

Lecture 3 – 描述数据 Describing data

- Outline for today
 - Recall L02 – Graphics 图像
 - Descriptive statistics 描述性统计量
 - Measurements & Comparison 测量和比较
 - Summary
 - R Lab

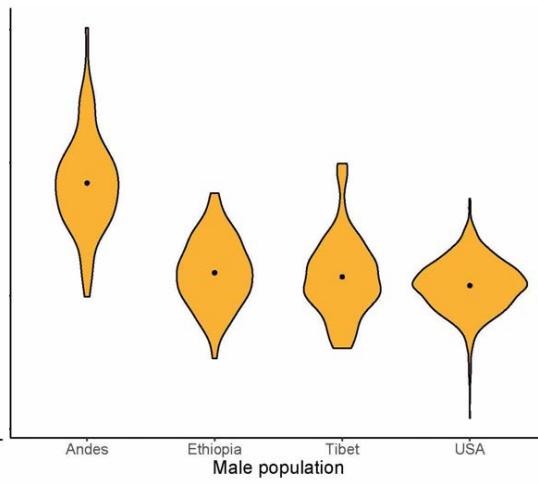
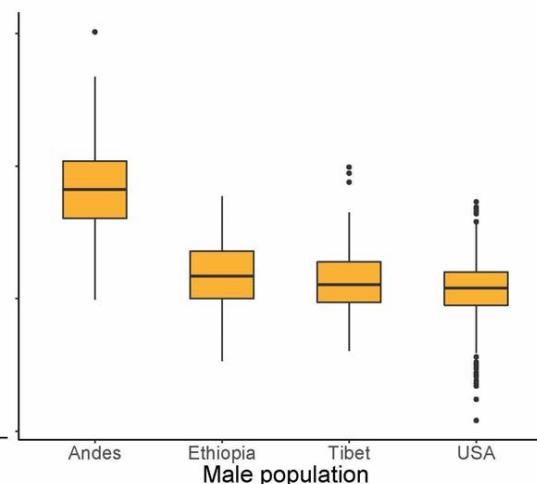
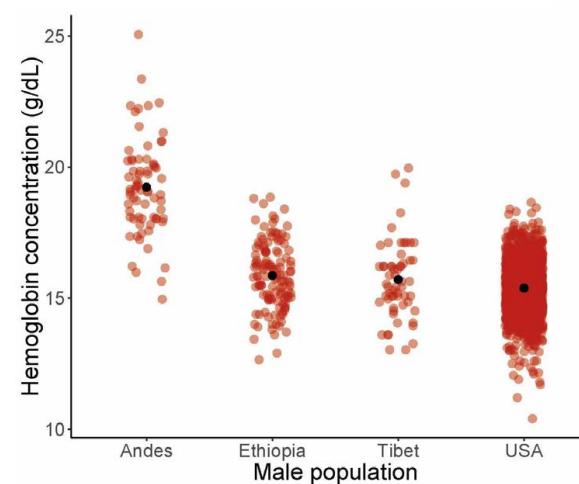
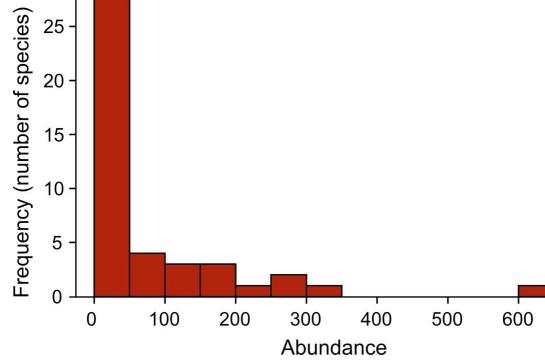
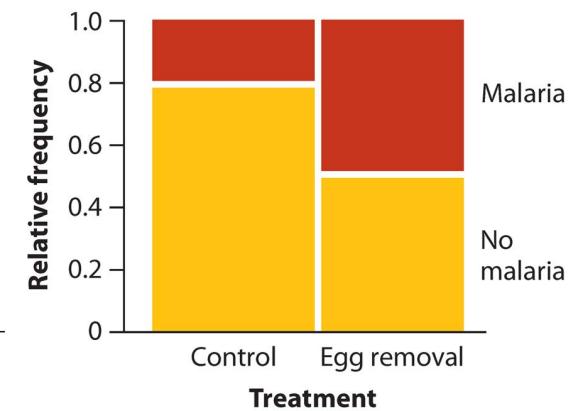
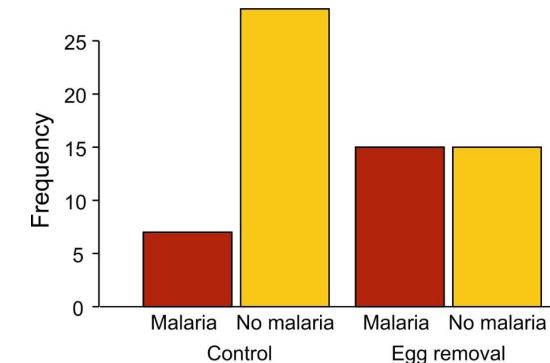
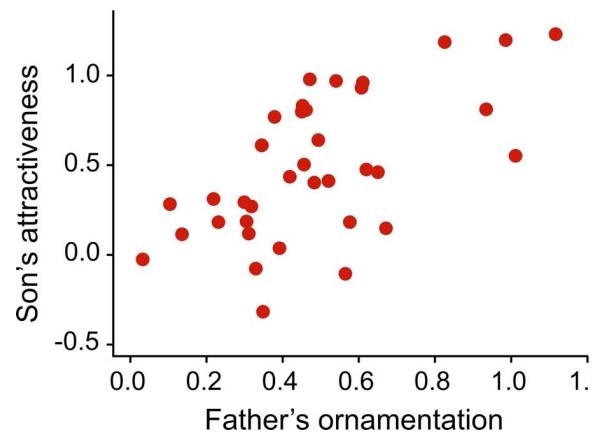
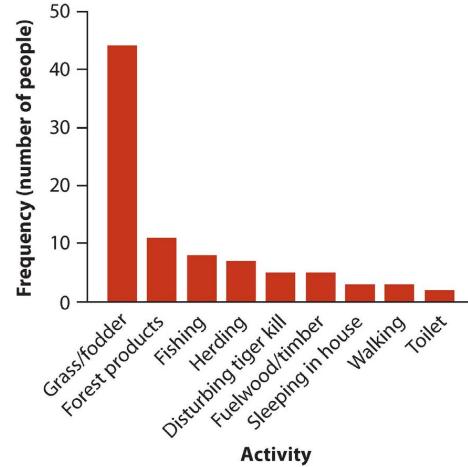
生物统计学

李 勤

生态与环境科学学院

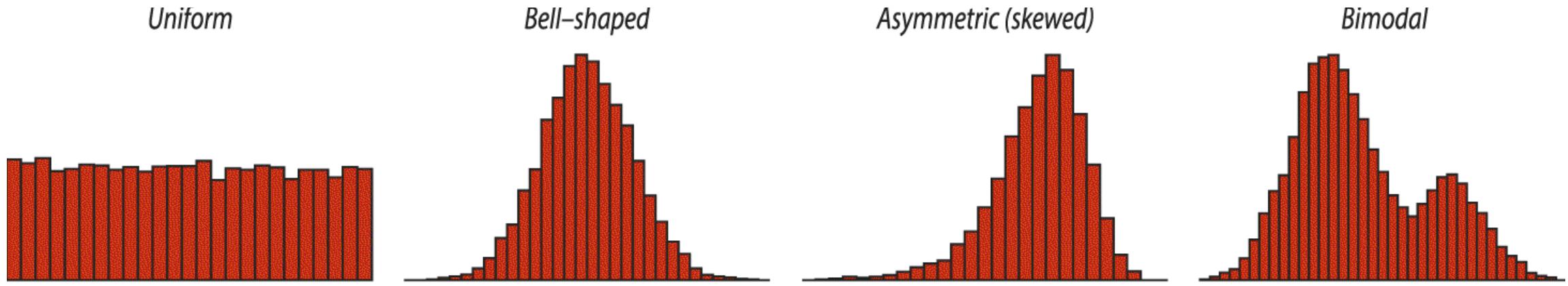
1. Recall L02 - Graphics

Types of data 数据类型	Graphical method 图像类型
One categorical variable 单个分类变量	Bar plot 柱状图
One numerical variable 单个数值变量	Histogram 直方图
Two numerical variables 两个数值变量	Scatter plot 散点图
Two categorical variables 两个分类变量	Grouped bar graph 分组柱状图 Mosaic plot 马赛克图
One numerical variable and one categorical variable 单个数值变量 ~ 单个分类变量	Strip chart 条形图 Boxplot 箱形图 Violin plot 小提琴图 Multiple histograms 分组直方图



1. Recall L02 - Graphics

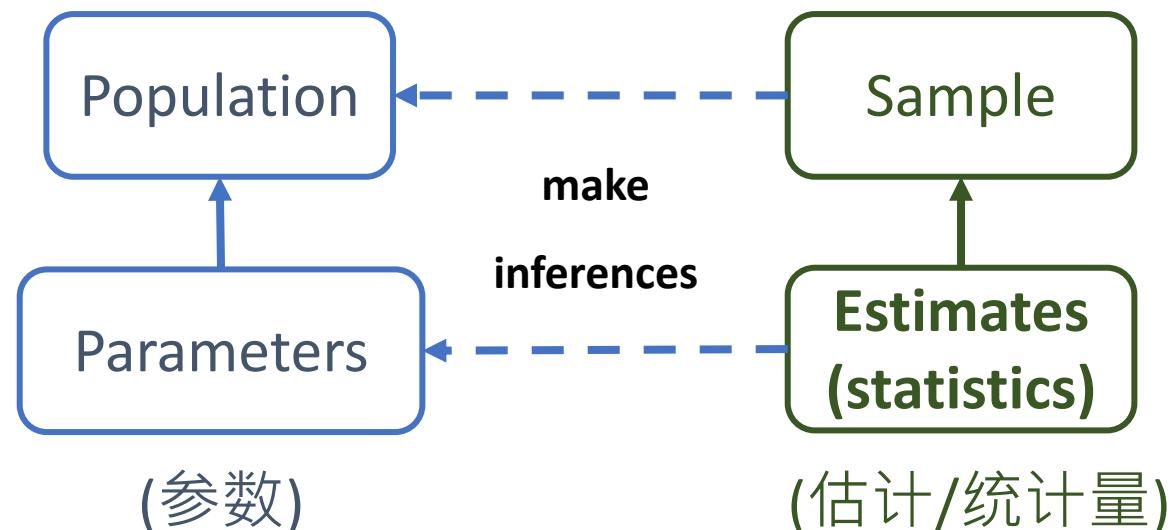
- The shape of a histogram (直方图的形状)
 - The peak number (峰的数量)
 - Symmetry (对称性)



(Whitlock & Schluter 2020)

1. Recall 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



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**)；

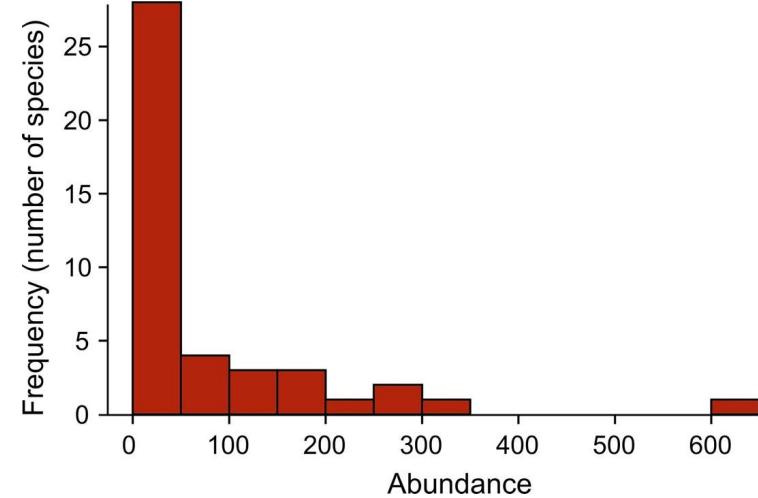
2. 描述性统计量

- Location (位置)

- 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)

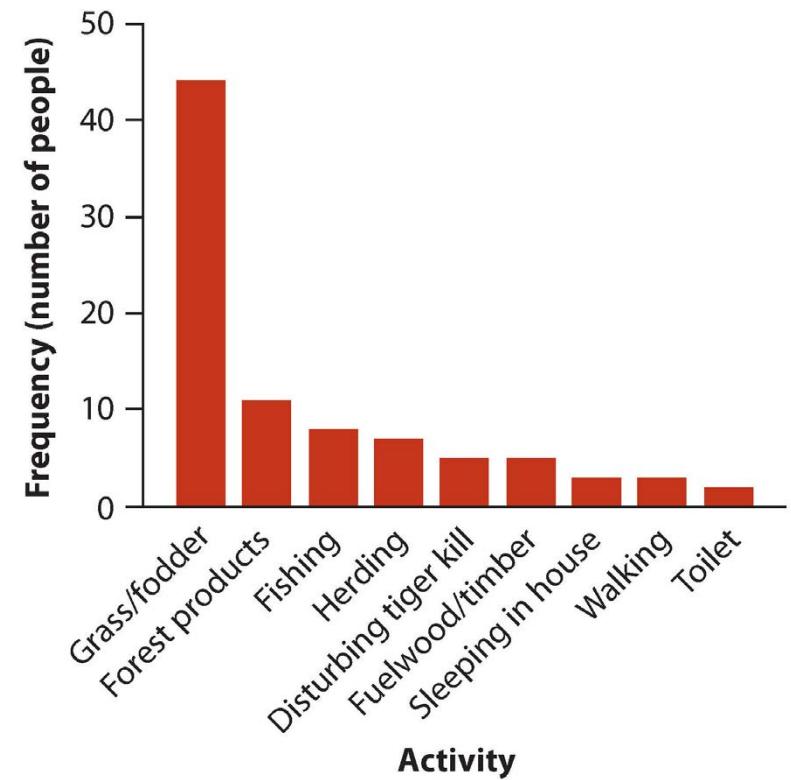
- Spread (散布程度)

- 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

- Proportion (比例)
 - 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

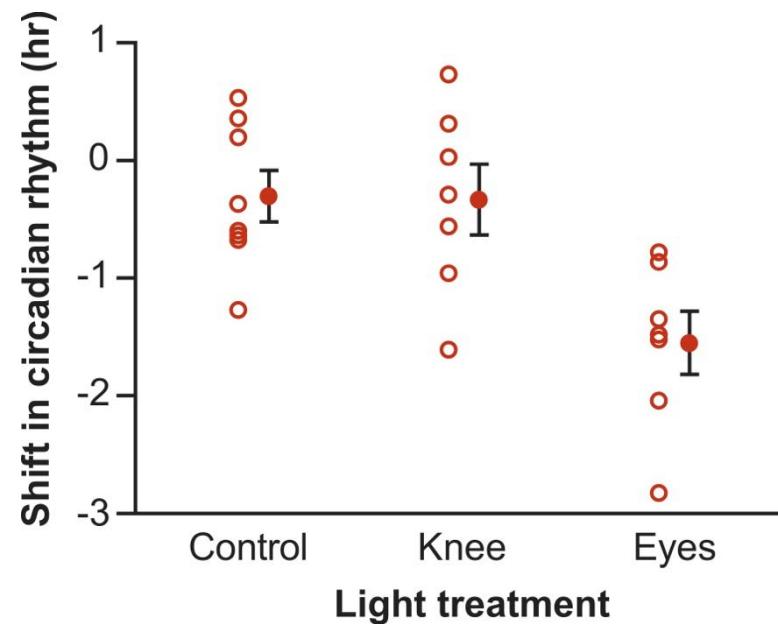


3. 描述性统计量的测量

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

3.1 算术平均值和标准差

- Arithmetic mean 算术平均值
 - 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。
 - Standard deviation 标准差
 - 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

- The sample mean 样本均值

- 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

- The sample mean 样本均值

- 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

$$\begin{aligned}Y_1 &= 0.9 \\Y_2 &= 1.4 \\Y_3 &= 1.2 \\Y_4 &= 1.2 \\Y_5 &= 1.3 \\Y_6 &= 2.0 \\Y_7 &= 1.4 \\Y_8 &= 1.6\end{aligned}$$

- 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。
- 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。

3.1 算术平均值和标准差

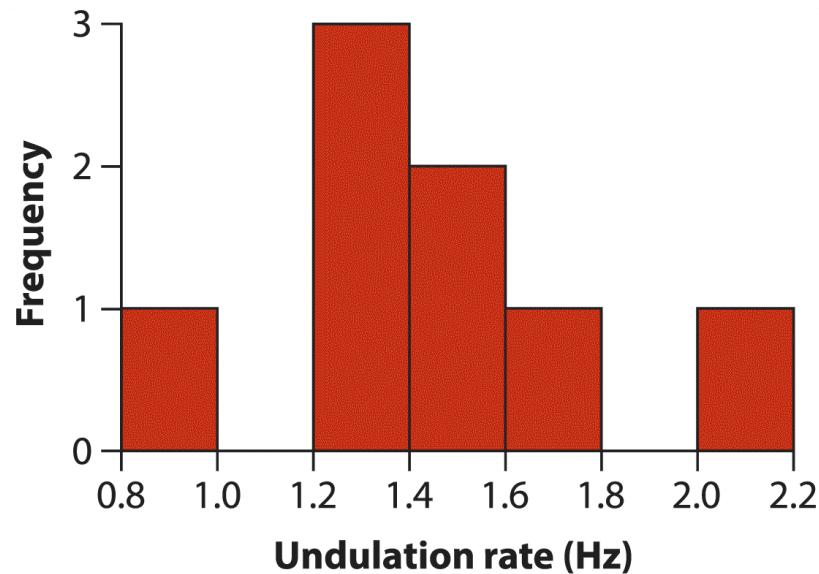
- The sample mean 样本均值

- 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

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



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

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

3.1 算术平均值和标准差

- Deviation 离均差

$$Y_i - \bar{Y}$$

- Squared deviation 离均差平方

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

- Variance 方差

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

- Standard deviation 标准差

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

\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)



3.1 算术平均值和标准差

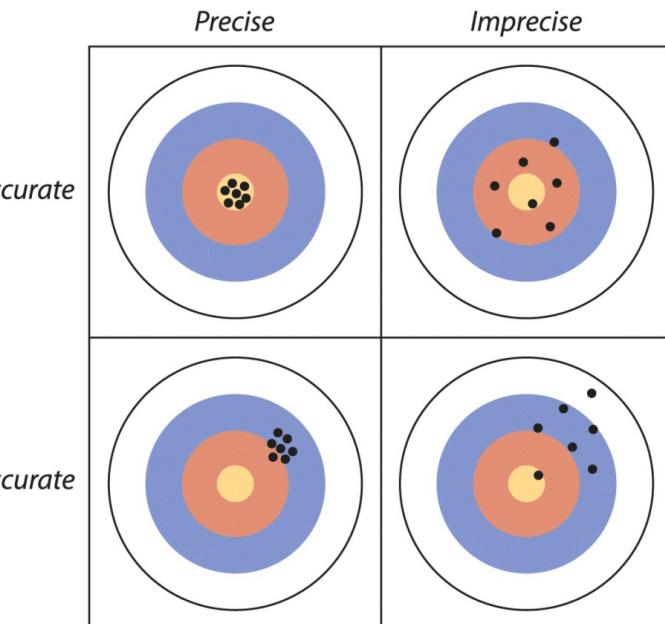
- Variance 方差

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

- Why $n - 1$?

- Standard deviation 标准差

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





3.1 算术平均值和标准差

- Variance 方差

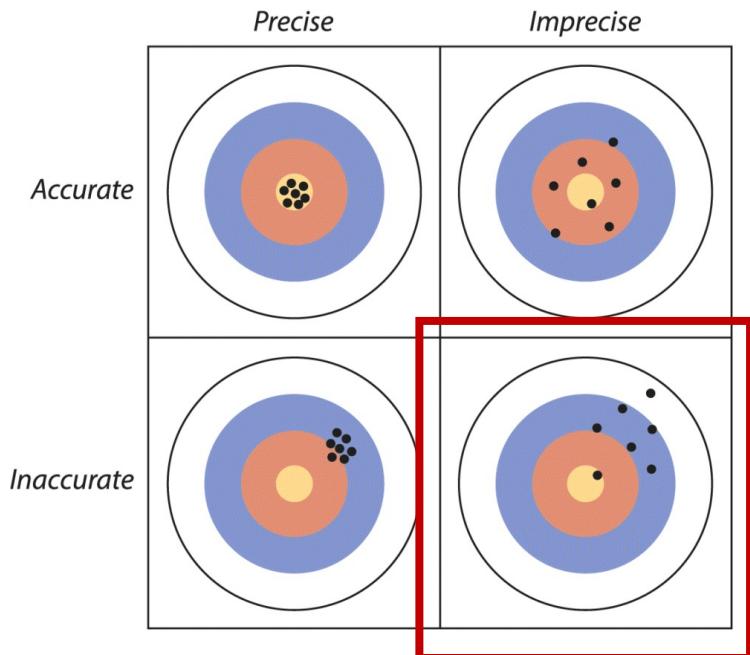
$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

- Standard deviation 标准差

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

- 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！

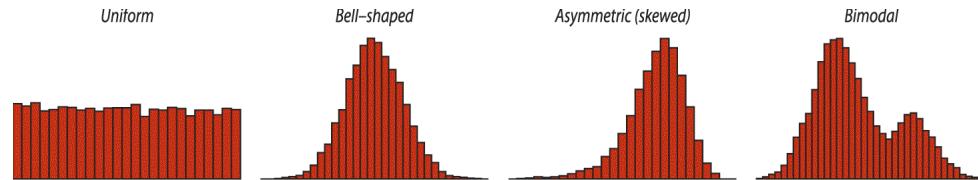
这是因为：

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

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

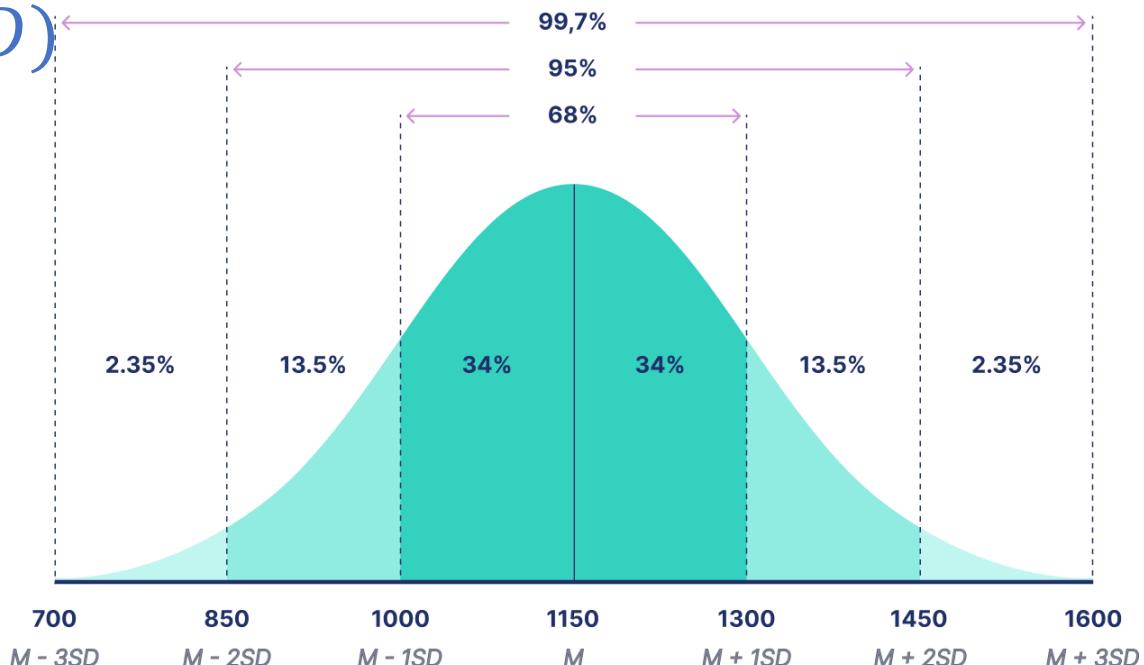


- 关联: 标准差 standard deviation (SD) ~ 频数分布 frequency distribution
 - 如果频数分布是钟形/正态分布 (bell shaped/a **normal distribution**)
 - 约有2/3的观测值将落在平均值两侧各1个标准差内:

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

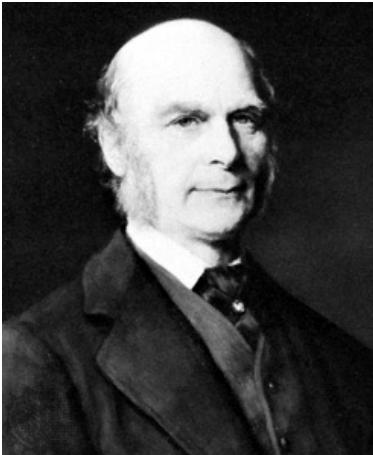
- 约 95% 的观测值将落在2个SD内:

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

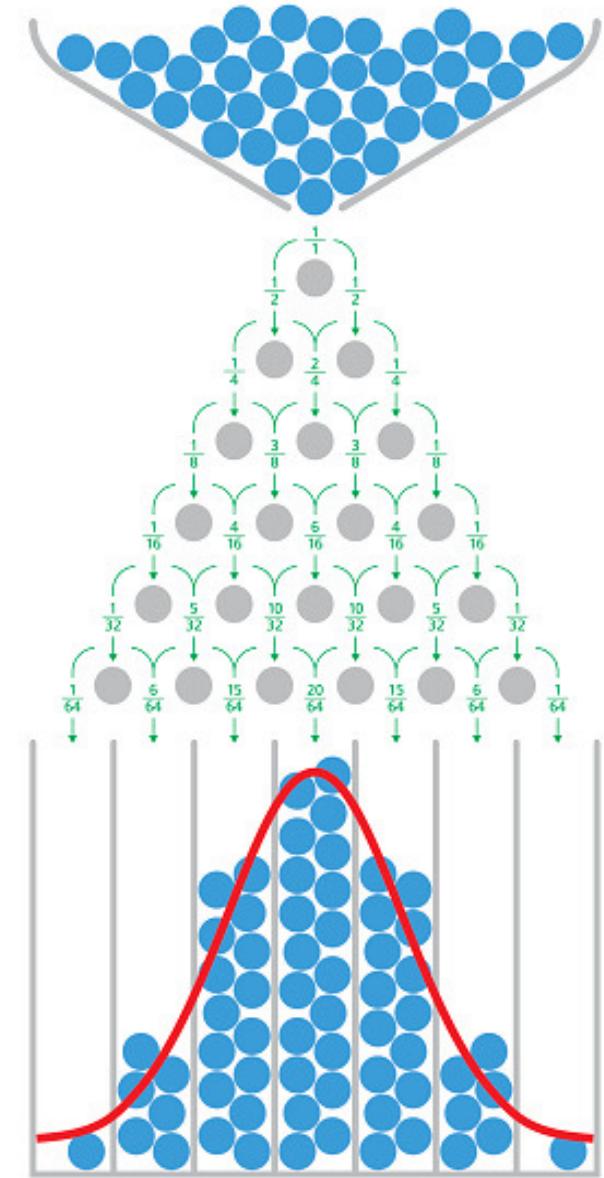


Normal distribution 正态分布

- 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



3.1 算术平均值和标准差

- Mean

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

- SD

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

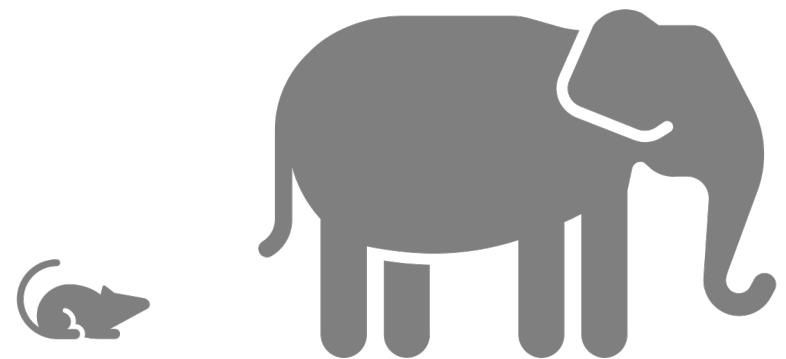
- Coefficient of variation (CV) 变异系数

- 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

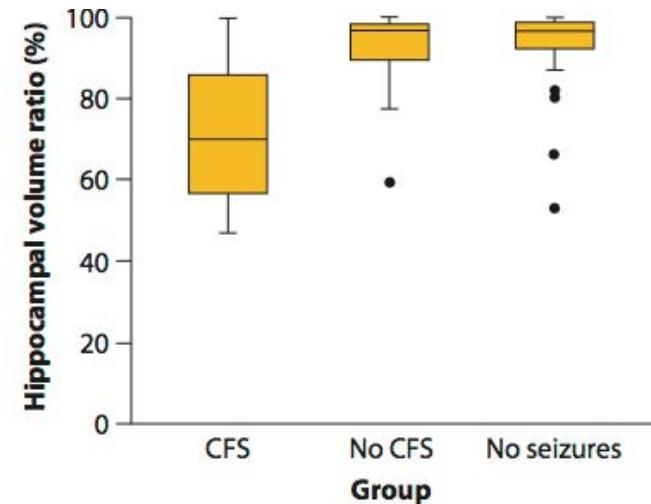
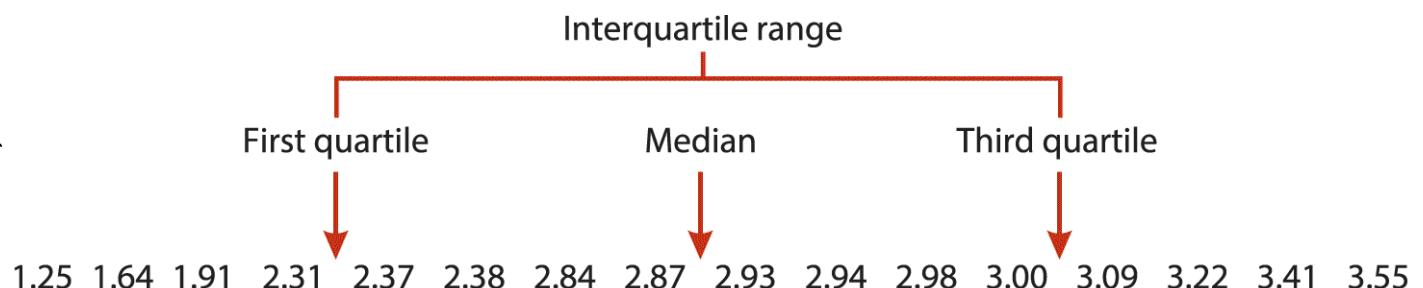
- Median 中值

- 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

- Interquartile range (IQR) 四分位距

- 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

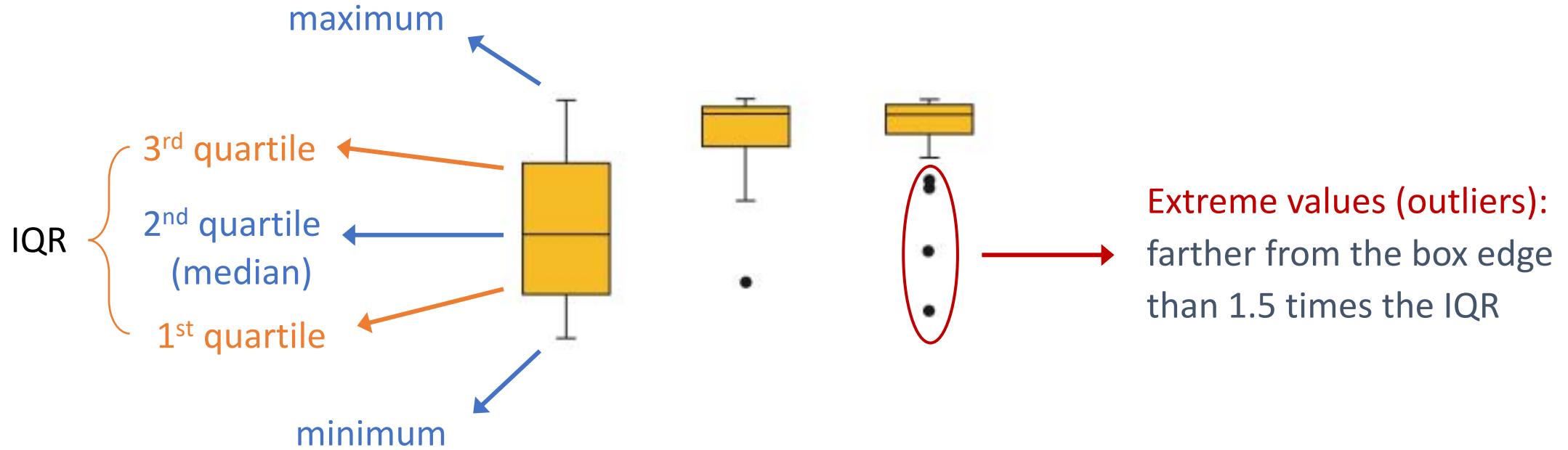


$$Y_{[n+1]/2}$$

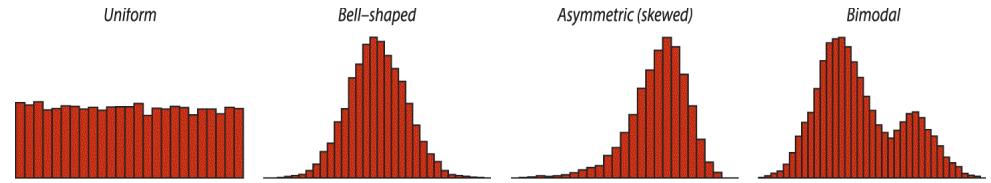
$$(Y_{[n/2]} + Y_{[n/2+1]})/2$$

3.2 中值和四分位距

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



3.3 和比较测量值



- Location (位置)
 - a central value: mean 均值 vs. median 中值
- How to compare? 如何比较?
 - 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)

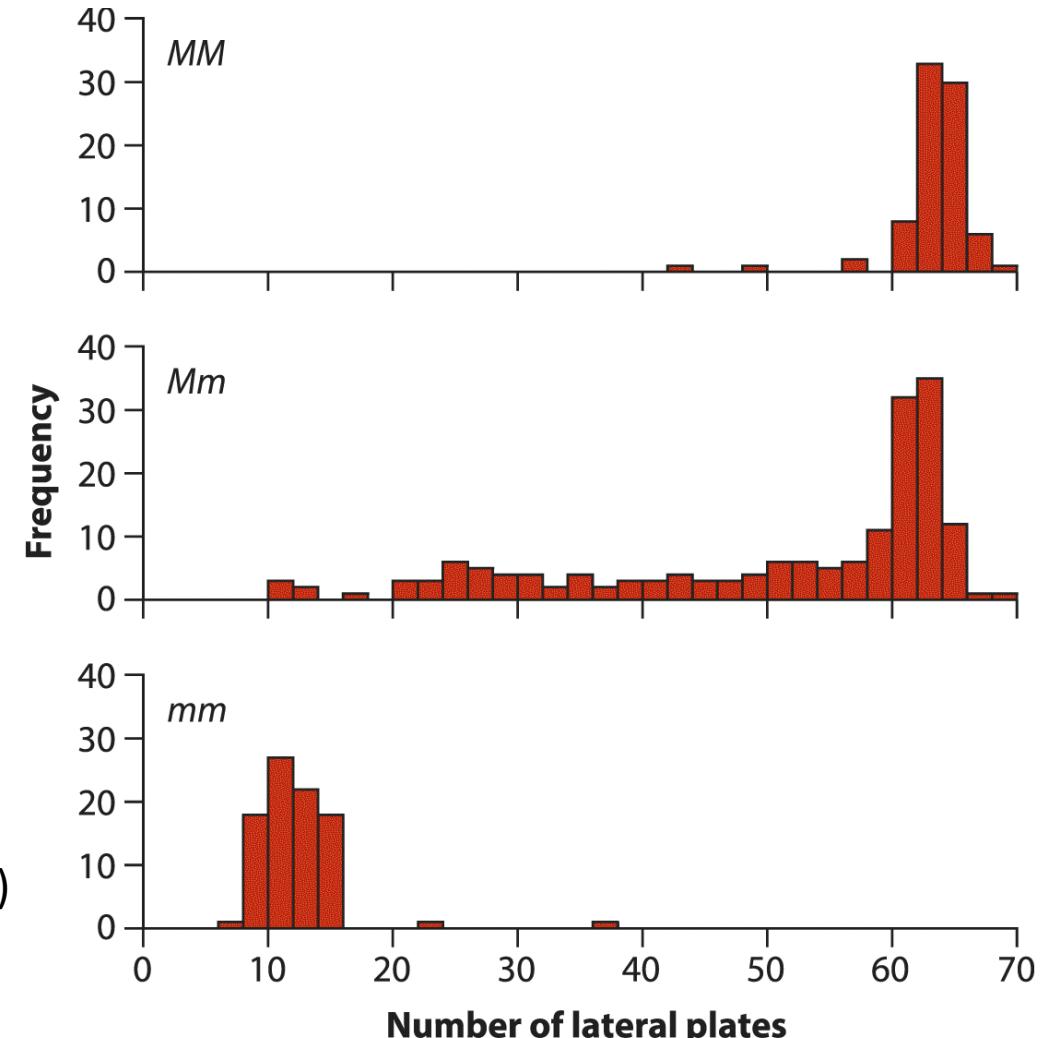
比较居中位置 the central value



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



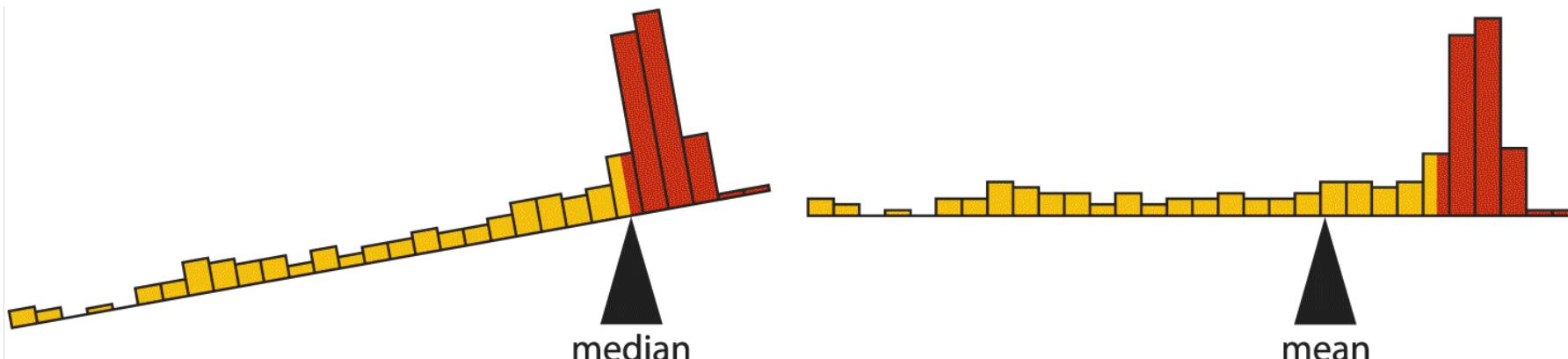
Threespine sticklebacks (三刺鱼)
reproduced with permission
from K. B. Marchinko and D.
Schluter (2007) [Evolution
61:1084–90, Wiley-Blackwell
Publishing Ltd.].



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;



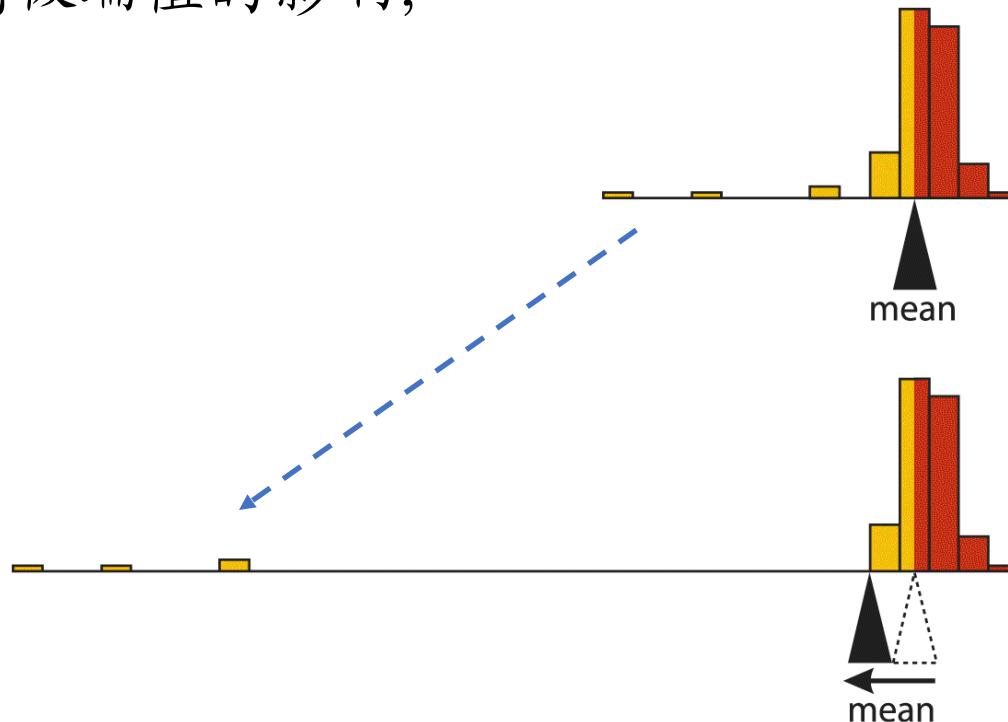
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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;

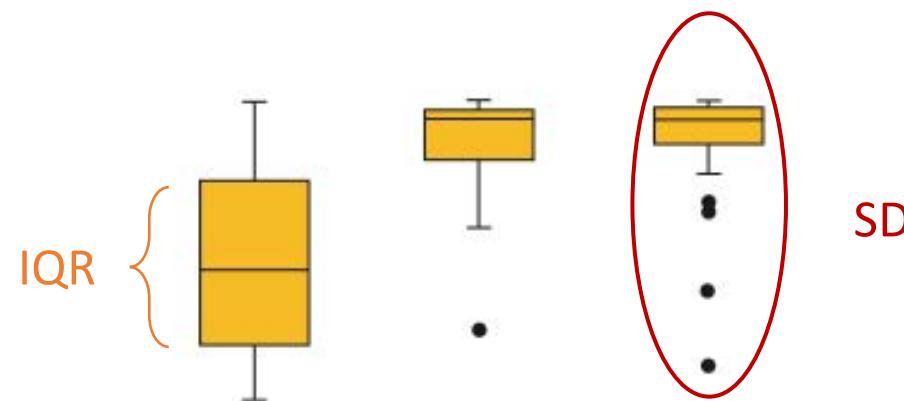
The balancing act

→ 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;

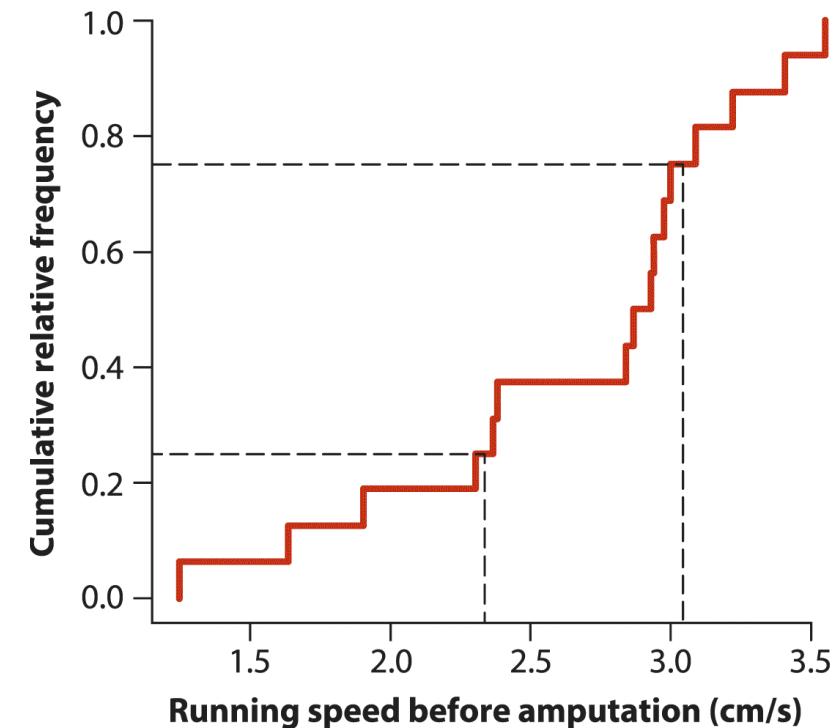


3.4 Cumulative frequency distribution 累积频率分布

- Percentile 百分位数
 - 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.
- Quantile 分位数
 - the 10^{th} percentile is the 0.10 quantile (第10个百分位数是0.01分位数)
 - 第1和第3四分位数 (quartiles) 分别是第0.25和第0.75分位数 (quantiles);
- Cumulative frequency distribution (CFD) 累积频率分布
 - 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;

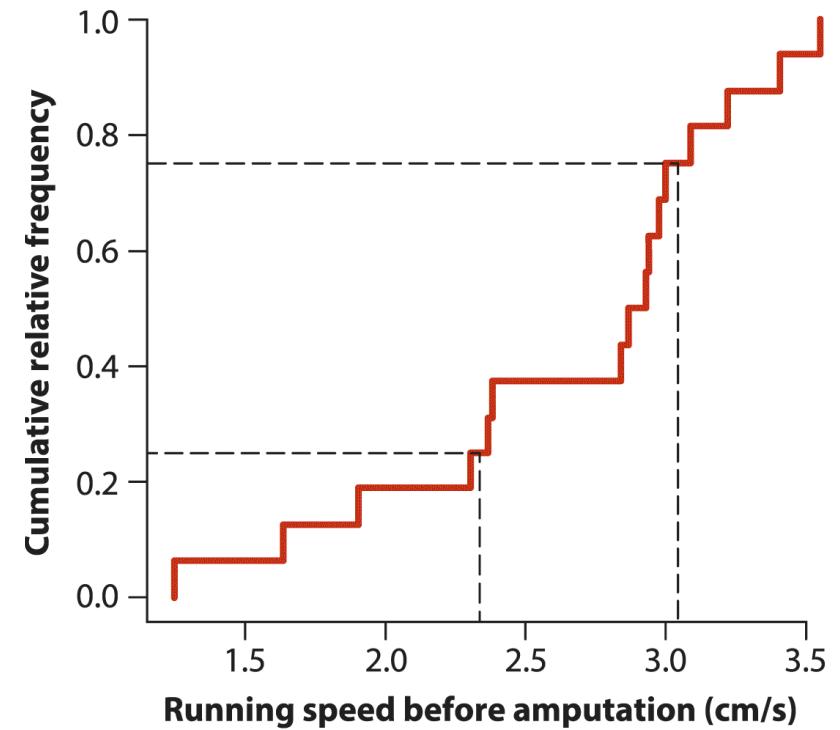
3.4 Cumulative frequency distribution 累积频率分布

- 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)
 - the fraction = the cumulative relative frequency
 - 累积相对频率: 曲线上该值对应的高度
 - indicated by the height of the curve at the corresponding data value



3.4 Cumulative frequency distribution 累积频率分布

- 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
 - box plot & histogram > CFD
 - But, CFD can be useful when comparing multiple groups (比较多组数据时, CFD可能会比较有用/直接);

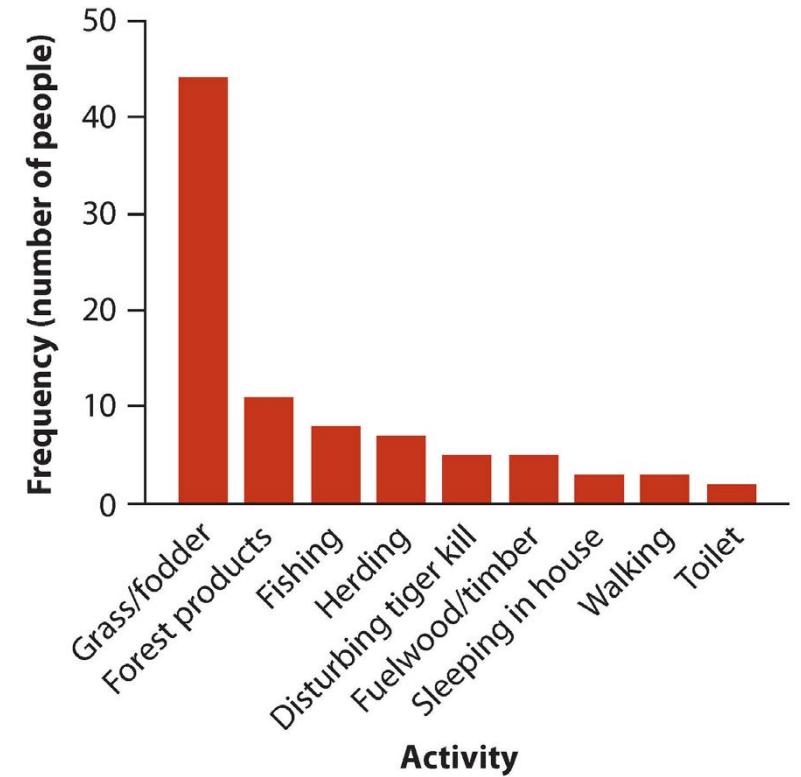


3.5 Proportion 比例

- 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???



4. Summary 总结

- 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. Summary 总结

- 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；

作业/考试范例

- 下列（ ）可以描述样本数据的变异程度。
 - A. P值, B. 样本方差, C. 标准误, D. 95%置信区间
- 假设xx(数据)服从正态分布，你随机抽样了10个个体，得到的数据分别为{44, 50, 55, ...}，请估计xx的平均值为（ ），及其95%置信区间（ ）。
- 题目中给出了具体的某一项研究，包括其研究目的（和要检验的假设）、数据采集方法、样本大小和数据类型，回答以下内容：
 - 研究的类型；通过绘图（手绘）来回答实验设置及可能的结果；
 - 假设检验的统计结果及生物学解释；

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

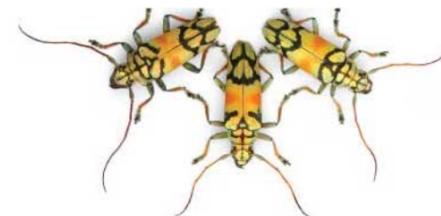
05. Fish fry survival

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

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...

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。