

Lecture 5 – Probability 概率

- Outline for today
 - Recall L04
 - Probability
 - Random trial
 - The event of interest
 - Summary
 - R Lab & Discussion

1. Recall from L04

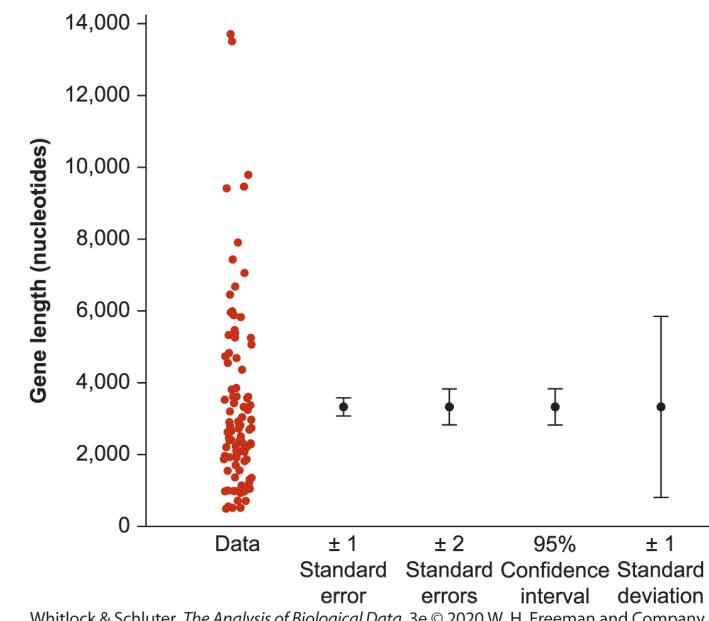
- Uncertainty of estimates: standard error & confidence interval

- 估计值都有一个抽样分布(多次抽样的估计值的分布)
- 估计值的标准误是其抽样分布的标准差

$$\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{Y}} \approx \text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

- 置信区间是可能包含目标参数数值的一个范围
 - 95% CI 的宽度可以很好地衡量我们对参数真实值的不确定性

- 通常在图中添加误差线(error bars)以说明标准误或置信区间



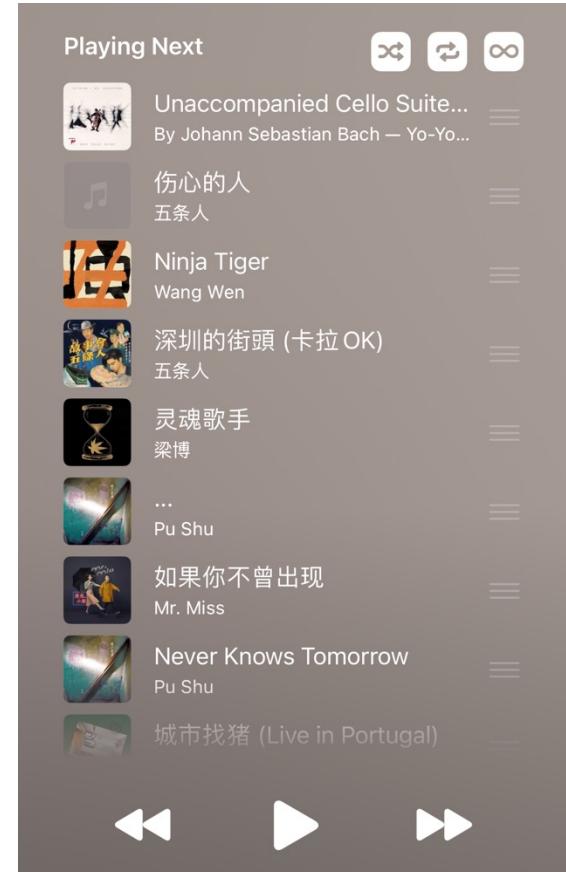
2. Probability 概率

- The value of an estimate calculated from data is almost never exactly the same as the value of the population parameter being estimated
 - because **sampling is influenced by chance.**
 - 因为抽样会受到偶然性的影响
- The crucial question is “In the face of chance, how much can we trust an estimate?”
 - 我们能在多大程度上相信一个估计值？
 - In other words, what is its **precision?**



2. Probability 概率

- Shuffle 1000 songs 随机播放 1000首歌 (有一首最爱)
 - The probability that the first song is your favorite song?
- Flip a coin 抛一次硬币 (一面图案一面数字)
 - The probability that you get the number?
- Roll a pair of dice 掷骰 (tóu) 子
 - The sum of their numbers?



2.1 The probability of an event 一次事件的概率

- A random trial 随机试验

- 随机试验是一种过程或实验它有两种或两种以上可能的结果，而这些结果的出现是无法准确预测的。
- 随机试验的每次重复只能观察到一种结果

- A random event 随机事件

- 随机试验的所有可能结果列表
 - roll a six-side die
 - roll a pair of six-side dice



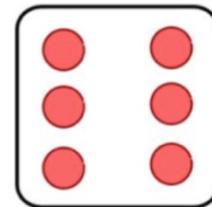
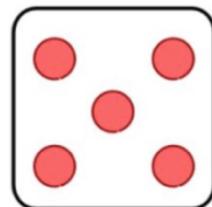
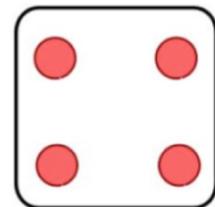
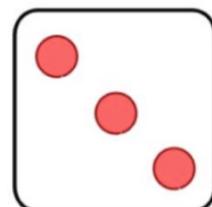
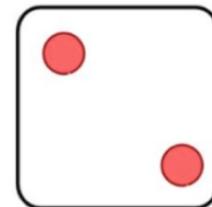
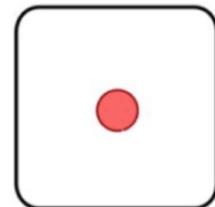
掷骰子 roll dice

2.1 The probability of an event 一次事件的概率



- A random event 随机事件

- 需要定义感兴趣的事件 (the event of interest)
- 事件是一次随机试验所有可能结果的任何潜在子集 (all possible outcomes)



- 数字是5
- 数字是偶数
 - 包括 ?
- 数字比3大
 - 包括 ?
- 其它事件 ?

掷骰子
roll a six-side die

2.1 The probability of an event 一次事件的概率

- A random event 随机事件
 - 需要定义感兴趣的事件 (the event of interest)
- The probability 概率
 - 我们一般是基于事件来定义概率
 - 一次事件的概率是指如果我们在相同的条件下反复进行随机试验，该事件发生次数的比例 (the probability of an event is the proportion of times the event would occur if we repeated a random trial over and over again under the same conditions.)
 - 概率介于 0 和 1 之间

2.1 The probability of an event 一次事件的概率



- The probability 概率

- 一次事件的概率是指如果我们在相同的条件下反复进行随机试验，该事件发生次数的比例。

- $\text{Pr}[A]$ means “the probability of event A.”

- $\text{Pr}[\text{rolling a four}] = 1/6$
- $\text{Pr}[\text{rolling an even number}] = ?$
- $\text{Pr}[\text{rolling a number} > 3] = ?$

- 数字是5
- 数字是偶数
 - 包括？
- 数字比3大
 - 包括？
- 其它？

2.1 The probability of an event 一次事件的概率

- The probability 概率
 - 一次事件的概率是指如果我们在相同的条件下反复进行随机试验，该事件发生次数的比例。
 - $\text{Pr}[A]$ means “the probability of event A.”
 - 概率介于 0 和 1 之间 [0, 1]
 - 如果一个事件从不会发生: $\text{Pr}[A] = 0$
 - 如果一个事件总是发生: $\text{Pr}[A] = 1$

2.1 The probability of an event 一次事件的概率

- The probability 概率

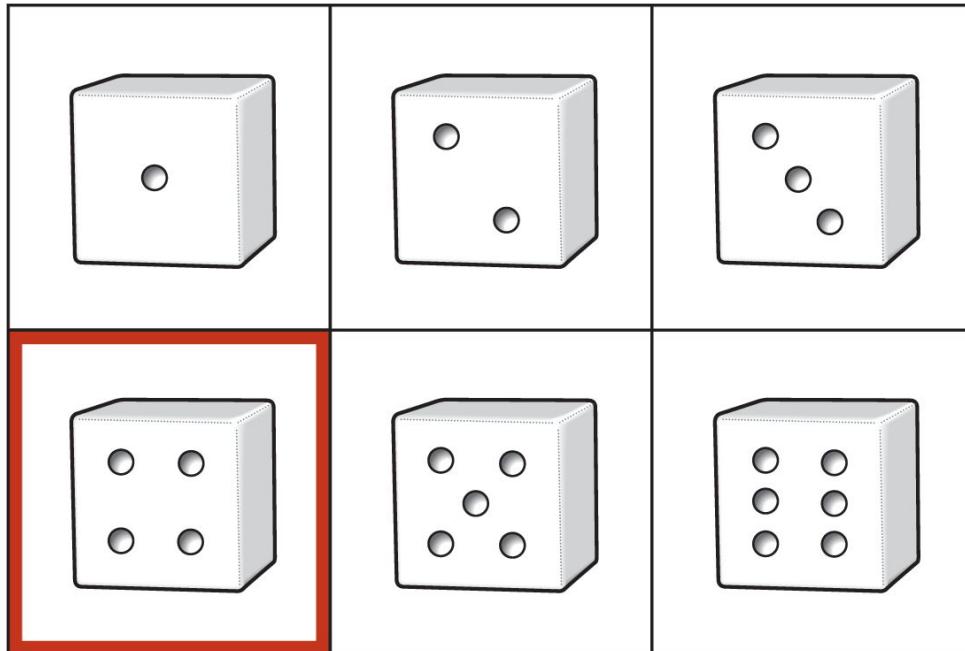
- 一次事件的概率是指如果我们在相同的条件下反复进行随机试验，该事件发生次数的比例。 $\text{Pr}[A]$
- 概率介于 0 和 1 之间 $[0, 1]$
- 抛硬币和掷骰子不是生物过程，但它们与生物学的相关性很高，因为它们模仿了取样过程 (the process of sampling)。
 - 随机抽样 100 个新生儿并计算其中男女的数量，就好比掷 100 次硬币并计算其中数字的数量：2 possible outcomes



2.2 Venn diagrams 维恩图

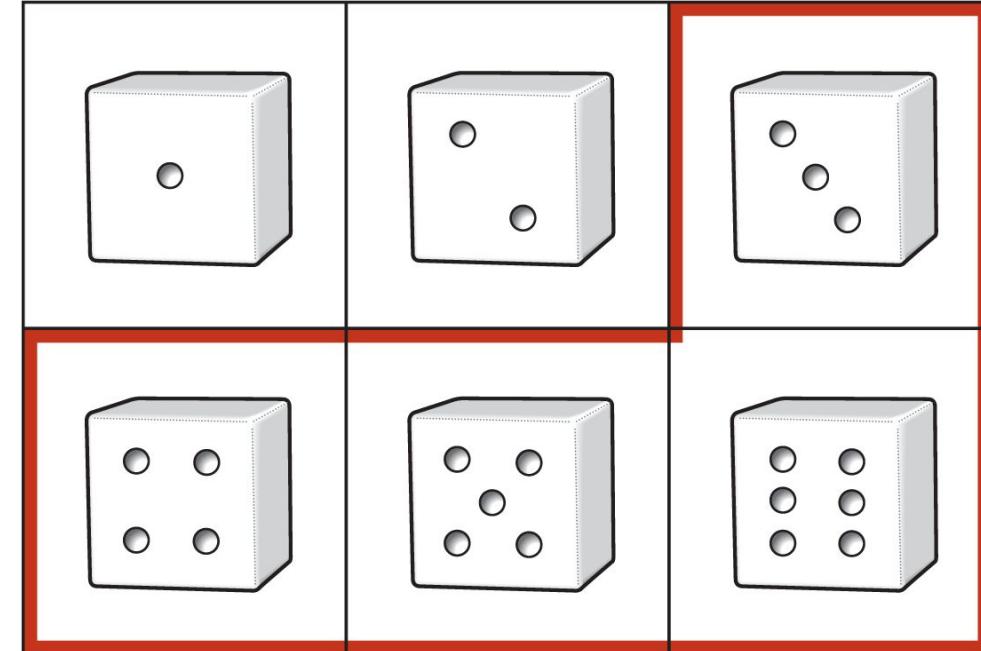


- 用集合的方式展示可能事件的概率：总面积 = 1
 - 部分面积 = 某一事件概率



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

$$\Pr[\text{rolling a four}] = ?$$

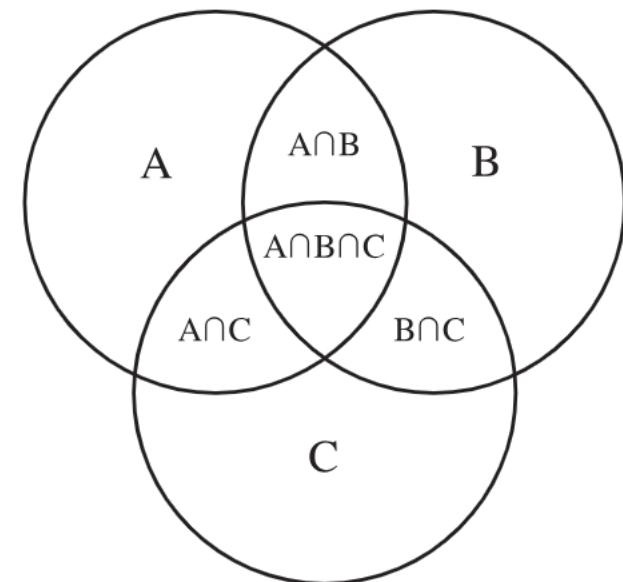
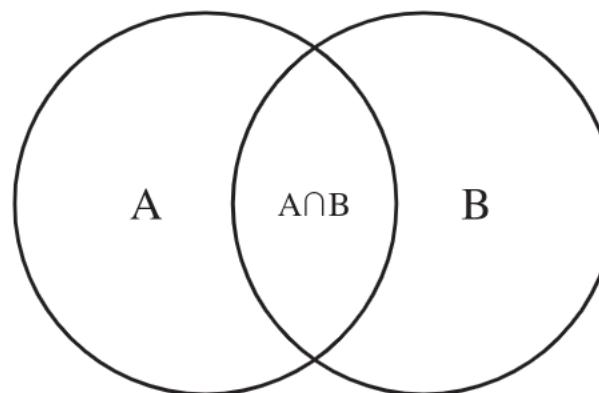
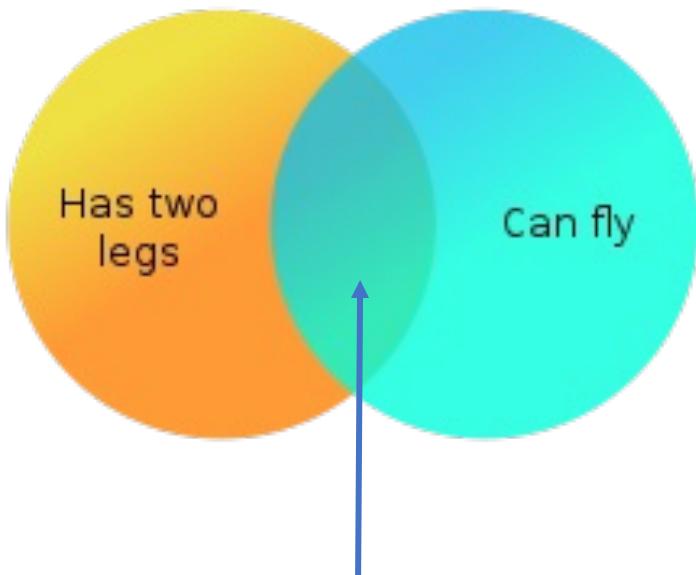


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

$$\Pr[\text{rolling a number} > 2] = ?$$

2.2 Venn diagrams 维恩图

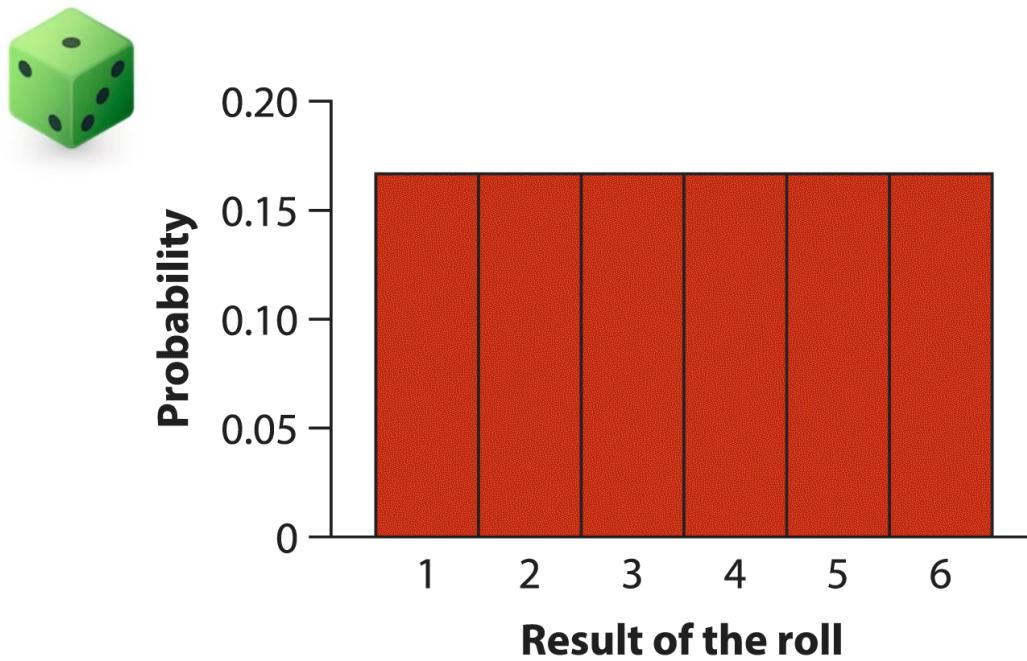
- 用集合的方式展示可能事件的概率：总面积 = 1
 - 部分面积 = 某一事件概率



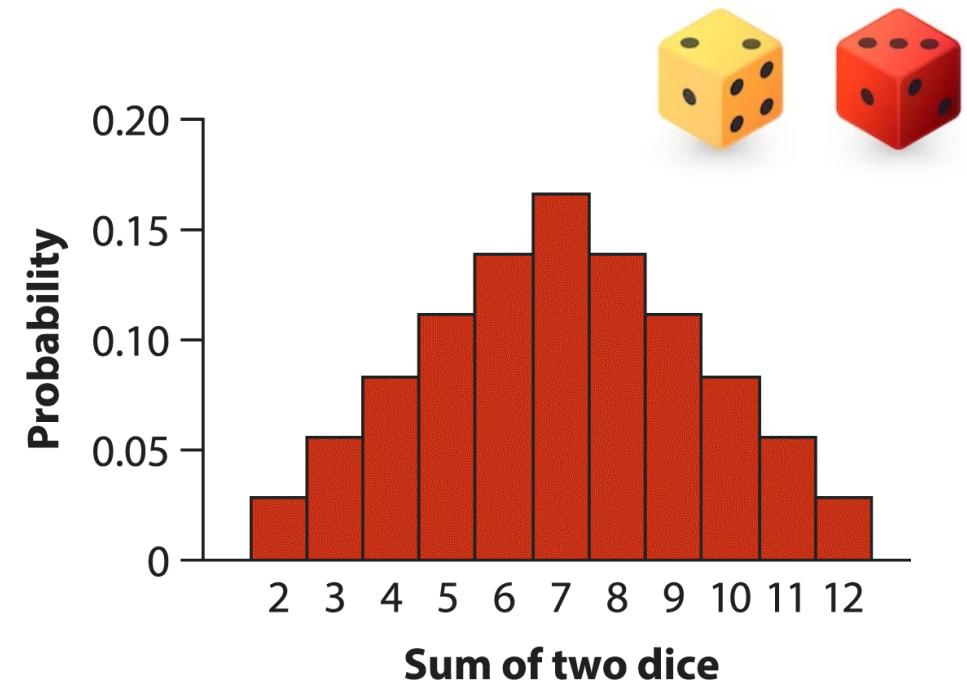
$$\Pr[2 \text{ legs} \& \text{ can fly}] = ?$$

2.3 Probability distribution 概率分布

- Discrete probability distributions 离散概率分布
 - For categorical and discrete numerical variables (分类变量/离散数值变量)



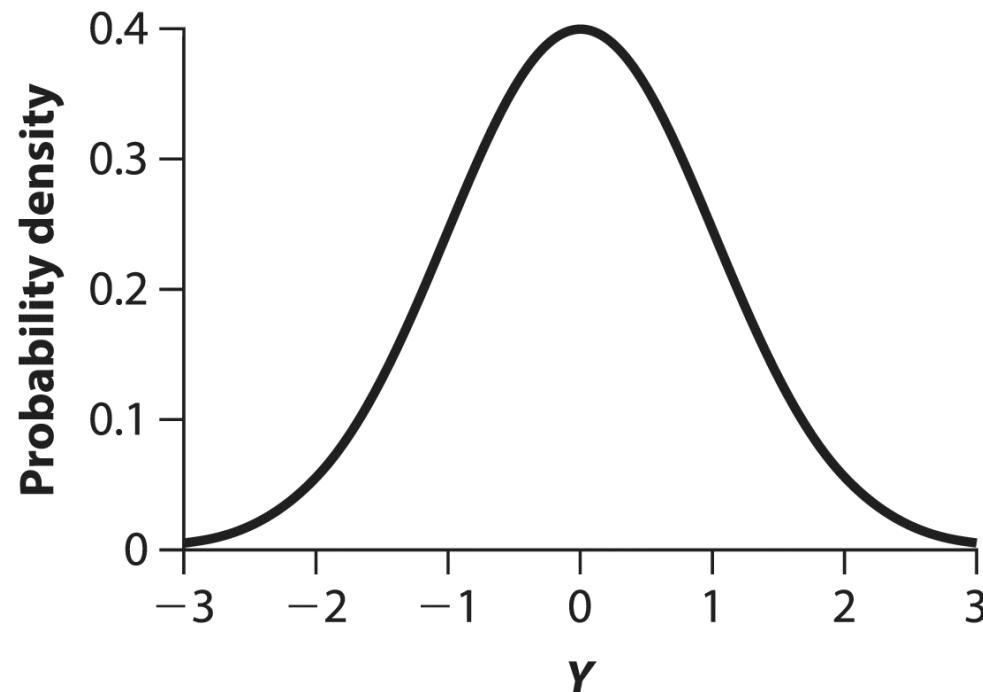
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

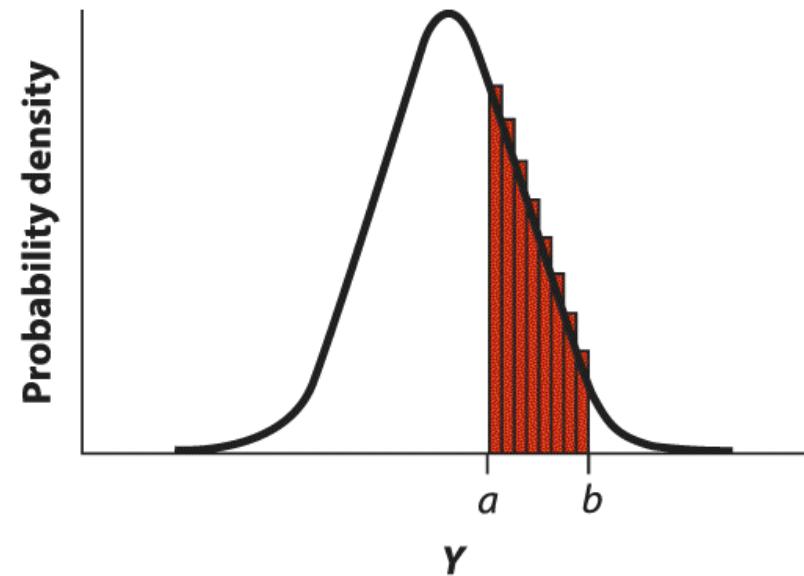
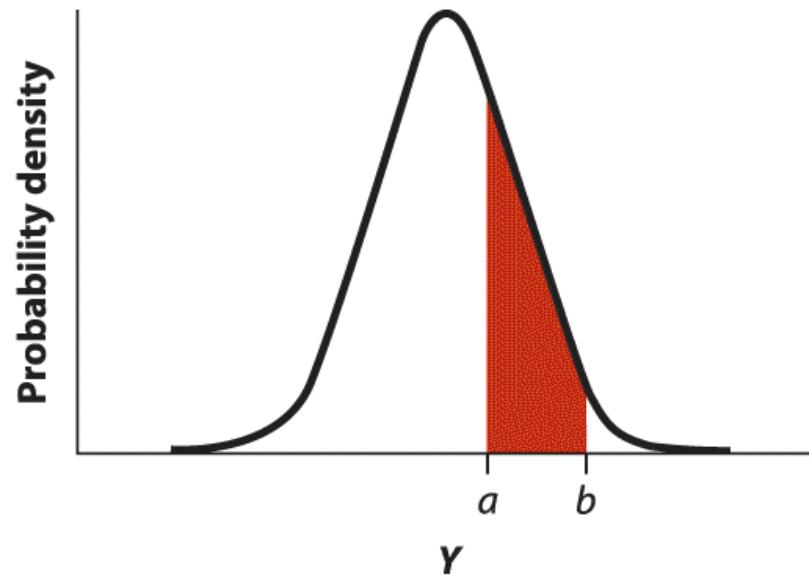
2.3 Probability distribution 概率分布

- Continuous probability distributions 连续概率分布
 - For continuous numerical variables (连续数值变量)
 - Y-axis: probability density 概率密度 (曲线高度)



2.3 Probability distribution 概率分布

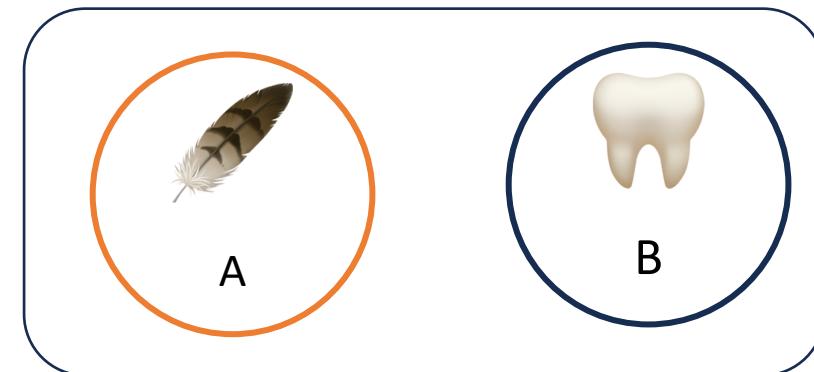
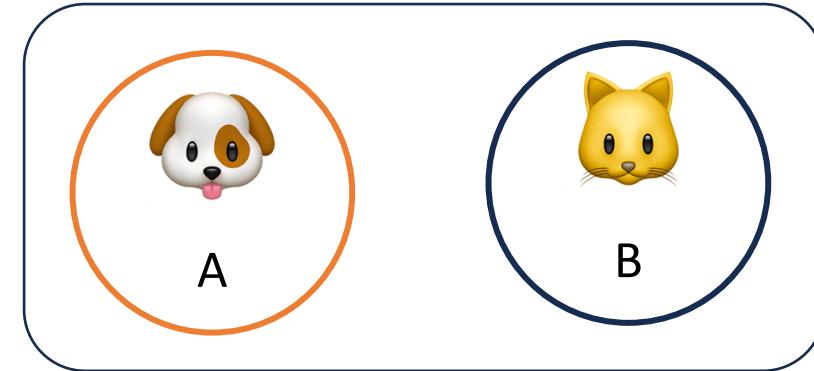
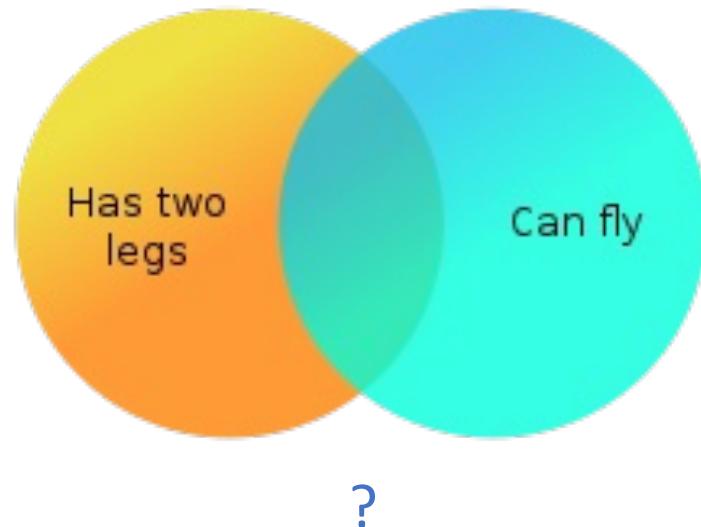
- Continuous probability distributions 连续概率分布
 - For continuous numerical variables (连续数值变量)
 - Y-axis: probability density 概率密度 (曲线高度)
 - The probability of obtaining a value of Y that falls within some range
 - 落在一定范围内的概率: $\Pr[a < Y < b]$



2.4 Mutually exclusive events 互斥事件

- 互斥事件

- 事件A和事件B不能同时发生: $\Pr[A \text{ and } B] = 0$
- They cannot both occur at the same time.



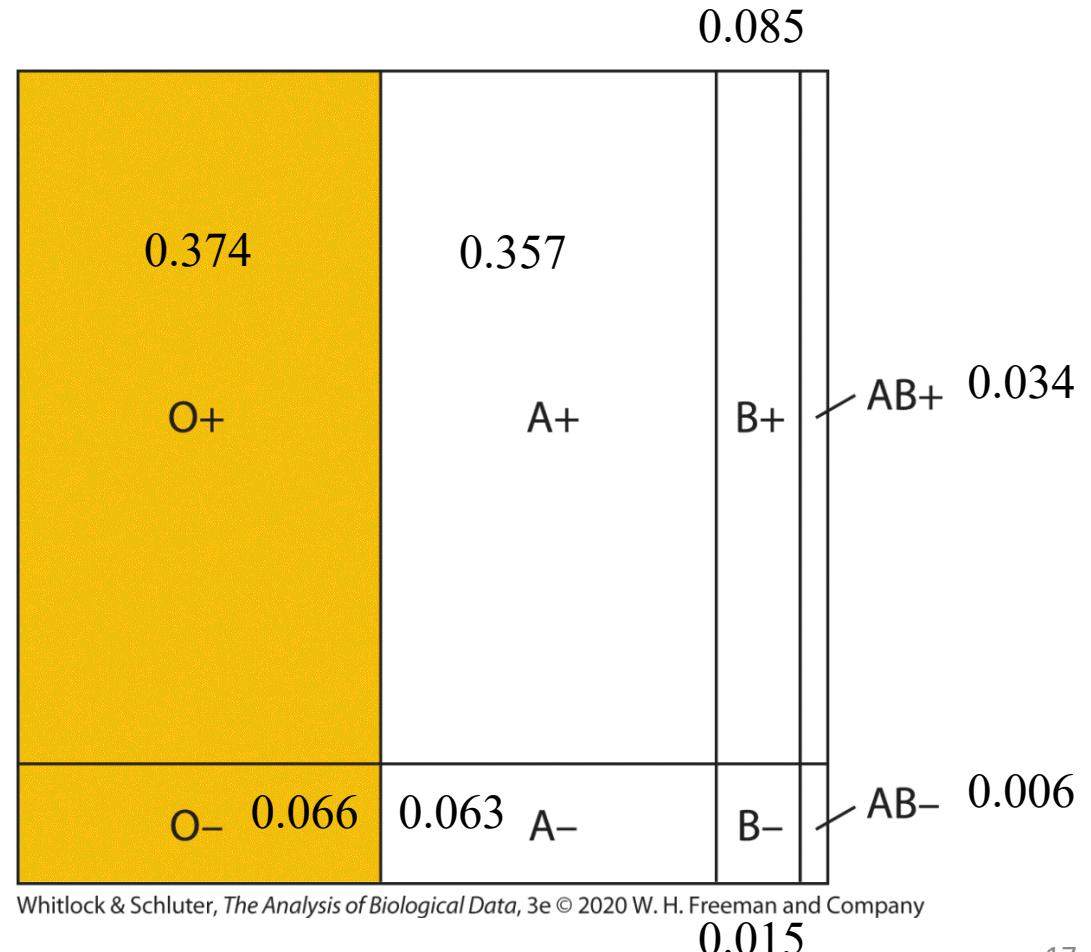
2.4 Mutually exclusive events 互斥事件

- 事件A和事件B不能同时发生: $\Pr[A \text{ and } B]=0$
- 事件A或事件B发生的概率?
- 互斥事件的概率加法公式:
 - $\Pr[A] + \Pr[B]$

• 血型是O的概率?

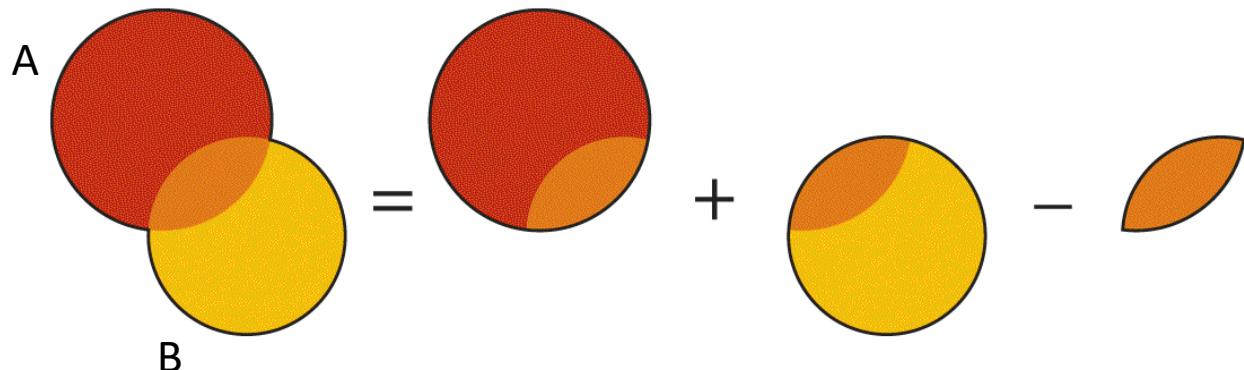
$$\Pr[O+] + \Pr[O-] = 0.374 + 0.066$$

• 血型是Rh+的概率?



2.4 Mutually exclusive events 互斥事件

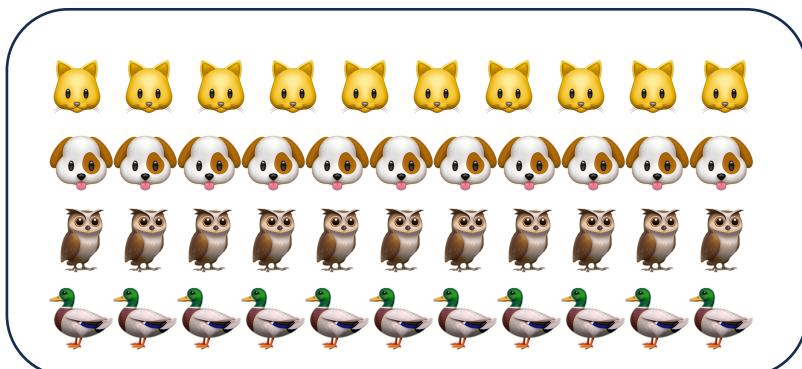
- 事件A和事件B不能同时发生: $\Pr[A \text{ and } B]=0$
- 互斥事件的概率加法公式: $\Pr[A] + \Pr[B]$
- 扩展的一般性加法公式: $\Pr[A] + \Pr[B] - \Pr[A \text{ and } B]$
 - 当A和B不一定是互斥事件时



$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \text{ and } B]$$

2.5 Independent events 独立事件

- 事件A和事件B的发生不依赖于彼此
- 独立事件的概率乘法表公式: $\Pr[A \& B] = \Pr[A] \times \Pr[B]$
 - 抽两次小动物



$$\Pr[\text{🐱} \& \text{🦉}] = \Pr[\text{🐱}] \times \Pr[\text{🦉}] = \frac{1}{4} \times \frac{1}{4}$$



Animal 1
or
Animal 2

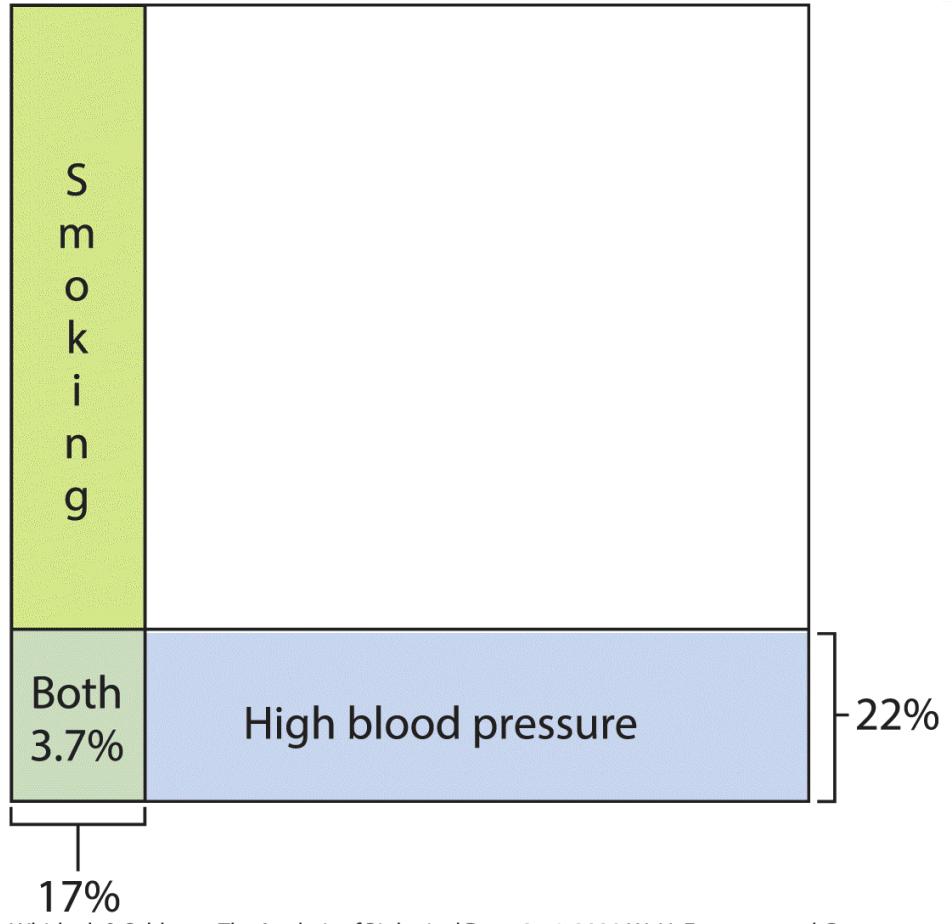


Animal 2
or
Animal 1



2.5 概率的“and” versus “or”

- 互斥事件 - 加法公式
 - $\Pr[A \text{ or } B] = \Pr[A] + \Pr[B]$
- 独立事件 - 乘法公式
 - $\Pr[A \text{ and } B] = \Pr[A] \times \Pr[B]$
- Q: 抽烟或高血压的概率 ?



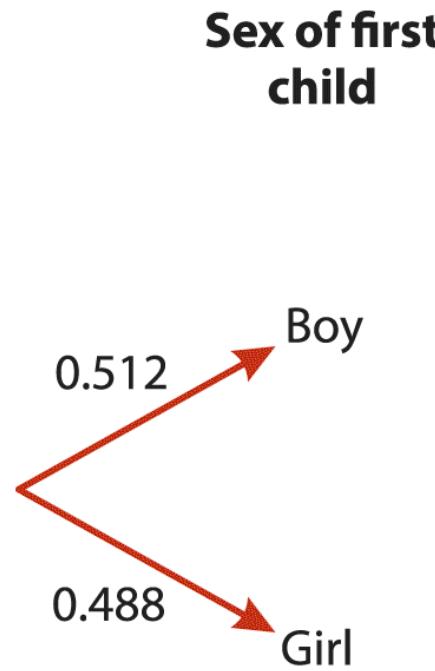
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.6 Probability trees 概率树

- 用于计算多次随机试验所产生的事件组合的概率

- 二胎家庭的子女

- 性别
 - 男女概率不一
- 顺序
 - 女女
 - 男男
 - 男女
 - 女男

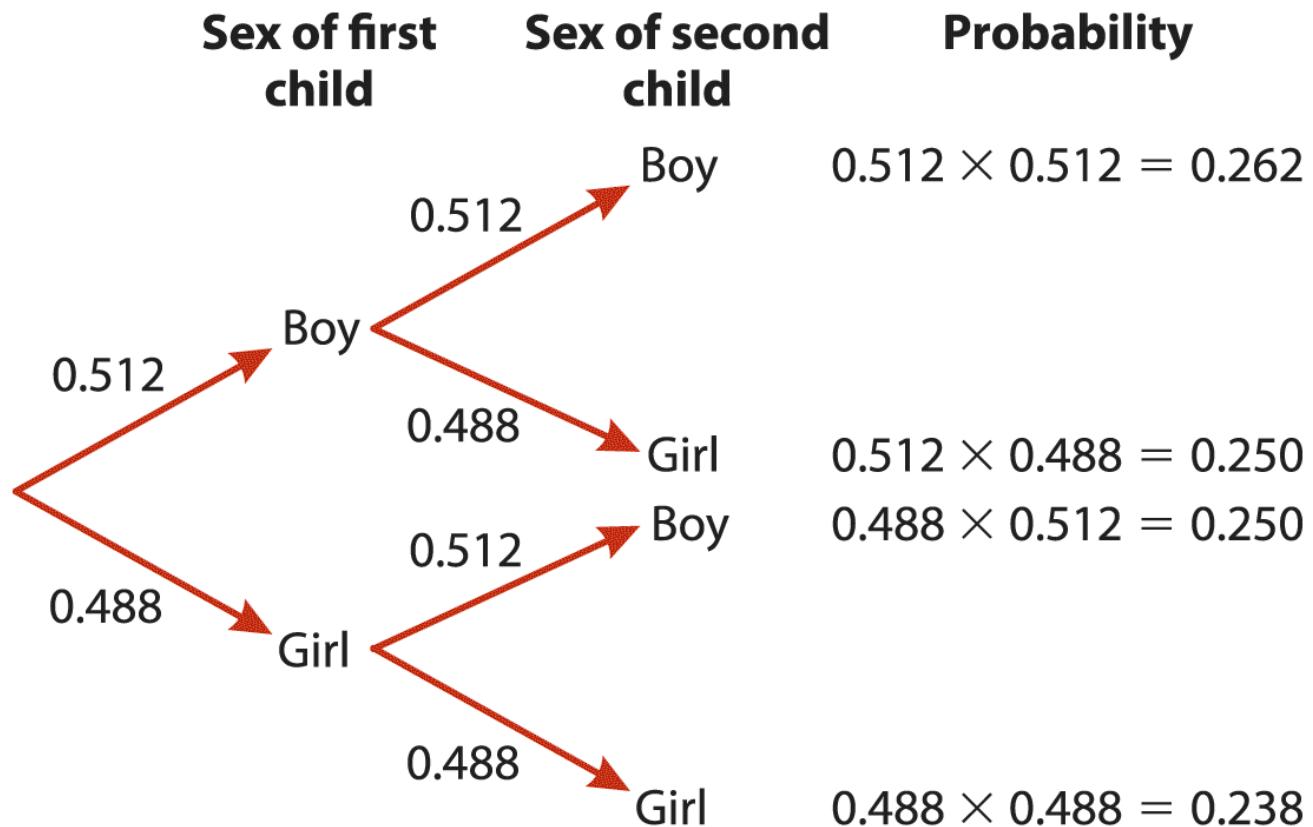


2.6 Probability trees 概率树

- 用于计算多次随机试验所产生的事件组合的概率

- 二胎家庭的子女

- 性别
 - 男女概率不一
- 顺序
 - 女女
 - 男男
 - 男女
 - 女男



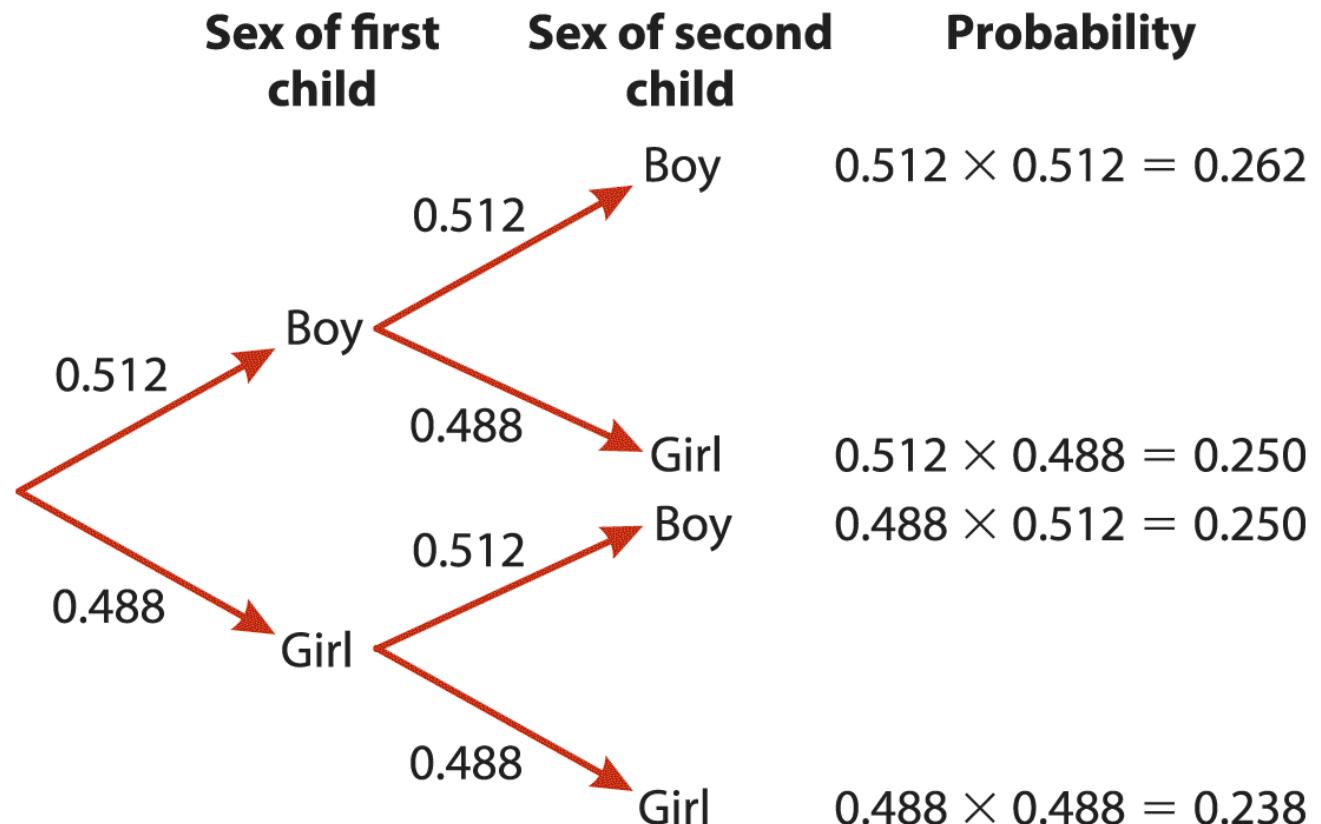
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



2.6 Probability trees 概率树

- 用于计算多次随机试验所产生的事件组合的概率
- 二胎家庭的子女

- 计算概率：
 - 至少一个女孩？
 - 至少一个男孩？
 - 孩子是相同性别？



2.6 其它内容

- Dependent event 非独立事件
 - 如果事件不是独立的，那么它们就被称为依存事件。两个从属事件同时发生的概率由一般乘法法则给出： $\Pr[A \& B] = \Pr[A]\Pr[B|A]$
- Conditional probability and Bayes' theorem
 - 条件概率和贝叶斯定理
 - 一个事件的条件概率是该事件在某种条件下发生的概率
- 参考阅读材料 Whitlock & Schluter-2020-Ch5.pdf

3. Summary - Probability

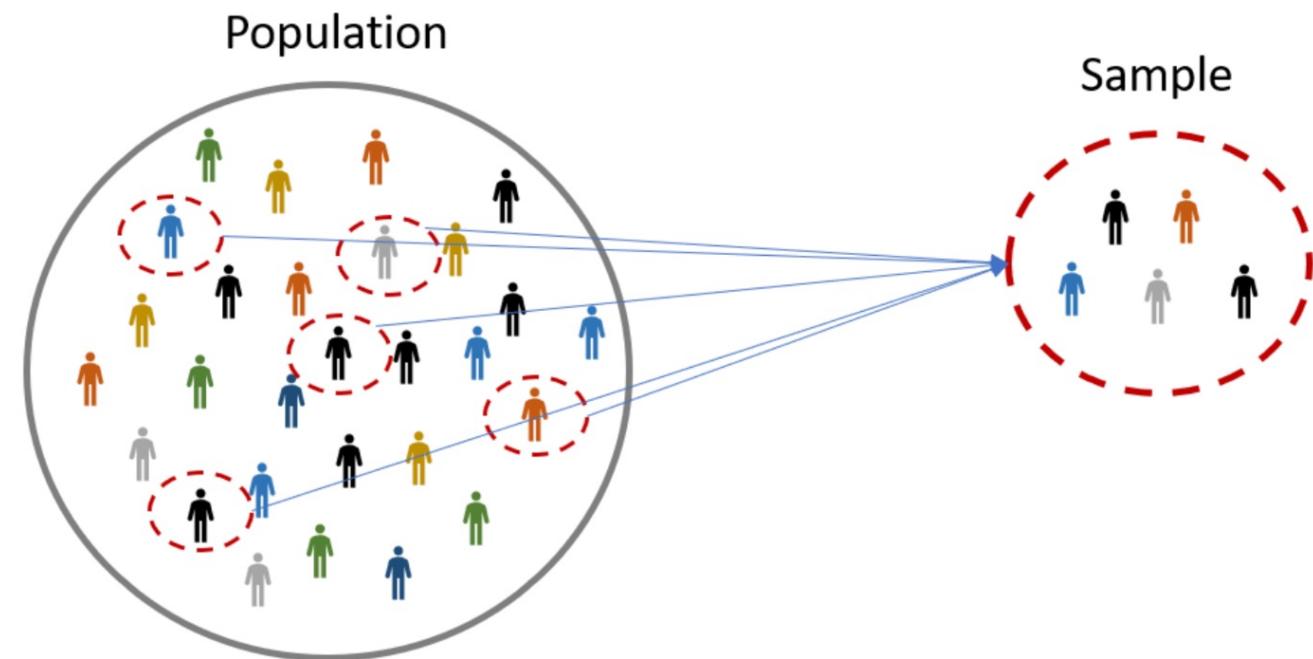
- 概率是生物学中的一个重要概念
 - 对一个总体进行随机取样即是一种随机试验 (random trial)，其结果受概率规则的制约。
 - 随机试验是一种过程或实验，它有两种或两种以上可能的结果，而这些结果的发生是无法准确预测的。
 - 事件的概率 (the probability of an event) 是指在相同条件下反复进行随机试验时，事件发生的次数比例。
 - 概率分布描述了随机试验所有可能结果的概率。

3. Summary - Probability

- 概率运算
 - 如果事件A和B不会同时发生，它们为互斥事件；那么A或B发生的概率参考加法公式： $\text{Pr}[A \text{ or } B] = \text{Pr}[A] + \text{Pr}[B]$
 - 扩展的一般加法公式为 $\text{Pr}[A \text{ or } B] = \text{Pr}[A] + \text{Pr}[B] - \text{Pr}[A \& B]$
 - 如果事件A或B的发生不依赖于彼此，它们互为独立事件；那么A和B同时发生的概率参考乘法公式： $\text{Pr}[A \& B] = \text{Pr}[A] \times \text{Pr}[B]$

4. 作业讲解

- 总体 A full set
- 样本 A subset
 - 通过随机抽样过程构建





4. 作业讲解

- 关键是要明白研究目的和研究对象是什么？
 - 总体 A full set versus 样本 A subset

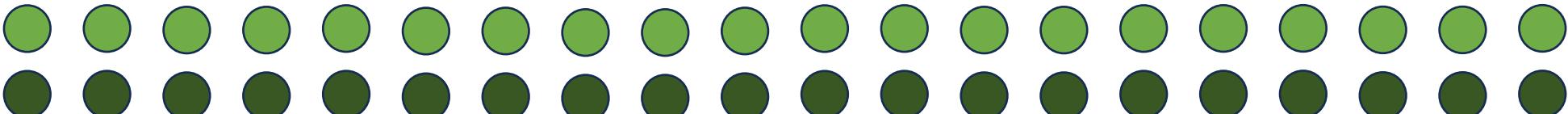
在一项关于食叶蚁饮食偏好的研究中，研究人员向 20 个随机选择的蚁群提供了周围森林中两种最常见树种的树叶。树叶以 100 片为一堆，每种树叶一堆，堆放在蚁群入口附近。树叶被裁剪得足够小，以便一只蚂蚁搬运。

24 小时后，研究人员返回计算每个树种 100 片树叶中剩余的数量。部分结果见下表。根据这些结果，研究人员估算出 *Spondias mombin* 的叶片比例为 0.65，并得出结论：蚂蚁偏爱这种叶片。



- Variable type / Data type
- Why do the 2412 leaves used in the calculation of the proportion not represent a random sample? 为什么计算比例时使用的 2412 片树叶不代表随机样本?

Tree species	Number of leaves removed
<i>Spondias mombin</i> (黄槟榔青)	1561
<i>Sapium thelocarpum</i> (乌桕属的某种)	851
Total	2412



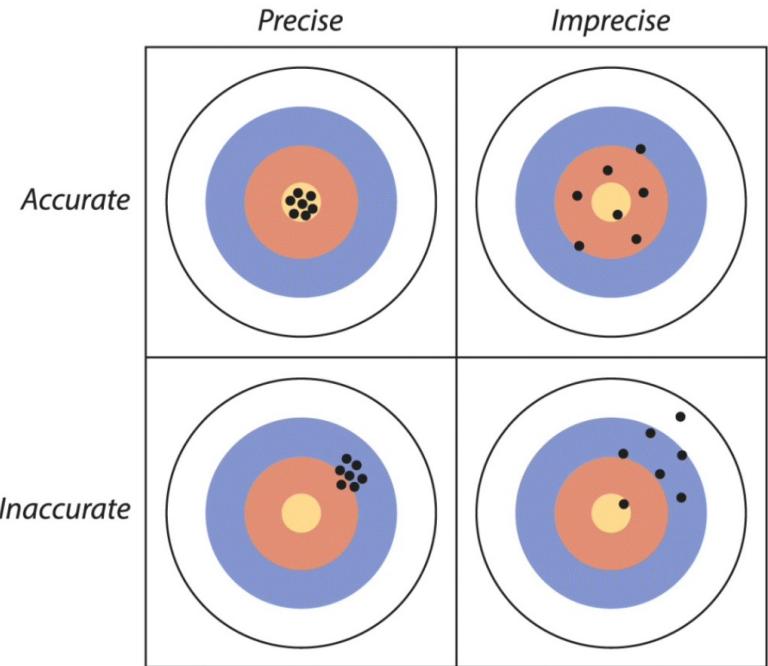
4. 作业讲解

- 抽样过程 sampling

- 对一个总体进行随机取样即是一种随机试验
- 多次取样构成一个样本。

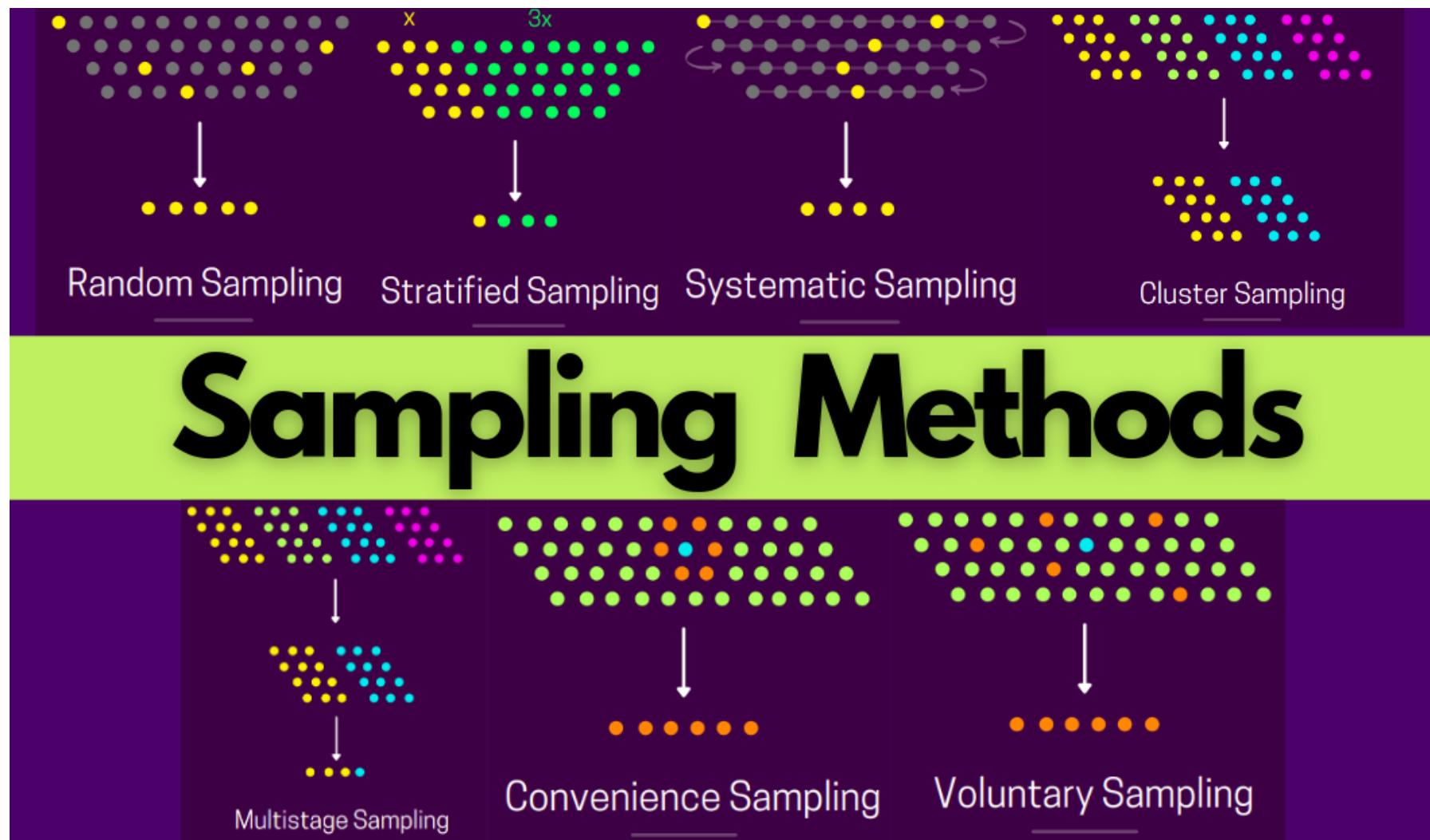
- 抽样误差 sampling error

- 抽样误差是指利用随机性原则从总体中抽取部分个体组成样本对总体进行研究时，样本的估计值与总体的真实值之间的差异。
- 抽样误差是由样本抽取的随机性 (chance) 造成的，而不是调查误差的结果，只要使用抽样调查，抽样误差就不可避免。
- 准确性 vs 精确性



4. 作业讲解

- 抽样方法



5. R-tips

- 读取文件
 - 路径: 绝对路径 vs 相对路径
 - Session – Set working directory – To source file location
 - 前提是最好code和file在一个文件夹下
 - 文件: 文件后缀名
 - 格式: csv or others?
- 新建 project
 - 免去每次设定路径的步骤

5. R-tips

- R calculations vs. report text
 - Numbers/values: decimals
 - Unit: different values, statistics
- Figures/Graphics
 - Graph types
 - one or two variables & data type

Recall L02 - Graphics

Types of data 数据类型

One categorical variable 单个分类变量

One numerical variable 单个数值变量

Two numerical variables 两个数值变量

Two categorical variables

两个分类变量

One numerical variable and

one categorical variable

单个数值变量 ~ 单个分类变量

Graphical method 图像类型

Bar plot 柱状图

Histogram 直方图

Scatter plot 散点图

Grouped bar graph 分组柱状图

Mosaic plot 马赛克图

Strip chart 条形图

Boxplot 箱形图

Violin plot 小提琴图

Multiple histograms 多组直方图

Questions

- The standard error of an estimate is the standard deviation of the estimate's sampling distribution.
 - 估计值的标准误是其抽样分布的标准差
 - <https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm>
- CI: confidence?
 - <https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm>