

Lecture 4 – Uncertainty and Probability

- Outline for today
 - Recall L03
 - Uncertainty of estimates 估计的不确定性
 - Standard error 标准误
 - Confidence interval 置信区间
 - Probability 概率
 - Summary
 - R Lab & Discussion

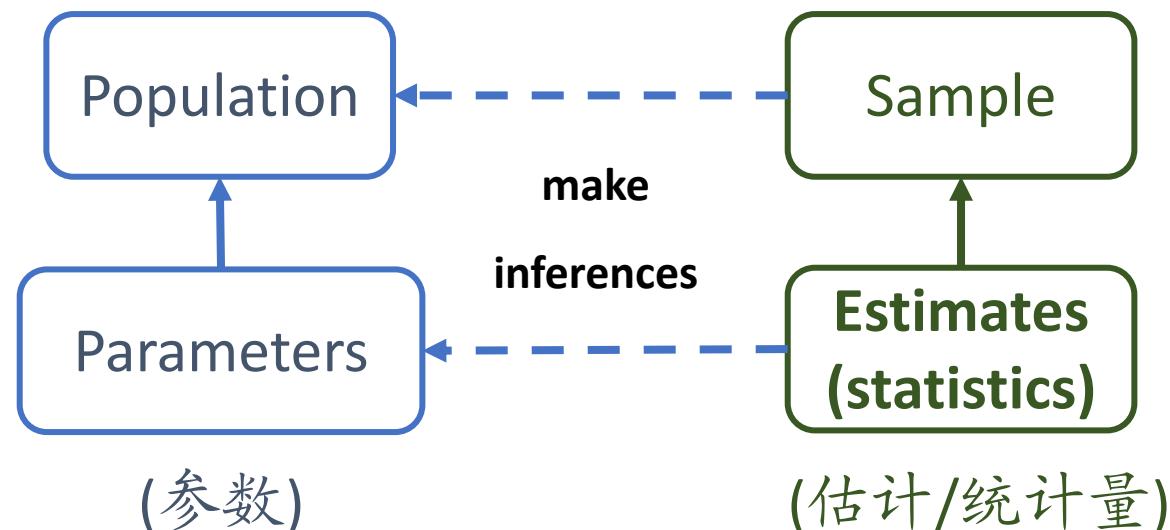
生物统计学

李 勤

生态与环境科学学院

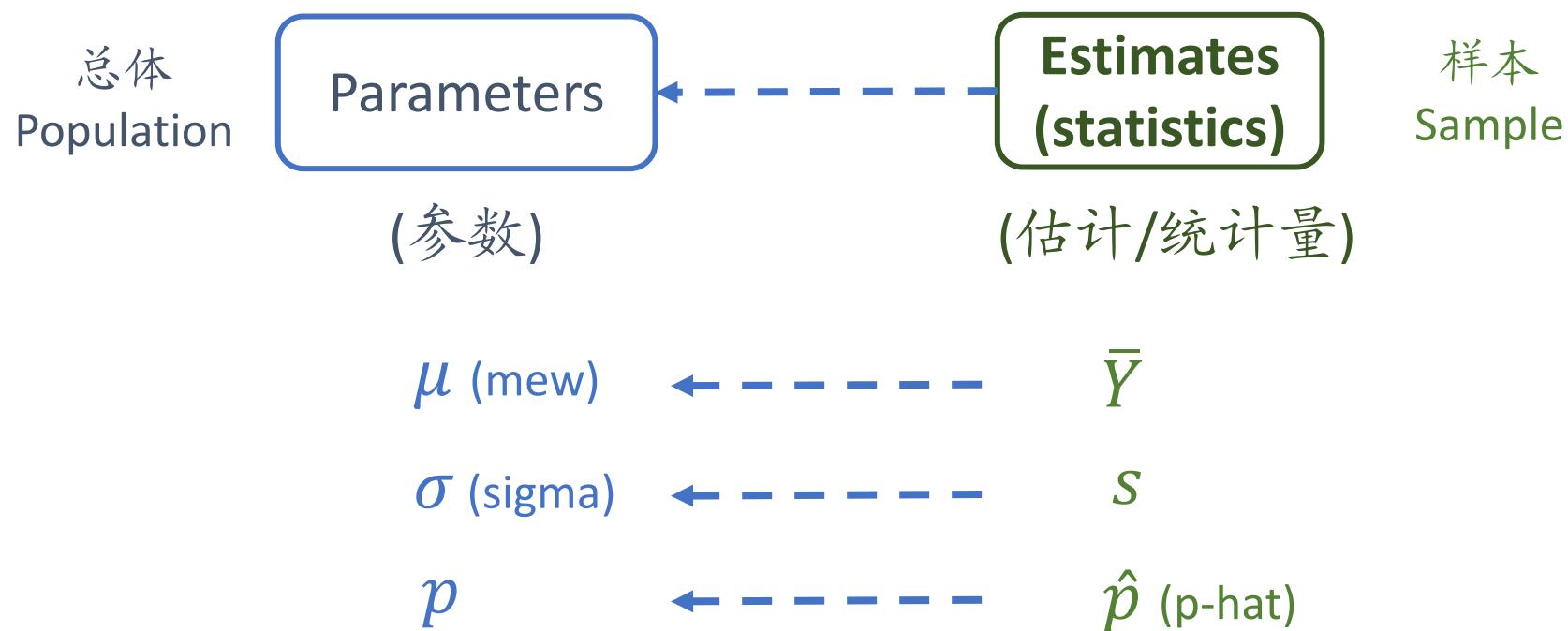
1. Recall from L03 - Descriptive statistics

- 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



1. Recall from L03 - Descriptive statistics

- The descriptive statistics measured on a **random sample** are used to estimate parameters of the population.





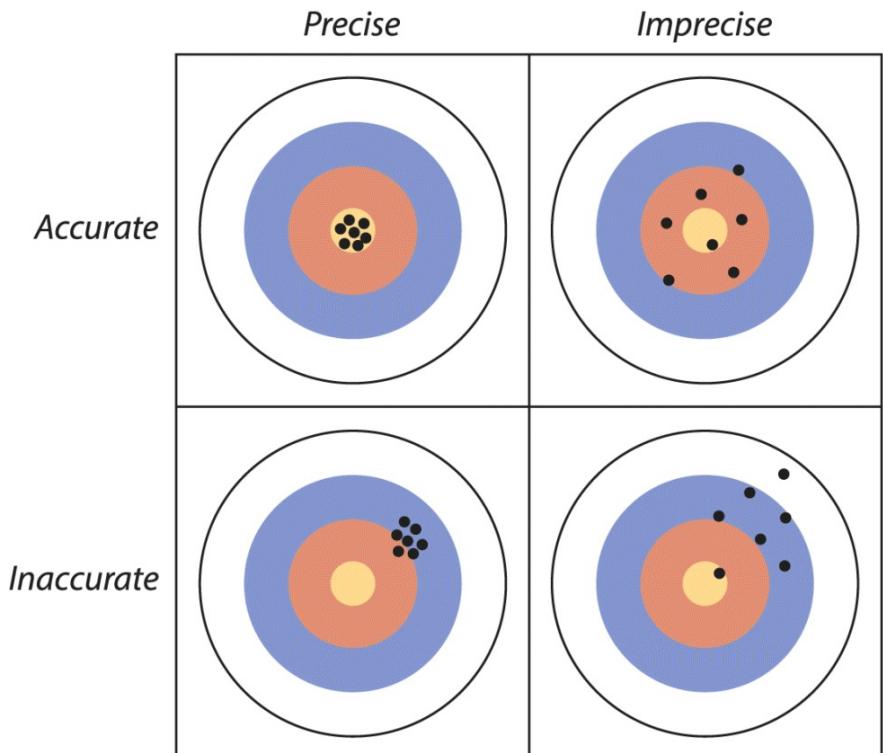
1. Recall from L03 - Descriptive statistics

- 描述性统计量及其计算
 - Mean 均值
 - SD 标准差
 - Median 中值
 - Quartile 四分位数
 - IQR 四分位距
 - Quantile 分位数
 - Proportion 比例
- When to use them?
- And how?



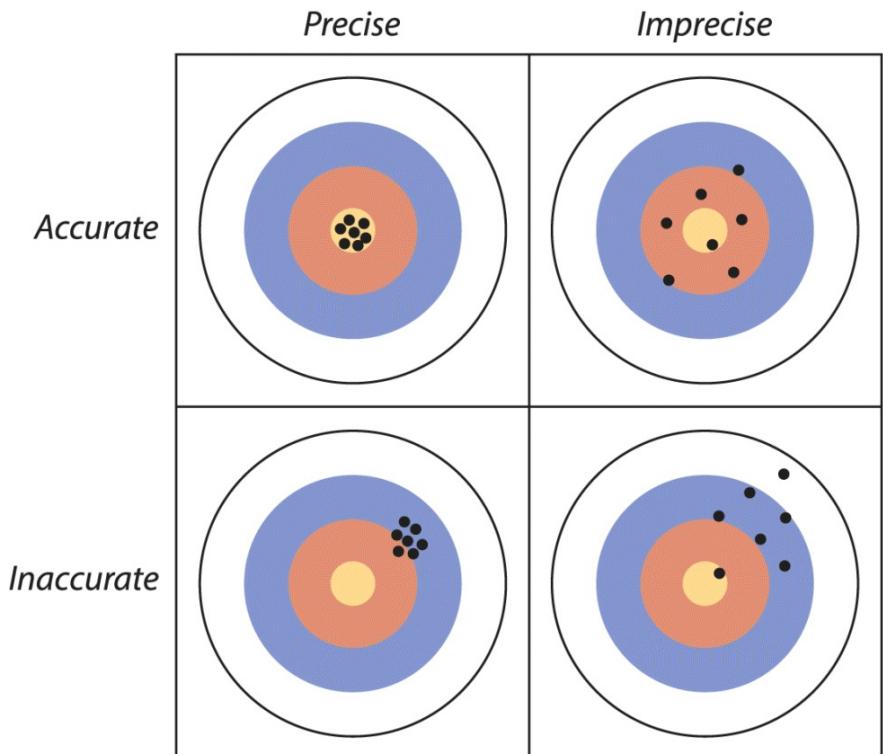
2. 估计的不确定性

- 针对总体参数的估计的应用，我们需要量化其精确度 (**precision**);
- 而从数据计算得到的估计值几乎不会与对应的总体参数完全相同 (**almost never exactly the same**);
- 为什么呢？ **Why?**



2. 估计的不确定性 Estimating with uncertainty

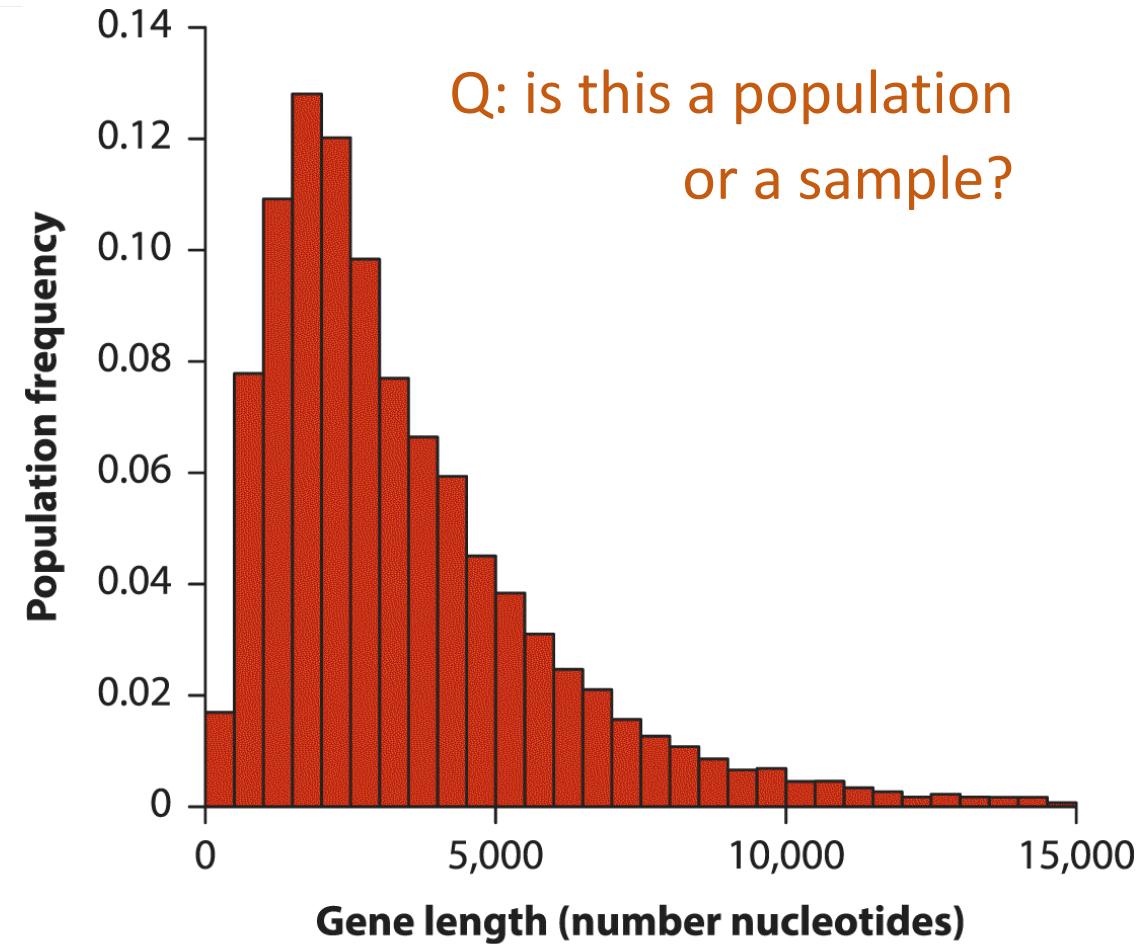
- 从数据计算得到的估计值几乎不会与对应的总体参数完全相同，因为抽样受到偶然性的影响(**sampling is influenced by chance**);
- 所以，关键的问题变成：
 - 面对偶然性，我们对估计值有多少把握？
 - 换句话说，它的精确度是多少？



2.1 抽样分布 Sampling distribution



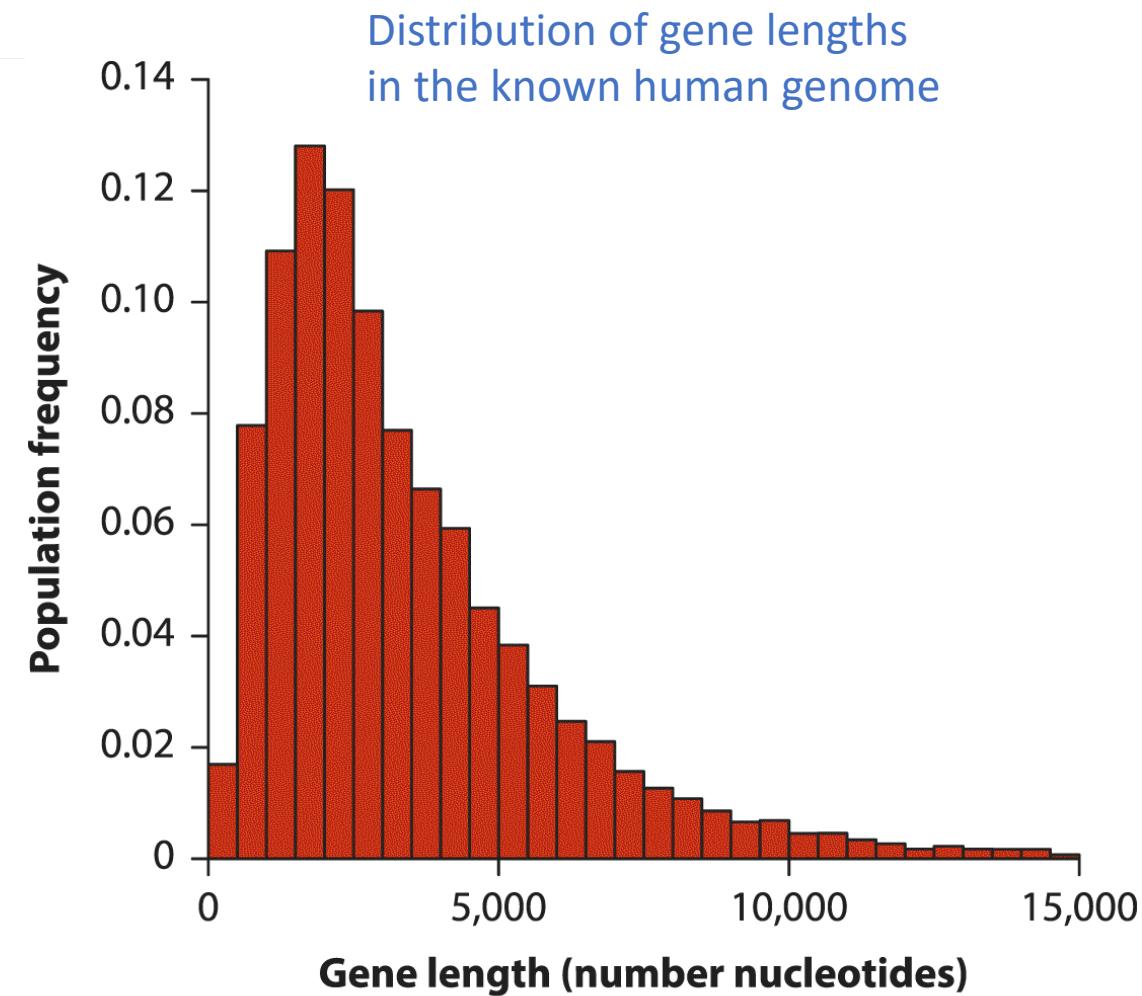
- 例子：人类基因片段的长度
 - 20,290个基因的长度
 - (来自已发表的基因组序列)
 - (Hubbard等人, 2005年; 已知/预测)
 - 基于通过国际人类基因组计划获得的所有23条人类染色体的DNA序列
 - 已知人类基因组中基因长度的分布
 - 在15,000个核苷酸处截断 (忽视了26个更长的基因)



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

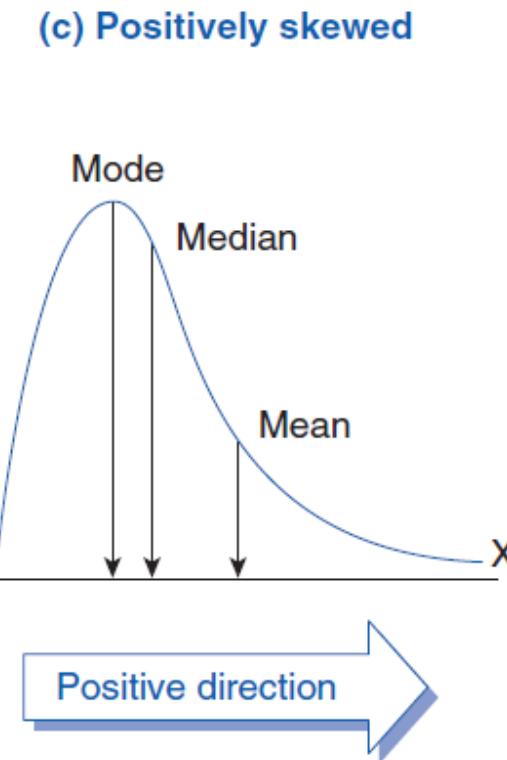
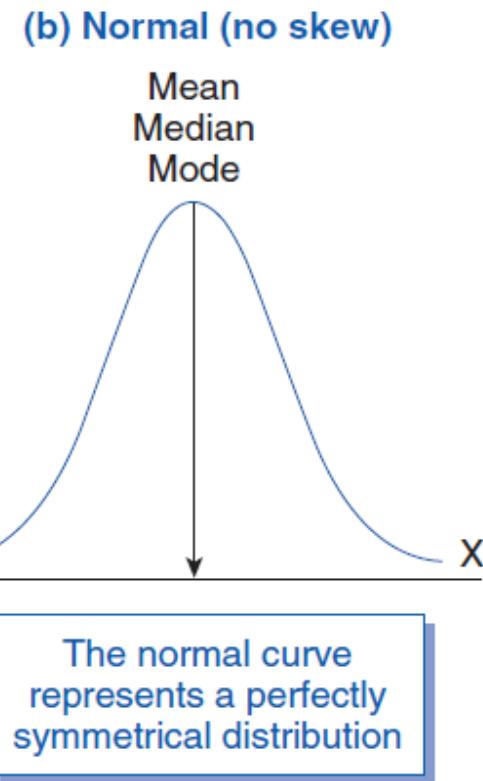
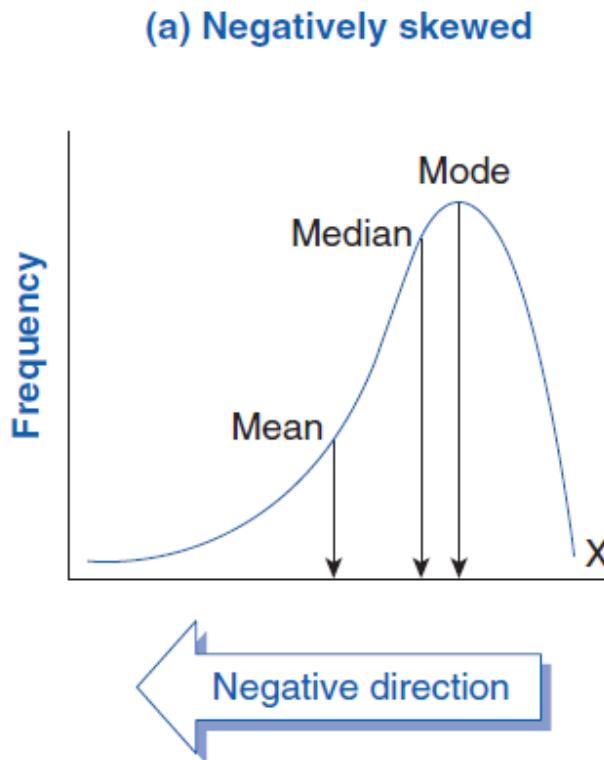
2.1 抽样分布 Sampling distribution

- 该直方图显示了基因片段总体 (**population of genes**) 的长度的频率分布；
 - 给定长度区间的基因的相对频率表示在随机抽取单个基因时获得该长度基因的概率。
 - 均值 mean (μ) : 3511.5
 - 标准差 Standard deviation (SD): 2833.2



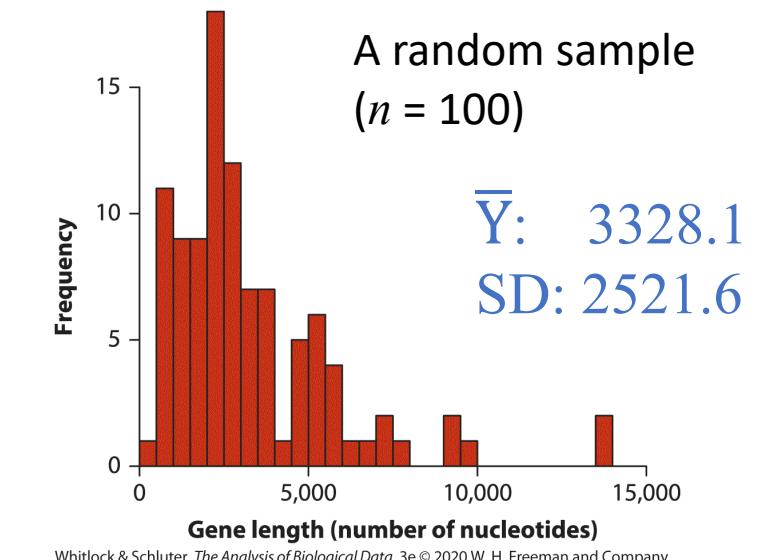
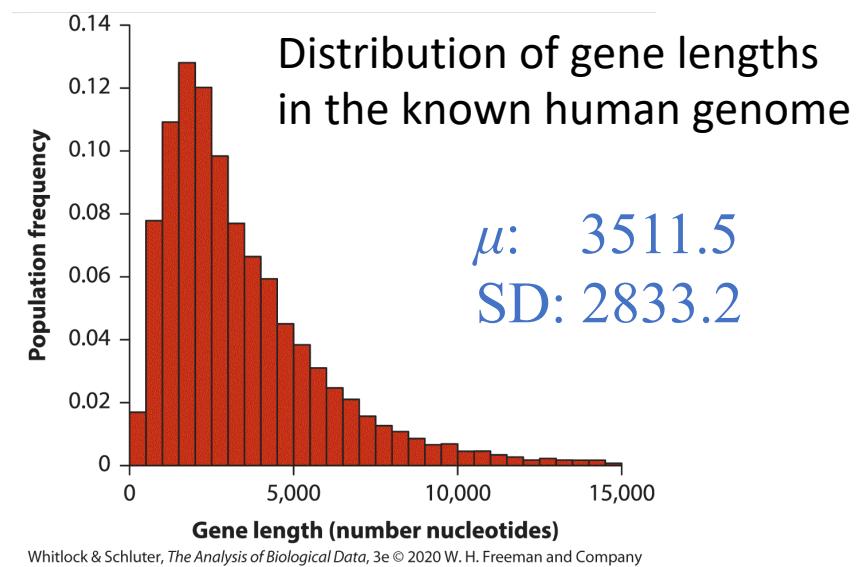
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Skewed distributions 偏斜的分布



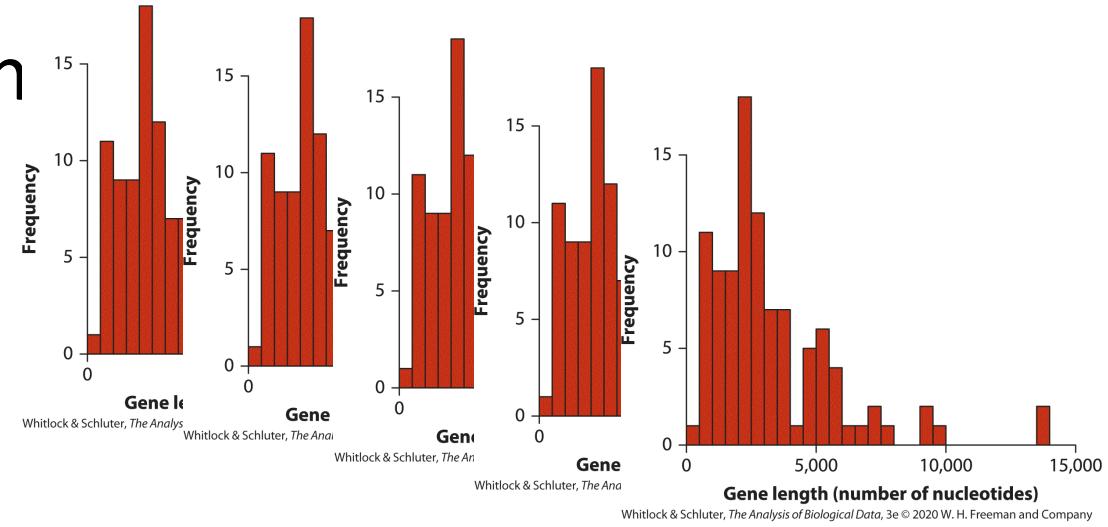
2.1 抽样分布 Sampling distribution

- 基于一个样本来估计基因片段长度的均值:
 - 样本大小: $n = 100$
- 这两个分布具有共同的重要特征:包括大致位置, 分布和形状;
 - The two distributions **share important features**, including approximate location, spread, and shape.

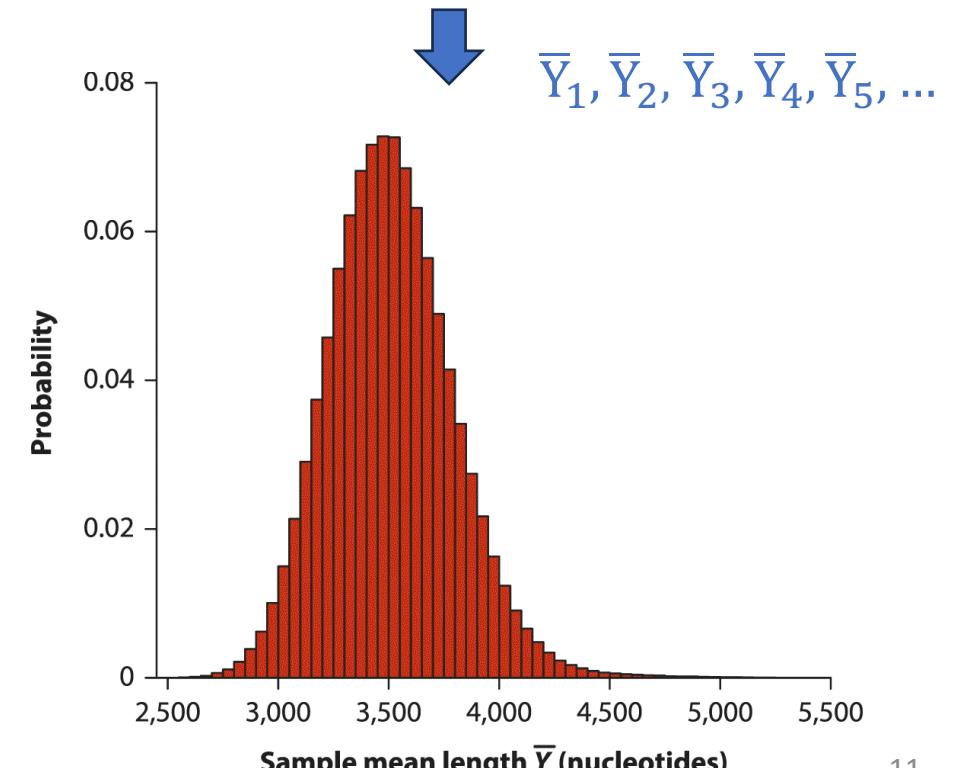


2.1 抽样分布 Sampling distribution

- 1) 重复取样
 - repeat this sampling an infinite number of times ($n = 100$)
 - *sample 1, sample 2, sample3, ...*
- 2) 估计值估计的概率分布
 - 多个样本各自计算了一个均值;
 - create the probability distribution of the estimate (mean)
- 估计值的概率分布构成了其抽样分布
 - The probability distribution of values we might obtain for an estimate makes up the estimate's **sampling distribution**.

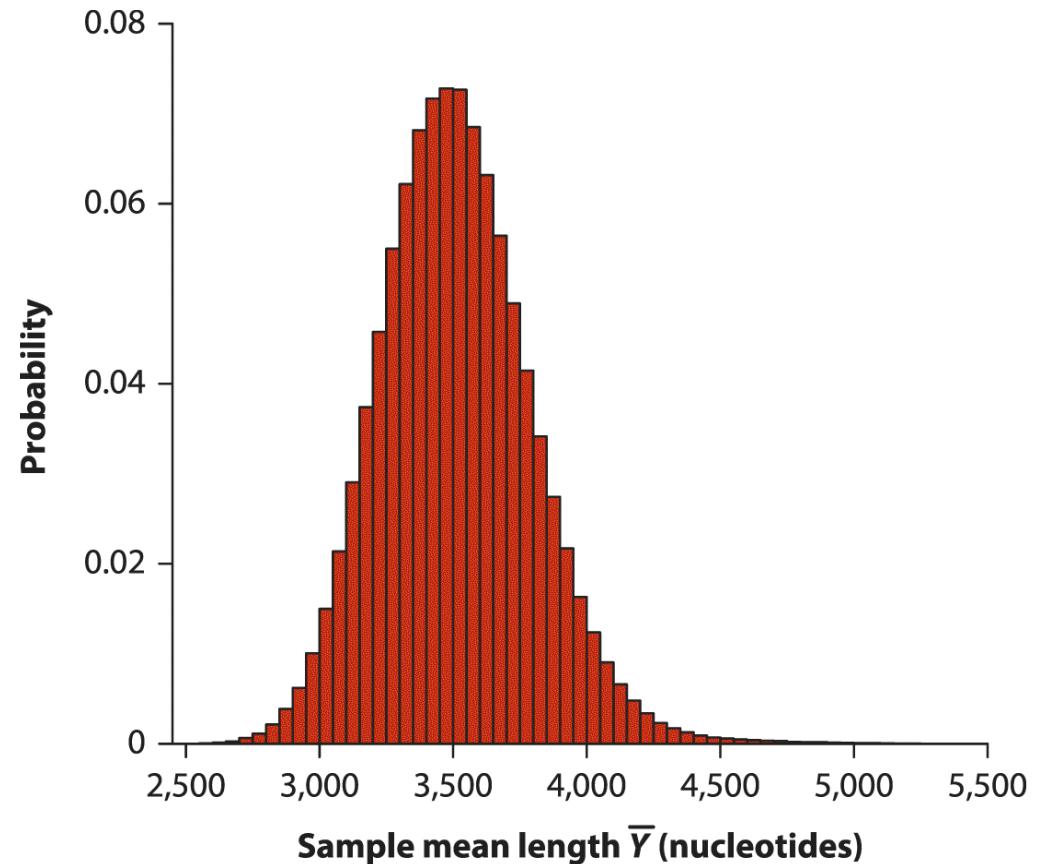


$\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4, \bar{Y}_5, \dots$



2.1 抽样分布 Sampling distribution

- 估计值的抽样分布代表其 "总体"
 - The sampling distribution represents the "population" of values for an estimate.
 - 它不是一个真实的总体;
- 总体均值 μ 是一个常数 (3511.5), 其估计值 \bar{Y} 是一个变量;
 - \bar{Y} 的抽样分布的正中恰好为真实总体的均值;
 - 这意味着 \bar{Y} 是 μ 的无偏估计。

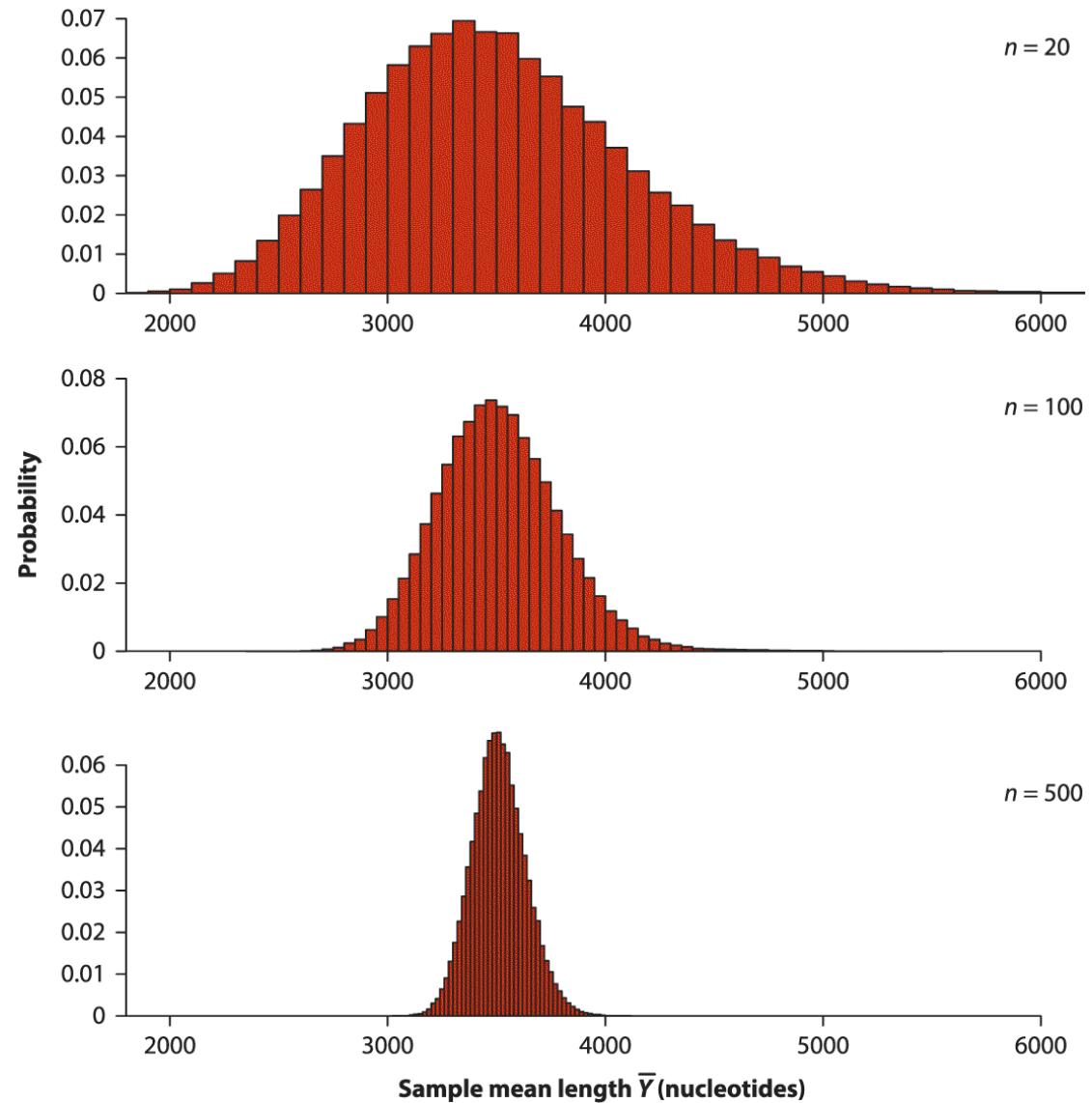


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

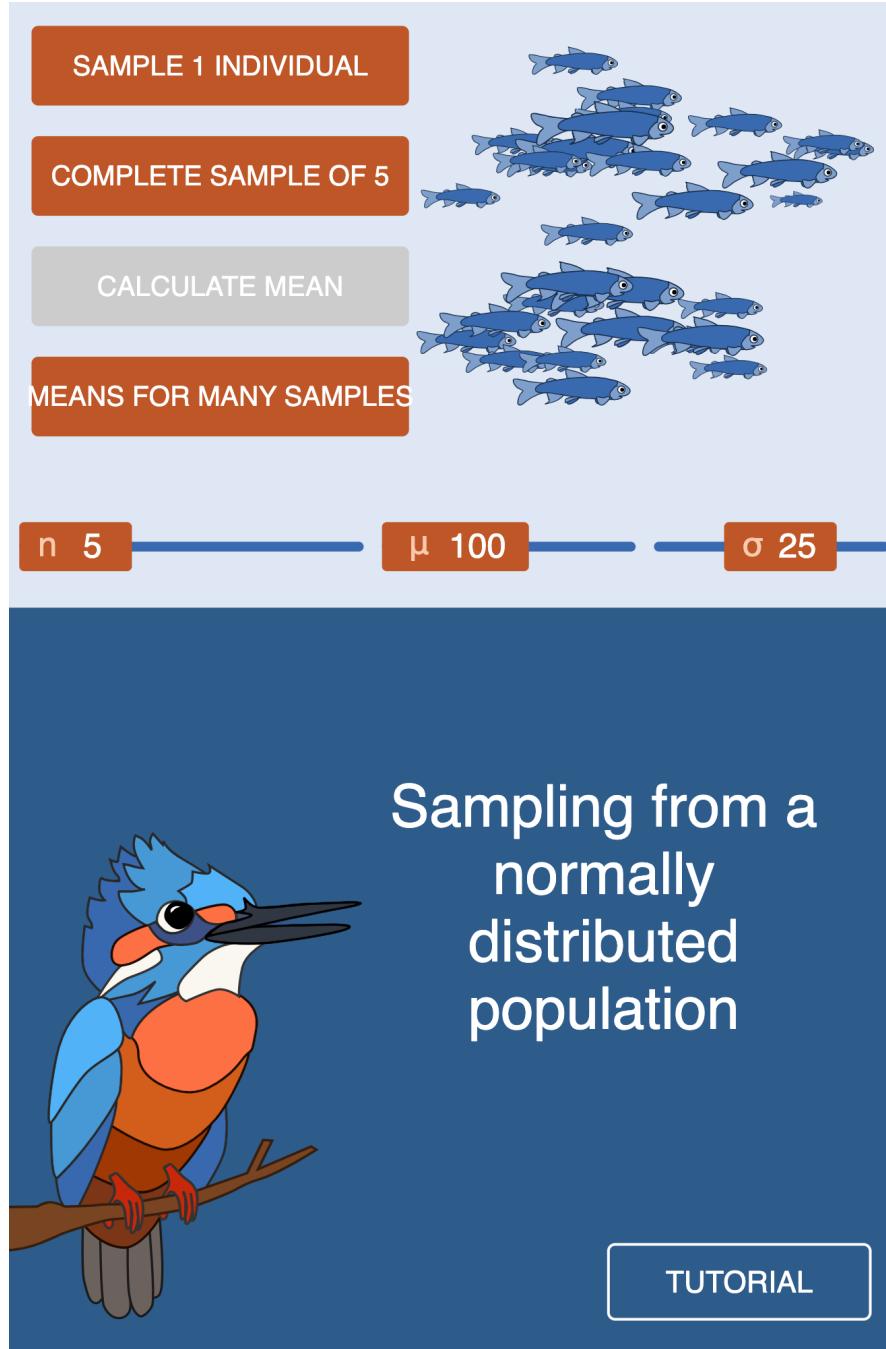
$$\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4, \bar{Y}_5, \dots$$

2.1 抽样分布

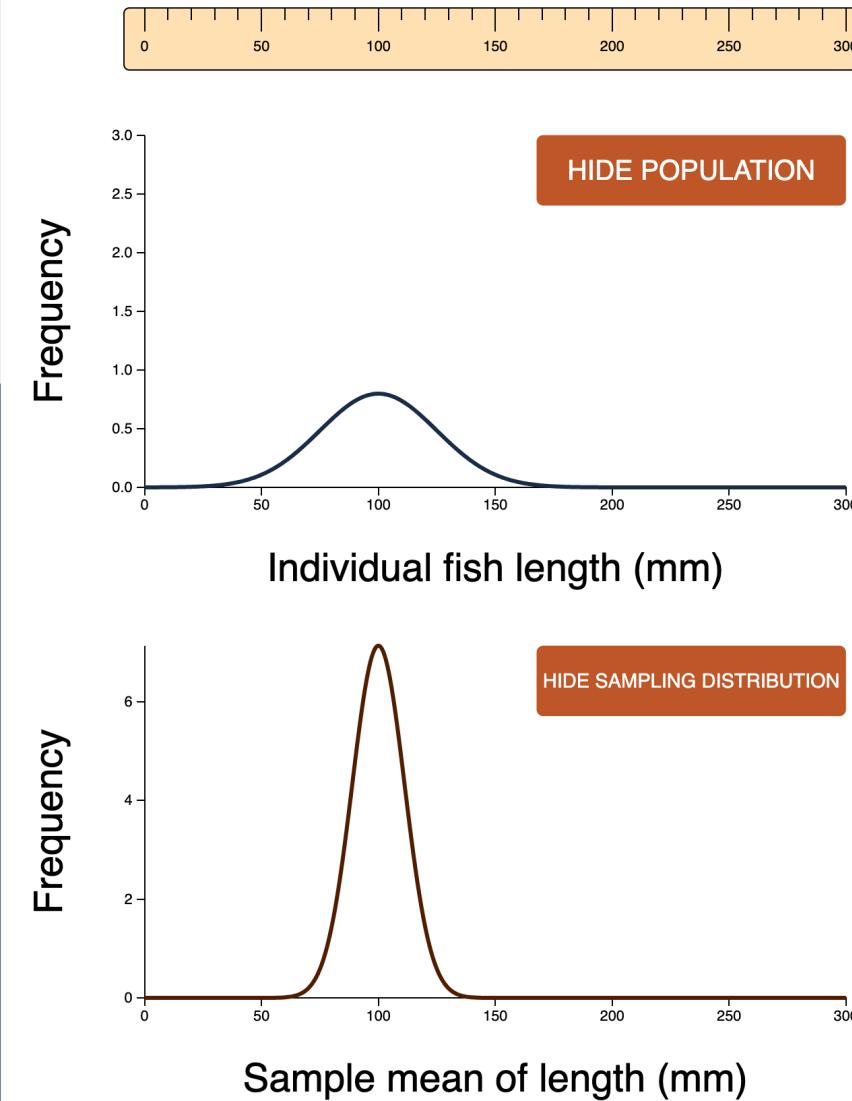
- 估计值的抽样分布的散布程度取决于样本大小；
 - The spread of the sampling distribution of an estimate depends on the **sample size**.
- 样本大小越大，抽样分布越窄；
 - 因此在统计学中需要尽可能获得更大的样本，因为它们产生的估计会更精确。

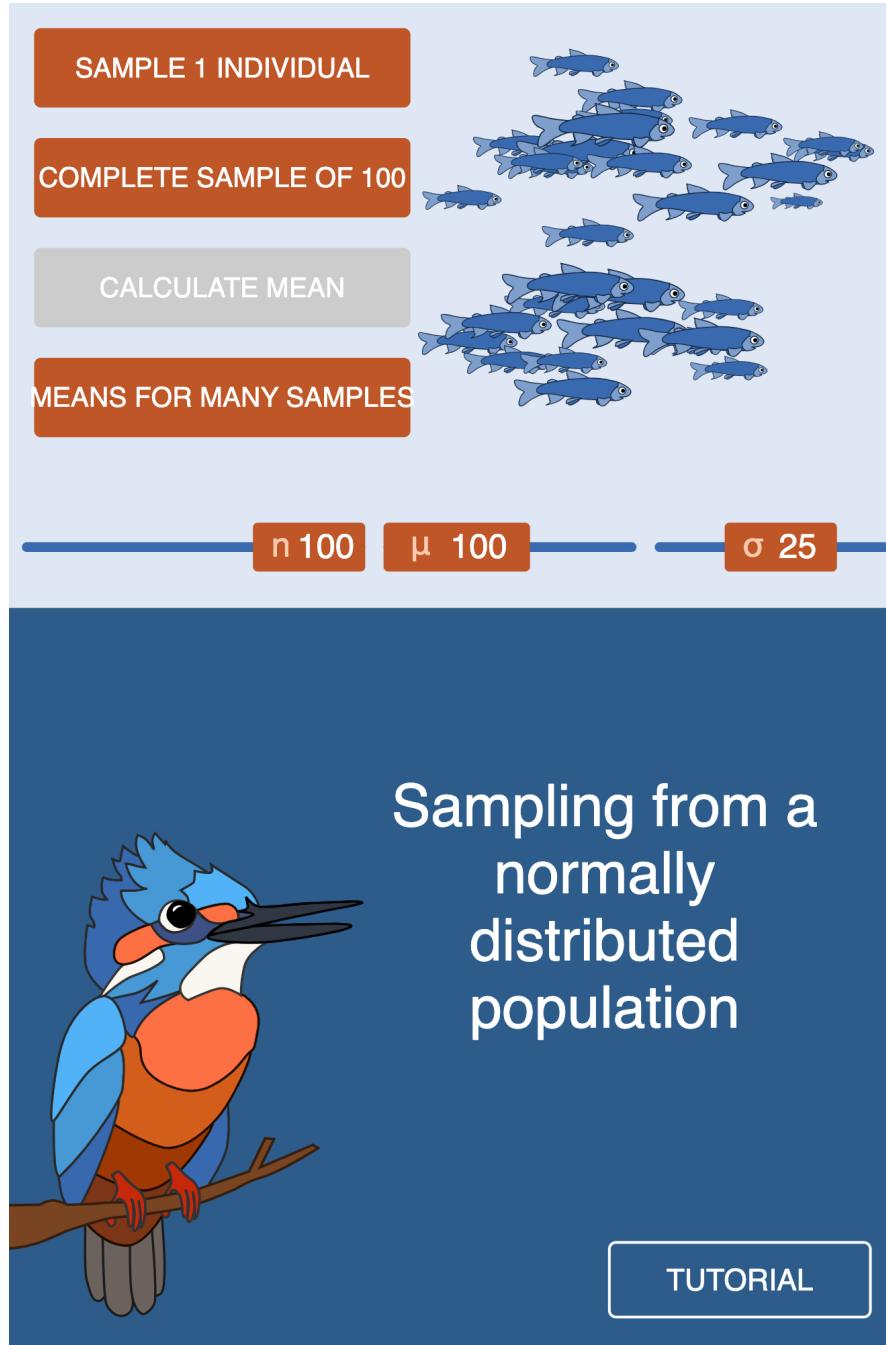


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

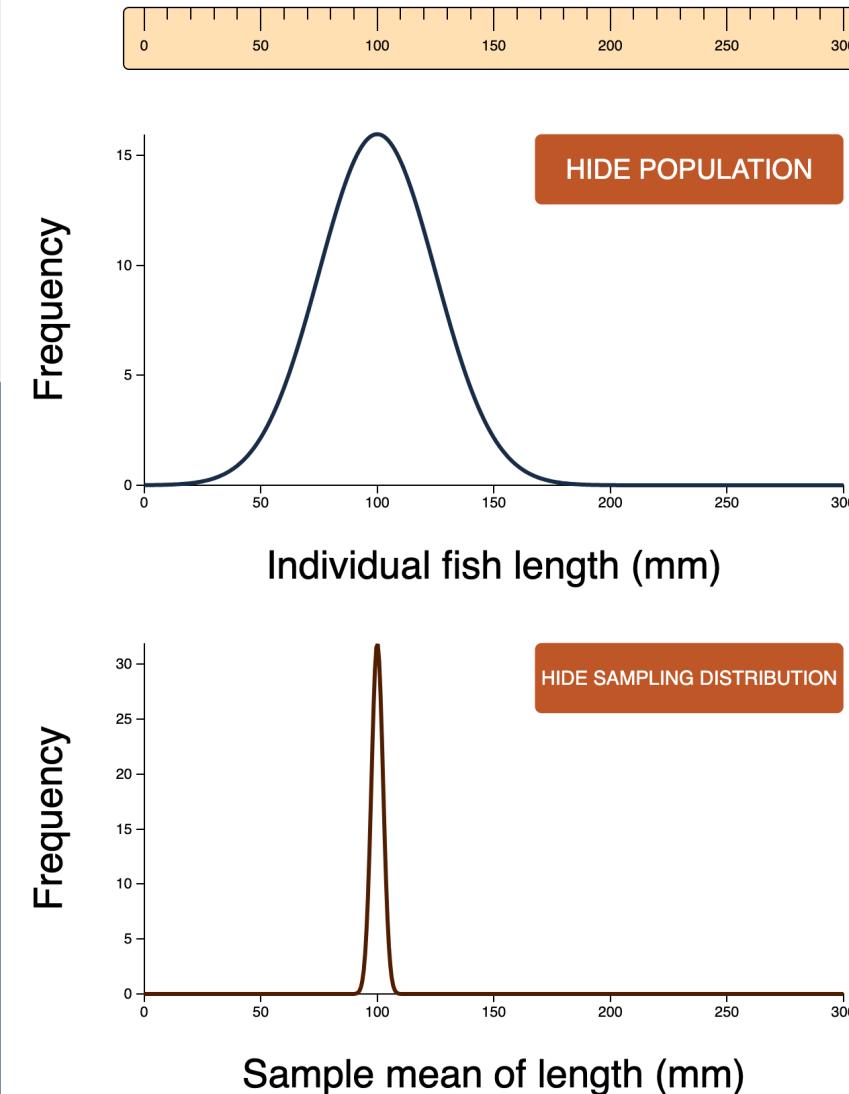


<https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm>





<https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm>



2.2 估计值不确定性的度量

- 标准误 Standard error

- 估计值的标准误是该估计值的抽样分布的标准差;
 - The standard error of an estimate is the standard deviation of the estimate's sampling distribution.
- 它反映了估计值与目标参数之间的差异，即估计值的精确度;
 - It reflects the differences between an estimate and the target parameter,
 - i.e., the precision of an estimate.
- 标准误差越小 → 目标参数的不确定性就越小;
 - The smaller the standard error, the less uncertainty there is about the target parameter in the population.

- SE的计算:

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

$\sigma_{\bar{Y}}$: SD of the mean (总体均值的标准差=标准误)

σ : SD of the population (总体的标准差)

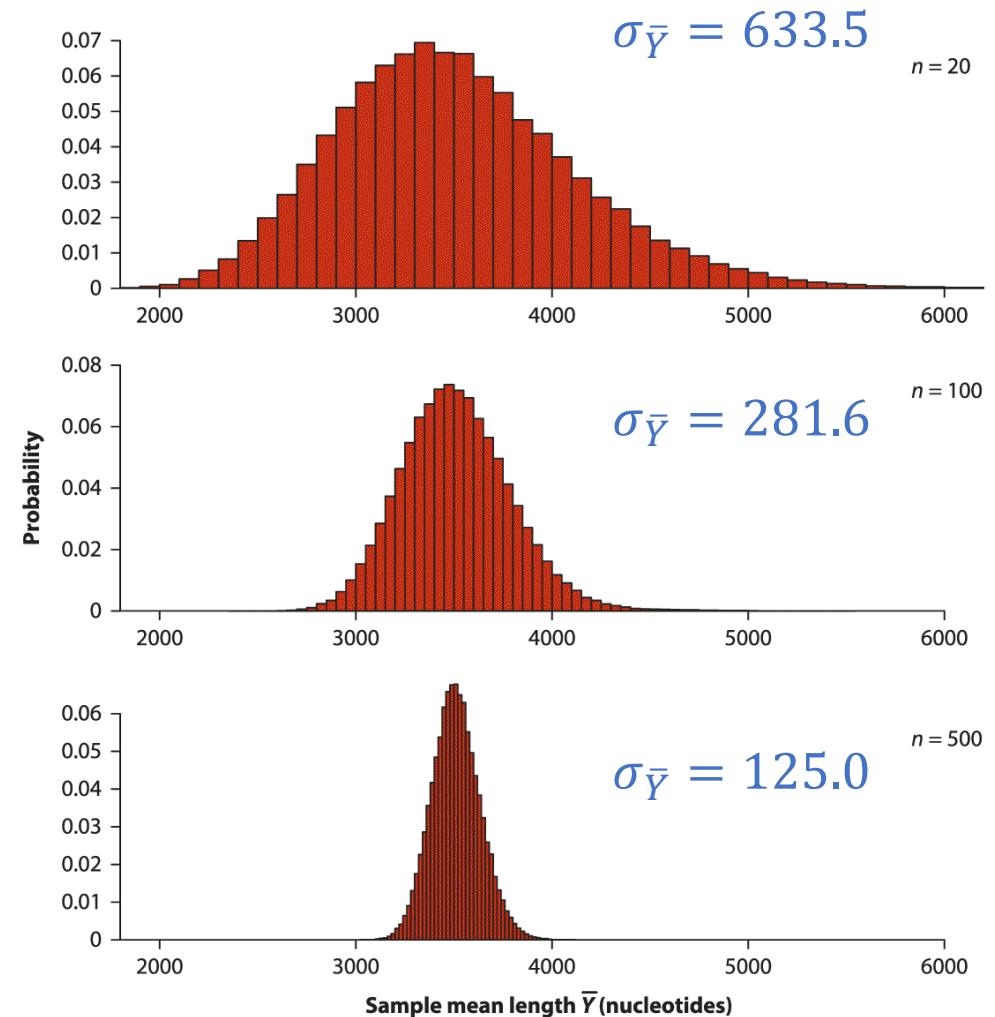
n : sample size (每次抽样的样本量)

2.2 估计值不确定性的度量

- 标准误 Standard error

- 估计值的标准误是该估计值的抽样分布的标准差；
- 随着样本大小的增加，标准误减小；

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$



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

2.2 估计值不确定性的度量

- 标准误 Standard error
 - 我们一般不知道总体估计值的标准差 σ ;
- 从样本数据中估计 \bar{Y} 的标准误
 - 对总体均值的标准误的估计等于样本的标准差 s 除以样本量的平方根

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} \approx SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{Y}}$: SD of the mean (总体均值的标准差=标准误)

σ : SD of the population (总体的标准差)

n : sample size (每次抽样的样本量)

$SE_{\bar{Y}}$: SE of the sample mean (样本均值的标准差=标准误)

s : SD of the sample (样本的标准差)

2.2 估计值不确定性的度量

- 从样本数据中估计 \bar{Y} 的标准误
 - 得出总体平均值的标准误的近似值;

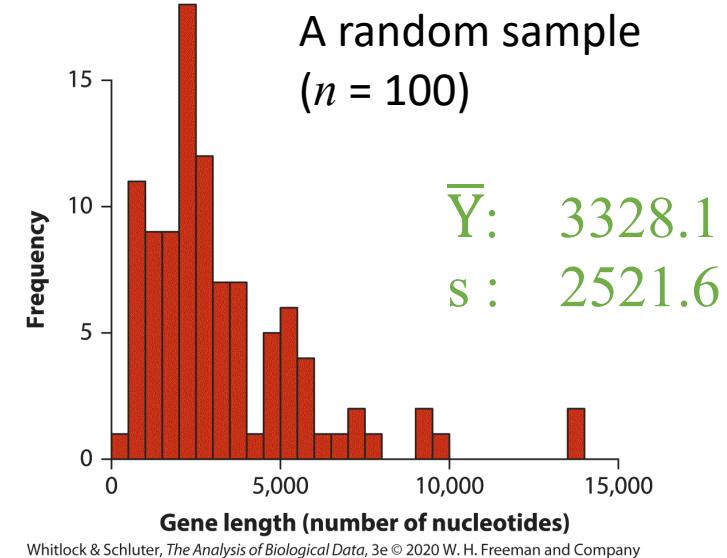
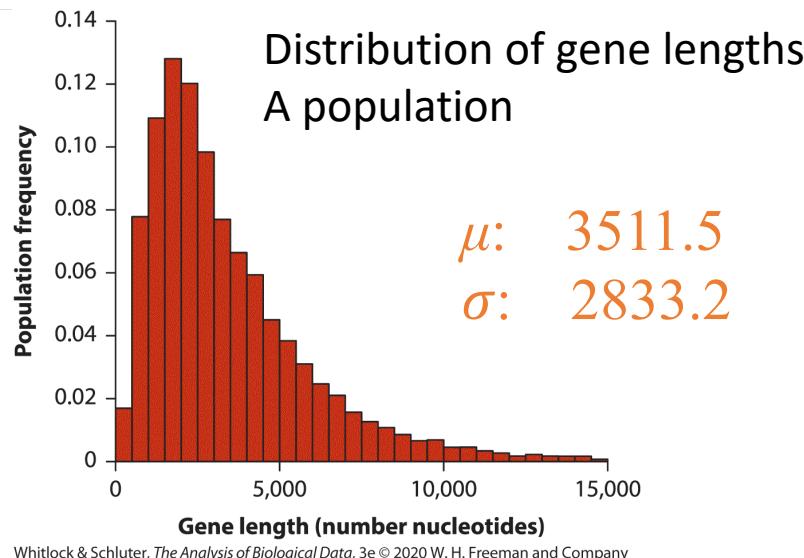
$$\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{Y}} \approx SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

- 从数据中估计均值的标准误 $SE_{\bar{Y}}$ 是生物学中的惯常;
 - 通常会在报告均值等估计值的同时需要报告标准误;
 - 报告 $SE_{\bar{Y}}$ 即是报告估计的不确定性，同时隐含了样本大小的信息。

$$\bar{Y} \pm SE_{\bar{Y}} (SE)$$

2.2 估计值不确定性的度量

$$\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{Y}} \approx SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$



$$\mu \pm \sigma_{\bar{Y}}$$

$$3511.5 \pm 281.6$$

$$\bar{Y} \pm SE_{\bar{Y}} (\text{SE})$$

$$3328.1 \pm 252.2 (\text{SE})$$

2.2 估计值不确定性的度量

- 每个估计值都有一个带标准误的抽样分布;
 - Every estimate, not just the mean, has a sampling distribution with a standard error, including the proportion, median, correlation, difference between means, and so on.
- 标准误是表示估计值不确定性的常用方法;
 - The standard error is the usual way to indicate uncertainty of an estimate.



2.3 置信区间 Confidence intervals (CIs)

- 另一种量化估计值不确定性的常见方式是置信区间（CI）
- 它是围绕样本估计的一个范围，可能包含总体参数；
 - CI是一个范围；
 - 表示总体参数可能落在这个范围内（通常使用95% CI）；
 - 例如，基因片段长度均值 μ 的置信区间为：
(the lower limit) $2827.8 < \mu < 3828.4$ (the upper limit)
- 如何解释？ How to interpret CI?

2.3 置信区间 Confidence intervals (CIs)



- 如何解释 CI?

- 例如，均值的95%置信区间是一个范围，可能包含真实均值 μ ;
 - The 95% CI of μ : (2827.8, 3828.4)

- 1) 我们有 95% 的把握认为: 真正均值介于 2827.8 和 3828.4 个核苷酸之间 (“We are 95% confident that the true mean lies between 2827.8 and 3828.4 nucleotides.”);
- 2) 总体均值在 2827.8 和 3828.4 个核苷酸之间的概率为 95% (“There is a 95% probability that the population mean falls between 2827.8 and 3828.4 nucleotides.”);

2.3 置信区间 Confidence intervals (CIs)



- 如何解释 CI?

- 例如，均值的95%置信区间是一个范围，可能包含真实均值 μ ;
 - The 95% CI of μ : (2827.8, 3828.4)

1) 我们有 95% 的把握认为: 真正均值介于 2827.8 和 3828.4 个核苷酸之间 (“We are 95% confident that the true mean lies between 2827.8 and 3828.4 nucleotides.”);

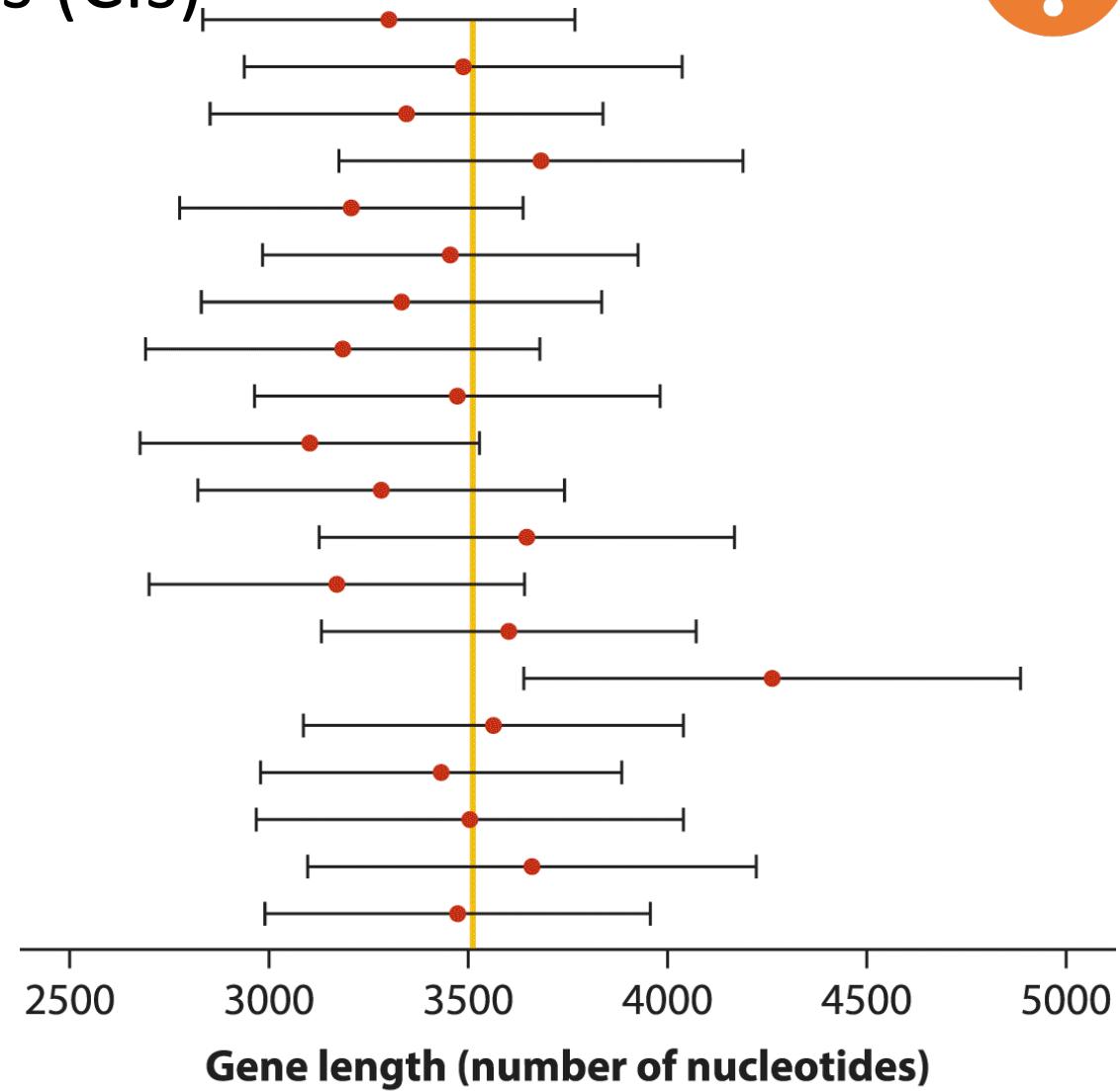
- 2827.8和3828.4都是常数 (**constants**)，并且真实均值要么在这两个数字形成的范围之间，要么不在这个范围中，因此没有涉及概率；
- 95%的置信区间将在95%的随机样本中捕获总体均值。



2.3 置信区间 Confidence intervals (CIs)

- 如何解释CI?
 - 一个围绕样本估计的范围，可能包含总体参数的值。
 - 我们有95%的把握认为真实均值在这个范围内。
 - 95%的置信区间在95%的随机样本中捕获总体均值。

e.g., 19 out of 20 (95%) of the researchers' intervals will contain the value of the population parameter.



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

2.3 置信区间 Confidence intervals (CIs)

- **95% confident**
 - 从总体中抽取100个不同样本，每个样本都用相同的统计量构造的置信区间（注意：由于样本不相同，这些置信区间的范围也不尽相同）
 - 那么有95个置信区间包含了总体参数的真值
- Parameter 参数（真值）vs. Estimate 估计值（随机变量）
 - 频率学派认为真值是一个常数，而非随机变量（后者是贝叶斯学派）
 - 所以我们不对真值做概率描述

2.3 置信区间 Confidence intervals (CIs)

Calculating the confidence interval

x

Sample properties

Sample mean, \bar{x} = ?

Sample standard deviation, s = ?

Sample size, n = 10

Degrees of freedom (df) = $n - 1 = 9$

$t_{\alpha/2}, df$ = ?

Formula for confidence intervals

Lower Bound:

$$\bar{x} - t_{\alpha/2}, df \frac{s}{\sqrt{n}} = ?$$

Upper Bound:

$$\bar{x} + t_{\alpha/2}, df \frac{s}{\sqrt{n}} = ?$$

(Click on variables to replace them with their current values.)

This confidence interval does? include the true mean.

95% confidence intervals for the mean

Successes: 40 Failures: 2 Success rate: 95.2%

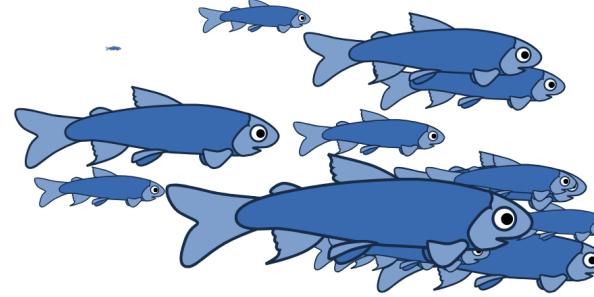
MAKE A SAMPLE



REPEATED SAMPLES

FASTER!

USE 99% CONFIDENCE



n 10

μ 100

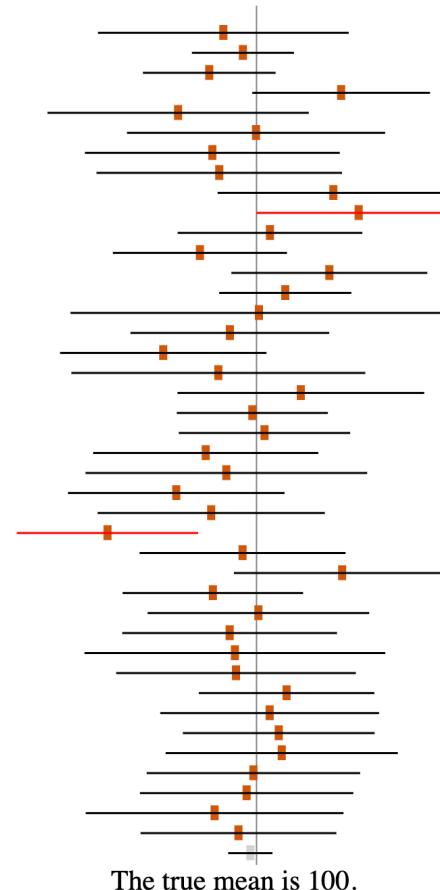
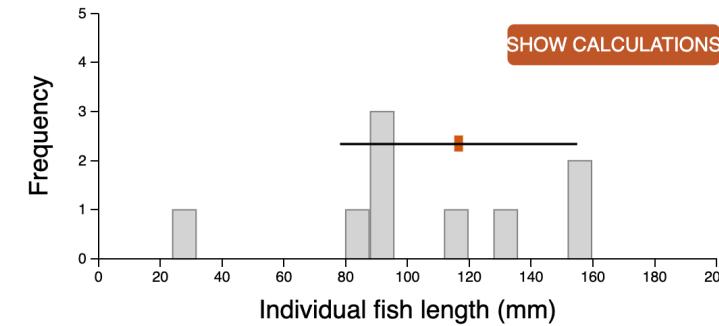
σ 50

<https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm>

Confidence intervals for
the mean



TUTORIAL



2.3 置信区间 Confidence intervals (CIs)

- 一般来说，95%置信区间的宽度可以很好地衡量我们对参数真实值的不确定性；
 - 如果置信区间很宽，则估计的不确定性很高；说明样本数据对于总体参数位置的信息量不足；
 - 如果置信区间很窄，我们可以相信估计值比较接近参数。
- 95%的置信区间提供了参数的最有可能的范围；
 - 基于数据来说，在区间内的值最有可能；
 - 而在区间外的值可信度较低；

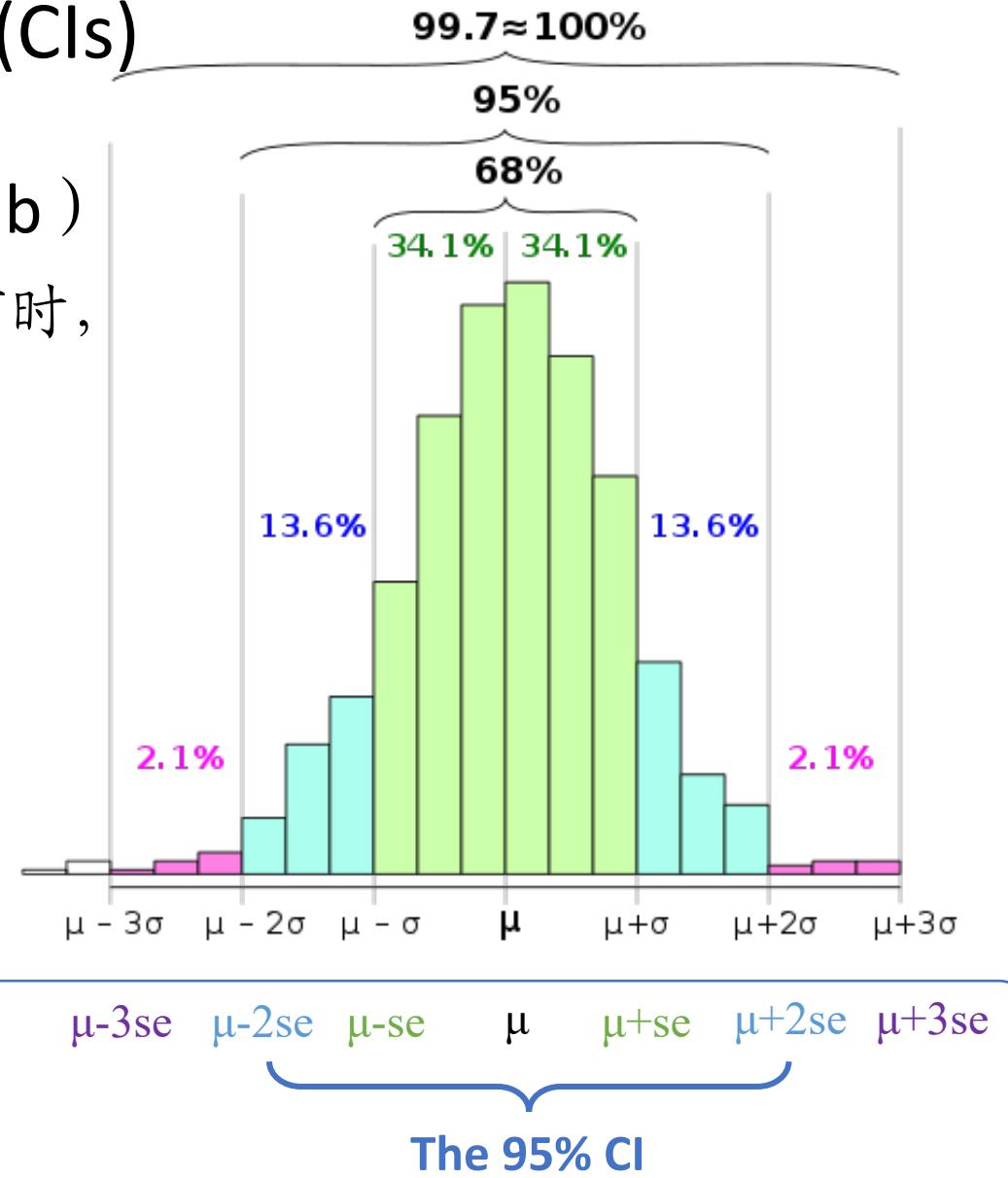
2.3 置信区间 Confidence intervals (CIs)

- 2SE的经验法则 (the 2SE rule of thumb)

- 当样本是随机样本，且数据呈正态分布时，
- 一个近似是：
 - $95\% \text{ CI} = 2 \text{ SE}$
 - The 95% CI of μ : (2827.8, 3828.4)
 - $\bar{Y} \pm 2\text{SE}_{\bar{Y}}$: (2827.7, 3832.5)

- The population (or a sample)

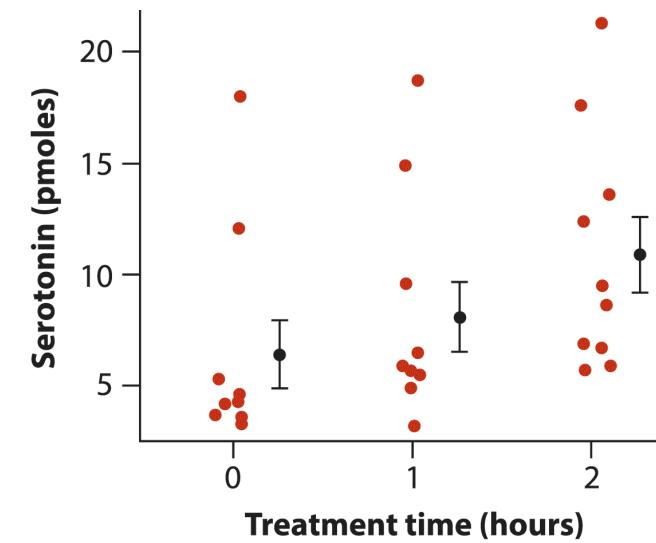
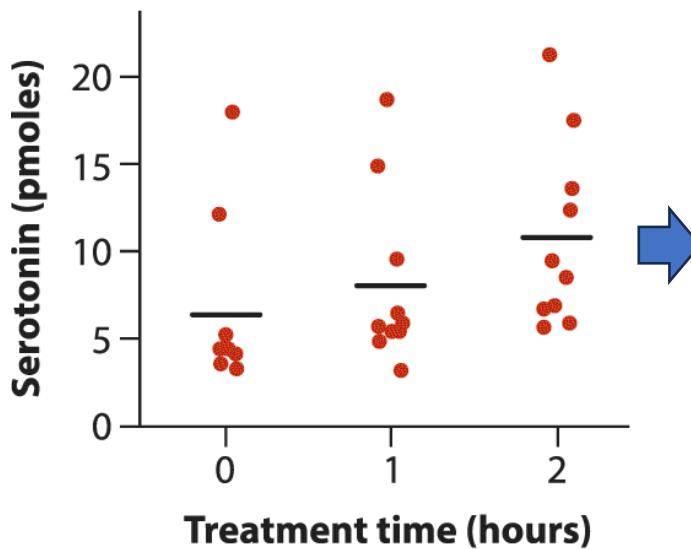
- the sampling distribution of
the mean (multiple samples)



2.4 误差线 Error bars

- 图像中，均值（和其它参数）的标准误或置信区间通常以“误差线”进行表示；
 - 以说明参数估计的精确度（不确定性）；
 - 而不是数据的变异性；
 - 不是数据的一部分；
- 例如， $\text{mean} \pm \text{SE}$
 - 均值的一个标准误差以上
 - 均值的一个标准误差以下

e.g., strip chart: mean \pm SE
the behavior change in locusts
(from solitary to gregarious)



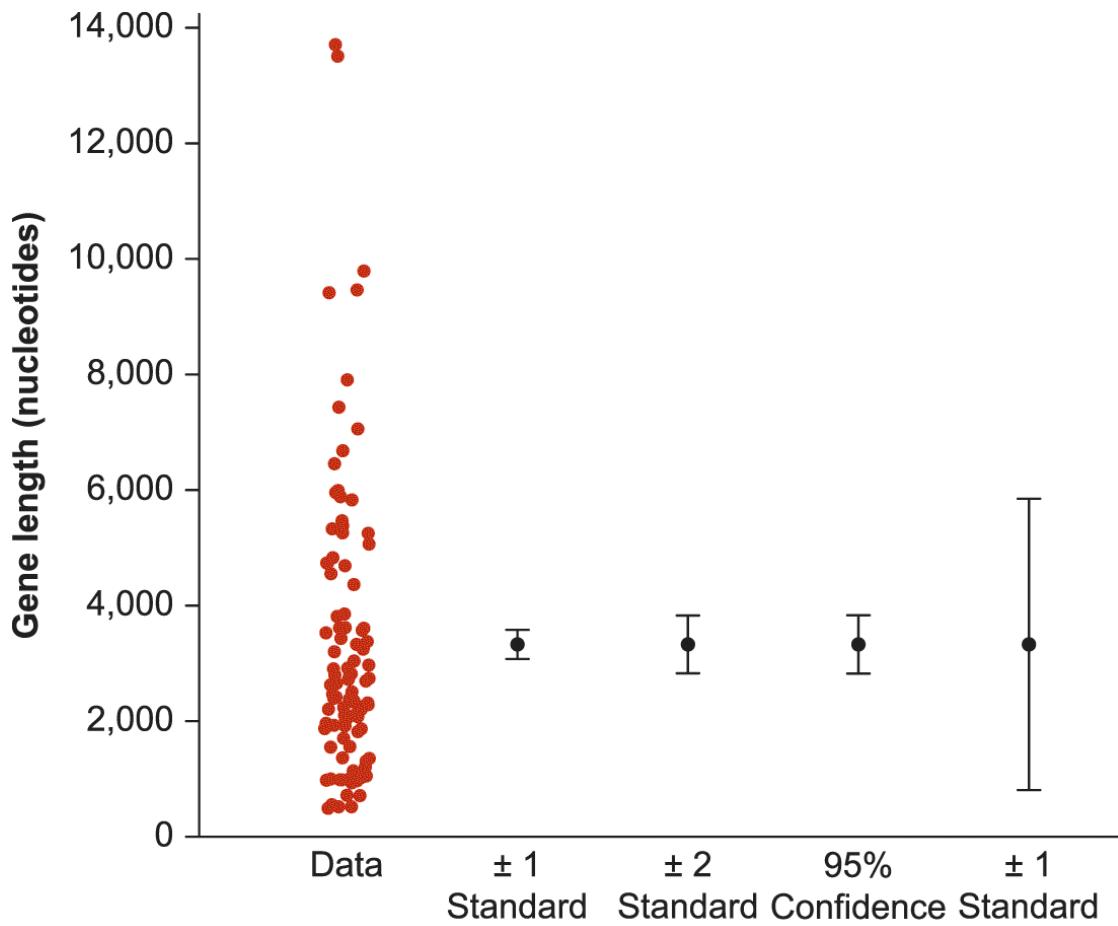
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.4 误差线 Error bars

- 误差线可以表示不同的量：

- 1 SE
- 2 SE = 95% CI
- SD (misleading)
 - 误差线用来显示SD是一种表示数据变异性的不佳方法；
 - 是多余的显示数据的方式，因为原数据点自身就可以表示散布程度/变异性；

e.g., gene lengths in a random sample
(n = 100)



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

3. Probability 概率

- 从样本数据计算得出的估计值几乎从不与总体参数值相同；
 - 因为抽样会受到偶然性的影响。
- 关键问题是：
 - “面对偶然性，我们在多大程度上能相信一个估计值？”
 - 换句话说，它的精确度是多少？

4. 小结——估计的不确定性

- 所有估计值都有一个抽样分布，即在给定样本量的随机抽样条件下可能得到的估计值的所有可能值的概率分布；
 - 即多次抽样的估计值的分布；
 - 通常标准误和置信区间的计算公式都假设抽样是随机的；
- 估计值的标准误是其抽样分布的标准差。
 - 标准误度量精确度；
 - 估计值的标准误随着样本量的增加而减小；

4. 小结——估计的不确定性

- 置信区间
 - 从样本数据计算出的一个数值范围;
 - 置信区间内可能包含目标参数的数值;
- 2SE的经验法则
 - 即样本估计值（如均值）加减两个标准误;
 - 为估计值的95%置信区间提供了一个粗略的近似;
- 通常在图中添加误差线以说明标准误或置信区间

4. 小结——估计的不确定性

R commands summary

Mean

Vector of numerical variable
Ignore missing data
`mean(titanicData$Age, na.rm = TRUE)`

Median

Vector of numerical variable
Ignore missing data
`median(titanicData$Age, na.rm = TRUE)`

Summary

`summary(titanicData$age)`

Variance

Vector of numerical variable
Ignore missing data
`var(titanicData$Age, na.rm = TRUE)`

Standard deviation

Vector of numerical variable
Ignore missing data
`sd(titanicData$Age, na.rm = TRUE)`

Coefficient of variation

`sd(titanicData$age, na.rm = TRUE) / mean(titanicData$age, na.rm = TRUE) * 100`

Interquartile range

Vector of numerical variable
Ignore missing data
`IQR(titanicData$Age, na.rm = TRUE)`

Confidence interval of mean

Vector of numerical variable
Calculate confidence interval (95% by default)
`t.test(titanicData$age)$conf.int`