Biostatistics (BIOL0031132104) - Assignment #1
(released on Oct. 18th, due on Oct 27th)

1. Hagen et al. (2011) estimated the home range sizes of four bumblebees (*Bombus*) by fitting them with tiny radio transmitters and tracking their positions by plane and ground surveys. They estimated the mean home range size to be 20.7±11.6 ha, where the number after the ± sign refers to standard error of the mean. [20 marks]

   Hagen 等人 (2011 年) 在四种熊蜂上安装了微型无线电发射器，并通过飞机和地面调查跟踪其位置，以此估算了它们日常活动领地的范围大小(home range size)。他们估计出平均领地大小为 20.7 ± 11.6 公顷（± 符号后面的数字为平均值的标准误）。请回答以下问题：[20 分]

   a. 请解释标准误，并回答标准误衡量的是什么特征？(Provide a description for the standard error. What does it measure?) [10 marks]

      The SE is the estimated standard deviation of the sampling distribution of the sample mean. It measures the uncertainty of the estimate of the mean.
      标准误是样本均值（或其它统计量）的抽样分布的标准差，它衡量了均值（或其它统计量）估计值的不确定性。

   b. 在计算均值的标准误时，我们使用了什么假设？(What assumption are we making when calculating the standard error of the mean?) [5 marks]

      Random sampling. 假设样本由随机抽样获得。

   c. 为了减小估算平均领地大小的标准误，你对研究人员后续操作有什么建议？(What would you recommend the researchers do next to reduce the standard error of their estimate of the mean home range size?) [5 marks]

      Increase their sample size. 增大样本大小。

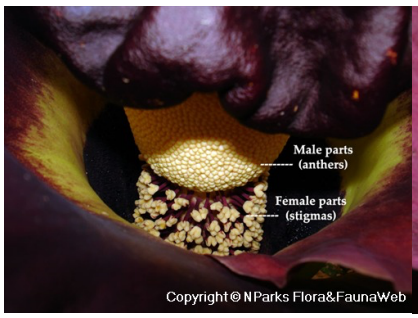
© David Pattemore @ Plant and Food Research

2. *Amorphophallus johnsonii* is a plant growing in West Africa, and it is better known as a "corpse flower." Its common name comes from the fact that when it flowers, it gives off a "powerful aroma of rotting fish and faeces" (Beath 1996). The flowers smell this way because their principal pollinators are carrion beetles, who are attracted to such a smell. Beath (1996) observed the number of carrion beetles (*Phaeochrous amplus*) that arrive per night to flowers of this species. The data are as follows:

Amorphophallus johnsonii（巨花魔芋，天南星科魔芋属）是一种生长在西非的植物，被称为"尸花"。它的俗名源于它开花时散发的一种"强烈的腐肉气味" (Beath 1996)。这种花之所以有这种气味，是因为它们主要的传粉者是会被这种气味所吸引的腐肉甲虫。Beath 观察了每晚到达巨花魔芋不同植株上花朵（访花）的腐肉甲虫的数量。数据如下：

51	45	61	76	11	117	7	132	52	149

© US Botanical Garden



(Male & Female parts)

(Carrion beetles on other *Arisaema sp.*)

请回答下面问题（每题 5 分共计 30 分）并给出 R 代码（20 分）。数值保留小数点后一位即可。（提示：R 可读取数据文件 'chap04q18Corpseflowers.csv'，或手动读取矢量数据，data = c(51, 45, 61, 76, 11, 117, 7, 132, 52, 149)）

a. 计算平均到达每朵花的甲虫数量的均值和标准差。(What is the mean and standard deviation of beetles per flower?)
Mean: 70.1; Standard deviation: 48.5.

b. 计算上述均值估计的标准误。(What is the standard error of this estimate of the mean?)
SE = 15.3

c. 计算均值的 95% 置信区间（含下限和上限）。(Give an approximate 95% confidence interval of the mean, with lower and upper limits.)
通过应用 2SE rule: $39.4 \le \mu \le 100.8$.
（有可能通过 t.test(data)$conf.int 获得，就是 $39.4 \le \mu \le 100.8$）

d. 如果给你 25 个数据点而不是题目中的 10 个数据点，你会预期新计算的访花甲虫数量的均值将大于、小于还是约等于现有样本的均值？(If you had been given 25 data points instead of 10, would you expect the mean to be greater than, less than, or about the same as the mean of this sample?)
About the same. 约等于。

e. 如果给你 25 个数据点而不是题目中的 10 个数据点，你会预期新计算的标准差将大于、小于还是约等于现有样本的标准差？(If you had been given 25 data points instead of 10, would you have expected the standard deviation to be greater than, less than, or about the same as this sample?)
About the same. 约等于。

f. 如果给你 25 个数据点而不是题目中的 10 个数据点，你会预期新计算的均值的标准误将大于、小于还是约等于现有样本的标准误？(If you had been given 25 data points instead of 10, would you have expected the standard error of the mean to be greater than, less than, or about the same as this sample?)
Less than this sample. 小于现有样本的标准误。

代码:

```
dt = read.csv("path to file")
dt_mean = mean(dt$numberOfBeetles) # 计算均值
dt_sd = sd(dt$numberOfBeetles)     # 计算标准差
dt_se = dt_sd/sqrt(nrow(dt))       # 计算标准误
CI_upper = dt_mean + 2*dt_se       # 计算置信区间上限
CI_lower = dt_mean - 2*dt_se       # 计算置信区间下限
```

3. The human genome is composed of the four DNA nucleotides: A, T, G, and C. Some regions of the human genome are extremely G–C rich (i.e., a high proportion of the DNA nucleotides there are guanine and cytosine). Other regions are relatively A–T rich (i.e., a high proportion of the DNA nucleotides there are adenine and thymine). Imagine that you want to compare nucleotide sequences from two regions of the genome. Sixty percent of the nucleotides in the first region are G–C (30% each of guanine and cytosine) and 40% are A–T (20% each of adenine and thymine). The second region has 25% of each of the four nucleotides. [30 marks]

人类基因组由四种脱氧核糖核苷酸组成：A、T、G 、C。基因组的某些区域极其富含 G-C（即鸟嘌呤和胞嘧啶所占比例较高），而其他区域则相对富含 A-T（即腺嘌呤和胸腺嘧啶所占比例较高）。设想比较基因组中两个区域的核苷酸序列：第一个区域 G-C 比例为 60%（G 和 C 各占 30%），40% 是 A-T（A 和 T 各占 20%）；第二个区域的四种核苷酸（A、T、G 、C）各占 25%。

请回答下列问题（给出解题思路及计算结果）：[30 分]

a. 如果从上述两个区域中各随机抽取一个核苷酸，它们是相同核苷酸的概率是多少？(If you choose a single nucleotide at random from each of the two regions, what is the probability that they are the same nucleotide?) [15 marks]

如果第一个区域抽取任何一个核苷酸，那么从第二个区域抽取相同核苷酸的概率为 25%，因为所有核苷酸在第二个区域的概率相同。因此，从每个区域随机抽取一个核苷酸的匹配概率为 0.25。

或者，计算针对每个核苷酸的概率：30%*25% + 30%*25% + 20%*25% + 20%*25% = 25% = 0.25。

b. 假设各个核苷酸在 DNA 单链上独立出现，那么从两个区域中各随机抽取一段连续三个核苷酸的短序列（3nt）。这两个三核苷酸序列相同的概率是多少？(Assume that nucleotides over a single strand of DNA occur independently within regions and that you randomly sample a three-nucleotide sequence from each of the two regions. What is the chance that these two triplets are the same?) [15 marks]
（提示：可基于 a 问题的回答来思考这个问题。）

即 a 中概率 0.25 的三次独立事件，$1/4^3 = 0.015625$。

Reference

1) Beath, D. D. 1996. Pollination of *Amorphophallus johnsonii* (Araceae) by carrion beetles (*Phaeochrous amplus*) in a Ghanaian rain forest. *Journal of Ecological Ecology* 12: 409–418.
2) Hagen, M., M. Wikelski, and W. D. Kissling. 2011. Space use of bumblebees (*Bombus* spp.) revealed by radio-tracking. *PLoS ONE* 6: e19997.