Chapter 4:
Estimating with uncertainty

1

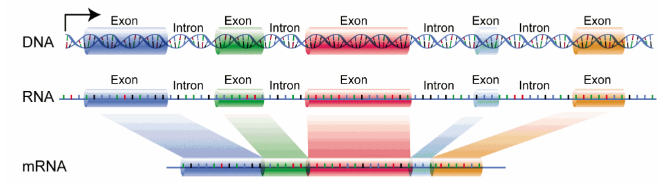
Sampling populations

- A **parameter** is a quantity describing a population, whereas an *estimate* or *statistic* is a related quantity calculated from a sample
- The parameter is the truth, whereas the estimate/statistic is an approximation of the truth that is subject to error

2

Ex 4.1: Length of human genes

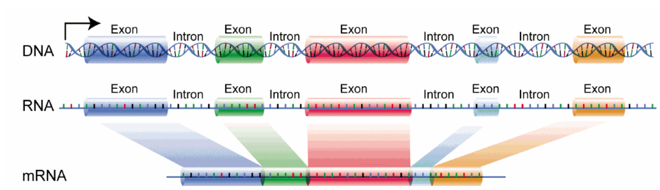
- Human reference genome has 20,290 known genes
  - population
- https://youtu.be/jEJp7B6u_dY?t=213



3

Ex 4.1: Length of human genes

- Human reference genome has 20,290 known genes
  - population
- We can calculate the lengths of all of these genes
- Since we can count **ALL** of gene lengths, we can calculate the true average (μ), standard deviation (σ), etc.
  - **parameters**



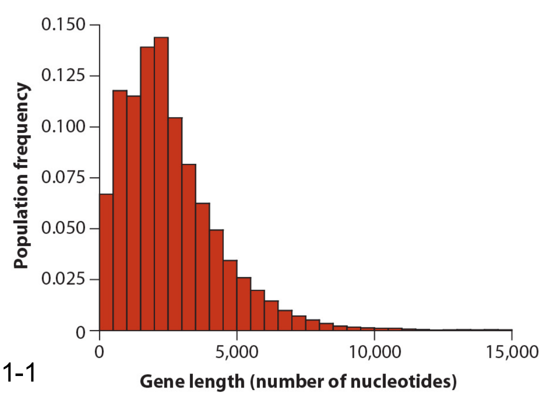
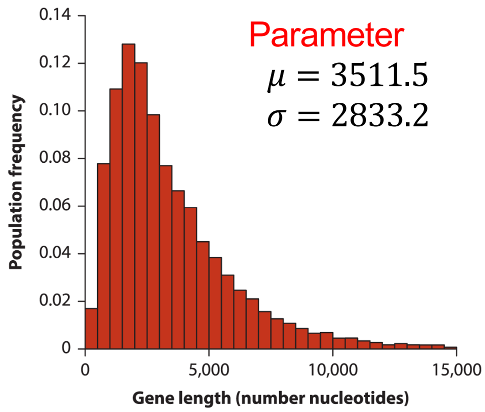


Fig 4.1-1

4

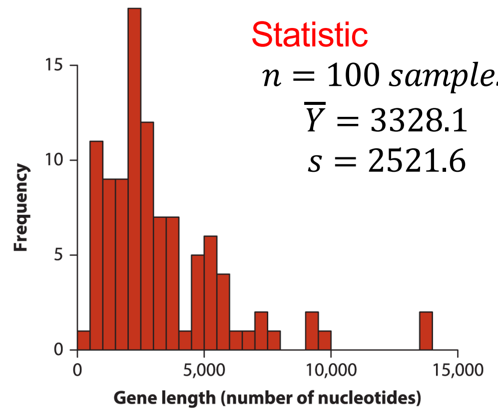
Parameter vs statistic

- In many cases we can't measure the whole population, so instead we sub-sample



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

Fig 4.1-1



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

Fig 4.1-2

5

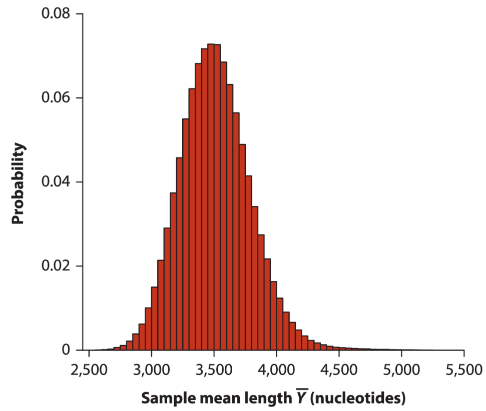
Sampling distribution

- Each random sample generates different statistic values
- Theoretical infinite number of samples (*n* = 100) will create an expected distribution of the statistic/estimate
- The **sampling distribution** is the probability of all values for an estimate that we might obtain when we sample a population
- The parameter μ is constant but the statistic $\overline{Y}$ is variable

6

Sampling distribution

- The **sampling distribution** is the probability of all values for an estimate that we might obtain when we sample a population
- The parameter μ is constant but the statistic $\overline{Y}$ is variable



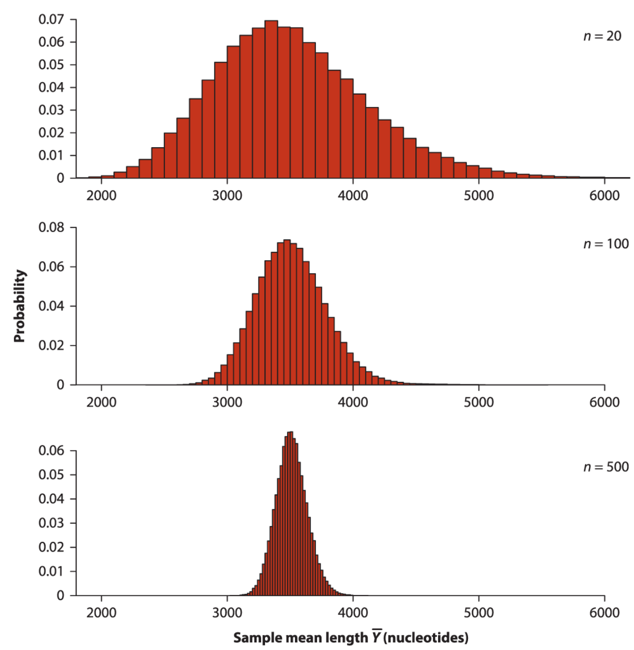
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Fig 4.1-3

7

Sampling distribution

- The spread of the sampling distribution depends on the number of samples
- $\uparrow observations/sample = \downarrow spread$



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

Fig 4.1-4

8

Standard error

- The **standard error** of an estimate is the standard deviation of the estimate's sampling distribution
- Reflects the **precision** of the estimate
- Smaller the standard error, the less uncertainty there is in the estimate of the target parameter

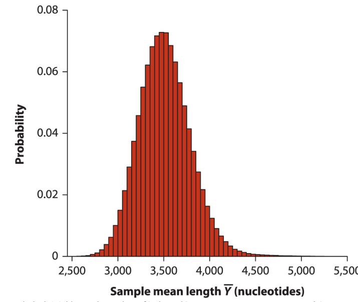


Fig 4.1-3

9

Standard error

- Standard error of the mean $$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$
- As sample size increases the standard error of the mean decreases
- Problem: we almost never know the population standard deviation (σ)
- So the standard error of the mean is approximated by the sample standard deviation (s) as an estimate of σ

10

Measuring uncertainty

- 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)

$$SE_{\overline{Y}} = \frac{s}{\sqrt{n}}$$

- Reflects the **precision** of the estimate
  - Smaller the standard error, the less uncertainty there is about the target parameter (mean in this case)

11

12

Measuring uncertainty

- A **confidence interval** is a range of values surrounding the sample estimate that is likely to contain the population parameter
- The **95% confidence interval** 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
- $2827.8 < \mu > 3828.4$
- Right: We are 95% confident that the true mean lies between 2827.8 and 3828.4.
- Wrong: There is a 95% probability that the true mean falls between 2827.8 and 3828.4

13

Measuring uncertainty

- $2827.8 < \mu > 3828.4$
- Right: We are 95% confident that the true mean lies between 2827.8 and 3828.4.
- Right: 95% of the time that we calculate a 95% confidence interval from a sample from a population, the true mean of the population is inside the calculated interval.
- Wrong: There is a 95% probability that the true mean falls between 2827.8 and 3828.4

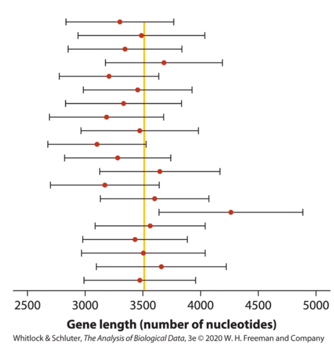


Fig 4.3-1

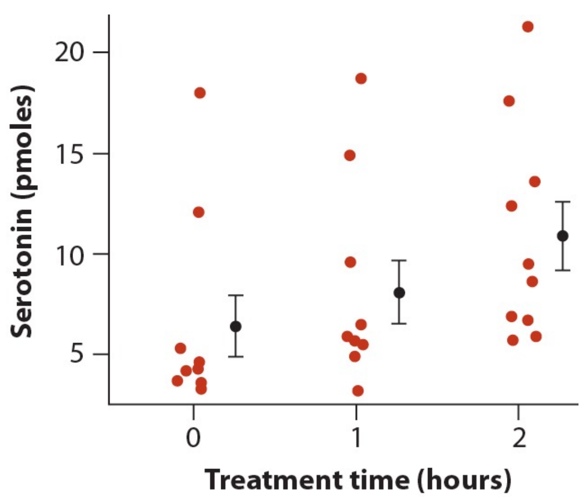
14

Measuring uncertainty

- A **confidence interval** is a range of values surrounding the sample estimate that is likely to contain the population parameter
- The **95% confidence interval** 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
- **2SE rule**: A rough approximation of the 95% confidence interval for a mean can be calculated as the sample mean plus and minus two standard errors

15

Error bars



- **Error bars** are lines on a graph extending outward from the sample estimate to illustrate uncertainty about the value value of the parameter being estimated

Fig 4.4-1

16

Error bars

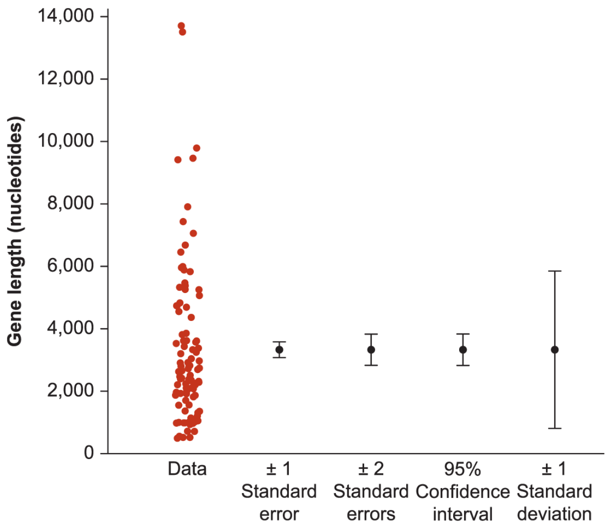


Fig 4.4-2

- Error bars are used to display uncertainty, not the spread of the data
- Therefore, SE or 95% confidence intervals should be used rather than the standard deviation

17