Lecture 4 – Uncertainty and Probability

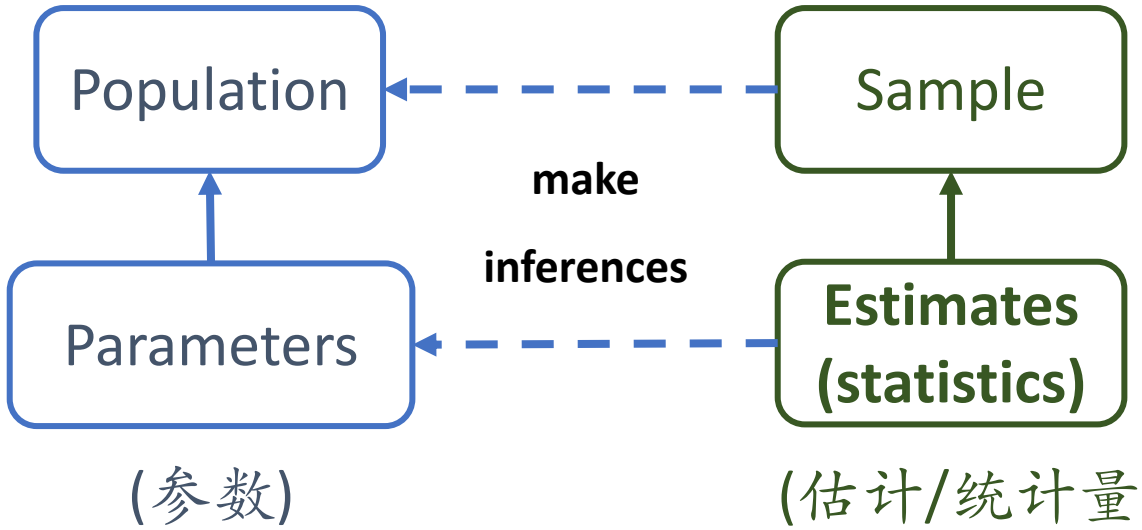
- Outline for today
  - Recall L03
  - Uncertainty of estimates
    - Standard error
    - Confidence interval
  - Probability
  - Summary
  - R Lab & Discussion

八、评分标准【请按照本门课程采用的课程考核方式选择下表之一填写】（具体分段可以根据实际情况调整）

课程目标	评分标准				
	90-100	80-89	70-79	60-69	0-59
目标 1：学生能够理解基本统计学概念与思想；	掌握抽样、参数估计、假说检验等统计学基本概念与方法；每次作业和考试能够独立完成，解答问题简洁明了，得分平均在 90 以上；熟练掌握 R 语言的使用。	掌握抽样、参数估计、假说检验等统计学基本概念与方法；每次作业和考试能够独立完成，解答问题思路清晰，得分平均在 80 以上；能较为熟练地使用 R 语言。	基本掌握抽样、参数估计、假说检验等统计学基本概念与方法；能在老师帮助或和同学讨论后独立完成作业，作业和考试答题思路清晰，得分平均在 70 以上；能较熟练地使用 R 语言。	基本了解抽样、参数估计、假说检验等统计学基本概念与方法；能在老师帮助或和同学讨论后独立完成作业，作业和考试得分平均在 60 以上；能使用 R 语言完成基本统计分析。	不理解抽样、参数估计、假说检验等统计学基本概念与方法；不寻求帮助或在老师同学帮助下仍无法独立完成作业，作业和考试得分平均在 60 以下；不能使用 R 语言完成基本统计分析。
目标 2：学生能够掌握并熟练应用常见的统计分析方法分析生物学数据，能够解释不同统计分析得到的结果。	熟练掌握常见生物学问题对应的统计分析方法，能借助 R 语言等工具独立完成数据分析，并精确解释分析结果。	掌握常见生物学问题对应的统计分析方法，能借助 R 语言等工具独立完成数据分析，并解释分析结果。	基本掌握常见生物学问题对应的统计分析方法，能借助 R 语言等工具独立完成数据分析，并解释关键分析结果。	了解常见生物学问题对应的统计分析方法，能借助 R 语言等工具完成基础数据分析，并解释关键分析结果。	不了解常见生物学问题对应的统计分析方法，不能借助 R 语言等工具完成基础数据分析，不会解释关键分析结果。

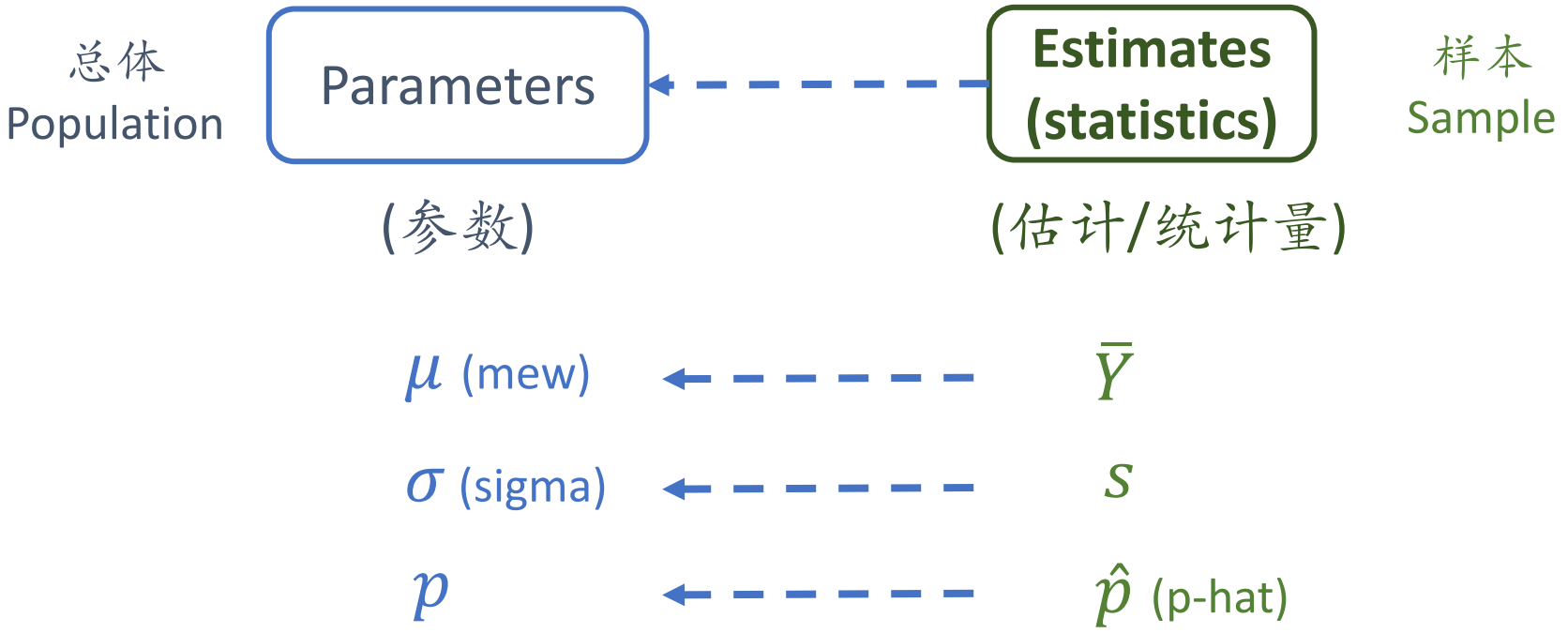
1. Recall from L03 - Descriptive statistics

- 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

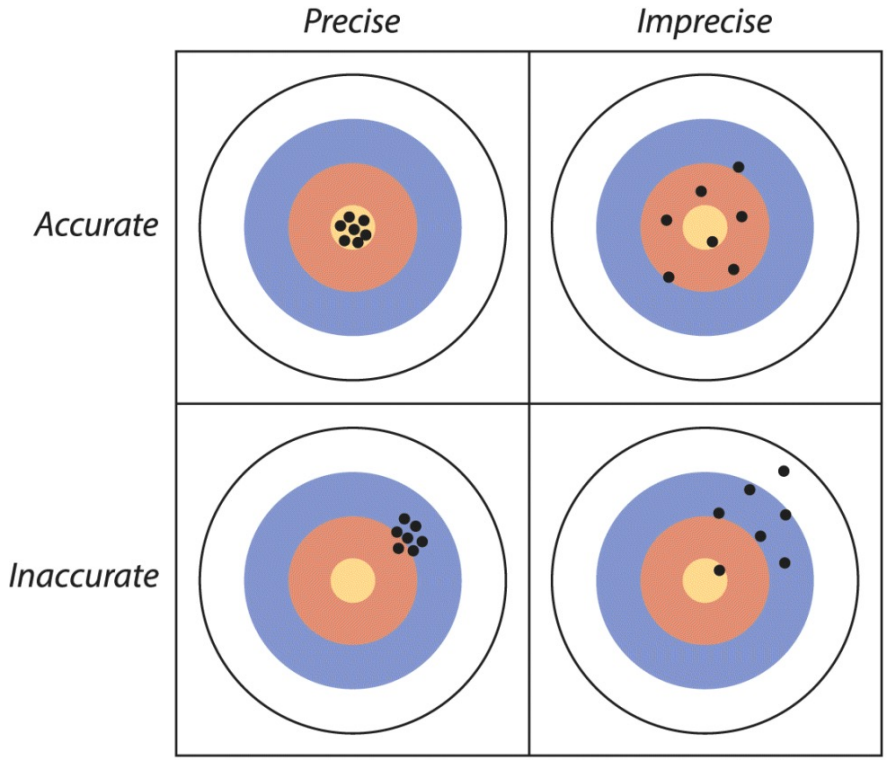
- Calculations
  - Mean 均值
  - SD 标准差
  - Median 中值
  - Quartile四分位数
  - IQR 四分位距
  - Quantile 分位数
  - Proportion 比例

- When to use them?
- And how?

2. Estimating with uncertainty

- For an estimate of a population parameter to be useful, we also need to quantify its **precision** (精确度).

- The value of an estimate calculated from data is almost **never** exactly the same as the value of the population parameter being estimated. **Why?**

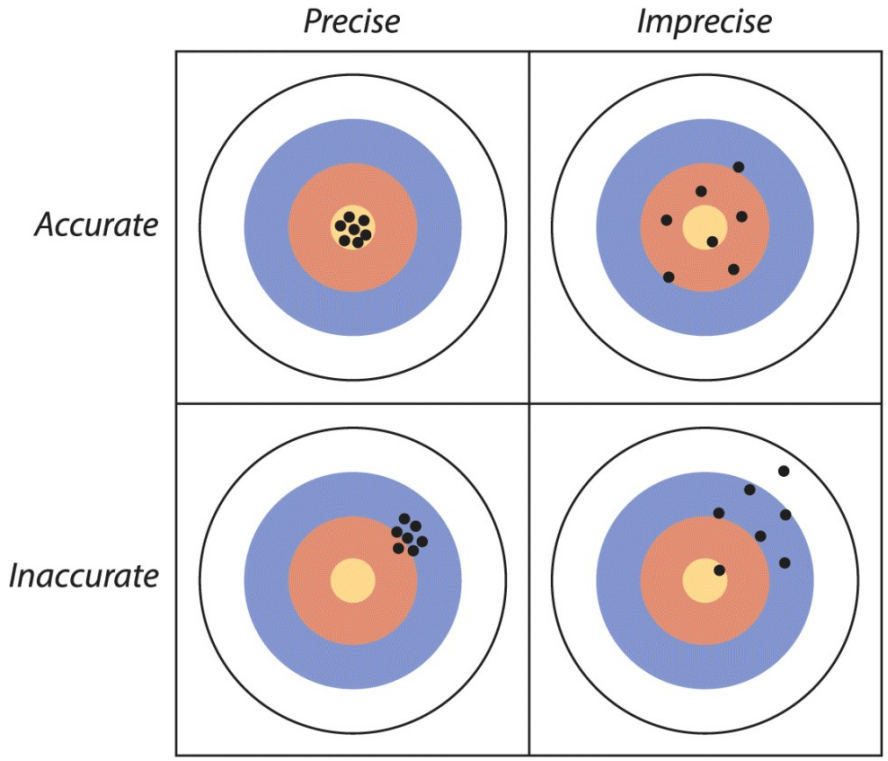


2. Estimating with uncertainty

- 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**.

  (抽样会受到偶然性的影响)

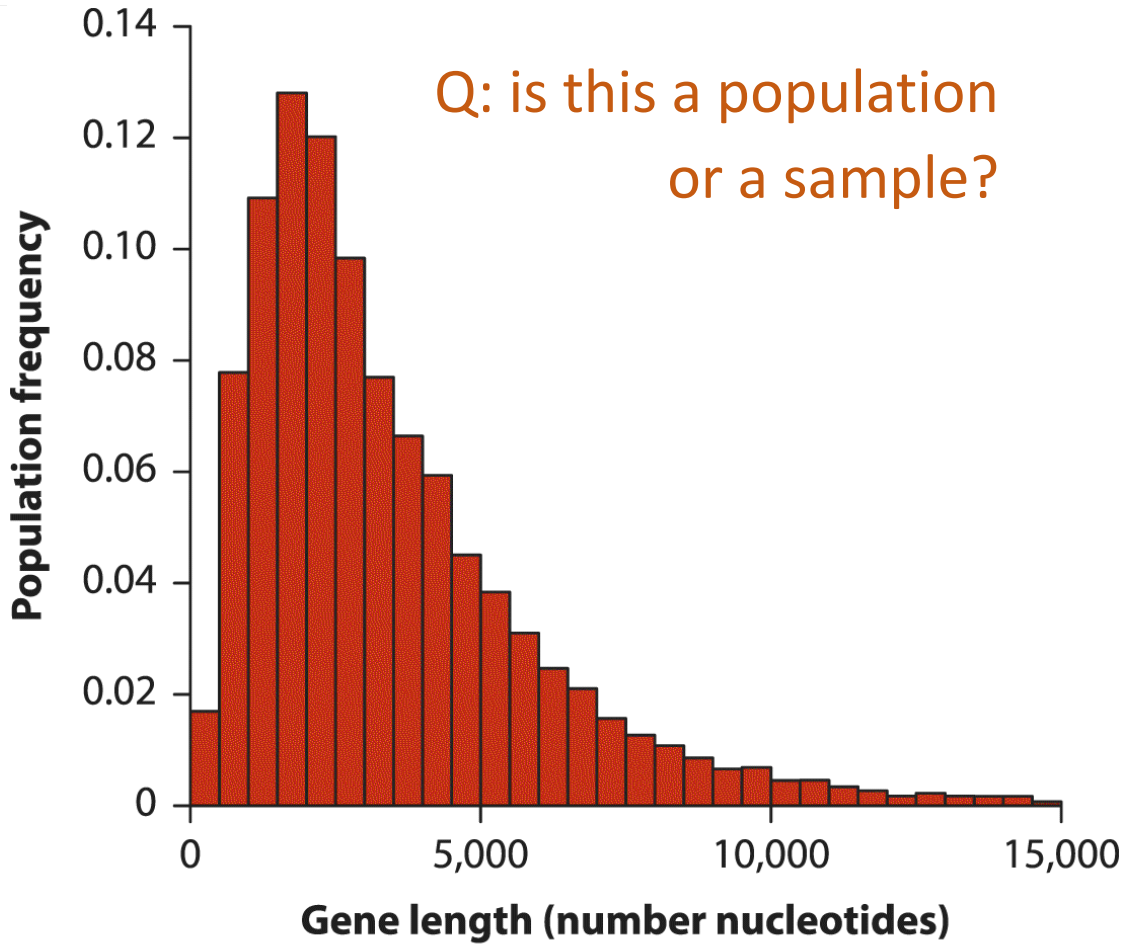
- The crucial question is “In the face of chance, how much can we trust an estimate?” In other words, what is its **precision**?



2.1 Sampling distribution (抽样分布)

- Example: The length of human genes
  - The **lengths** of all 20,290 known and predicted genes of the published genome sequence

    (Hubbard et al. 2005)
    - Based on the DNA sequence of all 23 human chromosomes via the international Human Genome Project
  - Distribution of gene lengths in the known human genome
    - truncated at 15,000 nucleotides
    - 26 larger genes are too rare to be visible here



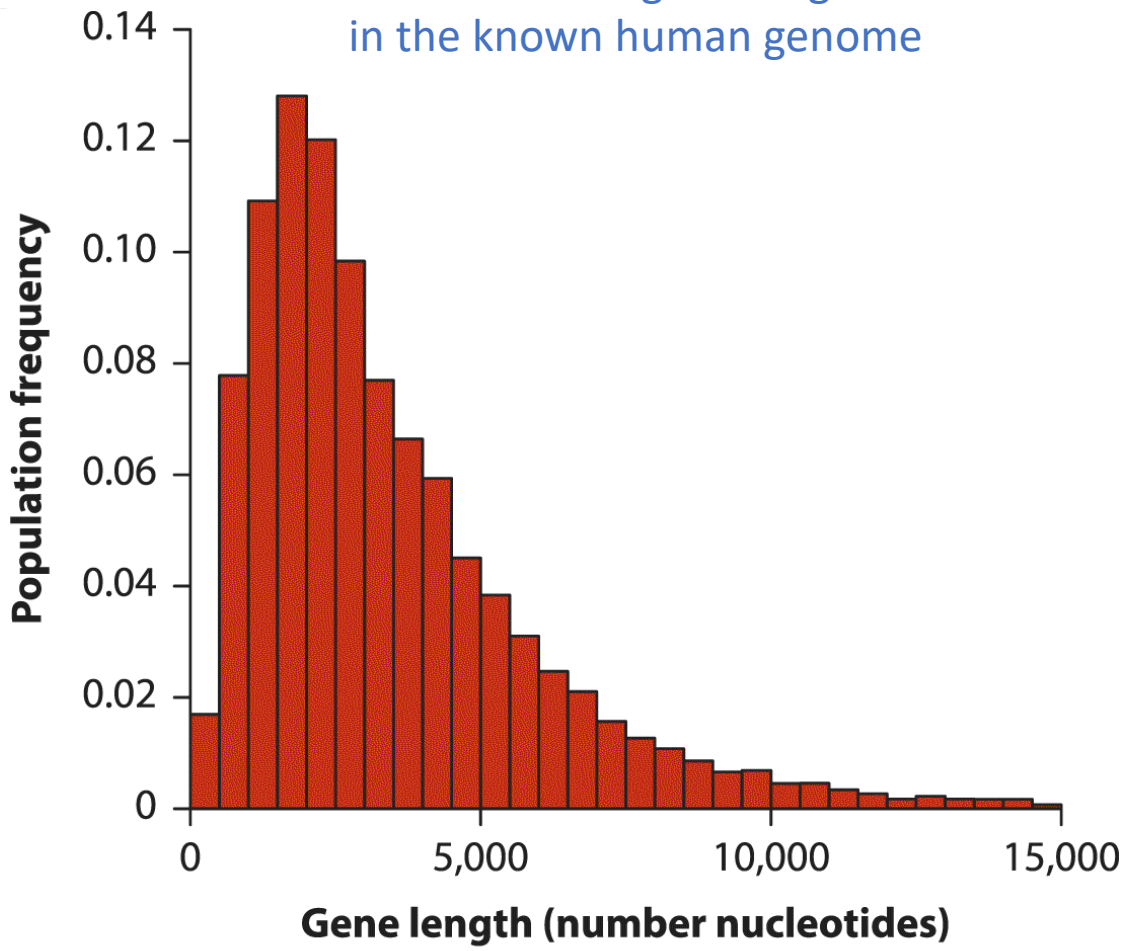
Q: is this a population or a sample?

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

2.1 Sampling distribution

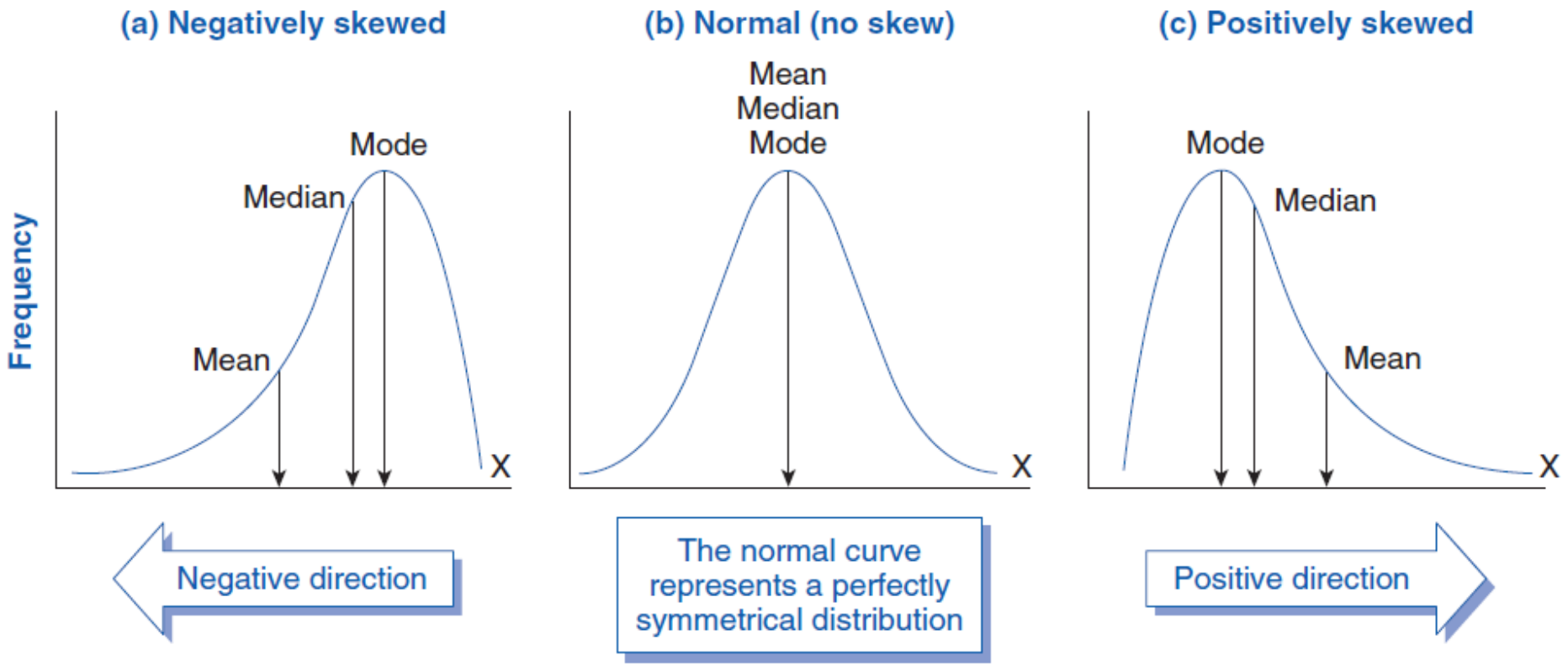
- This histogram shows the distribution of lengths in the **population** of genes.
  - the relative frequency of genes of a given length interval represents the probability of obtaining a gene of that length when sampling a single gene at random. (相对频率表示随机抽取单个基因时获得该长度基因的概率)
  - mean (μ) : 3511.5
  - Standard deviation (SD): 2833.2

Distribution of gene lengths in the known human genome



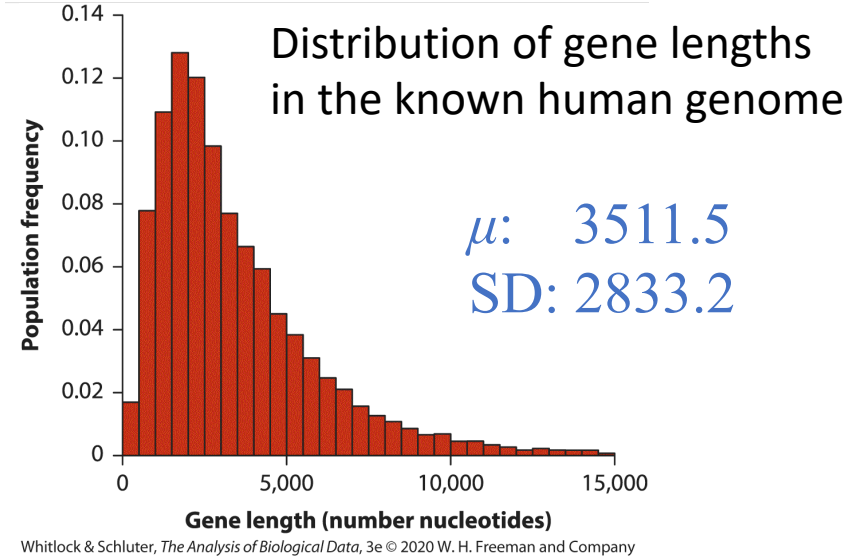
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Skewed distributions

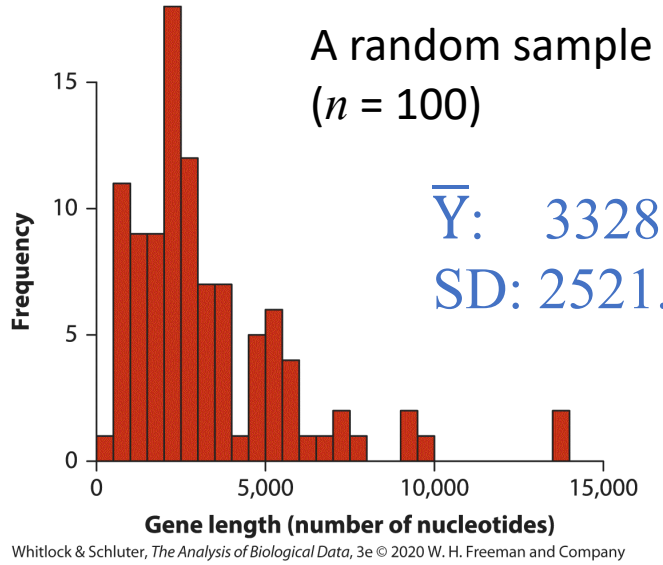


2.1 Sampling distribution

- Estimating mean gene length with a random sample
  - n = 100

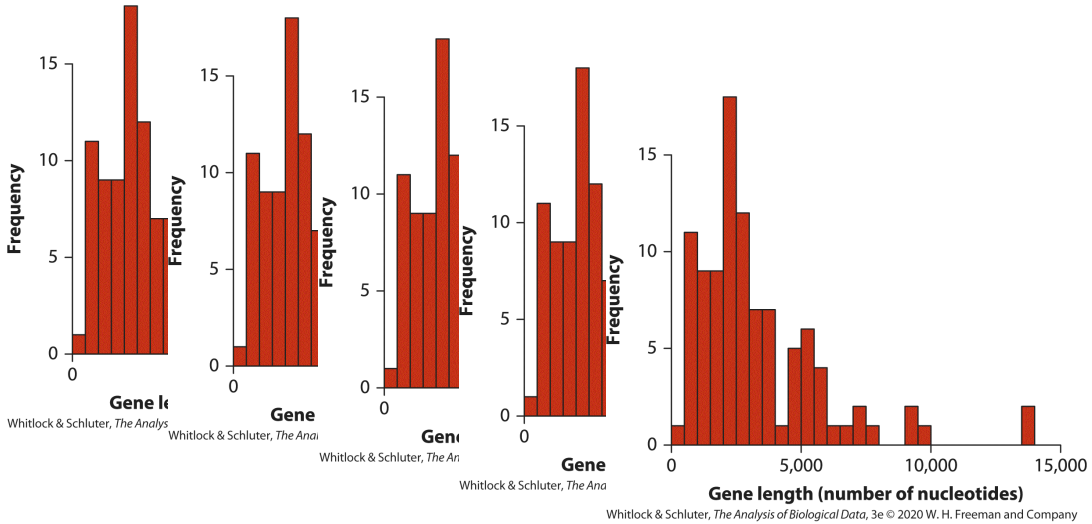


- The two distributions **share important features**, including approximate location, spread, and shape. (这两种分布具有共同的重要特征: 包括大致位置, 分布和形状)

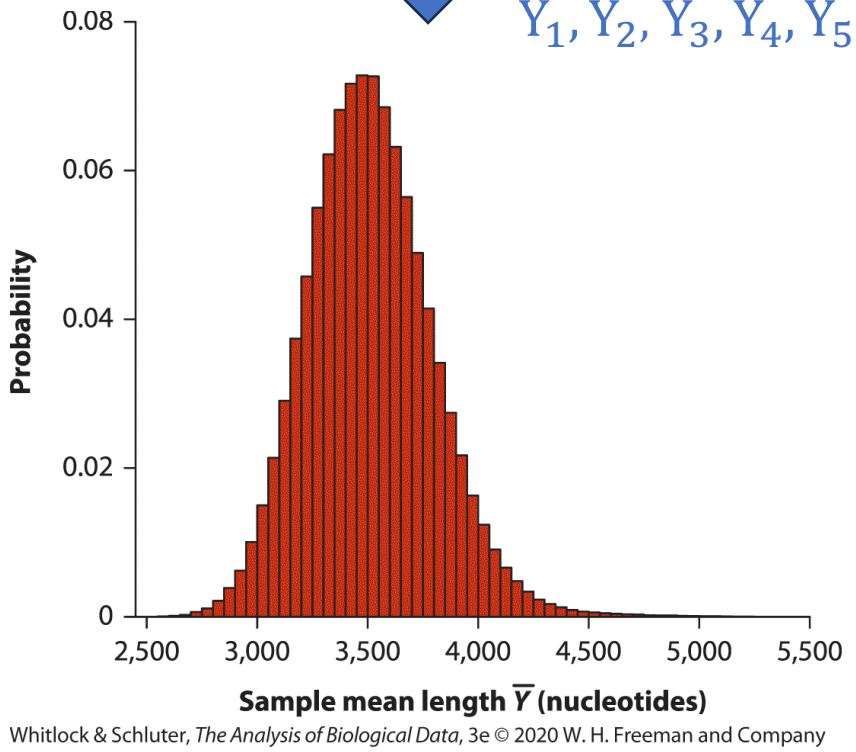


2.1 Sampling distribution

- 1) 重复取样 - repeat this sampling an infinite number of times (n = 100)
  - *sample 1, sample 2, sample3, …*
- 2) 概率分布 - create the probability distribution of the estimate (mean)
- The probability distribution of values we might obtain for an estimate makes up the estimate's **sampling distribution**. (估计值的概率分布构成了其抽样分布)

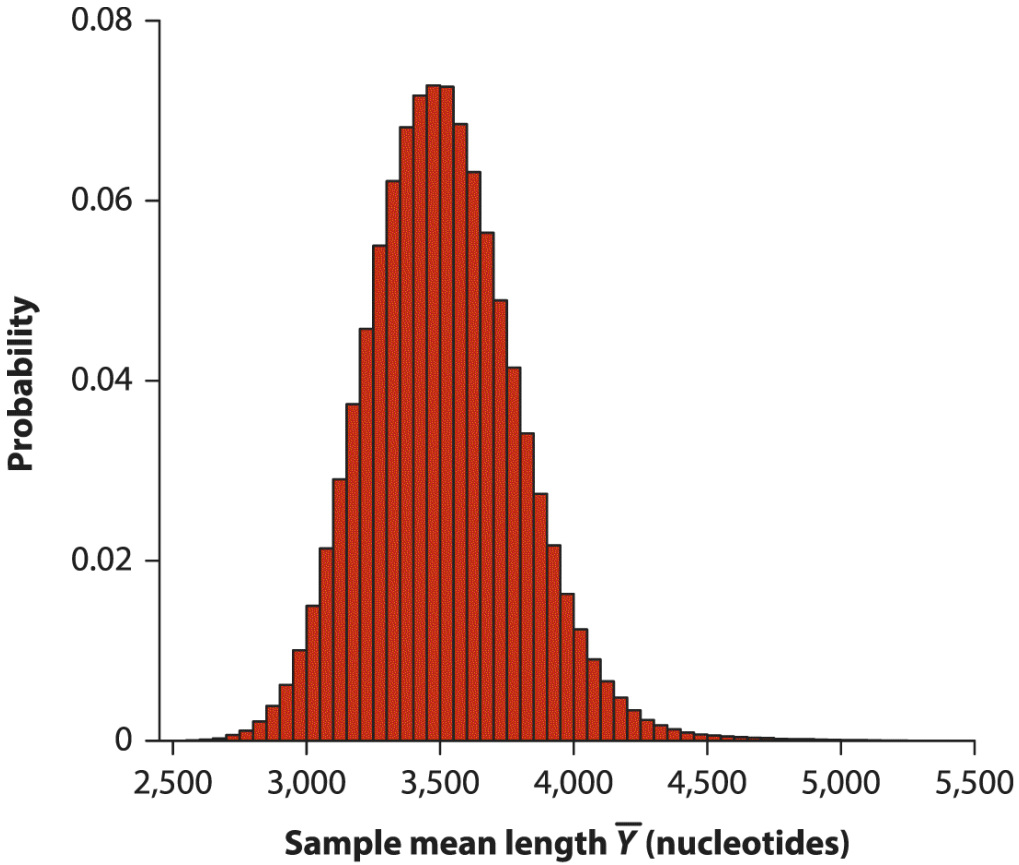


$\overline{Y}_1, \overline{Y}_2, \overline{Y}_3, \overline{Y}_4, \overline{Y}_5, \ldots$



2.1 Sampling distribution

- The sampling distribution represents the "population" of values for an estimate. (抽样分布代表估计值的 "总体")
  - It is not a real population.
- the population mean μ is a constant (3511.5), its estimate $\overline{Y}$ is a variable.
  - the sampling distribution for $\overline{Y}$ is centered exactly on the true mean.
  - This means that $\overline{Y}$ is an unbiased estimate of μ.



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

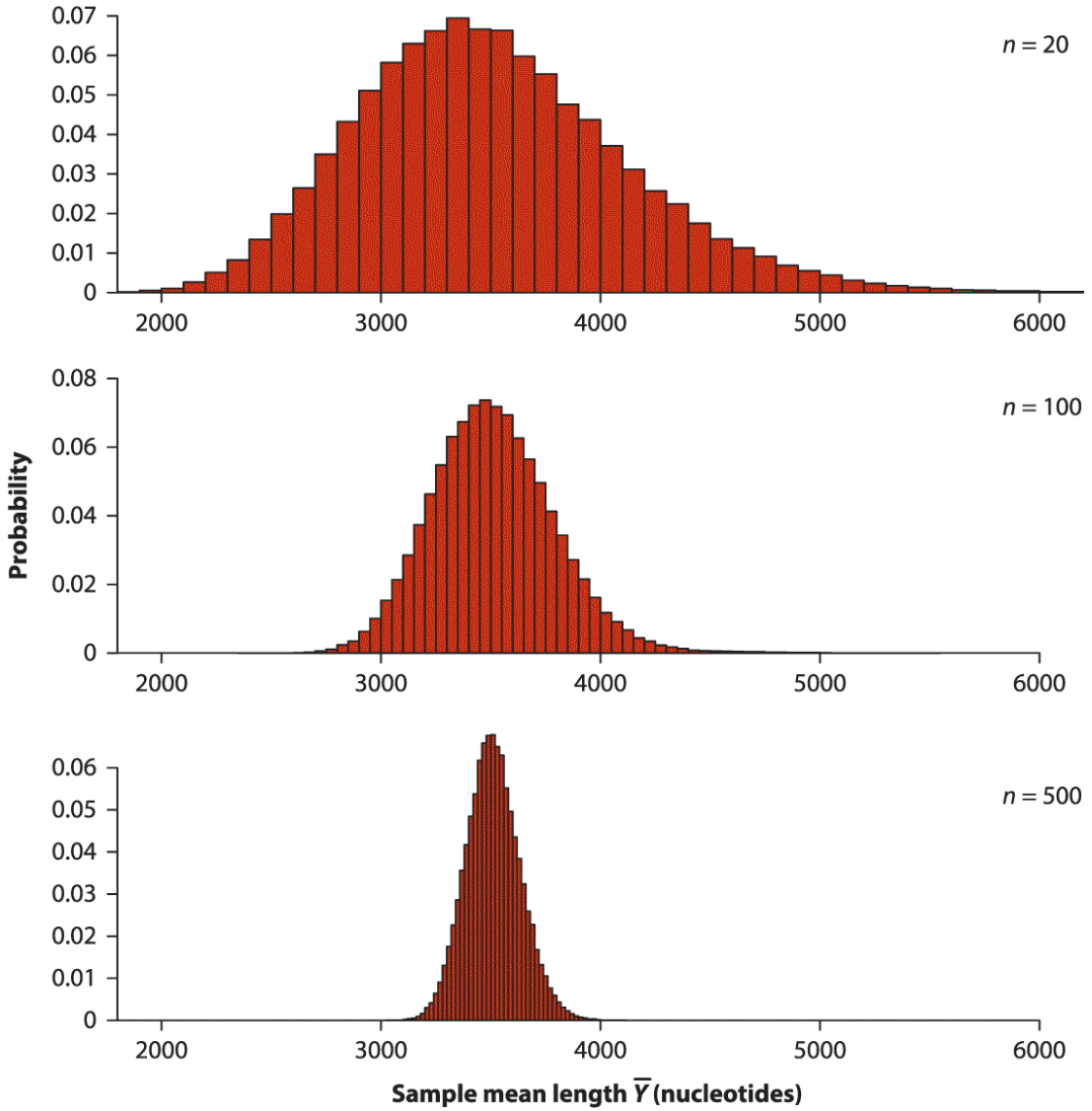
$\overline{Y}_1, \overline{Y}_2, \overline{Y}_3, \overline{Y}_4, \overline{Y}_5, \ldots$

2.1 Sampling distribution

- The spread of the sampling distribution of an estimate depends on the **sample size**.

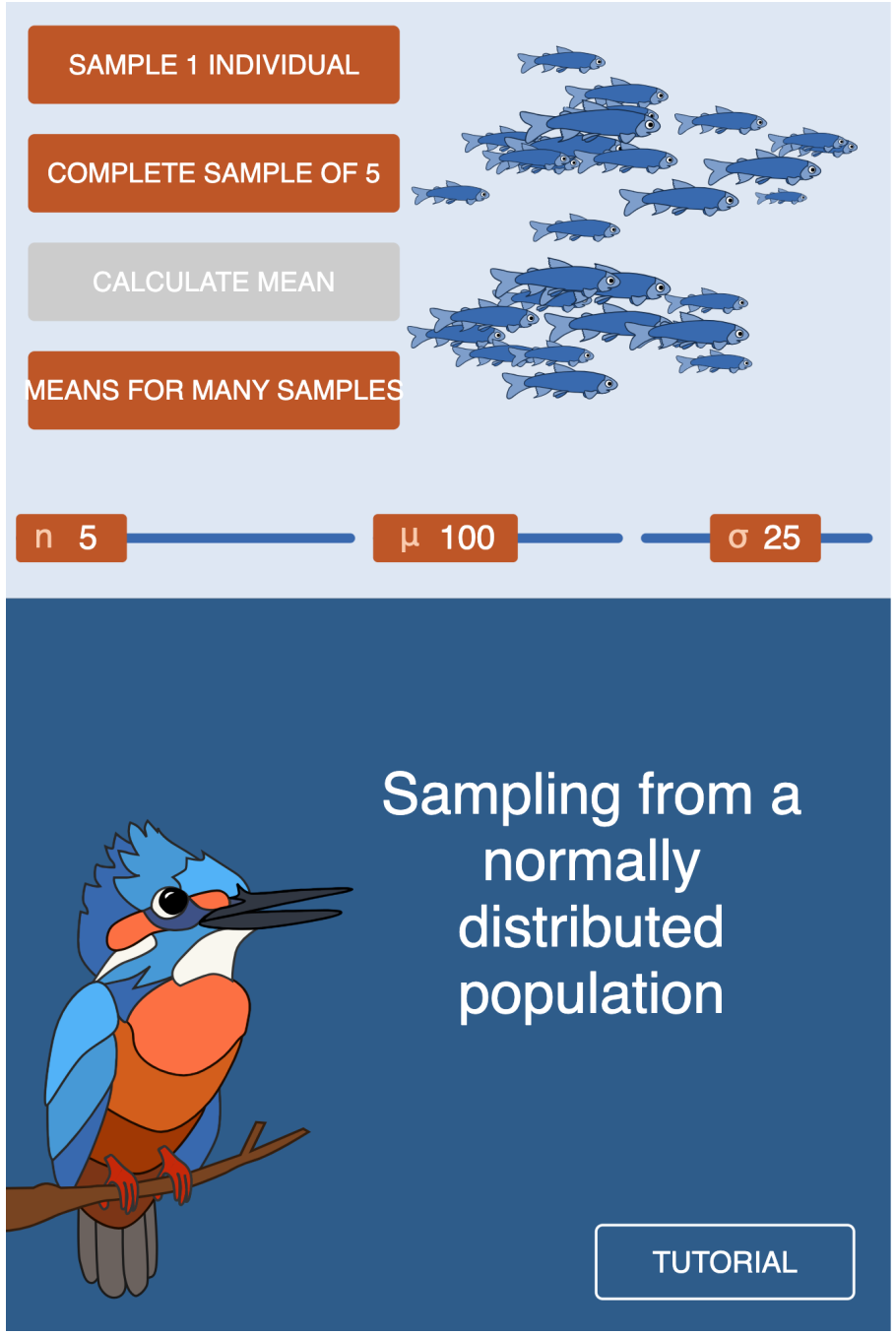
  (估计值的抽样分布的散布程度取决于样本大小)

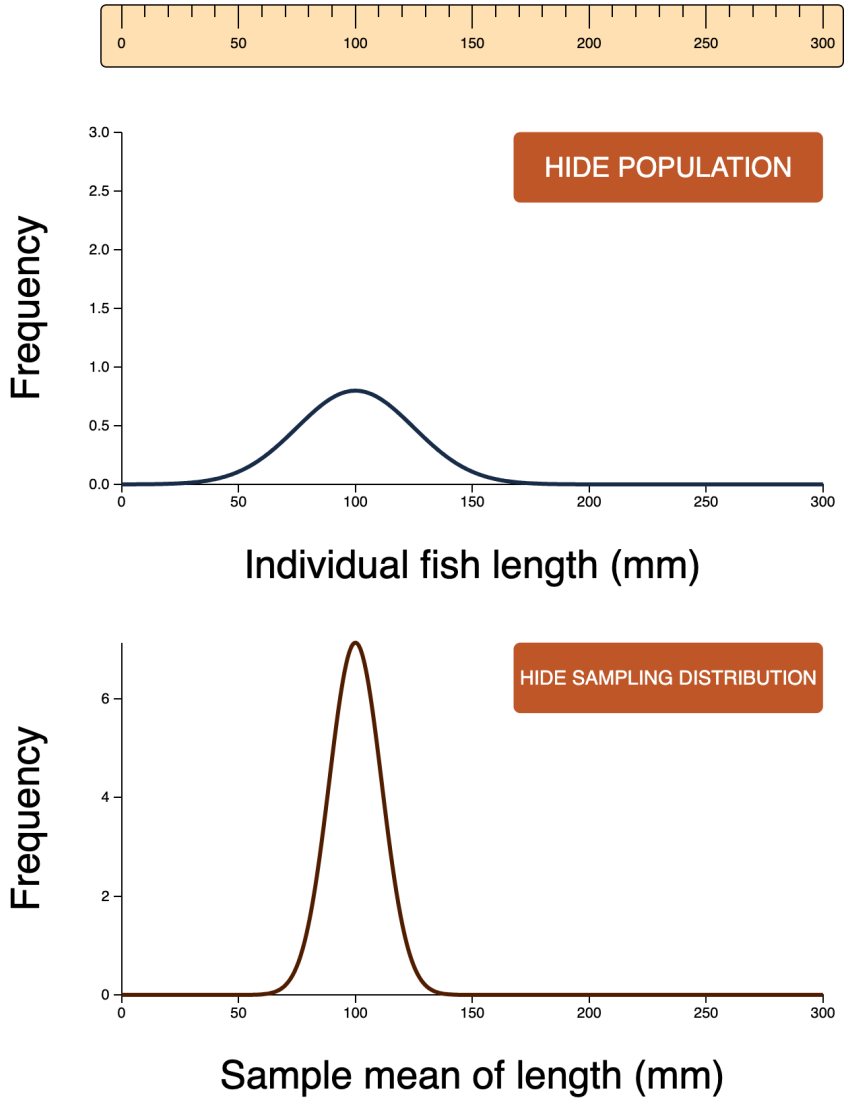
- The larger the sample size, the narrower the sampling distribution.
  - **larger samples are desirable** whenever possible because they yield more precise estimates.



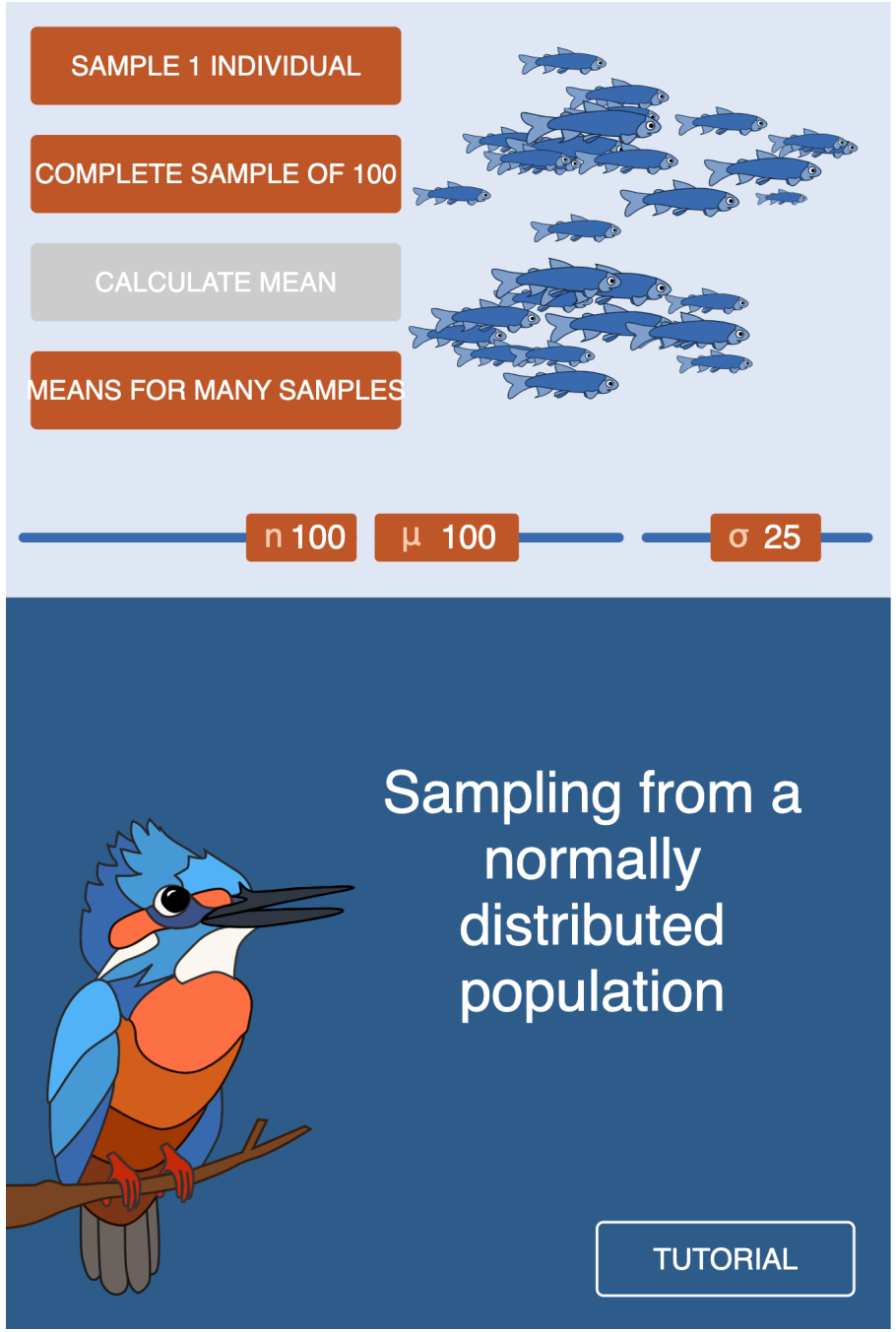
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

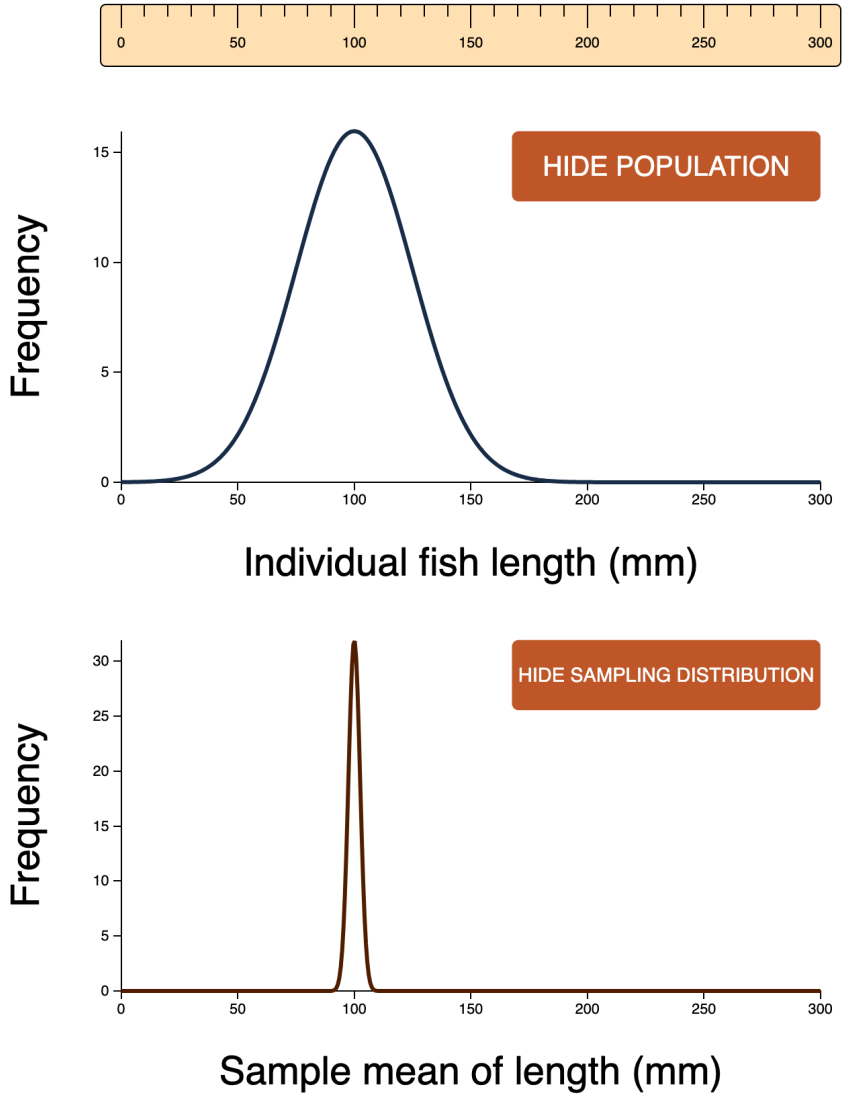
https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm





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





2.2 Measuring the uncertainty of an estimate

- Standard error 标准误
  - The standard error of an estimate is the standard deviation of the estimate's sampling distribution.

    (估计值的标准误是估计值抽样分布的标准差)
  - It reflects the differences between an estimate and the target parameter, i.e., the precision of an estimate. (反映了估计值与目标参数之间的差异)
  - The smaller the standard error, the less uncertainty there is about the target parameter in the population.

    (标准误差越小 → 目标参数的不确定性就越小)
- The calculation of SE

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

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

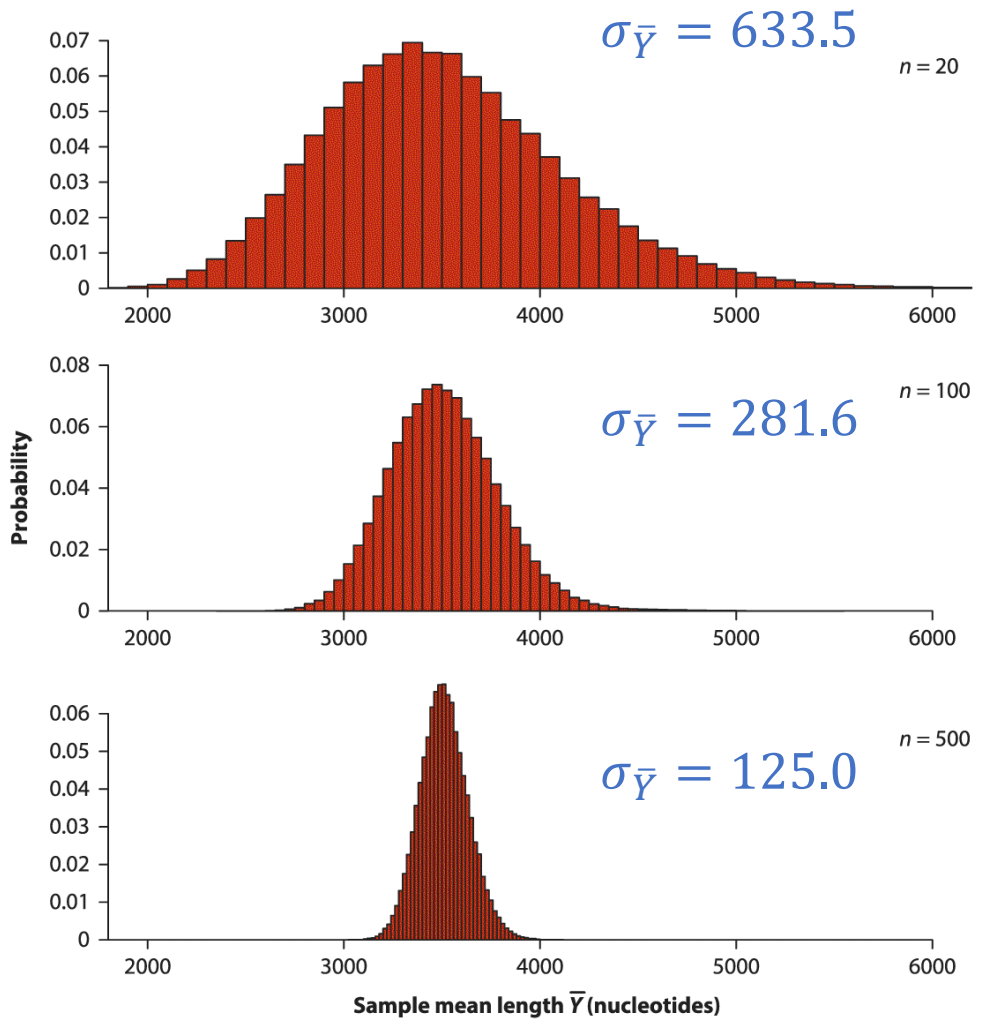
σ: SD of the population (总体的标准差)

n: sample size (每次抽样的样本量)

2.2 Measuring the uncertainty of an estimate

- Standard error 标准误
  - The standard error of an estimate is the standard deviation of the estimate's sampling distribution.
  - The standard error decreases with increasing sample size.

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



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

(Whitlock & Schluter 2020)

2.2 Measuring the uncertainty of an estimate

- Standard error 标准误
  - But, we almost never know the value of σ (总体的标准差)
- The standard error of $\bar{Y}$ from the data
  - The standard error of the mean is estimated from data as the sample standard deviation, s ,divided by the square root of the sample size, n.

  (对总体均值的标准误的估计等于样本的标准差除以样本量的平方根)

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

$\sigma_{\bar{Y}}$: SD of the mean (抽样均值的标准差)
σ: SD of the population (总体的标准差)
n: sample size (每次抽样的样本量)

s: SD of the sample (样本的标准差)

2.2 Measuring the uncertainty of an estimate

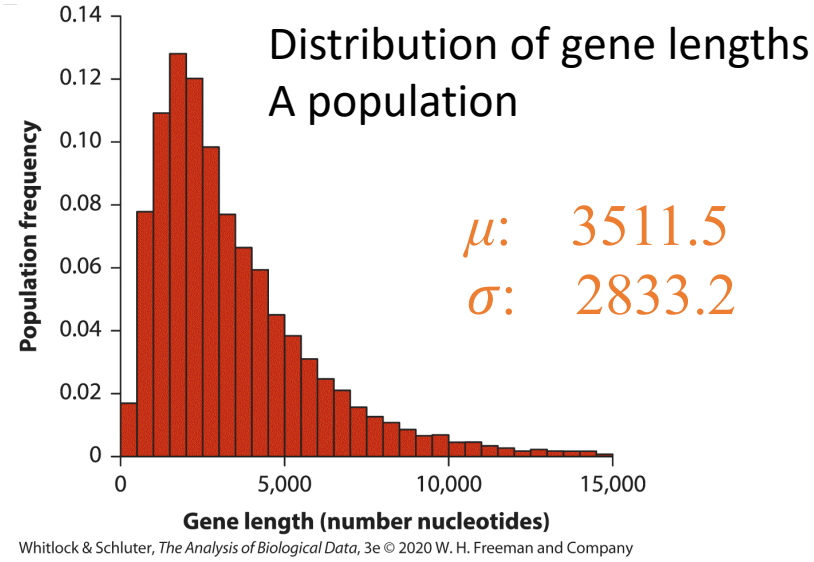
- The standard error of $\bar{Y}$ from the data
  - To **approximate** the standard error of the mean of the population (得出总体平均值的标准误的近似值)
  - Calculating $SE_{\bar{Y}}$ is so routine in biology that a sample mean should never be reported without it.
    - To report $SE_{\bar{Y}}$ is a way to report the uncertainty, and the implicit information of sample size

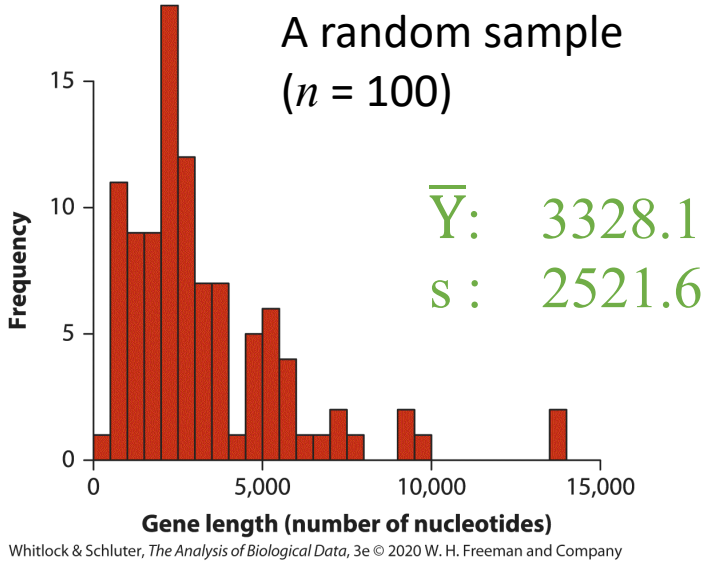
$$\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{Y}} \approx SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

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

2.2 Measuring the uncertainty of an estimate

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





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

$$3511.5 \pm 281.6$$

$$\bar{Y} \pm \mathrm{SE}_{\bar{Y}}\ (\mathrm{SE})$$

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

2.2 Measuring the uncertainty of an estimate

- 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.

  (每个估计值都有一个带标准误的抽样分布)

- The standard error is the usual way to indicate uncertainty of an estimate.

  (标准误是表示估计值不确定性的常用方法)

2.3 Confidence intervals (CIs) 置信区间

- Another common way to quantify uncertainty about the value of a parameter
- A confidence interval (CI) is a **range** of values surrounding the sample estimate that is **likely to contain** the population parameter.
  - CI 是一个范围, 表示总体参数可能落在这个范围内 (常用的是95% CI)
  - E.g. the 95% confidence interval for the mean is a range likely to contain the value of the true population mean μ.

$$(\text{the } \textit{lower limit}) \; 2827.8 < \mu < 3828.4 \; (\text{the } \textit{upper limit})$$

  - How to interpret CI?

2.3 Confidence intervals (CIs)

- How to interpret CI?
  - E.g. the 95% confidence interval for the mean is a **range** likely to contain the value of the true population mean μ. (可能包含真实均值μ的范围)

The 95% CI of μ: (2827.8, 3828.4)

1) "We are 95% confident that the true mean lies between 2827.8 and 3828.4 nucleotides." (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)
2) "There is a 95% probability that the population mean falls between 2827.8 and 3828.4 nucleotides." (总体均值在 2827.8 和 3828.4 个核苷酸之间的概率为 95%)

2.3 Confidence intervals (CIs)

- How to interpret CI?
  - E.g. the 95% confidence interval for the mean is a **range** likely to contain the value of the true population mean μ. (可能包含真实均值μ的范围)

The 95% CI of μ: (2827.8, 3828.4)

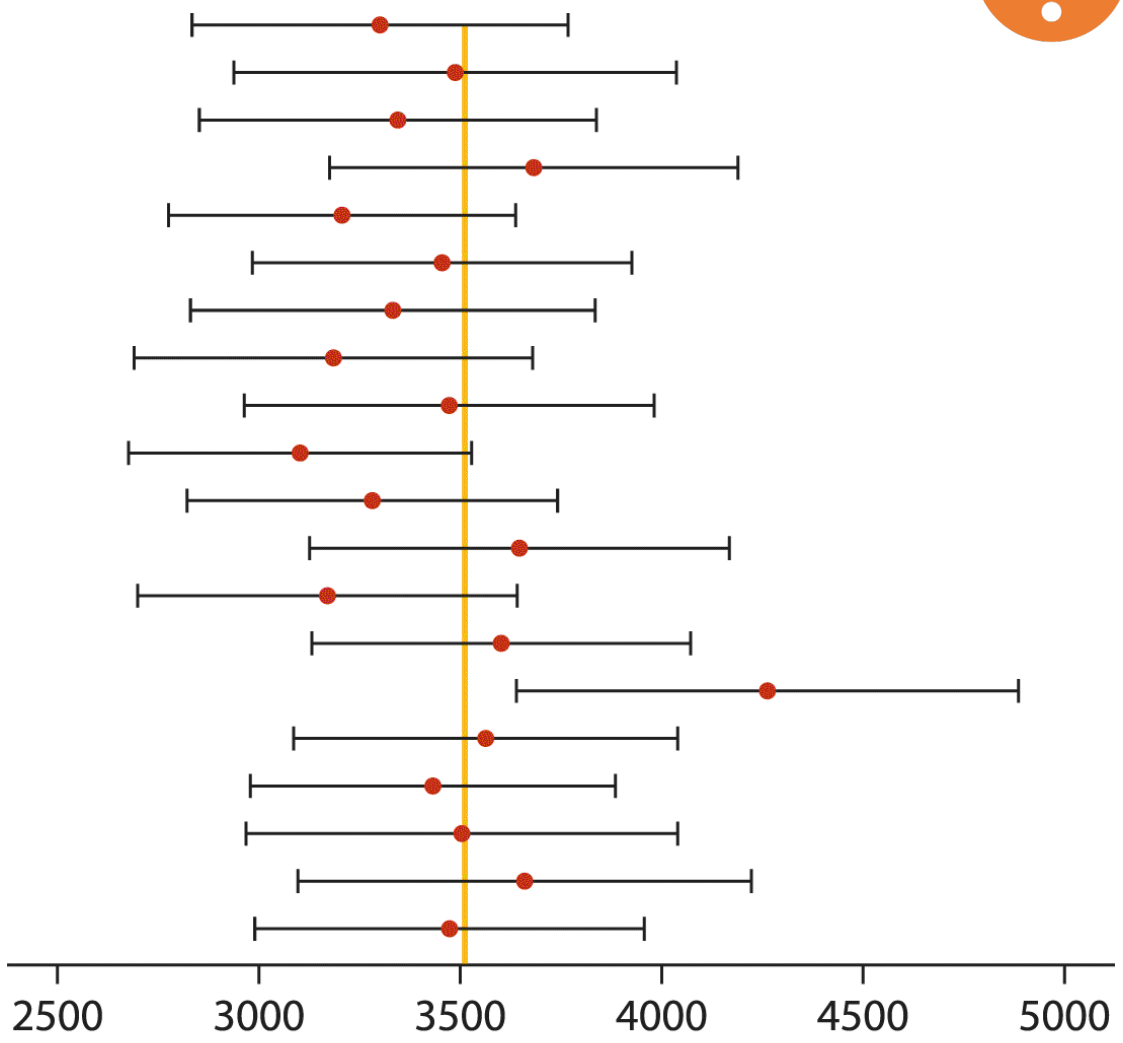
1) "We are **95% confident** that the true mean lies between 2827.8 and 3828.4 nucleotides."
   - 2827.8 and 3828.4 are both **constants** (常数), and the true mean either is or is not between them, so there's **no probability involved**.
   - the 95% CI will capture the population mean **in 95% of random samples**.

2.3 Confidence intervals (CIs)

- How to interpret CI?
  - A **range** of values surrounding the sample estimate that is **likely to contain** the population parameter.

    (样本估计值周围可能包含总体参数的数值范围)
  - We are **95% confident** that the true mean lies within the range.
  - The 95% CI captures the population mean **in 95% of random samples**.

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



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

2.3 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 Confidence intervals (CIs)

Calculating the confidence interval

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

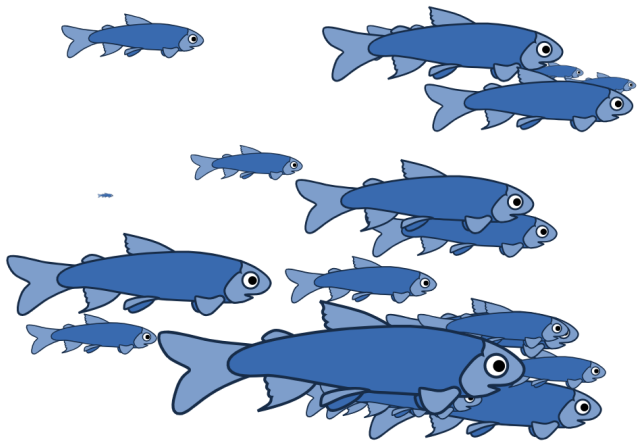
Successes: 40 Failures: 2 Success rate: 95.2%

MAKE A SAMPLE

REPEATED SAMPLES

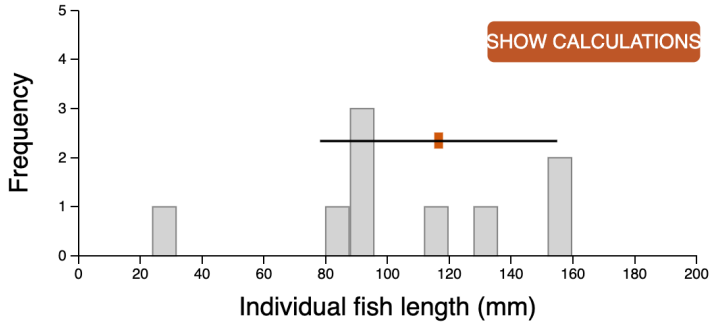
FASTER!

USE 99% CONFIDENCE



n 10 μ 100 σ 50

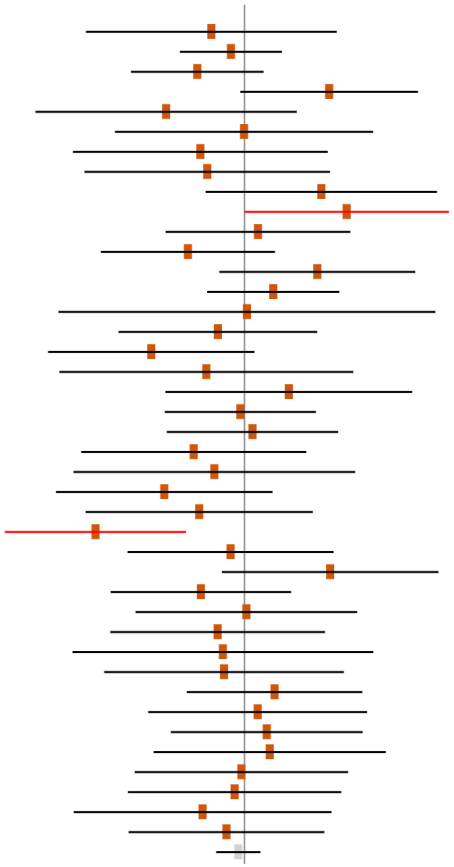
SHOW CALCULATIONS



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

Confidence intervals for the mean

TUTORIAL



The true mean is 100.

2.3 Confidence intervals (CIs)

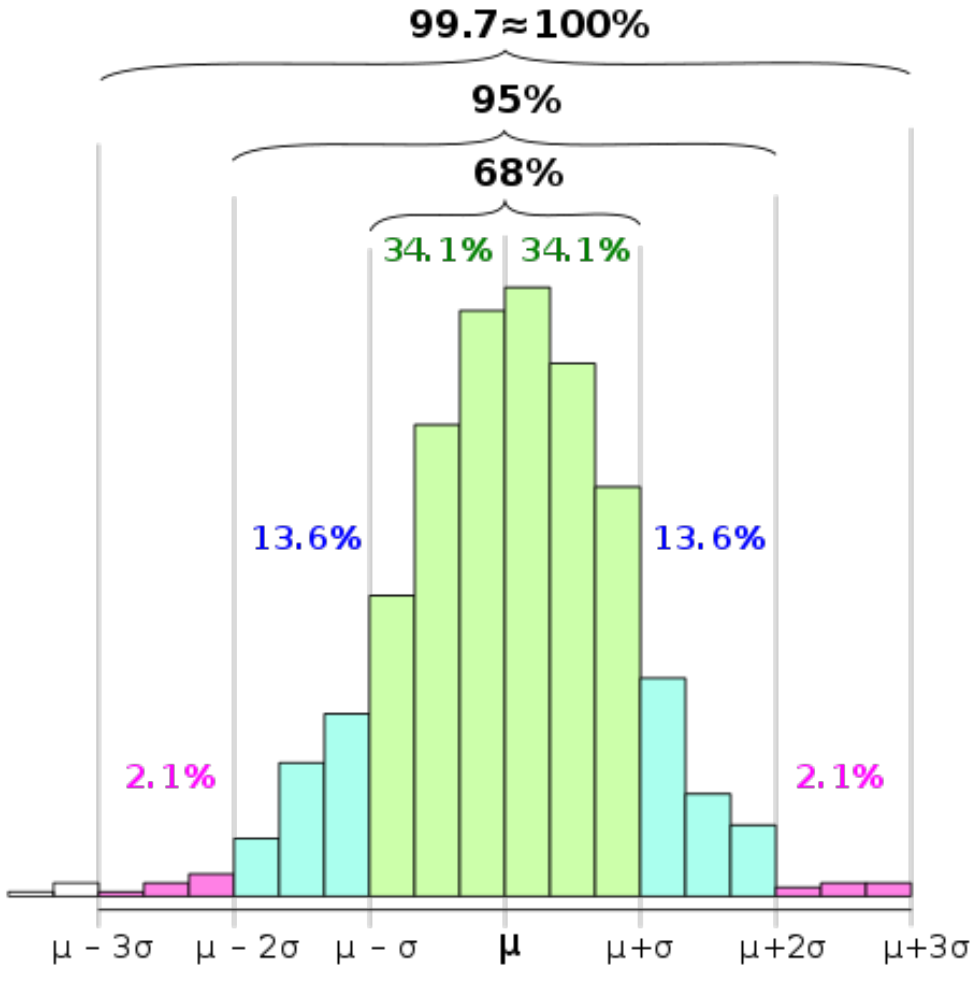
- In general, the width of the 95% CI is a good measure of our uncertainty about the true value of the parameter. (95% CI 的宽度可以很好地衡量我们对参数真实值的不确定性)
  - If the confidence interval is broad, then uncertainty is high and the data are **not very informative** about the location of the population parameter.
  - If the confidence interval is **narrow**, on the other hand, then we can be **confident** that the parameter is close to the estimated value.

- The 95% CI provides a most-plausible range for a parameter.
  - Values lying within the interval are most plausible, whereas those outside are less plausible, **based on the data**.

    (基于数据, 区间内的值最可信, 而区间外的值可信度较低)

2.3 Confidence intervals (CIs)

- The 2SE rule of thumb, when
  - the sample is a random sample;
  - data falls as a normal distribution;
- An approximate
  - The 95% CI of μ: (2827.8, 3828.4)
  - $\bar{Y} \pm 2SE_{\bar{Y}}$: (2827.7, 3832.5)



- The population (or a sample)

- the sampling distribution of the mean (multiple samples)

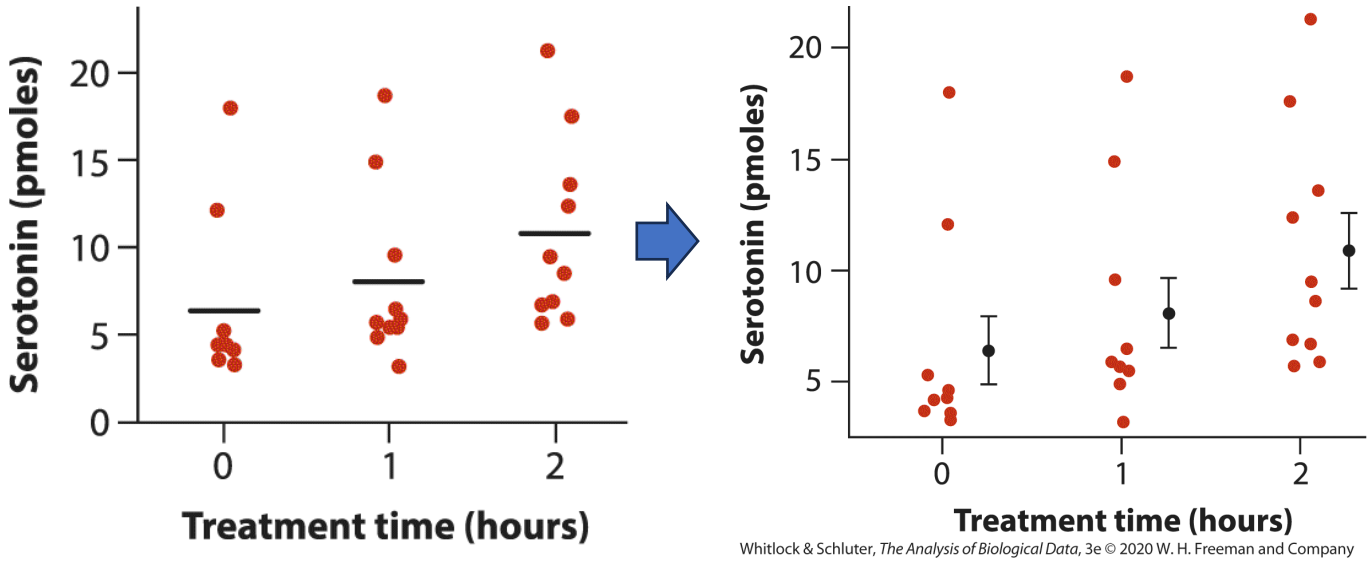
μ-3se μ-2se μ-se μ μ+se μ+2se μ+3se

The 95% CI

2.4 Error bars 误差线

- Standard errors or confidence intervals for the mean (and other parameters) are often illustrated graphically with “error bars.”
  - to illustrate the precision (uncertainty) of the estimated parameter
  - not variability in the data
  - not a fraction of the data
  - *one SE above the mean*
  - *one SE below the mean*

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

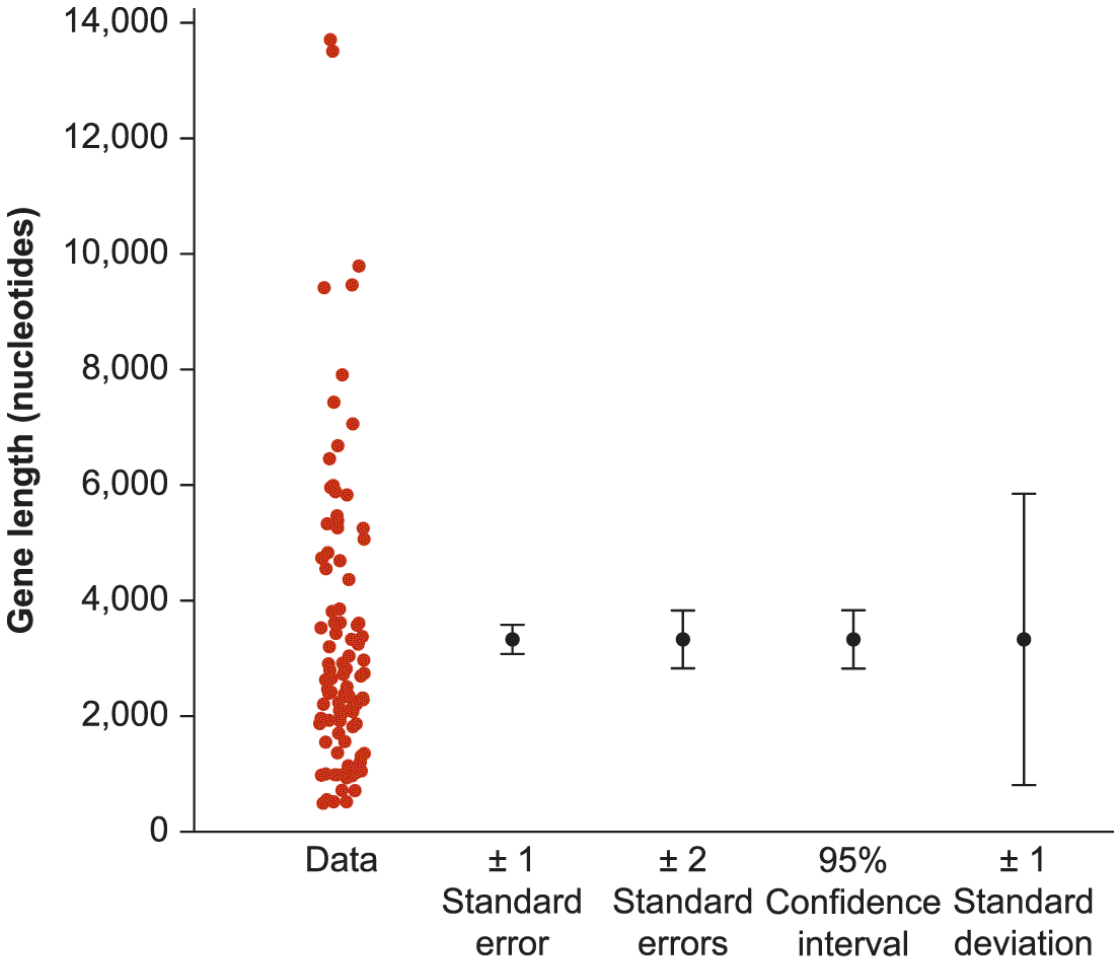


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.4 Error bars

- Error bars can be used for multiple purposes/different quantities:
  - 1 SE
  - 2 SE = 95% CI
  - SD (misleading)
    - Error bars are a poor method for illustrating variability in the data, and they are redundant if you show the data.

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

- 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**?

4. Summary - Estimating with uncertainty

- All estimates have a sampling distribution, which is the probability distribution of all the possible values of the estimate that might be obtained under random sampling with a given sample size.
  - (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)
  - The usual formulas for standard errors and confidence intervals assume that sampling is random.
- The standard error of an estimate is the standard deviation of its sampling distribution.
  - (估计值的标准误差是其抽样分布的标准偏差)
  - The standard error measures precision.
  - The standard error of an estimate declines with increasing sample size.

4. Summary - Estimating with uncertainty

- The confidence interval
  - A range of values calculated from sample data that is likely to contain within its span the value of the target parameter.
  - 置信区间内可能包含目标参数的数值
- The 2SE rule of thumb
  - i.e., the sample mean plus or minus two standard errors
  - provides a rough approximation to the 95% confidence interval for a mean.
- Add error bars to graphs to illustrate standard errors or confidence intervals.
  - 通常在图中添加误差线以说明标准误或置信区间

4. Summary - Estimating with uncertainty

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
```