1. Which of the following are required for sharing a data set?

- ○ The raw data
- ◉ All of these options
- ○ A code book describing each variable and its values
- ○ An explicit and exact recipe to go from the raw to the tidy data

2. Which of the following should be included in data tidying recipes?

- ◉ Parameter values for all functions
- ○ Units of variables
- ○ Power calculations
- ○ Preprocessed data

3. What is the central dogma of statistics?

- ○ Estimating parameters using frequencies of observed events
- ○ Using Bayes rule to calculate probabilities we care about
- ◉ Using measurements on a probabilistically selected sample to infer knowledge about a population
- ○ Using measurements on a population to infer knowledge about a sample

4. Which of the following are types of variability in all genomic data?

- ◉ Measurement error
- ○ Missing data variability
- ○ Geographic variability
- ○ Variation from changing technology

5. Which of the following will increase power in a statistical analysis?

- ◉ Increasing sample size
- ○ Using a new technology
- ○ Adjusting for confounders
- ○ Increasing measurement variation

6. If 100 p-values are calculated on a data set with no signal, how many p-values would we expect to be less than 0.05 on average?

- ○ 20
- ○ 0.05
- ◉ 5
- ○ 50

7. If we report 500 results as significant out of 10,000 tests while controlling the family-wise error rate at 5%, about how many false positives do we expect?

- ○ 200
- ○ 5
- ◉ 0
- ○ 500

8. What is the most common confounder in genomics?

- ○ Age
- ○ Genetic background
- ◉ Batch effects
- ○ Sex

9. Which of the following can be used to address potential confounders at the experimental design stage?

- (●) Blocking
- ( ) Improved technology
- ( ) Increasing sample size
- ( ) Measuring DNA instead of RNA

10. Which of the following are benefits of making big data as small as possible as soon as possible?

- ( ) Smaller data sets will decrease false discovery rates
- (●) Interactive analysis can improve our ability to make discoveries
- ( ) Reduced data will increase the power of statistical tests
- ( ) Reducing the data will reduce the number of hypothesis tests