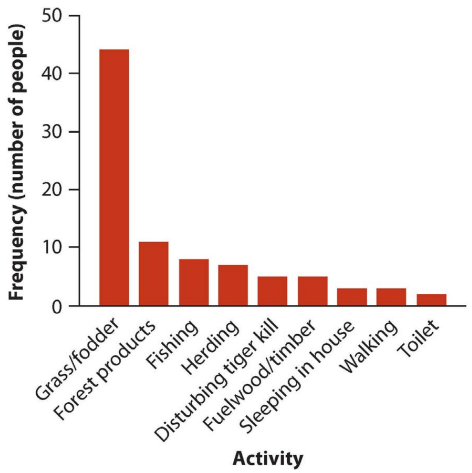
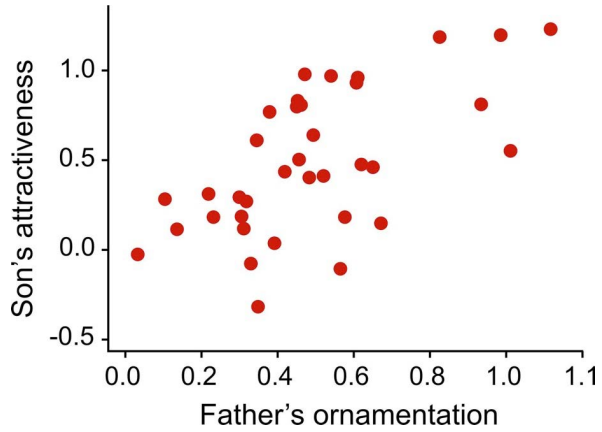
Lecture 3 – Describing data

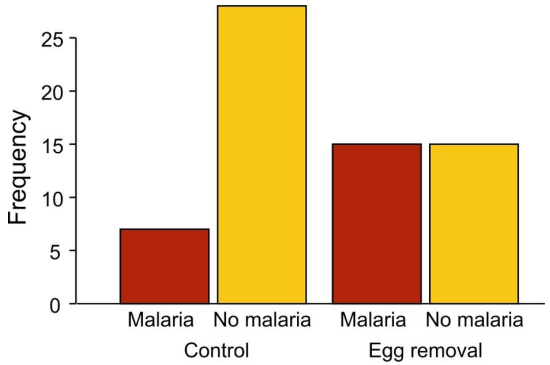
- Outline for today
  - Recall L02 – Graphics
  - Descriptive statistics
    - Measurements & Comparison
  - Summary
  - R Lab

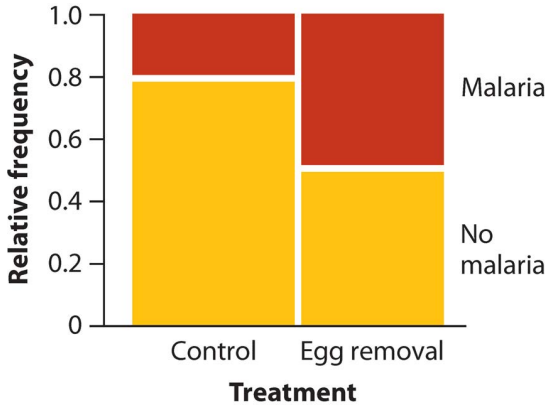
1. Recall L02 - Graphics

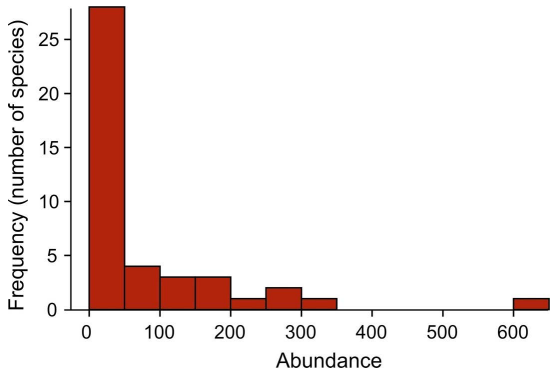
Types of data 数据类型	**Graphical method** 图像类型
One categorical variable 单个分类变量	Bar plot 柱状图
One numerical variable 单个数值变量	Histogram 直方图
Two numerical variables 两个数值变量	Scatter plot 散点图
Two categorical variables 两个分类变量	Grouped bar graph 分组柱状图 Mosaic plot 马赛克图
One numerical variable and one categorical variable 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 多组直方图

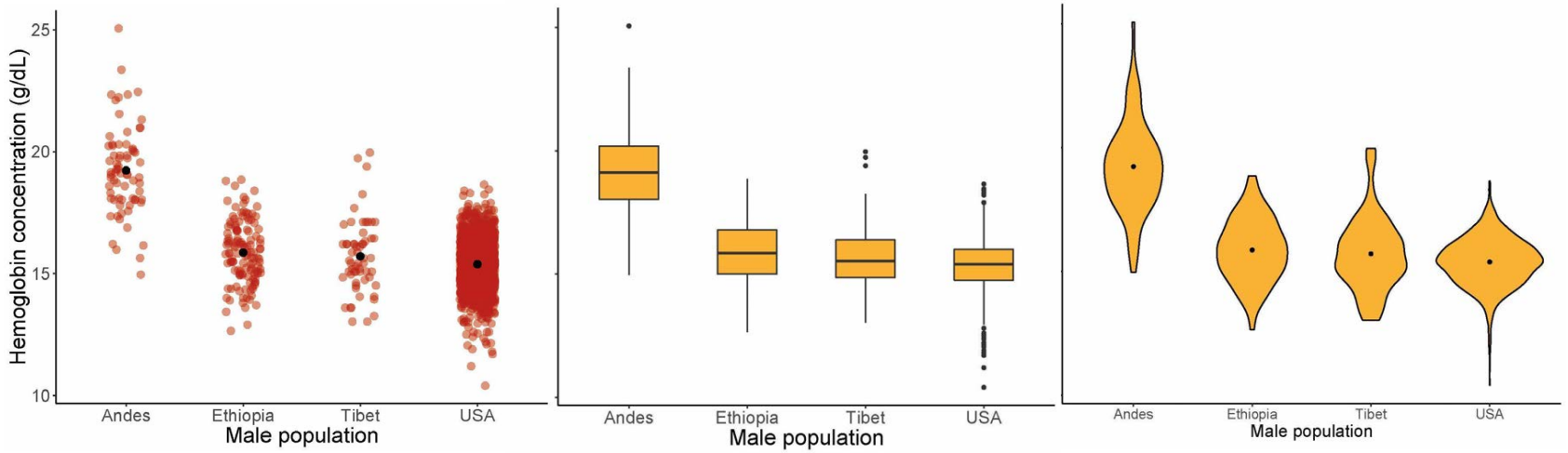








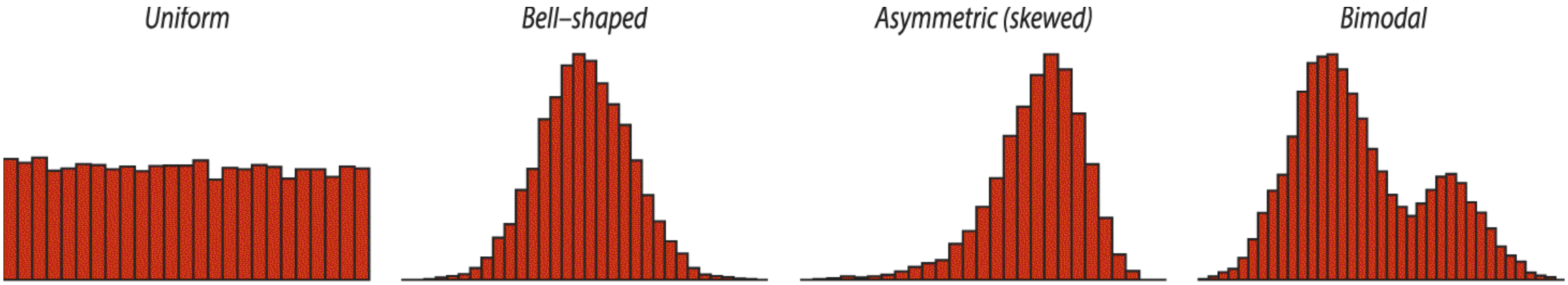




(Whitlock & Schluter 2020)

1. Recall L02 - Graphics

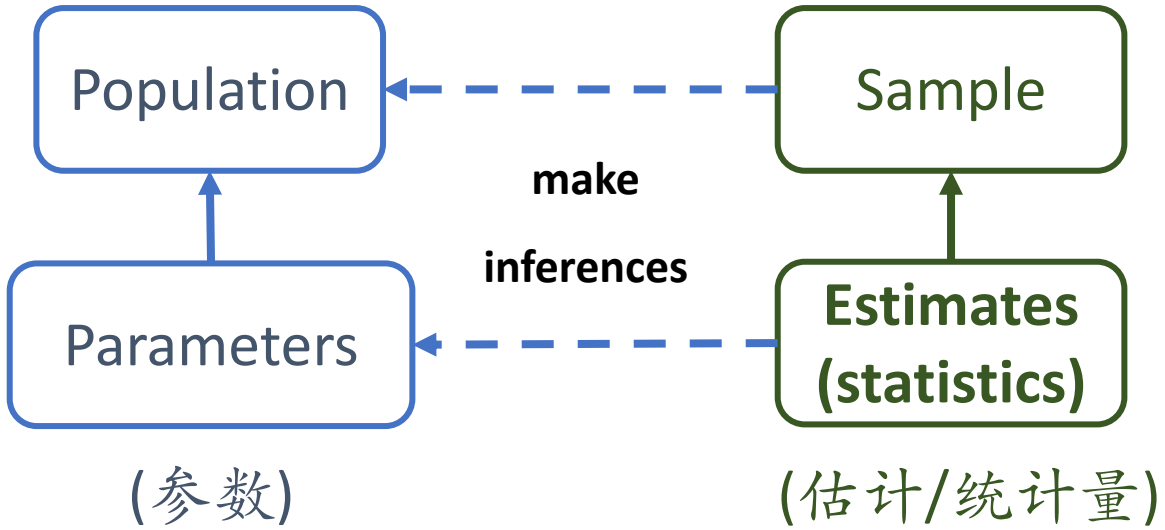
- The shape of a histogram (直方图的形状)
  - The peak number (峰的数量)
  - Symmetry (对称性)



(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. Descriptive statistics

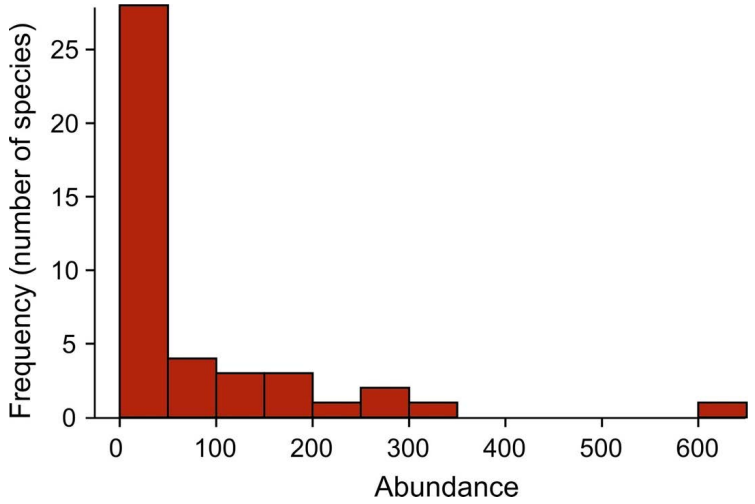
- **Descriptive statistics (描述性统计量)**, or summary statistics
  - are **quantities** that capture important features of frequency distributions.

    (频数分布的量化特征)
  - Whereas graphs reveal shapes and patterns in the data, descriptive statistics provide hard numbers (确切的数字).
    - For numerical data, the most important descriptive statistics are those measuring the **location** of a frequency distribution and its **spread**.

      (对数值变量而言，最重要的统计量包括其位置和散布程度)
    - For categorical data, the most important descriptive statistic is the **proportion**.

      (对类型变量而言，最重要的统计量是某一类的比例/fraction)

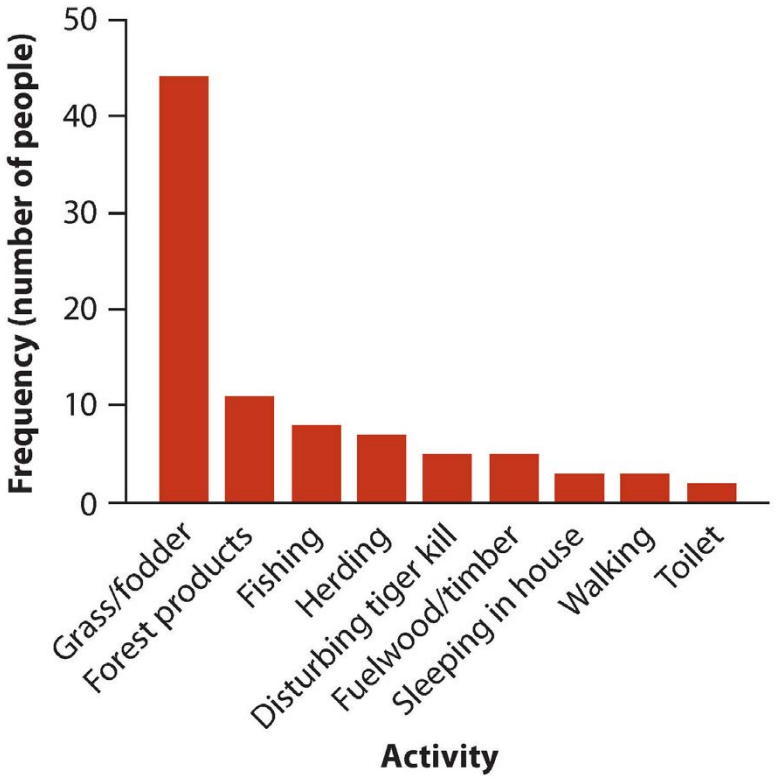
2. Descriptive statistics



- 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)
  - In some fields of science, variability around a central value is instrument noise or measurement error (仪表噪声/测量误差)
  - But in biology, much of the **variability** signifies real differences among individuals (个体间差异)
    - Biologists also appreciate variation as the stuff of evolution.

2. Descriptive statistics

- Proportion (比例)
  - The fraction of observations in a given category
  - The fraction differences between categories



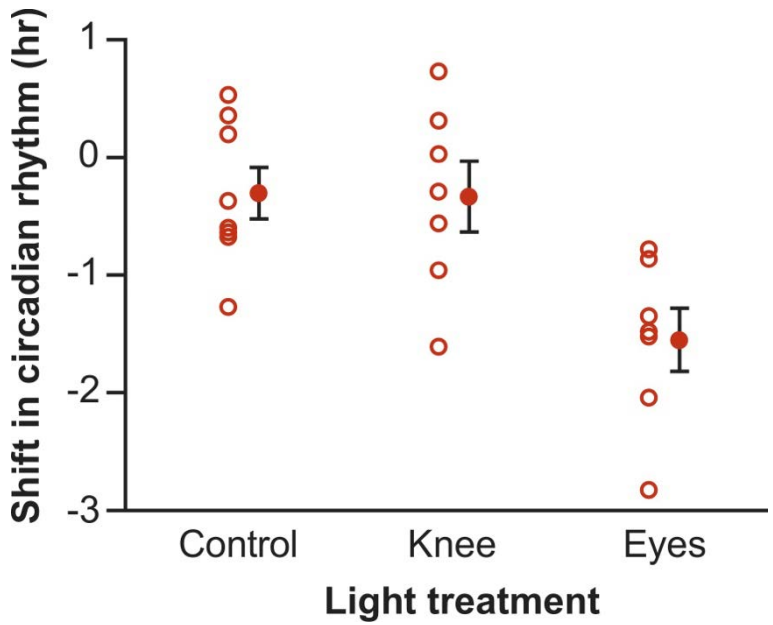
3. Measures of descriptive statistics

- 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)
- Cumulative frequency distribution 累积频率分布
- Proportion 比例

3.1 Arithmetic mean and standard deviation

- Arithmetic mean 算术平均值
  - The arithmetic mean is the most common metric to describe the **location** (位置) of a frequency distribution. It is the average of a set of measurements.

- Standard deviation 标准差
  - The standard deviation is the most commonly used measure of distribution **spread** (散布程度) .



(Whitlock & Schluter 2020)

Arithmetic mean 算术平均值

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

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

- $\bar{Y}$: 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)
- $\sum$: sum (求和符号)

Arithmetic mean 算术平均值

- The sample mean 样本均值
  - the average of the measurements in the sample
    - the sum of all the observations divided by the number of observations

      (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。

Arithmetic mean 算术平均值

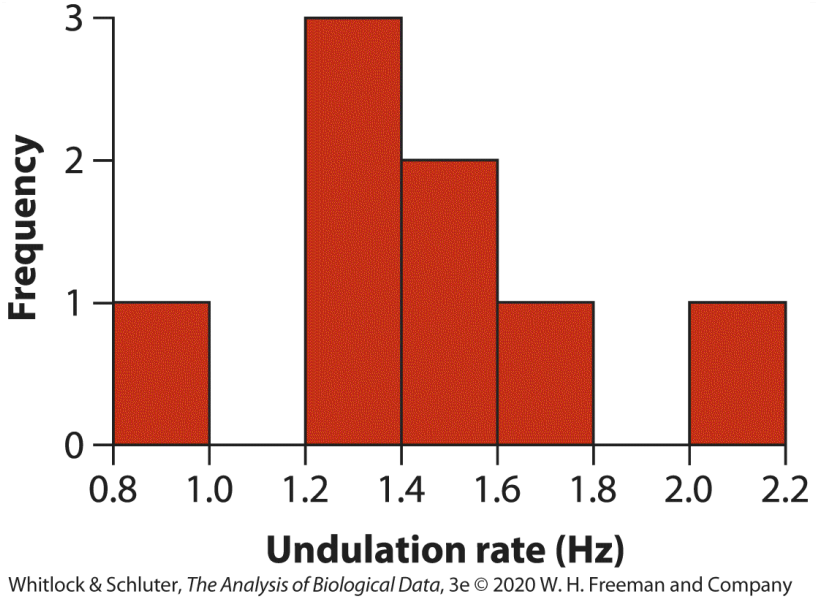
- The sample mean 样本均值
  - the average of the measurements in the sample
    - the sum of all the observations divided by the number of observations

    (观测值之和除以观测样本量)

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

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

$$\bar{Y} = \frac{0.9 + 1.4 + 1.2 + 1.2 + 1.3 + 2.0 + 1.4 + 1.6}{8}$$

$$= 1.375$$

Standard deviation 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)

Standard deviation 标准差

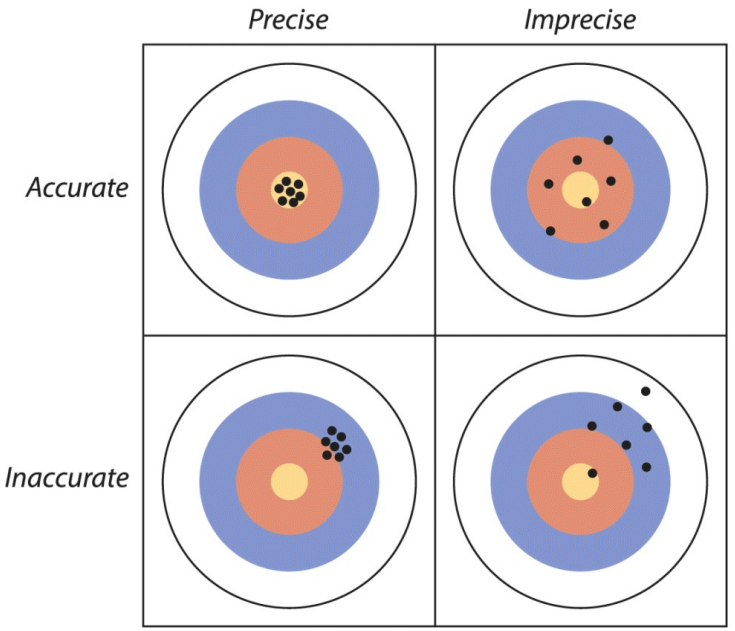
- Variance 方差

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

- Why $n-1$?

- Standard deviation

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



Standard deviation 标准差

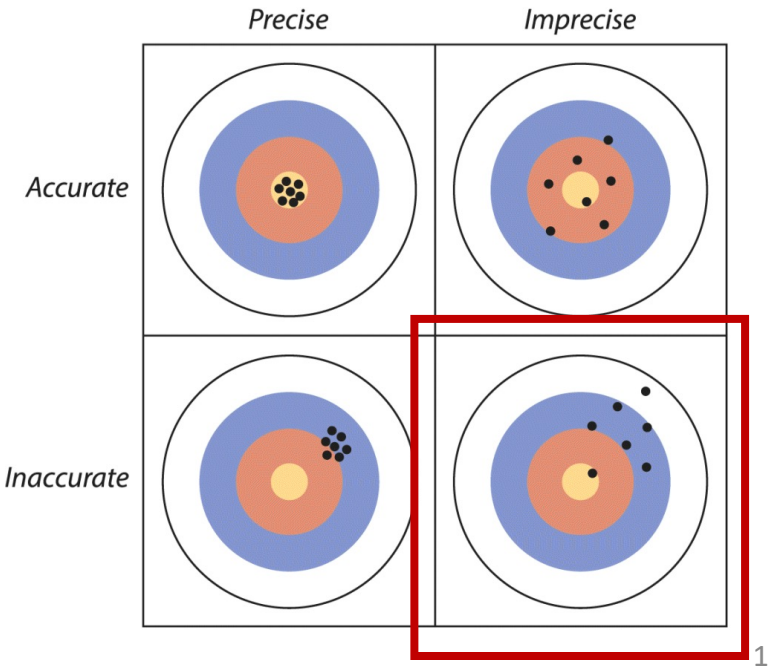
- Variance 方差

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

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

这是因为：

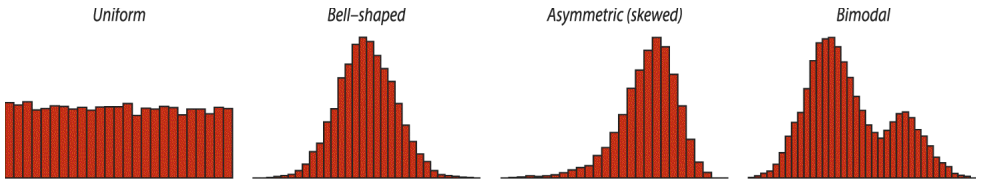
$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big[(X_i-\mu)+(\mu-\bar{X})\Big]^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-\bar{X})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{X})^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\bar{X}-\mu)(\mu-\bar{X})+(\mu-\bar{X})^2\\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

换言之，除非正好$\bar{X}=\mu$，否则我们一定有

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

https://www.zhihu.com/question/20099757

Standard deviation 标准差

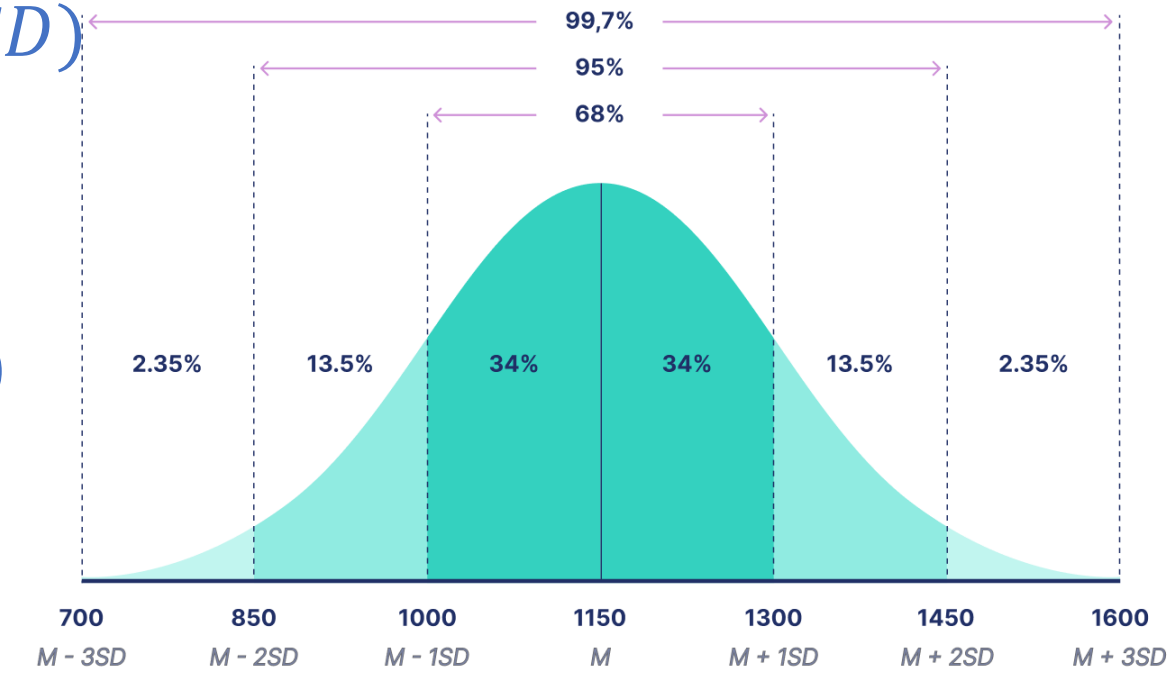


- The connection: standard deviation (SD) ~ frequency distribution
  - If the frequency distribution is bell shaped/**a normal distribution** (正态分布)
  - about 2/3 of the observations will lie within 1 SD of the mean

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

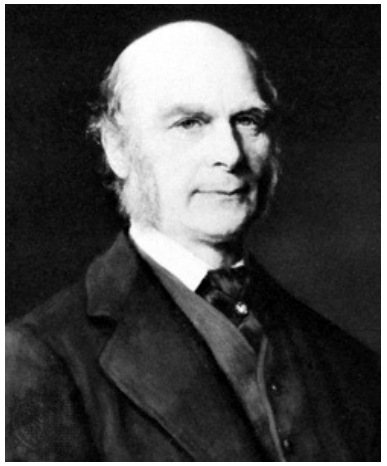
  - about 95% will lie within 2 SD

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

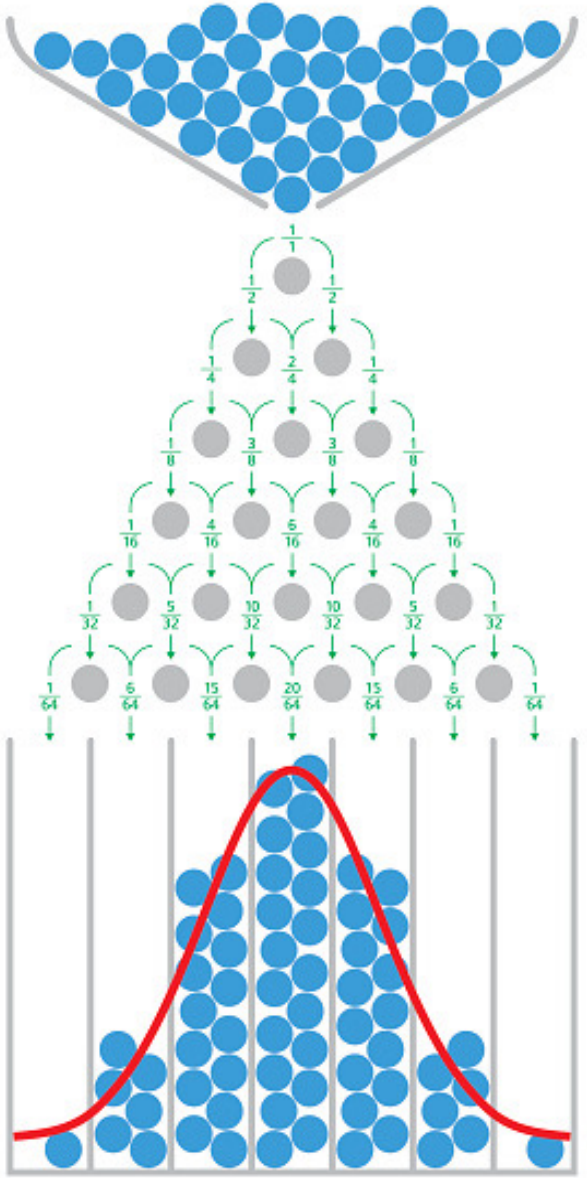


Normal distribution 正态分布

- 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



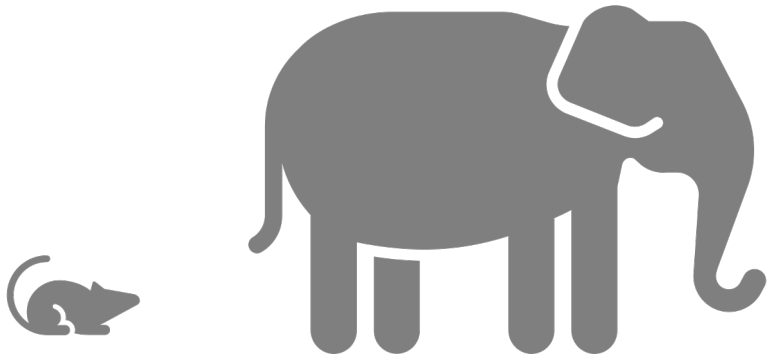
Sir Francis Galton



Arithmetic mean 算术平均值 & Standard deviation 标准差

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

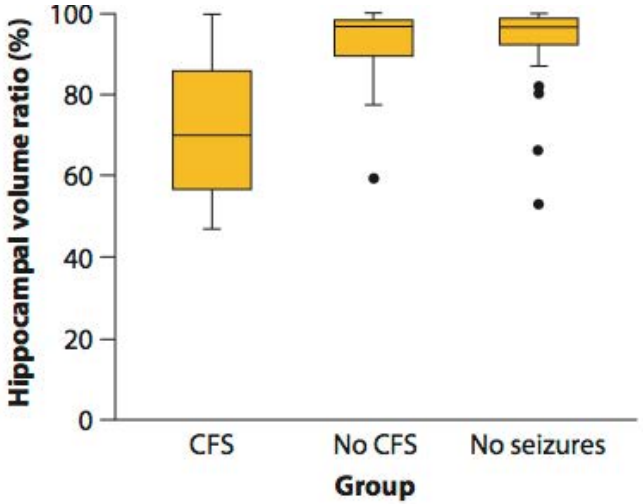
- SD $s = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$



- Coefficient of variation (CV) 变异系数
  - the standard deviation as a percentage of the mean
    - To express the **relative** variation among individuals
    - Because mean & SD may change together
    - E.g., biomass of elephant vs. mouse; 3000 kg vs. 20 g → 20% vs. 20%

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

3.2 Median and interquartile range

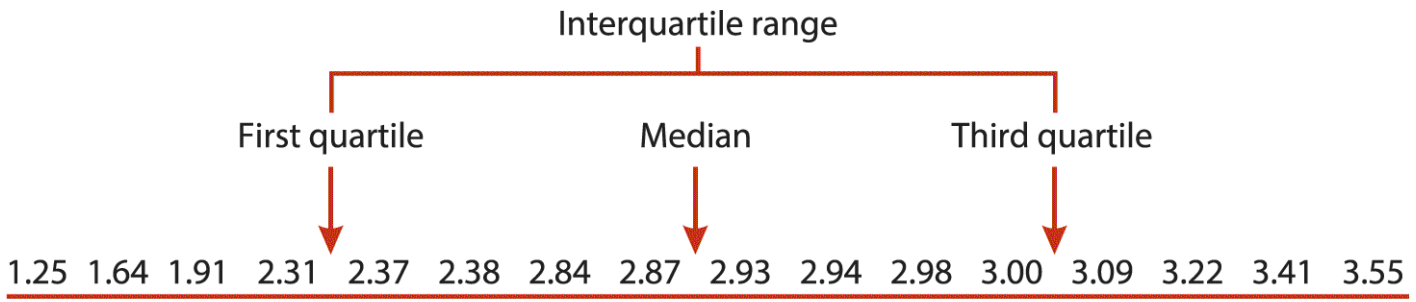


- Median 中值
  - the middle observation in a set of data
    - 1. sort the data (排序)
    - 2. obtain the middle one
      - Odd (奇数) sample size: the middle one
      - Even (偶数) sample size: the mean of the middle two

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

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

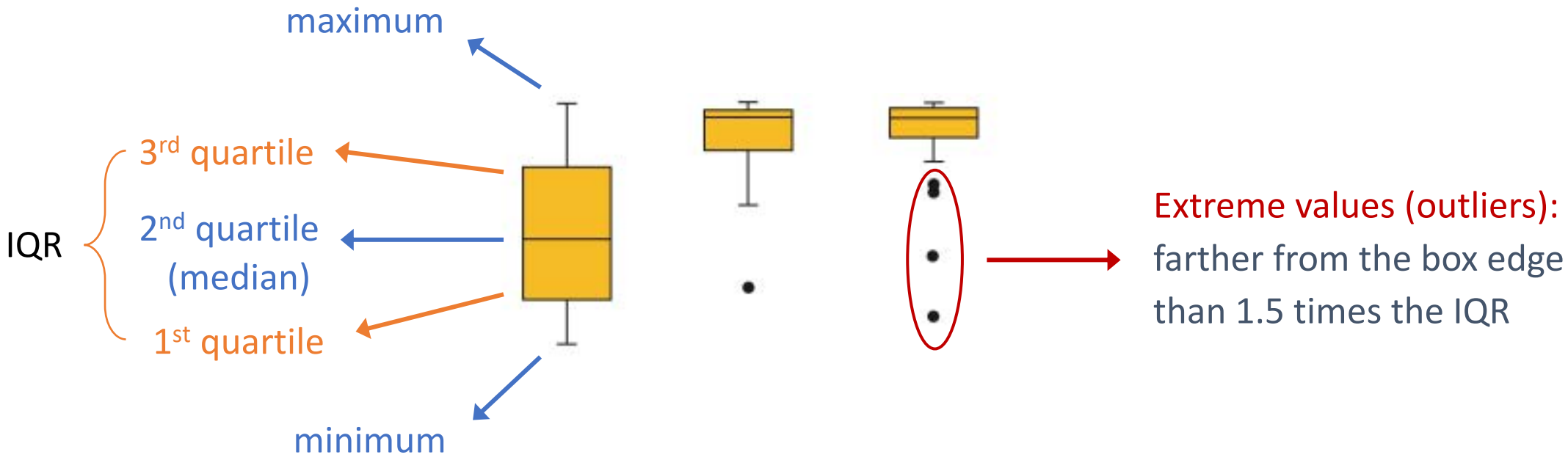
- Interquartile range (IQR) 四分位距
  - 1st quartile (四分位数)
  - 2nd quartile = median
  - 3rd quartile
  - IQR = 1st ~ 3rd Qu
    - the span of the middle half of the data



(Whitlock & Schluter 2020)

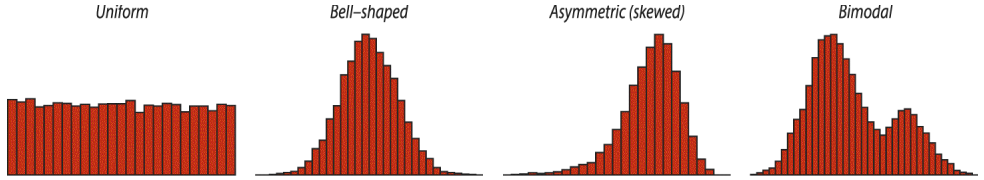
3.2 Median and interquartile range

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



(Whitlock & Schluter 2020)

3.3 Compare measures



- Location (位置)
  - a central value: mean 均值 vs. median 中值
- How to compare?
  - depend on the shape of the frequency distribution
  - E.g., when the data are strongly skewed or include extreme observations,

    mean & standard deviation **<<** *median & interquartile range*

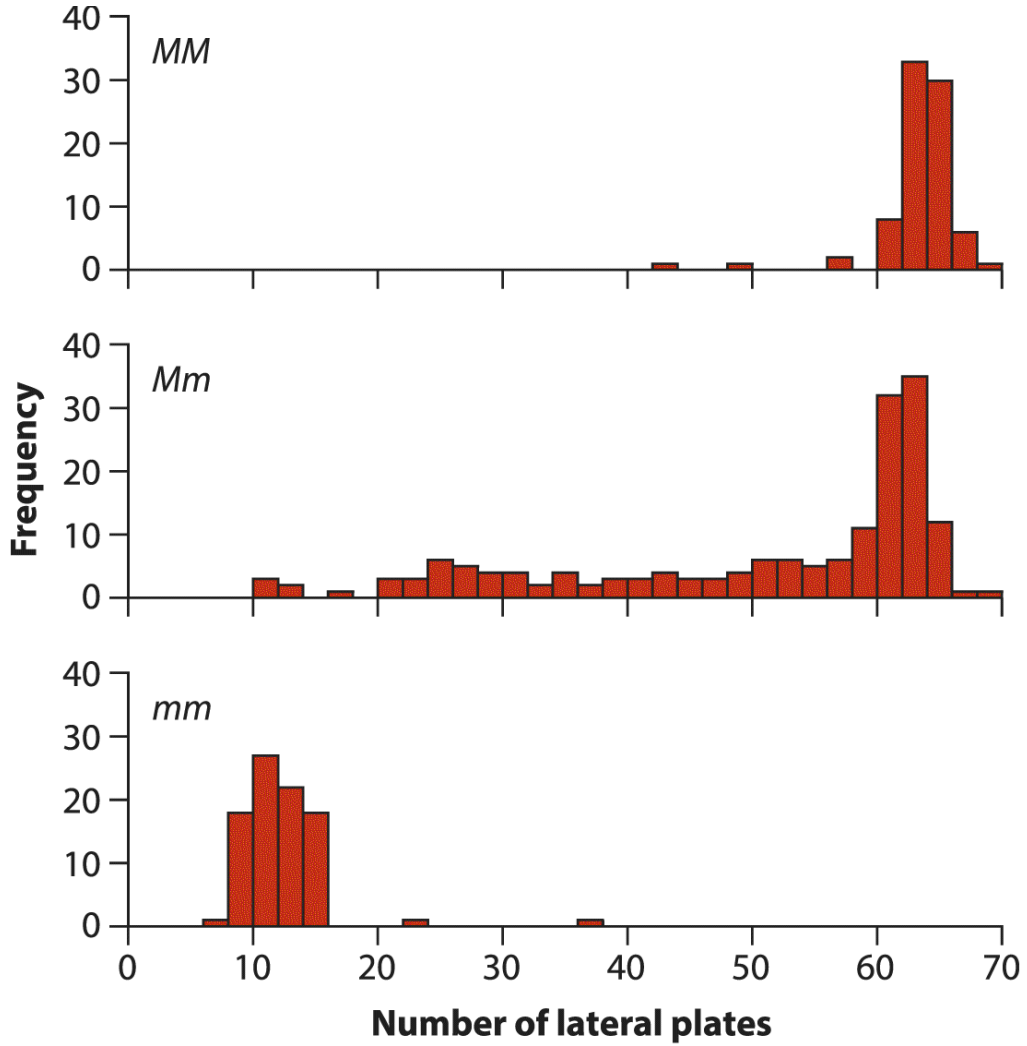
    (less informative)

    (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)

Compare the central value

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

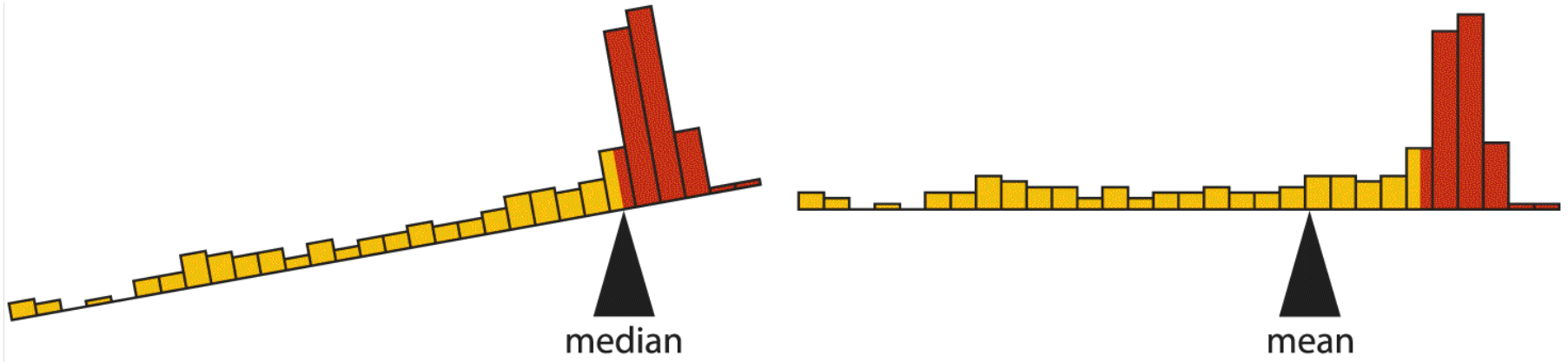
Threespine sticklebacks（三刺鱼）
reproduced with permission from K. B. Marchinko and D. Schluter (2007) [Evolution 61:1084–90, Wiley-Blackwell Publishing Ltd.].



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

Compare the central value

- How to compare? → depend on the shape of the frequency distribution
- when the distribution is **asymmetric/skewed** (非对称/倾斜)
  - median is the middle measurement of a distribution
  - the mean is the “center of gravity” → more sensitive to extremes



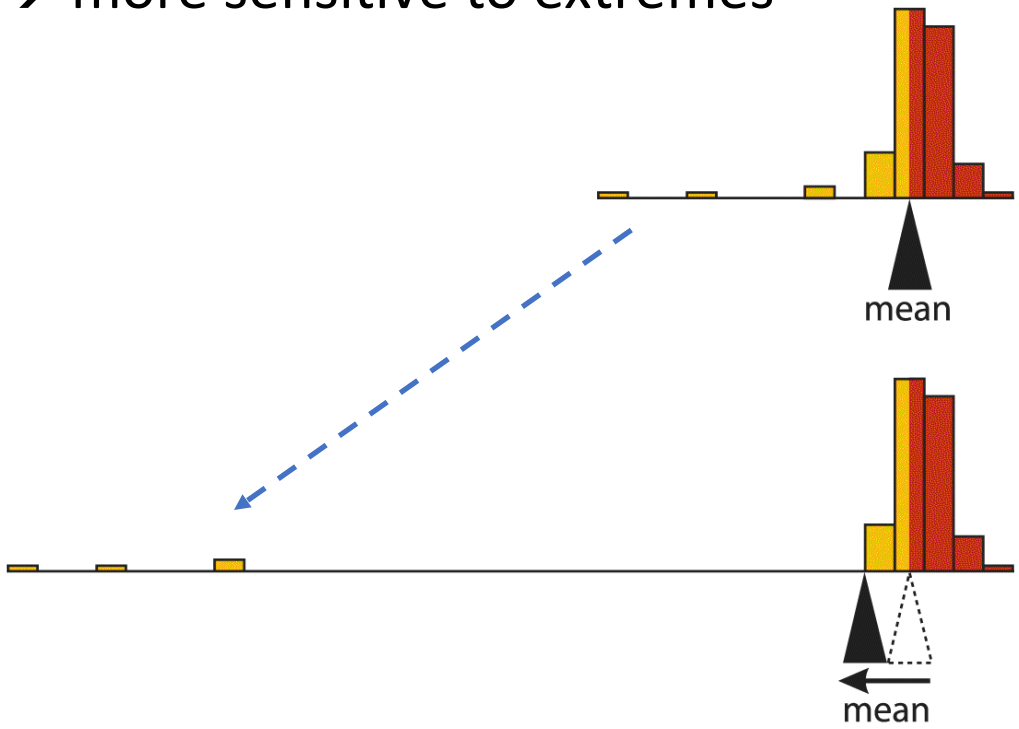
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Compare the central value

- when the distribution is **asymmetric/skewed** (非对称/倾斜)
  - e.g., add minimum extremes
  - median is the middle measurement of a distribution (unaffected 没影响)
  - the mean is the "center of gravity" → more sensitive to extremes

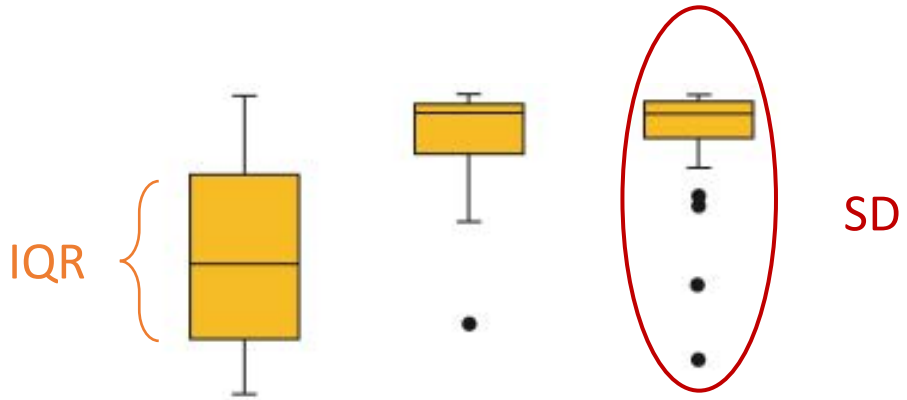
The balancing act

→ *The mean shifts leftward*



Compare the spread

- The variability: SD 标准差 vs. IQR 四分位距
  - SD: even **more sensitive** to extreme observations than is the mean
- when the distribution is **asymmetric/skewed** (非对称/倾斜)
  - the interquartile range (**IQR**) is a **better** indicator of the spread of the **main** part of a distribution than the standard deviation (SD);
  - the SD reflects the variation among **all** of the data points.



3.4 Cumulative frequency distribution 累积频率分布

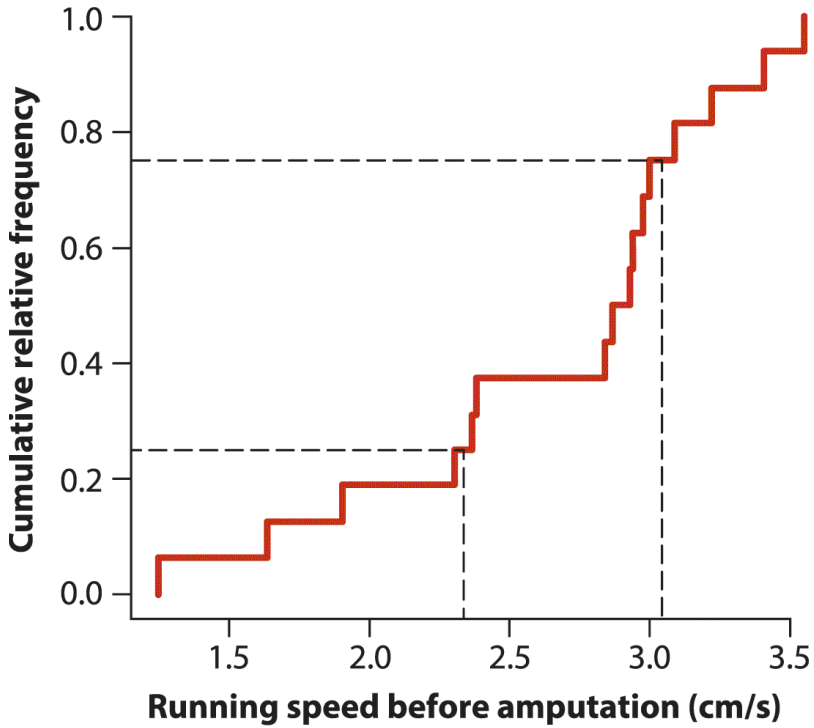
- Percentile 百分位数
  - The X^{th} percentile is the **value** below which X percent of the individuals lie
  - E.g., 第10个百分位数意味着样本中有10%的观测值低于该百分位数.
- Quantile 分位数
  - the 10^{th} percentile is the 0.10 quantile (第10个百分位数是0.01分位数)
  - quartile (四分位数): the 1^{st} and 3^{rd} **quartiles** are the 0.25 and 0.75 **quantiles**.
- Cumulative frequency distribution (CFD) 累积频率分布
  - Another way to compare the **shapes** and **positions** of frequency distributions
  - All the quantiles of a **numerical** variable can be displayed by CFD.

3.4 Cumulative frequency distribution 累积频率分布

- Displaying CFD
  - 1. sort all data from the smallest to the largest

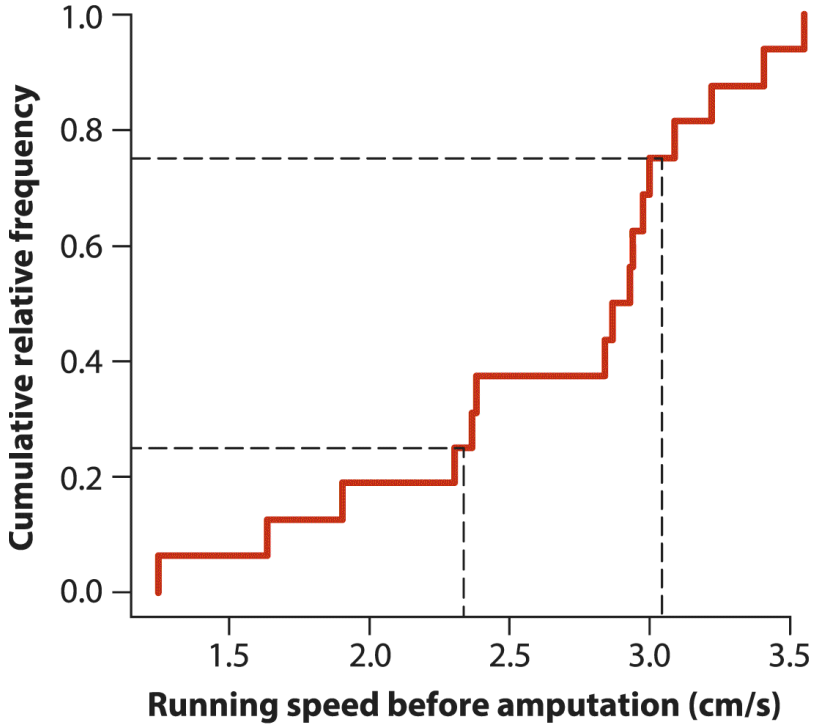
    (从小到大排序)
  - 2. calculate the **fraction** of observations less than or equal to each data value

    (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
  - Dashed lines 虚线
    - 1/4 observations < 2.34 (1st quartile)
    - 3/4 observations < 3.045 (3rd quartile)
- Compare
  - box plot & histogram > CFD
  - But, CFD can be useful when comparing multiple groups (比较多组数据时)

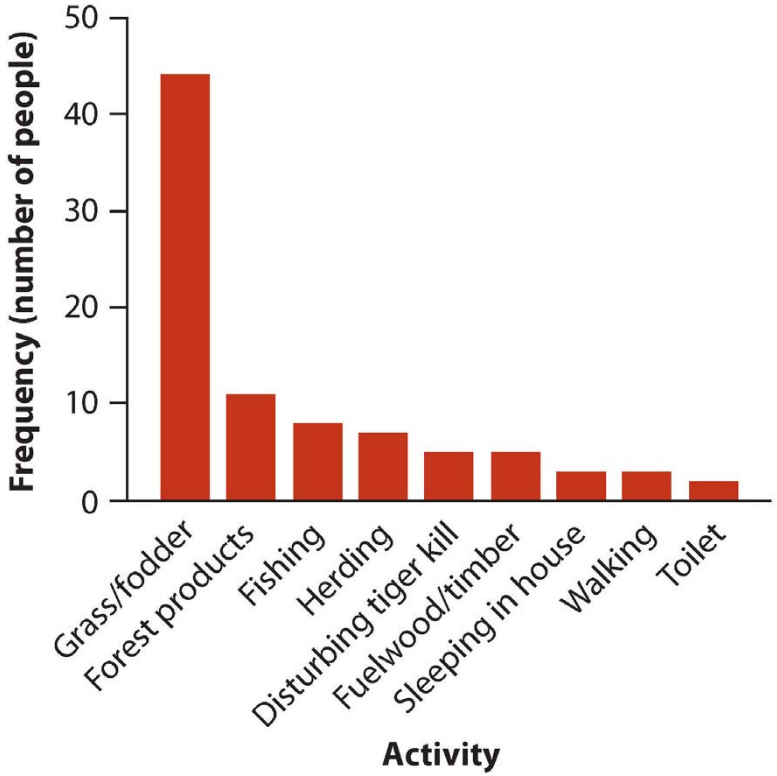


3.5 Proportion 比例

- The most important descriptive statistic for a categorical variable
- Calculation

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

- The proportion is like a sample mean???



4. Summary 总结

- The location of a distribution for a numerical variable can be measured by its mean or by its median.

  （均值或中值可以度量样本数值变量分布的位置）

  - The mean gives the center of gravity of the distribution;
  - The median gives the middle value.
- The standard deviation measures the spread of a distribution for a numerical variable.

  （标准差可以度量样本数值变量分布的散布程度）

  - It is a measure of the typical distance between observations and the mean;
  - The variance is the square of the standard deviation.

4. Summary 总结

- The quartiles break the ordered observations into four equal parts.

  （四分位数将样本观测值分为四等分, 其中四分位距也度量了其散布程度）

  - The interquartile range, the difference between the first and third quartiles, is another measure of the spread of a frequency distribution.
- The standard deviation measures the spread of a distribution for a numerical variable.

  （标准差可以度量样本数值变量分布的散布程度）

  - It is a measure of the typical distance between observations and the mean;
  - The variance is the square of the standard deviation.

4. Summary 总结

- All the quantiles of a sample of data can be shown using a graph of the cumulative frequency distribution.

  （样本的所有分位数可以用累计频率分布图来表示）

- The proportion is the most important descriptive statistic for a categorical variable.

  （比例是类型数据最重要的描述性统计量）

  - It is calculated by dividing the number of observations *in the category of interest* by the total number of observations in all categories combined.

Assignment 01

- Due on Oct. 6
- Will be released on Sept. 28 (later this week)

- Include calculations and plotting figures in R

- Submit via Daxia Xuetang (elearning.ecnu.edu.cn)

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

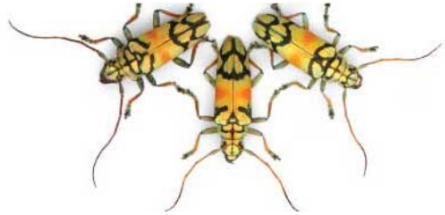
Miller, L. M., T. Close, and A. R. Kapuscinski. 2004. *Molecular Ecology*

06. Endangered species

U.S. Fish & Wildlife Service. 2018. Listed Animals. U.S. Fish & Wildlife Online System. Data link. Accessed August 27, 2018.

The Analysis of Biological Data by Whitlock and Schluter

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