Scientific and Engineering Projects Class
October 16, 2010

Experimentation and Data Analysis

Experimentation

- Projects involving the Scientific Method require experimentation
- Important consideration for designing and conducting an experiment are
  - Control conditions/environment
    - Avoid bias
  - Take accurate measurements
  - Do multiple trials for each set of conditions
  - Take all possible factors into account
  - **Safety!**

Data Collection and Analysis Introduction

- Many projects, but not all, require the collection and analysis of data
- There are various reasons for data collection and analysis
- Sound scientific techniques must be following in collecting data
- Appropriate analysis is required

What is Data?

- What do you think data is?
  - Distinct pieces of information. For our purposes, the information needs to be quantifiable
  - Factual information, especially information organized for analysis or used to reason or make decisions
- What data, if any, might you collect in your project? Class Discussion

Why Collect Data?

- Many experimental, social science, and product design projects involve collection of data
- Data may be collected for various reasons
  - To compare the effectiveness of different substances
    - Does Tide or Cheer clean grass stains better?
  - To assess the performance of a product
    - How fast can a Ford Mustang accelerate from 0-60 mph?
  - To ascertain the effect of a design choice
    - Do packing peanuts or bubble wrap better protect an egg dropped from 20 feet?
  - To determine correlation between various factors
    - Is there a relationship between lung cancer and emphysema?
- Data must be analyzed, often statistically, to answer the questions of interest

Collect Good Data

- To get good results from your analysis, you need good data, no matter how good your statistical techniques
  - GIGO (Garbage In – Garbage Out)
- If you're doing a survey, ask the right questions
- If you need to take measurements, make sure you do it as accurately as possible and that you are measuring what you think you are measuring

Exercise

- How would you determine the top speed of a model car?
- How might you characterize the performance of a safety device, like a seatbelt?
- What measurements would you take if you built a solar cooker?

Statistical Analysis

- This next section will serve as a high-level primer on a few statistical concepts
- We won't discuss specific techniques, but you will learn and apply what's appropriate for your project

Requirements

- Quantifiable data
  - Length
  - Weight
  - Speed
  - Time
  - Reliability
  - Percent
- Sufficiently large sample size (number of distinct measurements or data points)

Sample Mean

- Average value
  - The sum of the values of the data points divided by the number of data points
  - Ex. The sample mean of 5 s., 8 s., 13 s., and 7 s. is 8.25 s.
- An estimate of the Expected Value

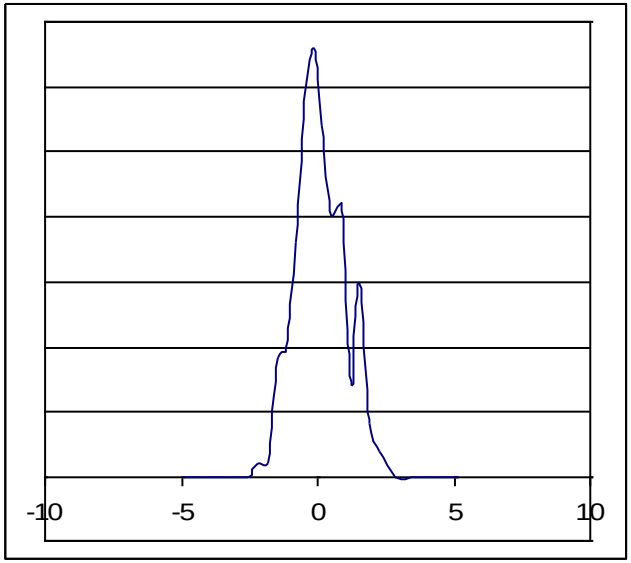
Standard Deviation

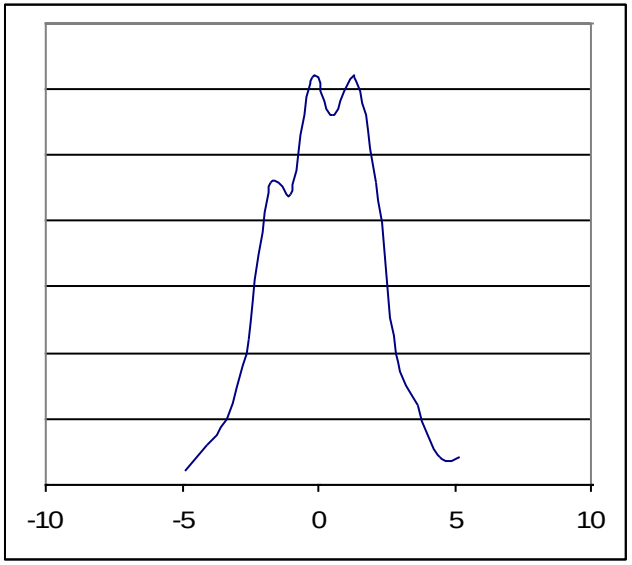
- A measure of how much the data is spread out
  - A small standard deviation means that the data is mostly clustered close to the sample mean
  - A large standard deviation means the data is more spread apart

Distribution

- A mathematical function that describes the probability that a data point will have a particular value
- Used to characterize the data in a statistical sense

Example of Distribution Functions





Independence

- Two events are independent if the occurrence of one does not affect the probability of the other occurring
- Example 1: Two flips of a coin
  - Event A: First flip of a coin is H
  - Event B: Second flip of a coin is H
- Example 2: Box with 3 red and 2 white balls; draw two balls (First ball is not replaced.)
  - Event C: First ball is white
  - Event D: Second ball is white

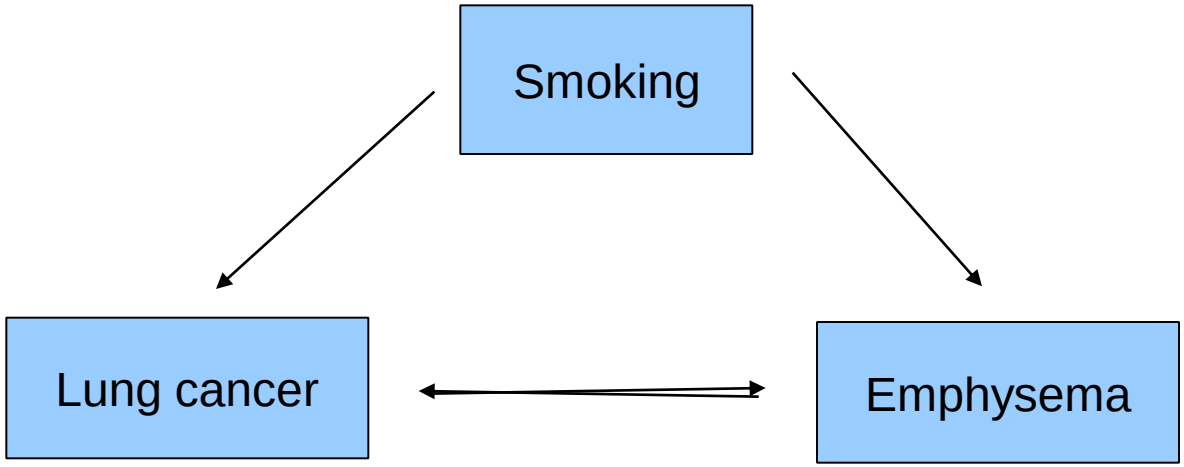
Correlation

- A measure of the dependence of two events.

Cause and Effect

- One question that is sometimes asked is whether one condition causes another. This is often the case in epidemiology.
- If two events are strongly correlated, you might investigate whether there is cause and effect relationship between them
- Be careful what conclusions you draw. See the example on the next slide

Example



There's a correlation between lung cancer and emphysema, but neither causes the other. Both can be cause by smoking.

Exercise

- Determine the average number of flips of a coin required to get the first head
- We will do a probabilistic "experiment" and collect several data points
  - Each data point is a sequence of coin flips which lasts as long as it takes to get one head.
  - The data value of the point is the number of flips that were required.
  - Here are three data points:
    - TTH (value = 3)
    - TTTTTH (value = 6)
    - H (value = 1)

Exercise (cont.)

- These are not data points for this experiment:
  - TTHT
  - THTTTTH
  - TTT
- Each of you, take out a coin and do ten runs of the experiment and record the results of each one (just the number of flips required)
- We'll pool our results and make some calculations

Summary

- Data collection is required in many projects
- There are standard statistical techniques that you should follow to analyze the data you collect
- This presentation provided an overview of data collection. You should discuss with your Science Forum advisor what is appropriate for your project

Appendix

- This next couple of slides has extra information for those that are interested and that have the background to appreciate it
- We will determine, through calculation, the expected value of the number of coin flips required to get the first head
- First we calculate a distribution function for the random variable N defined as
  - N = number of flips to get the first head

Appendix (cont.)

- $\Pr(N = 1) = \Pr(\text{H on the first flip}) = 1/2$
- $\Pr(N = 2) = \Pr(\text{TH}) = 1/2 \times 1/2 = (1/2)^2$
- $\Pr(N = 3) = \Pr(\text{TTH}) = 1/2 \times 1/2 \times 1/2 = (1/2)^3$
- In general, $\Pr(N = n) = \Pr(n{-}1 \text{ T's and } 1 \text{ H}) = (1/2)^{n-1} \times 1/2 = (1/2)^n$
- The value we're interested in is the expected value of N, $E(N)$
- $E(N)$ is the sum over all values of n of $n\Pr(N = n)$
- The rest of the derivation involves advanced mathematical techniques including Calculus. The answer is $E(N) = 2$. You can verify this by summing 30 or so terms in a spreadsheet