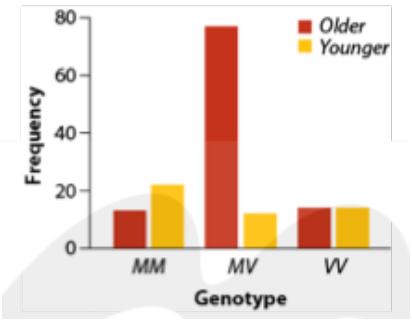
（observed frequency）	Genotypes at codon 129		
Age	MM	MV	VV
Elderly （老年人）	13	77	14
Young（年轻人）	22	12	14

a. 用分组条形图（grouped barplot）展示表中的数据。
b. 与年轻人相比，哪些基因型在老年人中特别普遍？
c. 针对年龄与基因型频率的相关性进行假设检验，计算期望频数（保留小数点后两位即可），并根据卡方检验给出“两个年龄组的基因型频率是否相同”的结论。同时请给出 R 代码。

（expected frequency）	Genotypes at codon 129		
Age	MM	MV	VV
Elderly （老年人）			
Young（年轻人）			

【参考答案】

(a) [2 marks]（如果以年龄组分组来看各年龄组里基因型的频率，酌情给 1 分）



(b) 与年轻样本相比，MV 基因型在老年样本中更为常见，也许这种基因型对库鲁病毒有很强的抵抗力。[2 marks]

(c) 假设检验及结论:

H_0: 年轻人和老年人的基因型比例没有差异。[2 marks]

H_A: 年轻人和老年人的基因型比例不同。[2 marks]

预期频数：（不要求给出总计这一行/列）[6 marks]

（expected frequency）	Genotypes at codon 129			
Age	MM	MV	VV	总计
Elderly（老年人）	23.95	60.89	19.16	104
Young（年轻人）	11.05	28.11	8.84	48
总计	35	89	28	152

χ^2 = 33.73 [2 marks]，df = 2，p-value = 4.73e-08 < 0.05 [2 marks]；
因此拒绝 H_0，说明年轻人和老年人的基因型比例不同[2 marks]。
（也可以根据临界值 $\chi^2_{0.05,df=2}$ = 5.99 来判断。）

R 代码： [8 marks]

```
# 生成 RxC 列联表的频数数据（这一步可能有各种形式）
genotype = data.frame(MM = c(13,22), MV = c(77,12), VV= c(14,14),
                      row.names = c("old","young"))
round(chisq.test(genotype)$expected,2) # 计算期望频数
chisq.test(genotype,correct = F) # 卡方检验
# X-squared = 33.733, df = 2, p-value = 4.73e-08 # 结果
```

5. 针对因外伤可能导致大量失血的患者，重要的治疗方法是在他们被送入急诊室后进行血浆输血。科研工作者研究了急救人员在此类患者到达医院前的急救过程中提早使用这种治疗是否有效。在一项随机对照试验中，外伤患者在空中医疗转运过程中接受或不接受血浆输血（转运之后的医疗中患者均接受标准护理）。在提早接受血浆治疗的 220 名患者中，有 53 人在受伤后 30 天内死亡；在只接受标准护理的 261 名患者中，有 89 人在 30 天内死亡。
请计算以下问题（小数点后保留两位），同时请给出 R 代码。[12 marks]

a. 与只接受标准护理（未提早接受血浆治疗）的患者相比，在转运过程中接受了血浆治疗的外伤患者的相对死亡风险是多少？

b. 在转运过程中接受血浆治疗后，相对死亡风险降低了多少？（提示：1-*RR*）

c. 这两组患者的死亡几率比例（比值比，the odds ratio of death），及其 95% 置信区间是多少？

【参考答案】

(a) RR = (53/220)/(89/261) = 0.71； [2 marks]
(b) 1-RR = 0.29； [1 marks]
(c) OR = 53*(261-89)/(89*(220-53)) = 0.61, 或 fisher.test 计算（见代码）； [2 marks]
95%CI: 0.41 < OR < 0.92。 [2 marks]

R 代码计算置信区间

```
death_table = data.frame(treat = c(53,220-53), control = c(89,261-89),
                         row.names = c("dead","alive")) [ 1 marks]
fisher.test(death_table)$estimate # odds ratio;   [ 2 marks]
fisher.test(death_table)$conf.int # 95%CI [ 2 marks]
```