

Lecture 2 – 描述数据 Describing data

- 内容

- 生物统计学 Bio-Statistics
- 描述性统计量 Descriptive statistics
 - 测量和比较 Measurements & Comparison
- 总结 Summary
- R Lab

生物统计学

李 勤

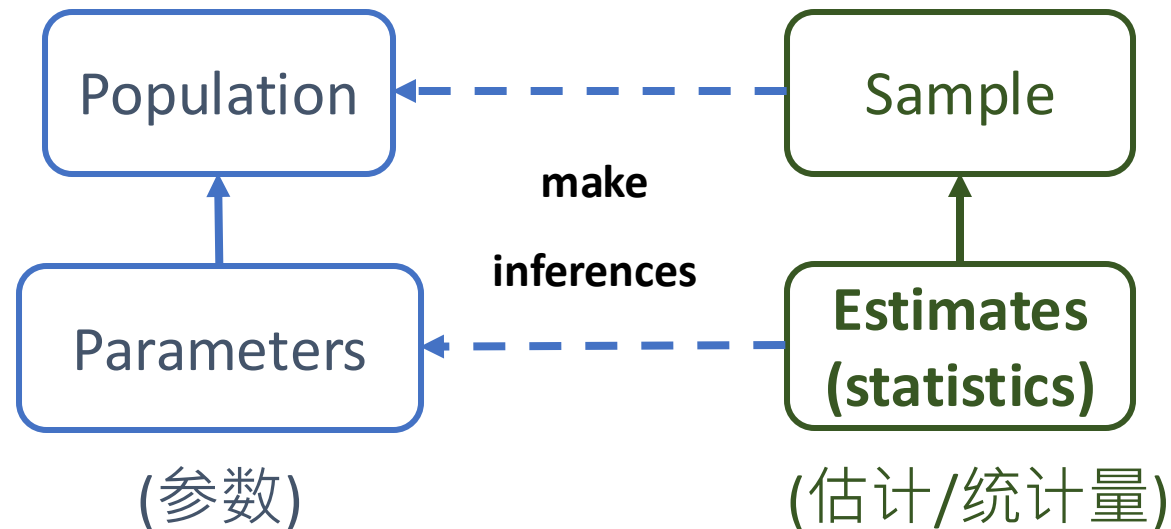
生态与环境科学学院

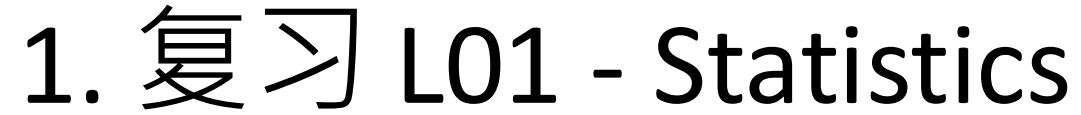
Lecture 2 – 学习重点

- 描述性统计量有哪些？
- 描述性统计量的适用场景是怎样的？
 - 针对不同的目的，如何选择合适的统计量？
 - 针对不同的数据类型，如何选择合适的统计量？
 - 针对数据的分布情况，如何选择合适的统计量？
- 在R中，如何进行描述性统计量的计算？

1. 复习 L01 - Statistics

- 统计学的目的是基于从**总体**中的**样本**所获得的信息,
- 对总体进行**推断**, 并且提供推断的准确性





- [illegible]



1. 复习 L01 - Statistics

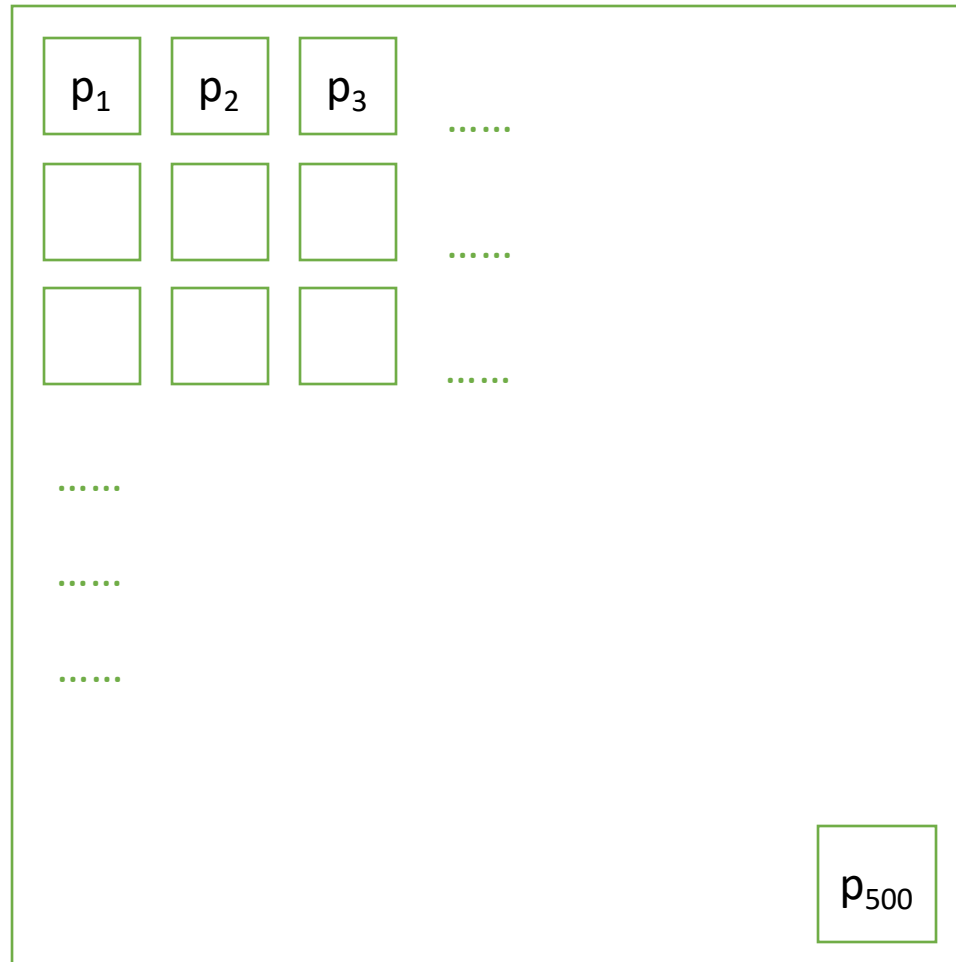
- 1. Which of the following numerical variables are continuous? Which are discrete?
 - a. Number of injuries sustained in a fall
 - b. Fraction of birds in a large sample infected with avian flu virus
 - c. Number of crimes committed by a randomly sampled individual
 - d. Logarithm of body mass



1. 复习 L01 - Statistics - Discussions

- 2. 为了调查一种松树的平均树龄，研究人员针对这种松树在加利福尼亚海岸山脉的地理分布图，利用计算机随机放置了 500 个 10 公顷的样方（plot）。然后，研究人员在实地找到每个随机样方的位置，并测量了每个 10 公顷样方内每棵松树的树龄。样方内的平均树龄被用作单位测量值（the unit measurement）。然后利用这些单位测量值估算出加利福尼亚地区这种松树的平均树龄。
 - 这个研究所感兴趣的总体是什么？
 - 为什么这个研究要先将每个样方内的多个松树的树龄进行平均，然后再针对 500 个样方再次进行平均？为什么不把所有的松树个体的树龄整合成为一个超大样本来计算平均值？

The average age of piñon/pinyon pine trees in the coast ranges of California
 - investigated by placing 500 10-hectare plots



$$\text{Age}_{p1} = \frac{a_1 + a_2 + \dots + a_{n_{p1}}}{n_{p1}}$$

$$\text{Age}_{p2} = \frac{a_1 + a_2 + \dots + a_{n_{p2}}}{n_{p2}}$$

.....

$$\text{Age}_{\text{mean}} = \frac{\text{Age}_{p1} + \text{Age}_{p2} + \dots + \text{Age}_{p_n}}{500}$$

2. 描述性统计量 Descriptive statistics

- **描述性统计量 Descriptive statistics** or summary statistics
 - 频数分布的量化特征
 - **quantities** that capture important features of frequency distributions.
- 虽然图表揭示了数据中的形状和模式 (shapes and patterns) , 但描述性统计提供了确切的数字 (hard numbers) ;
 - 对数值变量而言, 最重要的统计量包括其居中位置和散布程度;
 - 对类型变量而言, 最重要的统计量是某一类的比例 (**proportion** /fraction);

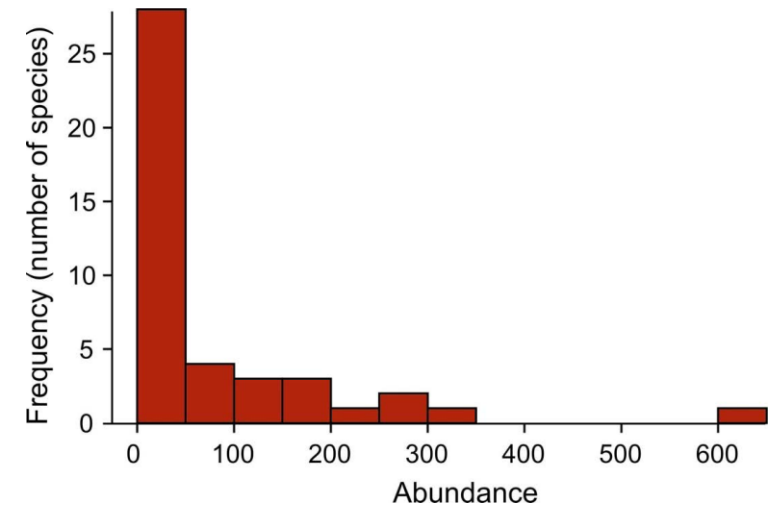
2. 描述性统计量

- 位置 Location

- A central value (mean/均值, median/中值)
- E.g., which species is larger? Which forest has more birds?
- Generally for comparison (两者及以上做对比)

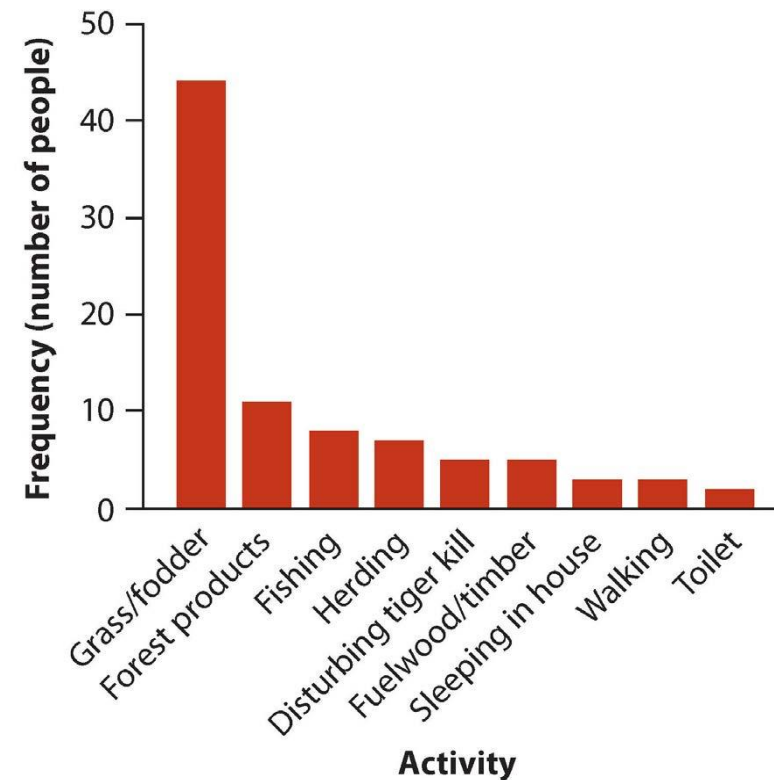
- 散布程度 Spread

- 在某些科学领域，围绕居中位置的值的变异性是仪器噪声或测量误差 (instrument noise or measurement error);
- 但在生物学中，很多变异性 (**variability**) 常常表示个体之间的真实差异。
 - Biologists also appreciate variation as the stuff of evolution.
 - 生物学家还将变异性视为进化的基础;



2. 描述性统计量

- 比例 Proportion
 - 类型变量中某一类所占的比例;
 - The fraction of observations in a given category
 - 类型变量中不同类的比例的差异;
 - The fraction differences between categories

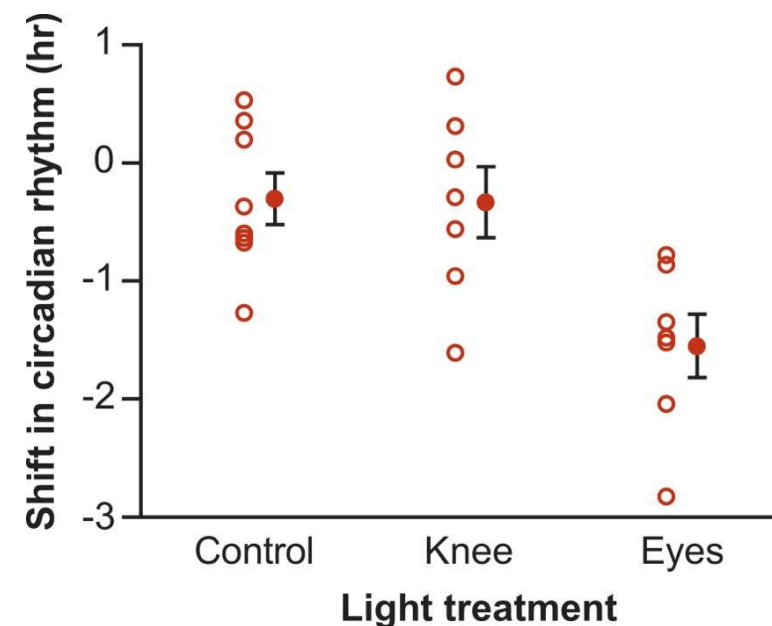


3. 描述性统计量的测量

- 算数平均值和标准差 Arithmetic mean and standard deviation
- 中值和四分位距 Median and interquartile range
- 比较位置及散布程度 Measures of location and spread compare
- 累积频率分布 Cumulative frequency distribution
- 比例 Proportion

3.1 算术平均值和标准差

- 算术平均值 Arithmetic mean
 - 算术平均值是最常见的描述频数分布位置 (**location**) 的度量——它是一组测量值的均值。
- 标准差 Standard deviation
 - 标准差是最常用的分布散布程度的度量 (**spread**)。



3.1 算术平均值和标准差

- 样本均值 The sample mean
 - 样本测量值的平均 (the average of the measurements in the sample)
 - 观测值之和除以观测样本量
(the sum of all the observations divided by the number of observations)

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

- \bar{Y} : the mean/average (平均值)
- Y_i : the value of the i^{th} observation (第*i*个观测值)
- n : the number of observations/sample size (观测值的个数/样本量)
- Σ : sum (求和符号)

3.1 算术平均值和标准差

- 样本均值 The sample mean
 - 样本测量值的平均 (the average of the measurements in the sample)
 - 观测值之和除以观测样本量



Cede Prudente/NHPA/
Photoshot

(*Chrysopelea paradisi* 天堂金花蛇)

Gliding snakes

$Y_1 = 0.9$
$Y_2 = 1.4$
$Y_3 = 1.2$
$Y_4 = 1.2$
$Y_5 = 1.3$
$Y_6 = 2.0$
$Y_7 = 1.4$
$Y_8 = 1.6$

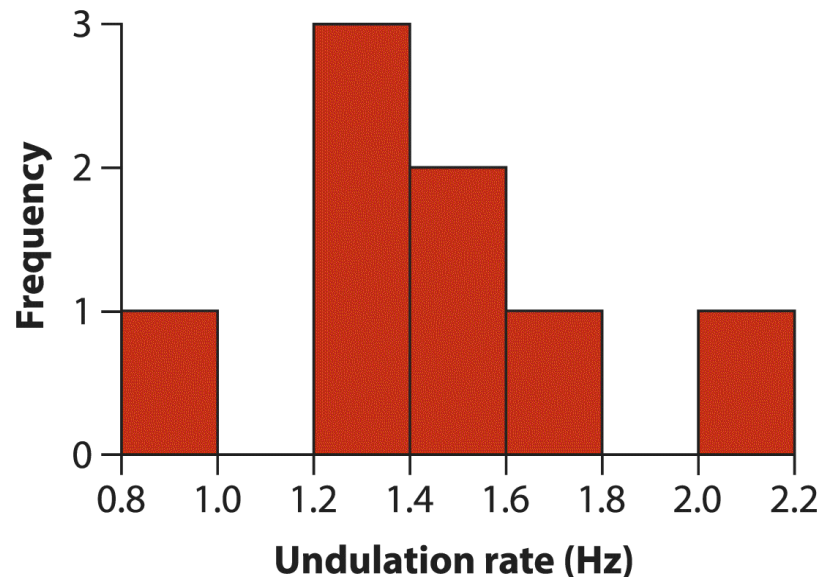
- 金花蛇俗称“飞蛇”
- 当它们要“飞翔”时，会先爬行到高处，压缩肌肉将身体压得扁平（其身体宽度可达身体水平高度的两倍），借此加强降落时的空气阻力，再将身体弹出，并滑翔至目的地处。
- 能在身处空中时，以摆动身体的方式（S形），稍微控制飞行方向。
- 视频记录其从10米高台“飞”下过程：
 - undulation rates
 - cycles per second (Hz)

3.1 算术平均值和标准差

- 样本均值 The sample mean
 - 样本测量值的平均 (the average of the measurements in the sample)
 - 观测值之和除以观测样本量

Gliding snakes

$Y_1 = 0.9$
 $Y_2 = 1.4$
 $Y_3 = 1.2$
 $Y_4 = 1.2$
 $Y_5 = 1.3$
 $Y_6 = 2.0$
 $Y_7 = 1.4$
 $Y_8 = 1.6$



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

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

$$\begin{aligned}\bar{Y} &= \frac{0.9 + 1.4 + 1.2 + 1.2 + 1.3 + 2.0 + 1.4 + 1.6}{8} \\ &= 1.375\end{aligned}$$

3.1 算术平均值和标准差

- 离均差 Deviation

$$Y_i - \bar{Y}$$

- 方差 Variance

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

\bar{Y} : the mean/average (平均值)

Y_i : the i^{th} observation (第 i 个观测值)

n : sample size (样本量)

- 离均差平方 Squared deviation

$$(Y_i - \bar{Y})^2$$

- 标准差 Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$



3.1 算术平均值和标准差

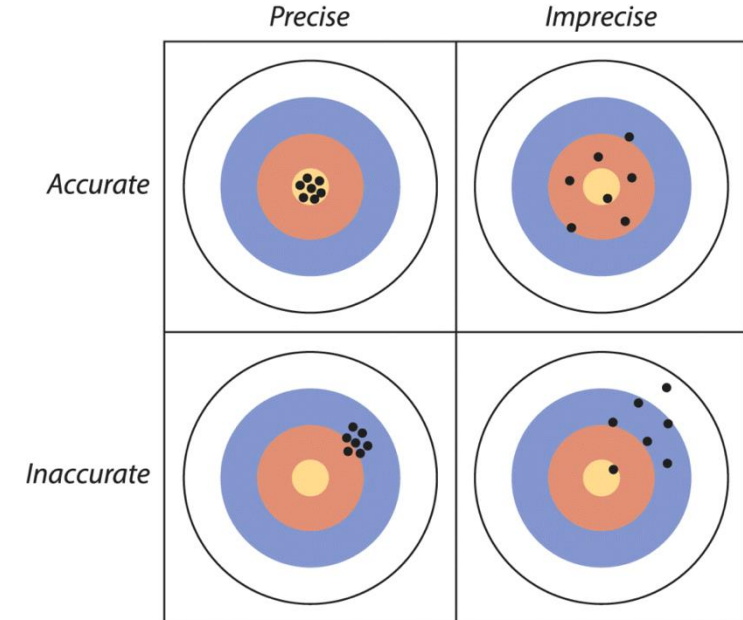
- 方差 Variance

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

- Why $n - 1$?

- 标准差 Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$





3.1 算术平均值和标准差

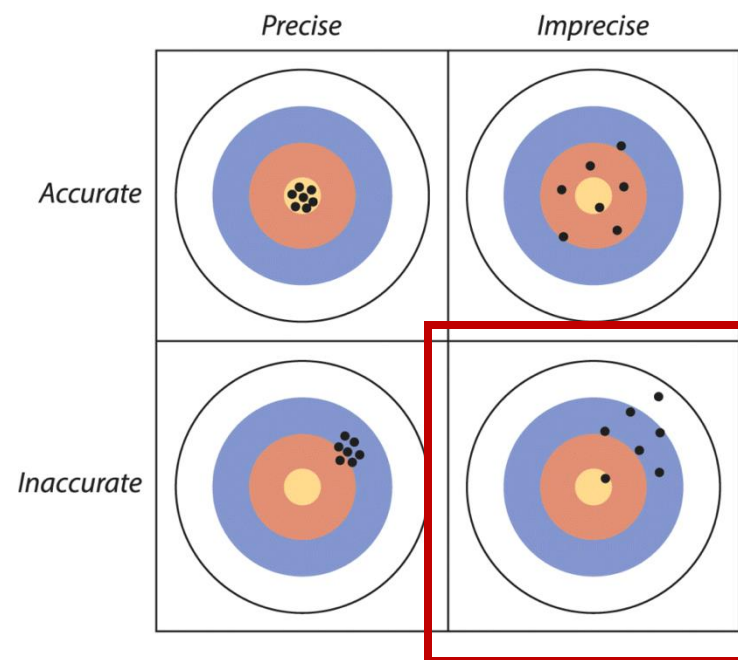
- 方差 Variance

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

- 标准差 Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

- Why $n - 1$?
 - We don't know the true mean of the population;
我们不知道真实的总体的均值
 - So, $\sum_{i=1}^n (Y_i - \bar{Y})^2$ is a bit smaller than what it would be;
样本的离均差平方和会相对偏小
 - Statisticians say there are $n - 1$ degrees of freedom.
 $n - 1$ 为自由度



如果直接使用 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 作为估计，那么你会倾向于低估方差！

这是因为：

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 &= \frac{1}{n} \sum_{i=1}^n \left[(X_i - \mu) + (\mu - \bar{X}) \right]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 + \frac{2}{n} \sum_{i=1}^n (X_i - \mu)(\mu - \bar{X}) + \frac{1}{n} \sum_{i=1}^n (\mu - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 + 2(\bar{X} - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\mu - \bar{X})^2 \end{aligned}$$

换言之，除非正好 $\bar{X} = \mu$ ，否则我们一定有

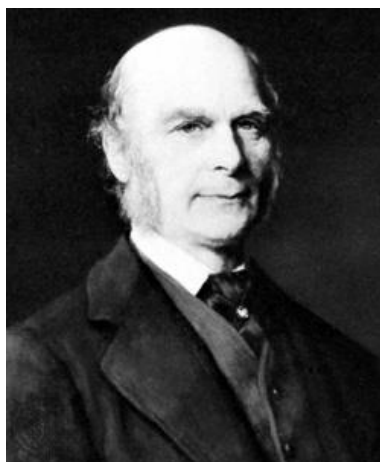
$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 < \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2,$$

3.1 算术平均值和标准差

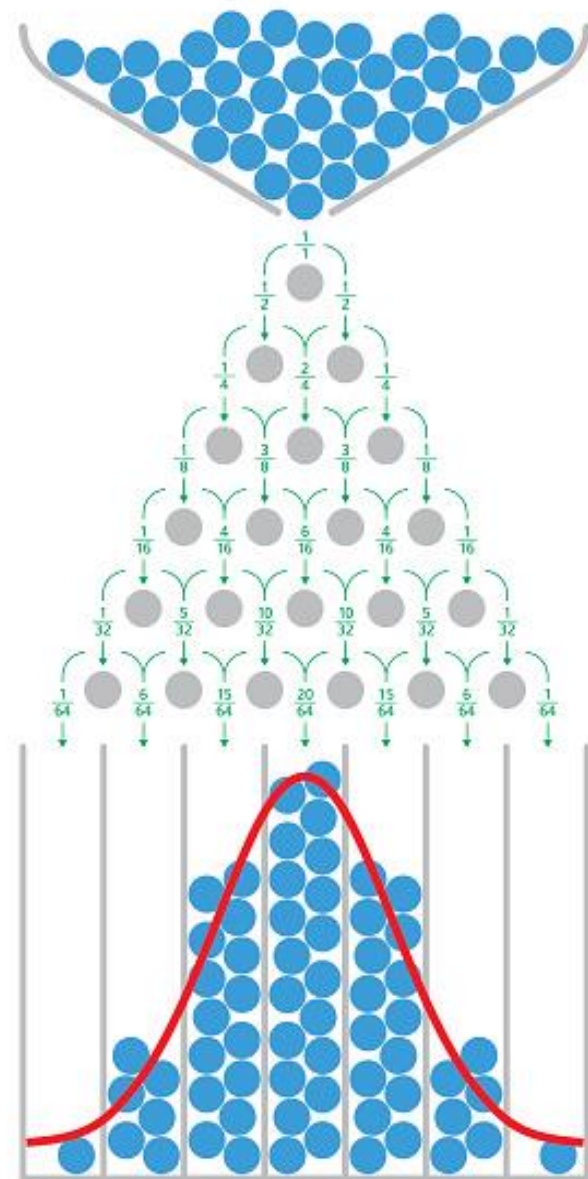
- 关联?

Normal distribution 正态分布

- 中心极限定理 the central limit theorem
 - 是概率论 (probability theory) 一个非常重要的结论
 - 它指出在一定条件下, 独立 (independent) 随机变量随样本量 (sample size) 变大会趋向正态分布 (normal distribution) [e.g., 抛硬币]
- 高尔顿版 Galton board



Sir Francis Galton (英)



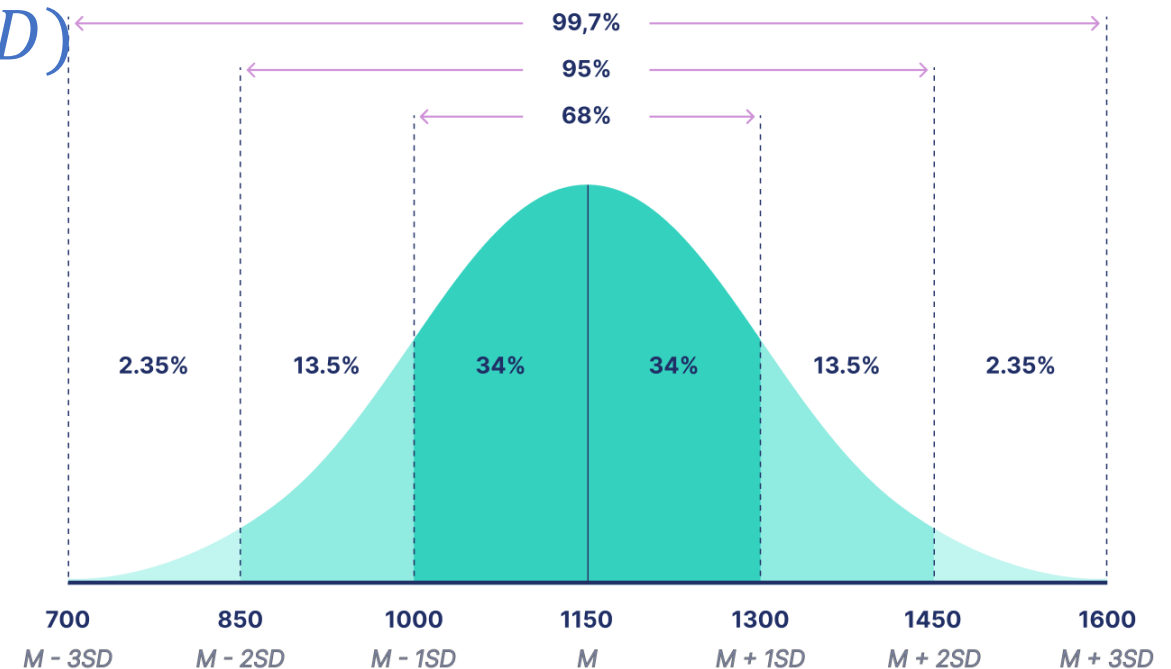
3.1 算术平均值和标准差

- 关联: 标准差 standard deviation (SD) ~ 频数分布 frequency distribution
 - 如果频数分布是钟形/正态分布 (bell shaped/a **normal distribution**)
 - 约有2/3的观测值将落在平均值两侧各1个标准差内:

$$67\%-68\% Y_i \in (\bar{Y} - SD, \bar{Y} + SD)$$

- 约 95%的观测值将落在2个SD内:

$$95\% Y_i \in (\bar{Y} - 2SD, \bar{Y} + 2SD)$$



3.1 算术平均值和标准差

- 均值 Mean $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$

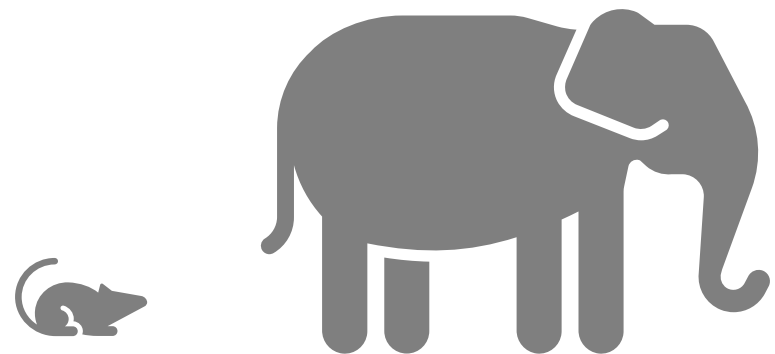
- 标准差 SD $s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$

- 变异系数 Coefficient of variation (CV)

- 将标准差表示为均值的百分比

- 以表达个体之间的相对变异 (the relative variation)
 - 因为均值和标准差可能一起变化:

- E.g., biomass of elephant vs. mouse; 3000 kg vs. 20 g → 20% vs. 20%

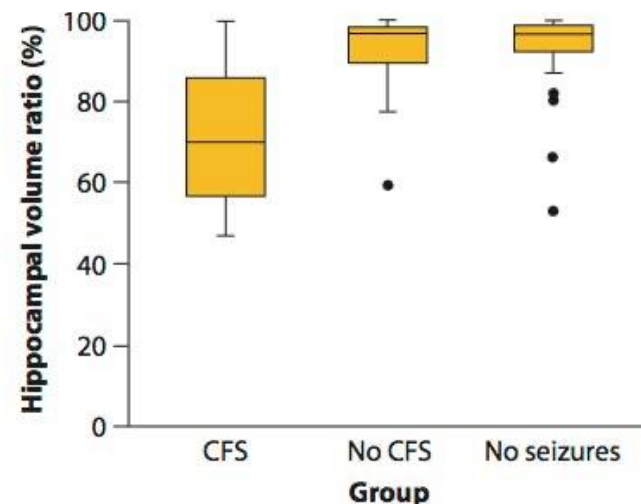


$$CV = \frac{s}{\bar{Y}} \times 100\%$$

3.2 中值和四分位距

- 中值 Median

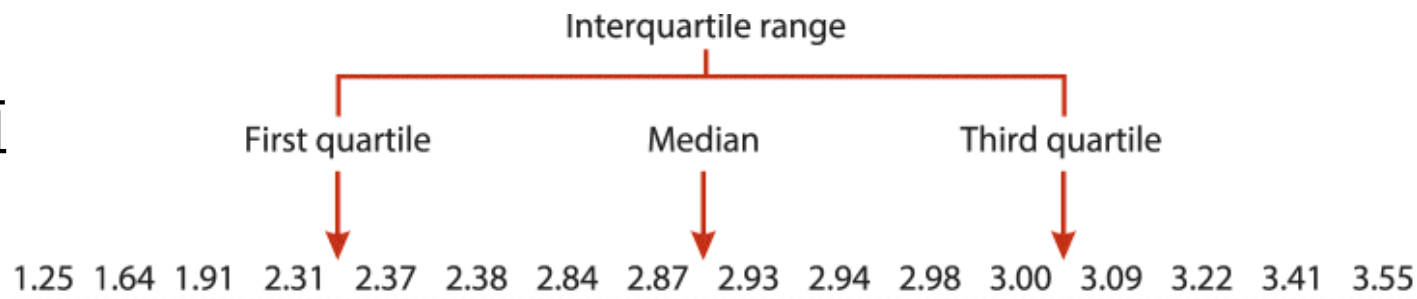
- the middle observation in a set of data
 - 1. sort the data (数据排序)
 - 2. obtain the middle one (获得居中的数值)
 - Odd (样本大小为奇数): the middle one
 - Even (样本大小为偶数): the mean of the middle two



$$Y_{([n+1]/2)}$$
$$(Y_{[n/2]} + Y_{[n/2+1]})/2$$

- 四分位距 Interquartile range (IQR)

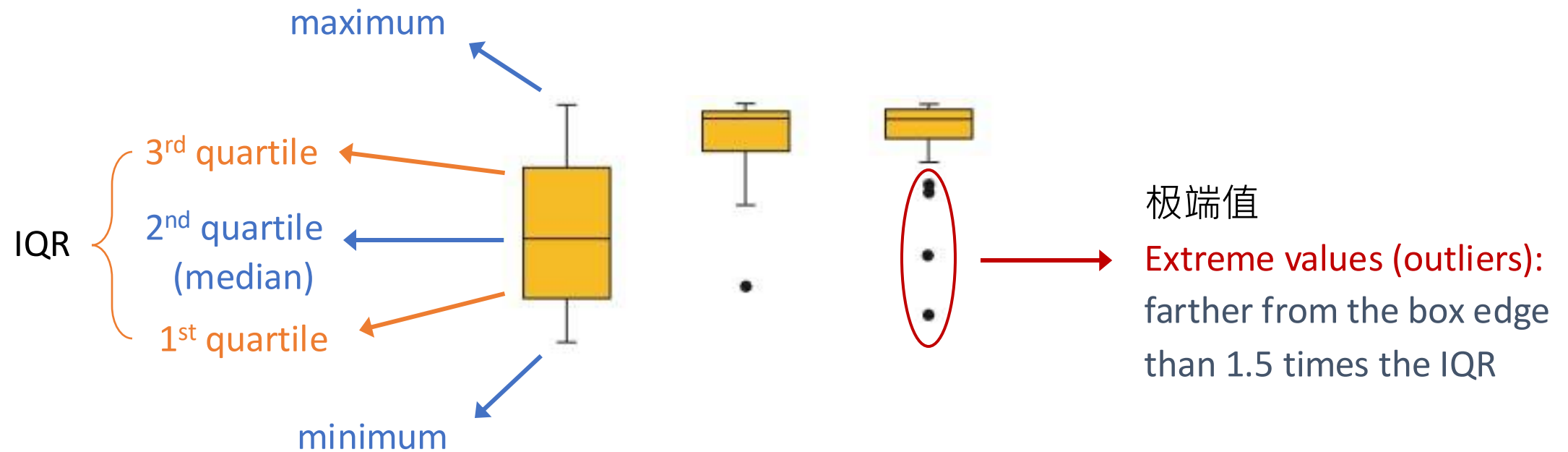
- 1st quartile (第一四分位数)
- 2nd quartile = median 中值
- 3rd quartile (第三四分位数)
- IQR = 1st ~ 3rd Qu



- the span of the middle half of the data

3.2 中值和四分位距

- 中值 Median
- 四分位距 Interquartile range (IQR)
- 箱形图 Boxplot



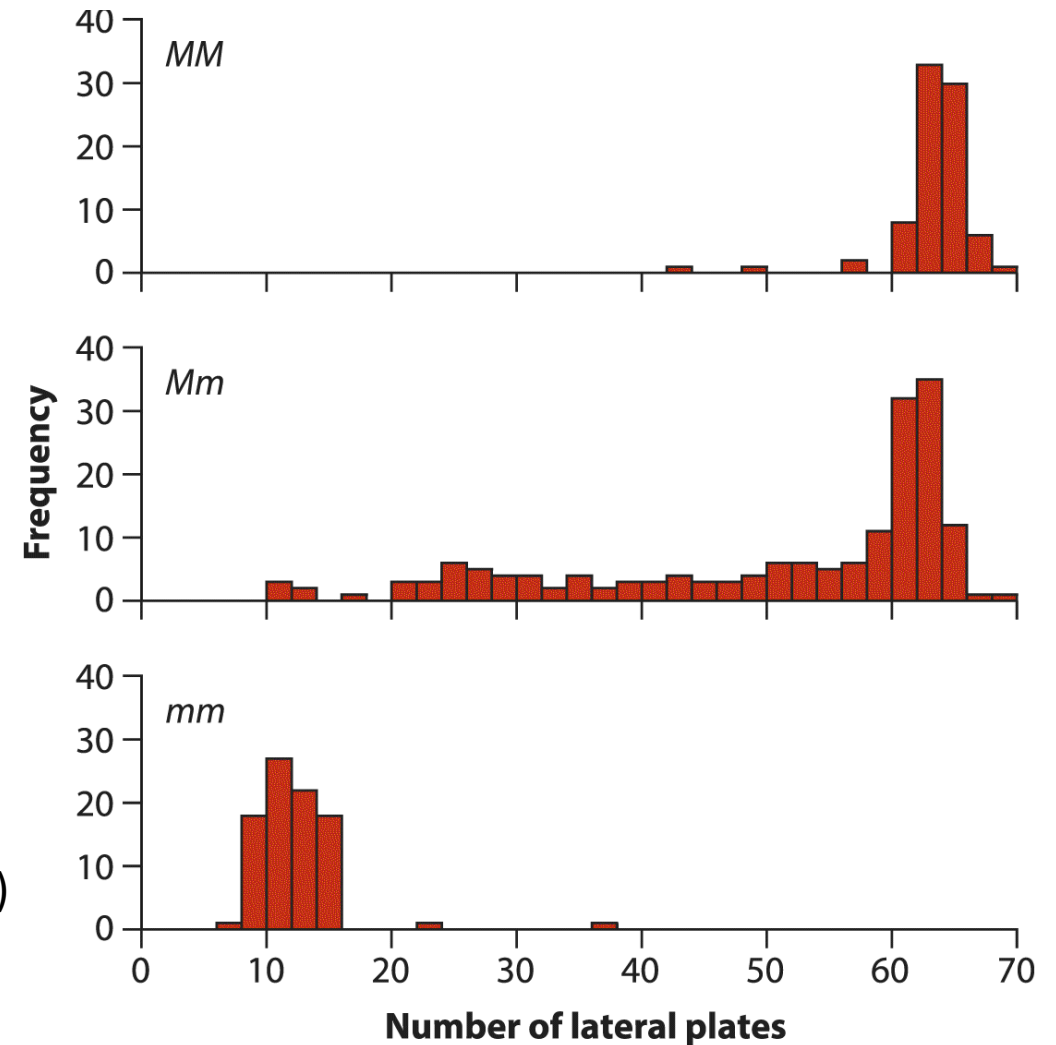
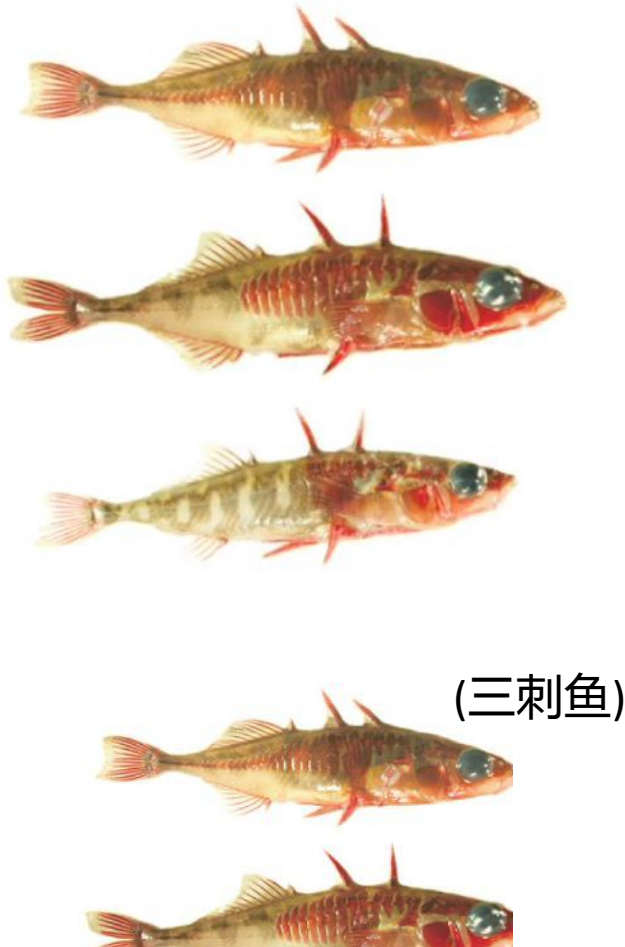
3.3 和比较测量值

- 位置 Location
 - a central value: mean 均值 vs. median 中值
 - 如何比较? How to compare?
 - 取决于频数分布的形状 (the shape of the frequency distribution)
 - 当分布是不对称/倾斜或包含极端值时(strongly skewed or include extreme observations):
 - 中位数是分布的中间测量值;而均值是 “重心” ——更容易受到极端值的影响;
mean & standard deviation << median & interquartile range
(less informative)
- (需要考虑频数分布的形状; 当形状偏离正态分布时, 均值和标准差所能提供的信息可能比不上中值和四分位距的信息)



比较居中位置 the central value

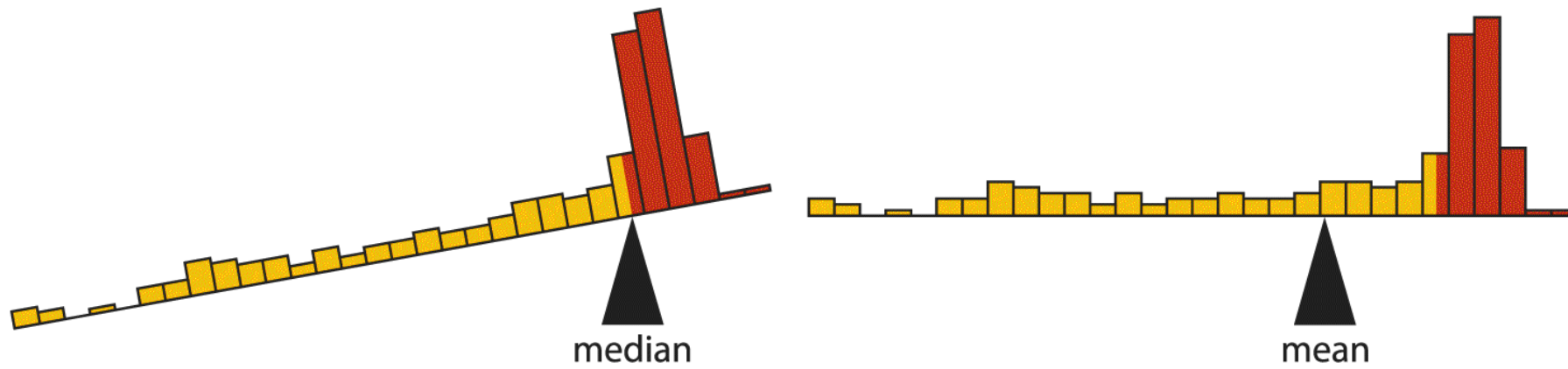
- lateral plates
 - 侧面甲板
 - a signal gene
- *M*
 - from marine
- *m*
 - from freshwater



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

比较居中位置 the central value

- 如何比较? → 取决于频数分布的形状
- 当分布是非对称/倾斜 (**asymmetric/skewed**)
 - 中位数是分布的中间测量值;
 - 而均值是“重心” → 更容易受到极端值的影响;



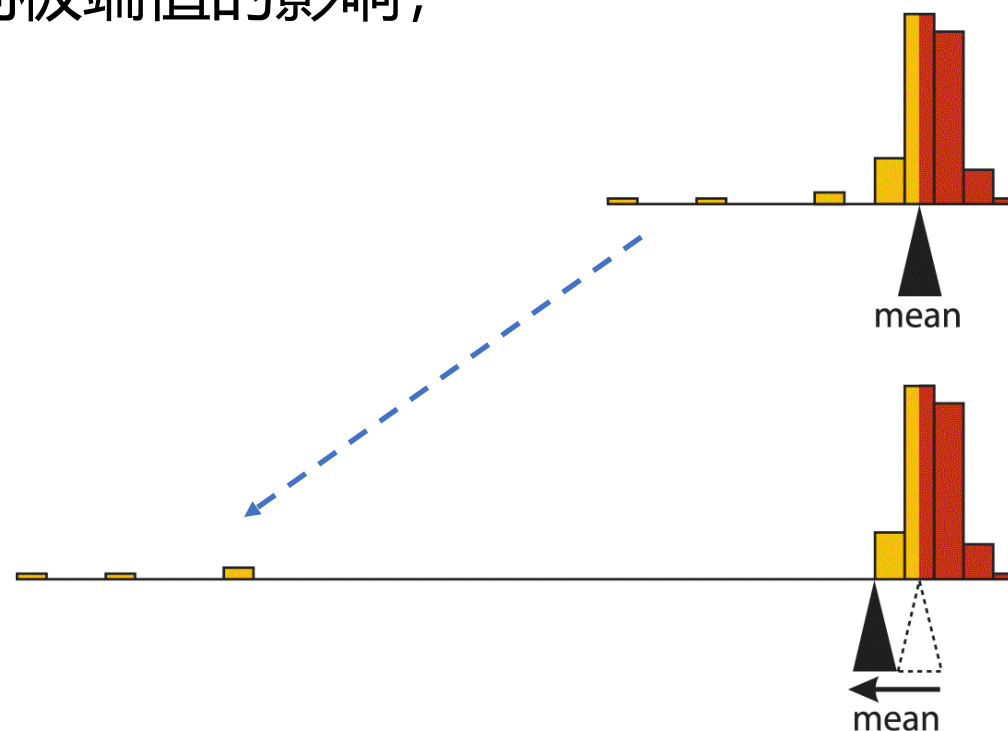
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

比较居中位置 the central value

- 当分布是非对称/倾斜 (**asymmetric/skewed**)
 - e.g., add minimum extremes 如果极端值更小时
 - 中位数是分布的中间测量值 (unaffected 没影响)
 - 而均值是“重心” → 更容易受到极端值的影响;

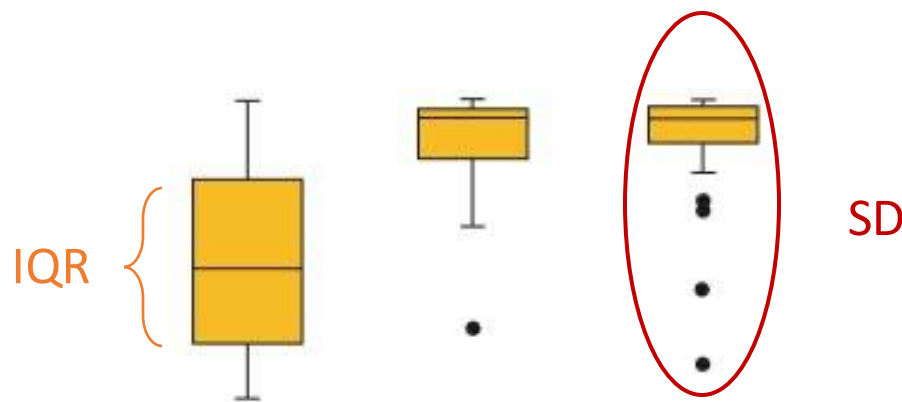
The balancing act

→ *The mean shifts leftward*



比较散布程度 the spread

- 散布程度 = 变异性 variability: SD 标准差 vs. IQR 四分位距
 - 标准差：对极端观测比均值更敏感（相比均值而言）；
- 当分布是非对称/倾斜 (**asymmetric/skewed**)
 - 四分位距 (IQR) 比标准差能更好地表示分布主体 (main part) 的变异程度；
 - 标准差反映了所有数据点 (all data points) 之间的变异；



3.4 累积频率分布 Cumulative frequency distribution

- 百分位数 Percentile

- The X^{th} percentile is the **value** below which X percent of the individuals lie
- E.g., 第10个百分位数意味着样本中有10%的观测值低于该百分位数

- 分位数 Quantile

- the 10th percentile is the 0.10 quantile (第10个百分位数是0.01分位数)

- 四分位数 Quartile

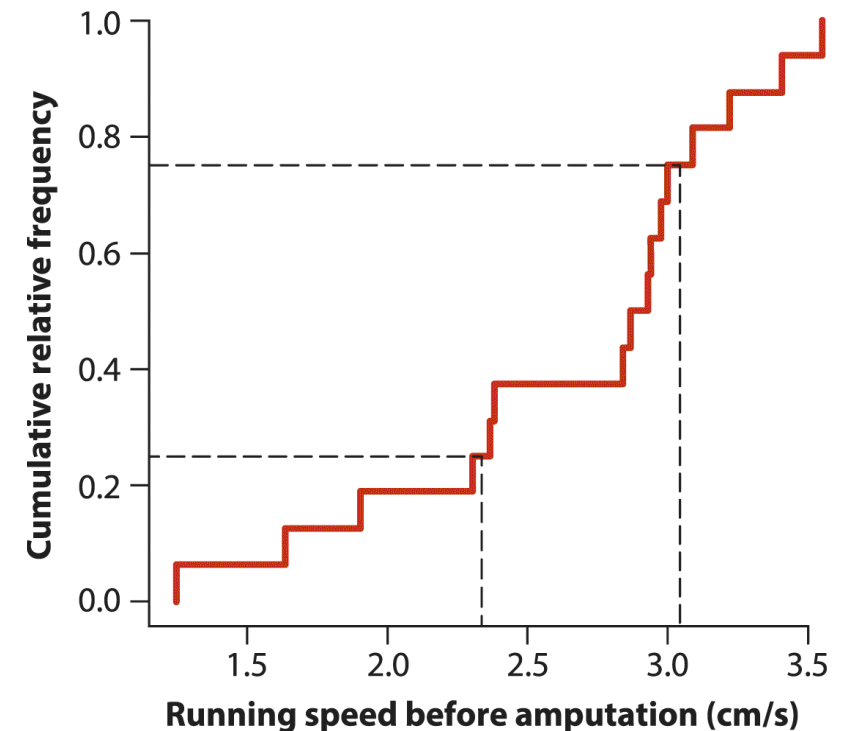
- 把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值
- 第1和第3四分位数 (quartiles) 分别是第0.25和第0.75分位数 (quantiles)
- 第2四分位数是中值

3.4 累积频率分布 Cumulative frequency distribution

- 累积频率分布 Cumulative frequency distribution (CFD)
 - 用于比较频数分布的形状和位置的另一种方法;
 - 数值变量的所有分位数可以通过累积频率分布来显示;
 - 简单来说, 它告诉我们有多少数据点的值小于或等于某个特定值;

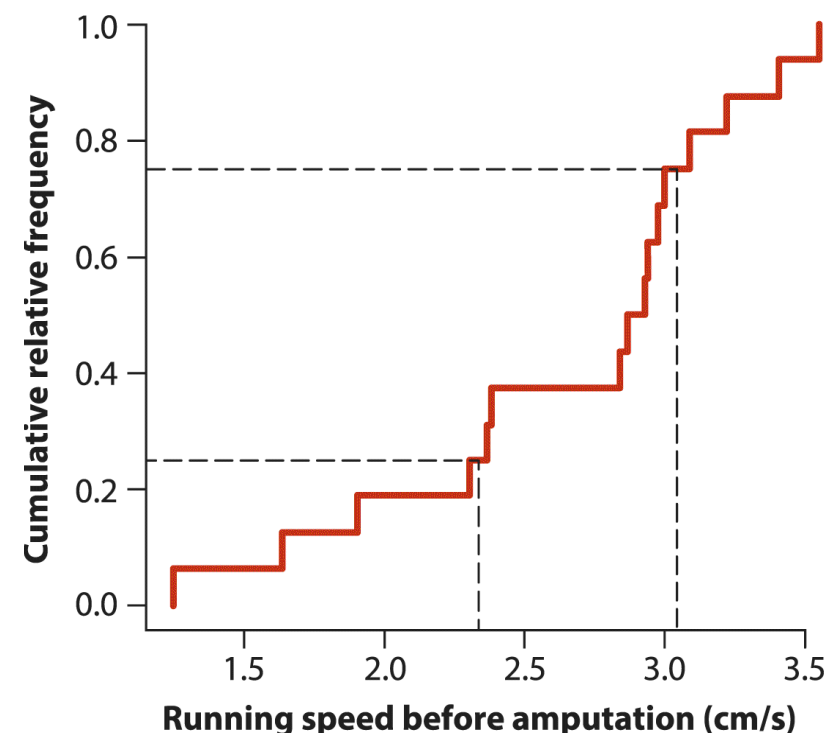
3.4 累积频率分布 Cumulative frequency distribution

- Displaying CFD
 - 1. sort all data from the smallest to the largest
把数据从小到大排序
 - 2. calculate the **fraction** of observations less than or equal to each data value
计算样本中小于或等于某一个值的**比例**
 - 3. draw the distribution 绘制分布图
 - the fraction = the cumulative relative frequency
 - 累积相对频率: 曲线上该值对应的高度



3.4 累积频率分布 Cumulative frequency distribution

- Displaying CFD 展示
 - Dashed lines 虚线
 - 1/4 observations < 2.34 (1st quartile)
 - 3/4 observations < 3.045 (3rd quartile)
- Compare 比较不同的图像展示
 - 通过这个分布，我们可以清楚看到数据的分布情况和百分比的变化；
 - 有助于比较不同组之间的分布差异；

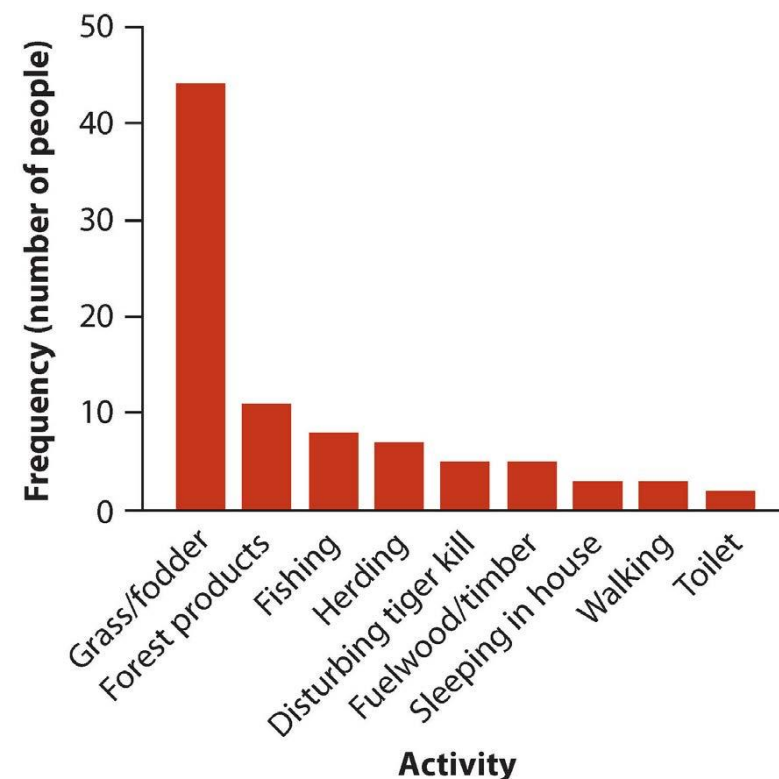


3.5 比例 Proportion

- 针对类型变量 (categorical variable) 最重要的一个描述性统计量;
- Calculation 计算

$$\hat{p} = \frac{n_{category}}{n}$$

- The proportion is like a sample mean???



3.5 比例 Proportion

- 针对类型变量 (categorical variable) 最重要的一个描述性统计量;
- Calculation 计算

$$\hat{p} = \frac{n_{category}}{n} = \frac{\sum_{i=1}^n \text{类别中的观测值}(0 \text{ or } 1)}{n} = \text{Mean}$$

- The proportion is like a sample mean???
- 在比例的计算中，某个类别的观测数可以看作是1，其他类别的观测数为0。因此，比例是这些“0”和“1”的均值；所以从数学的角度来看，比例的计算和样本均值的计算是类似的。

4. 总结 Summary

- 均值或中值可以度量样本数值变量分布的居中位置
 - The location of a distribution for a numerical variable can be measured by its mean or by its median.
 - 均值 (mean) 给出了分布的重心; 中位数 (median) 给出了中间值。
- 标准差可以度量样本数值变量分布的散布程度
 - The standard deviation measures the spread of a distribution for a numerical variable.
 - 它是观测值与均值之间距离差异的度量;
 - 方差 (variance) 是标准差 (standard deviation) 的平方;

4. 总结 Summary

- 四分位数将样本观测值排序后分为四等分, 其中四分位距IQR也度量了数据的散布程度 (第一和第三四分位数间的范围) ;
- 样本的所有分位数可以用累计频率分布图来表示;
- 比例是类型数据最重要的描述性统计量;
 - 通过某一类别中的观测数量除以所有类别的总观测数量来计算;

R references

- <https://whitlockschluter3e.zoology.ubc.ca>

Chapter links

R lab

R code for book examples

Data

Data for examples

Data for problem sets

Data for examples

Example 2.2A. Deaths from tigers

Gurunga, B., et al. 2008. *Biological Conservation* 141: 3069–3078.

Example 2.2B. Effects of Zika virus

Brasil, P., et al. 2016. *New England Journal of Medicine* 375: 2321–2334.

Figure 2.2-5. Salmon body mass

Data for problem sets

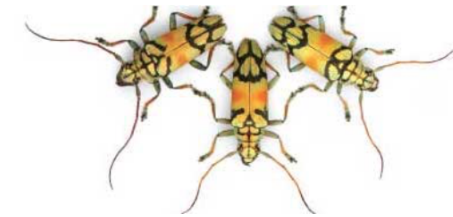
05. Fish fry survival

Miller, L. M., T. Close, and A. R. Kapuscinski. 2004. *Molecular Ecology*

06. Endangered species

U.S. Fish & Wildlife Service. 2018. Listed Animals. U.S. Fish & Wildlife Online System. [Data link](#). Accessed August 27, 2018.

Data & resources by chapter



Resources for *The Analysis of Biological Data*

Welcome to the resource pages for the *The Analysis of Biological Data*, 3rd edition, by Michael Whitlock and Dolph Schluter. The book is an introduction to statistics for biologists, available from Macmillan [here](#).

Below, you'll find links to [data sets](#) and [other resources for each chapter](#).

On these pages, you will find a variety of learning resources, including:

- **R labs:** Learn basic statistical analyses and core concepts using the statistical package R.
- **R code for examples:** We used R to analyze all examples in the book. We put the code here so that you can too.
- **Interactive visualizations:** Concept visualizations to develop intuition about some of the trickier concepts in statistics.
- **Data sets:** Download a .zip file with all data sets in the book [here](#).

R references

- https://bookdown.org/qiyuandong/_book/
- https://www.math.pku.edu.cn/teachers/lidf/docs/Rbook/html/_Rbook/index.html
- <https://bookdown.org/hezhijian/book/>

零基础学R语言

1 前言

1.1 R 的前世今生

1.2 R 的安装

2 Basics

2.1 R的数学运算:

2.2 赋值给变量

2.3 R 的基本数据类型

3 向量Vectors

3.1 创建一个向量 Create a vector (2)

3.2 Create a vector (3)

3.3 Naming a vector

3.4 Naming a vector (2)

3.5 Calculating total winnings

3.6 Calculating total winnings (2)

3.7 Calculating total winnings (3)

3.8 Comparing total winnings

3.9 Vector selection: the good times

3.10 Vector selection: the good time...

R 语言入门，给一心只有学习的你

Chris Qi from Data Maniac

2018-09-06

Chapter 1 前言

想直接上手的同学，可以跳过这一部分，从安装软件开始。如果软件已经安装了，可以跳到第二章。对于喜欢把书从头读到末的同学，欢迎从这里开始。

1.1 R 的前世今生

看到这个题目，你以为我会跟你絮絮叨叨讲一个软件的发展史？这种东西听一耳朵就可以了，写出来都浪费纸墨，噢，这是电子书，不用纸也不用墨，但是打字也费劲儿呀。所以在这里，我就做个大概介绍吧：

R是一门用于统计计算和作图的语言，由S语言发展而来，以统计分析功能见长。

- S语言由贝尔实验室1976年开发，是一个内部使用的统计分析工具。
- R是新西兰的罗斯·伊哈卡 (Ross Ihaka)和罗伯特·金特elman (Robert Gentleman) 基于S语言开发，1993年面世。1995年采用通用公共许可协议，使之成为免费软件。

前言

I 介绍

1 R语言介绍

2 R语言入门运行样例

II 数据类型与相应的运算

3 常量与变量

4 数值型向量及其运算

5 逻辑型向量及其运算

6 字符型数据及其处理

7 R向量下标和子集

8 R数据类型的性质

9 R日期时间

10 R因子类型

11 列表类型

12 R矩阵和数组

13 数据框

14 工作空间和变量赋值

III 编程

15 R输入输出

R语言教程

李东风

2023-07-27

前言

这是李东风开设《统计软件》等课程的讲义， 也可以用作《数据科学》的入门教材。本书使用了其它教材的例子和讲法， 仅供学生内部使用， 不是公开出版图书。鉴于本人水平有限， 错漏之处难免， 欢迎指出错误或提出改进意见。

尽管本书中包含假设检验、回归分析等统计方法内容， 但是侧重点是这些方法的使用， 不能当作统计学的入门教材使用。

相关下载：

- [Rbook-data.zip](#) : 一些配套数据的打包文件
- [bookdown-template-v0-6.zip](#) : R Markdown和bookdown的模板

网页版的书中数学公式使用MathJax库显示， 下面是数学公式测试。 如果数学公式显示不正常， 在浏览器中用鼠标右键单击公式， 在弹出的菜单中选择 Math Settings--Math Renderer 选HTML-CSS或SVG即可。

R Lab

- R语言

- R是一门用于统计计算和作图的语言，1993年面世，1997年成为GNU计划；
- R 是新西兰奥克兰大学的Ross Ihaka和Robert Gentleman基于S语言开发（S语言由贝尔实验室1976年开发，是一个内部使用的统计分析工具）；
- 在统计系，R语言几乎是一门必修的语言；
- 免费，开源！

- 软件下载和安装：R & Rstudio

- RStudio 是一个集成开发环境 (IDE)，专门为 R 编程语言设计；提供了一个图形用户界面，使得编写、运行和调试 R 代码更加方便；
- 它包括一个代码编辑器、工作区管理工具、控制台窗口、图形输出窗口和帮助文档等；其它功能还包括代码自动补齐、颜色区分、方便查看数据集等；

R Lab

R version 4.3.1 (2023-06-16) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

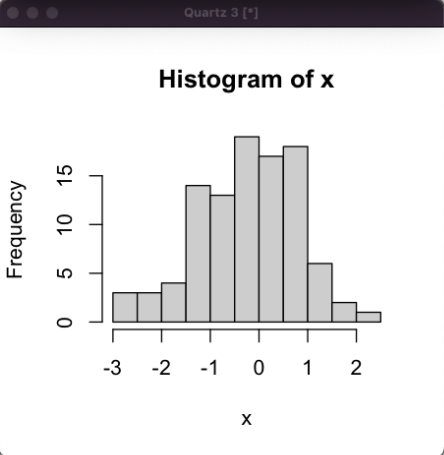
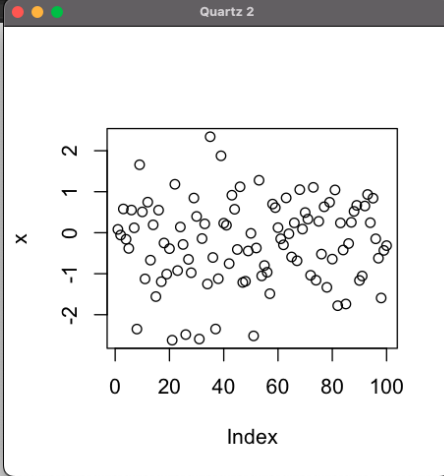
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.79 (8238) x86_64-apple-darwin20]

```
> x = rnorm(n = 100, mean = 0, sd = 1)
> plot(x)
> quartz()
> hist(x)
>
```



Biostats_ECNU - main - RStudio

```
1 x = rnorm(n = 100, mean = 0, sd = 1)
2 hist(x)
3
```

Environment History Connections Git Tutorial
R Global Environment 146 MiB

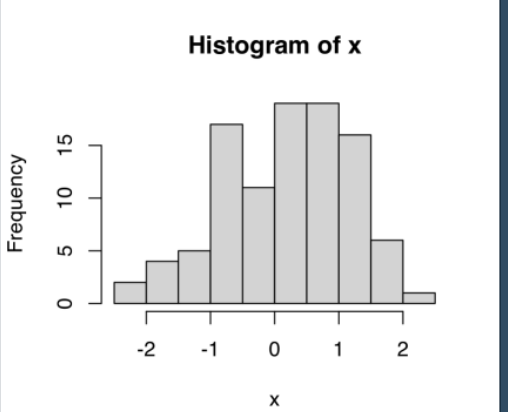
Values
x num [1:100] -0.725 ...

Files Plots Packages Help Viewer Presentation
Zoom Export

Console Terminal Background Jobs
R 4.3.1 ~/Mirror/Teaching/Biostats_ECNU/

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

```
> x = rnorm(n = 100, mean = 0, sd = 1)
> plot(x)
> hist(x)
>
```



R Lab

- R语言
- 软件下载和安装
 - R <https://www.r-project.org/>
 - Rstudio <https://posit.co/download/rstudio-desktop/>
- R的数据对象：character, vector, matrix, list, function, etc.
- R的数据导入和导出
 - 有的可能要借助特定的软件包 package，例如openxls;