



Module 4: Drawing Conclusions From Data

4.3 Pseudoreplication, Multiple Testing, Meta-Analysis, and Bayesian Methods



Pseudoreplication

- ◆ Pseudoreplication is an issue when data points are correlated in space or time or when the data collected are not representative of the entire population
- ◆ In these cases, you may have less information than you think
- ◆ You should adjust your degrees of freedom downward

Module 4.3





Pseudoreplication

- ◆ Example:

- In a study I was involved with, a graduate student was to collect data on lead contamination in homes in the Bunker Hill Superfund area
- The plan was to drive to a neighborhood and go door-to-door asking if they would participate. If yes, then information and samples were collected and they went next door and continued.
- At the end of the week, the data collection would end.

Module 4.3



Pseudoreplication

- ◆ Example:

- But, homes in a neighborhood tend to be alike in age, value, condition
- People who live there also tend to be alike
- Also, some neighborhoods would be expected to be more contaminated than others
- So, data points collected in this way are correlated.
- Also, some neighborhoods would be well covered and others may be skipped

Module 4.3






Pseudoreplication

- ◆ Example:
 - The solution was to change the design
 - Randomly pick a neighborhood
 - Go door-to-door until someone agrees to participate. Collect that data.
 - Then randomly pick another neighborhood
 - The difference is that each data point is chosen at random from the entire population and should be uncorrelated with the others

Module 4.3




Multiple Testing

- ◆ This problem occurs when many tests of significance are carried out on a data set
- ◆ Some are likely to be significant by chance
- ◆ In fact, $\alpha\%$ of the null hypotheses will be incorrectly rejected. This is, after all, the definition of α .
- ◆ So, some results will appear to be significant when they aren't

Module 4.3





Multiple Testing

- ◆ The solution is to take the number of tests into account and adjust the procedure for rejecting the null hypothesis when multiple tests are performed
- ◆ Beware of blindly running many many tests on a data set searching for the one that is significant! This could lead to a declaration of COLD FUSION!!!

Module 4.3



Meta-Analysis

- ◆ Meta-Analysis involves combining the information from a number of studies together to see if the studies, as a group, support or reject a hypothesis
- ◆ There are many methods for doing this
- ◆ We won't explore the details of these methods.

Module 4.3





Bayesian Methods

- ◆ There are two types of statisticians in the world, Frequentists and Bayesians
- ◆ Frequentists view probability as completely objective. They look at all statistical methods from a standpoint of what would happen in the long run if a sample were taken over and over
- ◆ Statistics is often taught by them beginning with coin flips, pulling balls from an urn, or using a deck of cards

Module 4.3



Bayesian Methods

- ◆ Bayesians, on the other hand, view probability as subjective
- ◆ Probability can be expressed as a degree of belief that an event will occur
- ◆ That belief could be based purely on the data in hand or could involve past experience, other data, expert judgement, theory, etc.

Module 4.3





Bayesian Methods

- ◆ Bayesians express their degree of belief about a parameter as a prior probability distribution
- ◆ Then they incorporate new data into the analysis using a likelihood function, the likelihood that those data occurred given a particular parameter value
- ◆ The result is another probability distribution called a posterior distribution

Module 4.3




Bayesian Methods

- ◆ The Reverend Thomas Bayes created this approach, called Bayes Theorem
- ◆ Θ = the parameter being estimated
- ◆ For simplicity assume that Θ can take on a small number of values $\Theta_1 \Theta_2 \dots \Theta_n$
- ◆ You, as the investigator, may have some guess as to the probabilities of these being the best estimate of Θ
- ◆ If you don't, then they are equally likely

Module 4.3



Bayesian Methods

- These probabilities are your priors:

$$P(\Theta_1), P(\Theta_2), \dots, P(\Theta_n)$$

(Note: these probabilities must sum to 1)

- Then you collect some new data
- You can determine the probability that those data occurred given that $\Theta = \Theta_i$
- That's called a likelihood and denoted by $P(\text{data} | \Theta_i)$ which is the probability you would have collected that data given $\Theta = \Theta_i$

Module 4.3



Bayesian Methods

- Then your prior beliefs and the data are combined to give a new, posterior, set of probabilities using Bayes Theorem

$$P(\Theta_i | \text{data}) = \frac{P(\text{data} | \Theta_i) P(\Theta_i)}{\sum_{k=1}^n P(\text{data} | \Theta_k) P(\Theta_k)}$$

Module 4.3





Bayesian Methods

- ◆ If, instead of a small number of possibilities for the parameter, it could fall anywhere in a range then the summation is replaced by an integral
- ◆ These methods are being used more and more in environmental science since events are rarely reproducible, data sets are limited, and expert judgement often must be used.

Module 4.3

