Although statistics and applied mathematics are typically viewed as separate disciplines with their own set of courses, this course seeks to unify the two areas by considering both deterministic and statistical modeling, with an emphasis as well on the computational aspects involved with both types. Additionally, applications will focus on infectious diseases, through both the data sets considered and scientific papers assigned for reading.

Because we are trying to cover much more material than would normally fit in a semester course on any one of these topics, much of the material presented (particularly the statistical tools) will be done so with an eye at exposing you to different tools as opposed to an in-depth treatment of topics. The goals are to leave you in a position where you can read papers using these methods and understand in broad terms what analysis is being done, to provide you the background necessary to decide what methods might be appropriate for a given problem, and to understand the utility of statistical tools in dynamical models.

In this second week of the course, we will discuss some basic exploratory tools for summarizing data by examining several data sets involving parasites and disease. In doing so, attention will be given to issues specific to the study of infectious disease systems. In particular, we will discuss:

1. Data collection issues
   (a) Understanding sources of variation
   (b) Recognizing bias, opportunistic sampling

2. Graphical tools for displaying data
   (a) Standard graphs (histograms, scatterplots, boxplots, etc.)
   (b) Spatial & temporal graphs (time plots, greyscale maps, etc.)

3. Numerical tools for summarizing data
   (a) Basic descriptive statistics (mean, SD, median, etc.)
   (b) The normal, Poisson, and negative binomial distributions
   (c) Measures of aggregation
   (d) Sensitivity and specificity

Most of the material discussed in this first week comes from the text *The Ecology of Wildlife Diseases* by Hudson et. al. Copies of this text will be on reserve in the M-EID student office in DHC 21, and Chapter 2 from that text, from which most of the material in this handout arises, will be distributed in class.
The term exploratory data analysis (EDA) was originally coined by John Tukey (Exploratory Data Analysis, Addison-Wesley, 1977). It refers to the initial exploration of data usually by means of graphical tools before analysis or modelling of data takes place.

- EDA is all about "getting to know" your data.

- The last part of this definition cannot be overemphasized. It is critically important to investigate the intricacies of a set of data before fitting models or performing inferences based on the data. There may be outlying data values or other patterns present that may guide how analysis is to take place.

- An often overlooked aspect of EDA is an evaluation of not just the data themselves, but the methods by which the data were collected.

There is no substitute for honest, thorough, scientific effort to get correct data (no matter how much of it clashes with preconceived ideas). There is no substitute for actually reaching a correct claim of reasoning. Poor data and good reasoning give poor results. Good data and poor reasoning give poor results. Poor data and poor reasoning give rotten results.


- There are entire courses on sampling methods, but there are effectively two types of samples taken in practice:
  1. **Probability samples** - samples based on some random or probabilistic mechanism (simple random samples, stratified samples, systematic samples, etc.)
  2. **Opportunistic samples** - samples done in an opportunistic or convenient way potentially leading to bias

- Much of the data collected on parasites in animal populations is opportunistic in nature, relying on road kills, beach strandings, harvested animals, or discovered dead animals. If we are studying the presence of some parasite in a host population, what types of bias might this introduce?

- Even seemingly unbiased sampling methods can lead to biased samples. If there are differences in behavior in parasitized and healthy individuals which impact whether or not they are sampled, this can lead to bias. Examples?

- Sensitivity and specificity
Some Basic Descriptive Measures - One Variable: For data describing a single variable, EDA typically focuses on three aspects of the data: **center**, **spread**, and **shape**. Some common measures of each are illustrated with the following example.

Example: Data were collected to examine the parasite-host interaction for the nematode parasite *Chandlerella quiscale* on the host gnat *Culicoides crepuscularis*. Specifically, 143 gnats were collected from a single infected bird and the number of nematodes counted for each gnat. The data appear in the table below along with a histogram displaying the distribution of counts. [Data taken from Schmid & Robinson 1972. The pattern of a host-parasite distribution, *The Journal of Parasitology*, 58: 907-910.]

<table>
<thead>
<tr>
<th>Number of Nematodes</th>
<th>Number of Gnats</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>≥ 15</td>
<td>3</td>
</tr>
</tbody>
</table>

Measures of Center: Let $x_1, \ldots, x_n$ be the $n$ data values.

1. **Mean**: the average of the data values: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
   (Gnat Ex: $\bar{x} = 2.615$ nematodes).

2. **Median**: the middle data value in the ordered list (Gnat Ex: $M = 1$ nematode).

3. **Mode**: the data value occurring with the greatest frequency (Gnat Ex: Mode = 0 nematode).

4. Others: Midrange, Trimmed Mean

**Issues?**
Measures of Spread

1. **Standard Deviation**: \( s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \) (Ex: \( s = 4.176 \) nematodes).

2. **Interquartile Range (IQR)** = the distance across the middle 50% of the data: IQR = \( Q_3 - Q_1 \), where \( Q_1 \) and \( Q_3 \) are the first and third quartiles of the distribution (25th & 75th percentile).
   (Ex: The 25th and 75 percentile occur approximately in the 143/4 \( \approx 36th \) and 143(3/4) \( \approx 108th \) positions in the ordered data. Here \( Q_1 = 0 \) & \( Q_3 = 3 \) for an IQR = 3 nematodes.

3. **Range**: maximum - minimum (Ex: Range = 25 - 0 = 25 nematodes).

**Issues?**

Distribution Shape: Although measures of center and spread give a numerical description of the data, the most important aspect of the distribution for purposes of statistical inference is the **shape**.

- Some common shapes are shown below.

  ![Symmetric](image1.png) ![Skewed right](image2.png) ![Skewed left](image3.png)

- Traditional statistical inference procedures (confidence interval estimation, tests of significance) rely on a particular type of symmetric data known as **normal** data. When data exhibit a normal shape (as indicated by the left curve above), there is well-developed theory for finding confidence intervals or test statistics for a population mean. [See Chapter 2 of *The Statistical Sleuth* by Ramsey & Schafer.]

- Unfortunately, a common issue with infectious disease data is the lack of normality in host-parasite distributions. Some papers illustrating this point are:
The Normal, Poisson, and Negative Binomial Distributions

Motivation

- Aggregation of parasites across host populations (most parasites found in a few individuals). Aggregation leads to highly right-skewed parasite distributions, creating analytical problems.

- Differences in susceptibility (due to age, sex, genetics, etc.) among host individuals.

What do we do with highly right-skewed data?

1.

2.

3.

The Normal Distribution: Let $X$ be a random variable from a normal distribution with mean $\mu$ and standard deviation $\sigma > 0$. The probability density function (pdf) for $X$ is given by:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}, -\infty < x < \infty, \sigma > 0.$$  

- We don’t actually use this form in practice, but it defines the bell-shaped normal distribution shown earlier.

- When working with parasite counts (or any type of count data), the data themselves are discrete (finite or countable). The normal distribution describes the shape of continuous data for which any real value is possible.

- So we recognize that the use of a normal model with discrete data is just an approximation of reality; for bell-shaped data, it is often a very good approximation.

The Poisson Distribution: Let $X$ be a random variable from a Poisson distribution with mean $\theta > 0$. The probability mass function (pmf) for $X$ is given by:

$$f(x|\theta) = P(X = x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \theta > 0, x = 0, 1, 2, \ldots$$

- The Poisson($\theta$) distribution has the interesting property that the mean and variance are the same: $E(X) = \text{Var}(X) = \theta$.  

5
• It can be shown (through Taylor series expansions) that for data where the mean is proportional to the variance, taking the square root of the data values makes the distribution more symmetric, and removes the dependence of the variance on the mean.

• With Poisson count data, analysts generally do one of the following:
  1. Perform a square root transformation and use normal-based inferences on the transformed data.
  2. Use generalized linear models with a Poisson error structure. For overdispersed Poisson data, additional parameters may be defined.
  3. For count data with many zero counts, a zero-inflated Poisson (ZIP) regression model can be used. A paper illustrating the effects of zero counts on parasite count data is: [Cox, Heyse, and Tukey, 2000. Efficacy estimates from parasite count data that include zero counts, Experimental Parasitology, 96: 1-8.]

• Some examples of Poisson distributions are shown below.

![](poisson_distributions.png)

• Why do we care about the Poisson distribution? Poisson models represent what we expect to see if parasites are randomly distributed! In what sense?

• Unfortunately, parasite count data generally do not follow Poisson distributions, but tend to be more skewed and better represented by the negative binomial distribution.

**The Negative Binomial Distribution**: Before defining the negative binomial distribution, we need some background on Bernoulli trials.

• A **Bernoulli trial** is any experiment resulting in one of two outcomes, generically referred to as a success or a failure. Examples?

Consider a sequence of independent Bernoulli trials where each trial has success probability \( p \). Let \( X \) be the number of failures we get before the \( r^{th} \) success. Then \( X \) has a negative binomial distribution with parameters \( r \) and \( p \) (written \( X \sim \text{NegBin}(r,p) \)), with probability mass function given by:

\[
f(x|r,p) = P(X = x|r,p) = \binom{r + x - 1}{x} p^r (1 - p)^x, \quad 0 < p < 1, \quad x = 0, 1, 2, \ldots
\]

• Negative binomial distributions tend to be more highly right-skewed than Poisson distributions, with the variance greater than the mean.
Some biologists use a different parametrization for the negative binomial distribution, as given below.

\[ f(x|m, k) = P(X = x|m, k) = \binom{k + x - 1}{x} \frac{m^x k^k}{(m + k)^{x+k}}, m, k > 0, x = 0, 1, 2, \ldots \]

Instead of parameters \( r \& p \), they use the mean \( m = r(1 - p)/p \) and \( k = r \), where \( k \) gives information about the degree of aggregation in the parasite distribution.

**Note:** Since \( \binom{k + x - 1}{x} = \frac{(k + x - 1)!}{x!(k - 1)!} \), we must have \( k \geq 1 \) as defined. However, for many applications, estimates of \( k \) (to be explored later) are often smaller than 1, indicating severe right skewness. This problem can be circumvented using a more general form of the negative binomial that makes use of the gamma function:

\[ \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt. \]

The Gamma function has the property that \( \Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \) for \( \alpha > 0 \). Thus, for \( \alpha \) a positive integer, \( \Gamma(\alpha) = (\alpha - 1)! \). We can use this function to more generally express the negative binomial distribution as:

\[ f(x|m, k) = P(X = x|m, k) = \frac{\Gamma(k + x)}{x! \Gamma(k)} \frac{m^x k^k}{(m + k)^{x+k}}, m, k > 0, x = 0, 1, 2, \ldots \]

Because of this greater degree of skewness (than a Poisson), negative binomial distributions sometimes cannot be adequately transformed to approximate a normal model. Some examples of negative binomial distributions and the corresponding log-transformed data are shown below (all have \( m = 10 \)).

---

![Graph 1](image1)

![Graph 2](image2)

![Graph 3](image3)

![Graph 4](image4)

---

![Graph 5](image5)

![Graph 6](image6)

---

Parasite Load

Log(Parasite Load + 1)
Example: Recall the gnat-nematode data from page 3 of this handout. The histogram of nematode counts is shown again below. To investigate whether these data agree with a Poisson or negative binomial model, these models were fit using maximum likelihood estimation, and the resulting observed counts plotted on the histogram. [For more on the use of maximum likelihood to estimate the $k$-parameter in a negative binomial distribution, see Shaw & Dobson, 1995, Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. Parasitology 111: 111-133.]

![Histogram of Nematode Counts](image)

- **Aside**: We will study maximum likelihood (ML) and some of its variants later in this course. To give a non-mathematical definition, ML is an optimization technique for finding those parameters in a model that maximize the probability of having observed a particular realization of the data.

  - For the Poisson($\theta$) case, ML finds that value of $\theta$ (denoted $\hat{\theta}$) which makes it most likely to have observed the data shown in the histogram if these data come from a Poisson model.

  - For the negative binomial($k, m$) case, ML finds those values of ($k, m$) which give maximum probability to having observed the nematode counts actually observed, if these data come from a negative binomial model.

  - Example: Suppose I take 10 free throws and make 4. What does maximum likelihood estimate my proportion made to be? What do you estimate?
• In viewing the expected counts according to the two models in this histogram, it is clear that the negative binomial model provides a superior fit.

• For this model, the maximum likelihood estimates for $k, m$ were $k = 0.466$ and $m = 2.615$ nematodes/gnat.

• We know what $m$ means. What meaning does the value of $k$ have?

**Measures of Aggregation**: Of interest in studying the relationship between a pathogen or parasite and its host is the level of aggregation of the parasite in individual hosts. We have looked briefly at an example of a parasite distribution and some ideas behind modelling the observed distribution. Here, we consider some potential causes of aggregation and some of the measures used to quantify the degree of aggregation in a given set of data.

• What is aggregation? Aggregation is the phenomenon where parasites tend to be clumped or *aggregated* in a few individuals in a host population, where most individuals harbor few or no parasites.

• What are some sources of aggregation in parasite distributions?

1. Genetic differences in individual host susceptibility
2. Males tend to be more heavily affected (mammals)
3. The infection process itself
4. Age differences in parasite loads
5. Host condition or habitat influences
6. Others?

**The Statistics of Aggregation**: Statistically, aggregation is said to occur whenever the variance/mean ratio of parasite counts per host exceeds 1.

• Keep in mind that a variance/mean ratio of 1 indicates the presence of a Poisson distribution (since the mean = variance in this case). Poisson distributions would describe the parasite counts if they were randomly distributed among individuals. So, ratios larger than 1 indicate that counts tend to be more clumped than expected by chance, indicating some mechanism for parasite aggregation.

• Letting $s^2$ and $m$ indicate the variance and mean of the parasite distribution respectively, some common measures of aggregation (based on the mean/variance ratio and $k$) are outlined below:

1. $s^2/m$: If the ratio $→ 0$, the distribution is uniform.
   - If the ratio $= 1$, the distribution is Poisson (random distribution).
   - If ratio $> 1$, the distribution is aggregated.

2. Standardized variance (SV): $s/m$. Sometimes, this quantity is multiplied by 100 to give the **coefficient of variation** (CV).
   - These measures are useful but simulation studies indicate that they should only be used when the number of uninfected hosts (the “0-class”) is large [Scott, M.E., Temporal changes in aggregation: a laboratory study. *Parasitology* 94:583-595].
3. \[ k \]: For the negative binomial distribution: \( s^2 = m + m^2/k \). This implies that:

For large \( k (k \geq 20) \): \( s^2 \to m \) and the distribution is Poisson.

For smaller \( k \), the distribution becomes more aggregated.

Typically, \( k < 1 \) for parasite distributions.

How do we estimate \( k \)? 3 basic ways:

(a) Method of moments: \( k = m^2/(s^2 - m) \). This method is simple, but unreliable when \( m \) is large, \( k \) is small, or \( n \) is small! [Hudson et.al., The Ecology of Wildlife Diseases]

(b) Corrected moment estimator: \( k = (m^2 - s^2/n)/(s^2 - m) \). This estimate provides an improvement for small sample sizes.

(c) Maximum likelihood: This method produces reliable estimates for all cases.

4. Several other measures of aggregation - see Box 2.1, pages 8-10 of Chapter 2 from *The Ecology of Wildlife Diseases*.

**Relationships between Two Variables**: In addition to tools for describing the distribution of a single variable, it is often of greater interest to examine tools for investigating the relationship between two or more variables.

- We will look at the use of a scatterplot to explore an issue of bias due to small sample sizes and the effect of sample size on a regression model.

Sample size bias issues: A simulation study by Gregory & Woolhouse [1993: Quantification of parasite aggregation - a simulation study, *Acta Tropica* 54: 131-139] found that for small sample sizes, estimates of the mean parasite burden and measures of aggregation tend to be biased. Some of their findings are summarized in the plots below.

![Sample Size Simulation Results](image)

- So when the statistician says to take more samples, do it! It is not just to increase the power of inferences, but for these type of data actually leads to less biased estimates of data attributes.
Example: Data were collected on 602 arctic charr from a small lake in Norway for the presence of the parasite *Diphyllobothrium ditremum*. In addition to counts of the number of cestodes on each fish, the age and sex of the fish were recorded. Of interest here is the relationship between cestode count and age. [Data taken from Halvorsen & Andersen (1983), The ecological interaction between arctic charr and the plerocercoid stage of *Diphyllobothrium ditremum*, *Journal of Fish Biology* 25: 305-316.]

A scatterplot of mean cestode count vs. charr age group is shown to the right below.

- Why plot the mean counts? Why not plot the data for each fish individually?

- If we fit a regression model to these data, what model seems reasonable?

- Any time you decide to exhibit mean values in a plot instead of the raw data values, you must be cognizant of the fact that each point in the plot does NOT carry the same weight for any analysis performed.

- As an example, there are 244 fish in the “5+” age class whereas only 4 fish are of age “10+” years. Hence, the single point plotted for the “5+” fish carries much more influence in describing the relationship between age and cestode count than the three points defining fish greater than 10 years of age.

- Pacala & Dobson [1988: The relation between the number of parasites/host and host age: population dynamic causes and maximum likelihood estimation, *Parasitology*, 96: 197-210] used maximum likelihood to fit a three-parameter nonlinear model of the form:

\[
M(a) = V_1 \left[1 - \frac{\exp\{-V_2(a - V_3)\}}{V_2}\right],
\]

where \(a\) is the age class, \(V_1\) is the mean parasite invasion rate for the host population, \(V_2\) is the mortality rate of parasites, and \(V_3\) represents a threshold age where charr are not susceptible until this age. This model weighted the plotted values according to sample size.

- The resulting fit of this model to the charr data is plotted through the mean cestode counts above. Note how the modeled pattern differs from what we might fit “by eye” if we ignore the sample sizes.
Time Series Plots: Time series plots are useful graphical descriptions for quantitative variables collected over time. Time series data are generally modeled in one of two ways:

1. Time domain approach: autoregressive, moving average models
2. Frequency domain approach: Fourier (harmonic) analysis

Example: Data were collected on the number of measles-related deaths occurring in England and Wales between 1940-2004, with an indication of when the measles vaccine and measles-mumps-rubella (MMR) vaccine were introduced. These data are shown in the time series plot below.

Patterns or trends?
What do you see?

**Note:** The time scale on the x-axis must be correctly spaced. One of the major deceptions that occur with time plots is when the axis is distorted in some way to make a trend appear stronger or in some way different than it really is.

**Note also that if only data from 1956-2004 were shown, the effect of the measles vaccine appears to be stronger, whereas with all of the data, one can see a general downward trend in the number of deaths preceding the vaccine.
Spatial Plots: In more recent years with the advent of Geographic Information Systems (GIS) and advances in spatial analysis techniques, spatial plots of one sort or another have become more commonplace as exploratory tools. Maps with extraordinary detail can be constructed for multiple layers of information over a given area. It is only recently that we are developing means by which to account for data which may have an underlying relationship, a temporal relationship, and perhaps a spatial relationship as well.

As a simple example to show an initial spatial exploration of a disease system, consider the following data on the spread of the West Nile virus from 1999-2006 in the United States. The sequence of greyscale maps shown indicates the number of human cases of West Nile virus reported to the Centers for Disease Control (CDC) in the 50 states and DC. These data were obtained from the website: http://westnilemaps.usgs.gov/.

**Number of US Human Cases of West Nile Virus**

- What patterns do you see?
- What is missing from these maps?
Early Example of Disease Mapping: Dr. John Snow’s map of dots from the 1855 mapping of the London Cholera Epidemic (1854) is one of the first maps with epidemic data (each dot represents a cholera death) in relation to the local water pump. Snow observed that cholera occurred almost entirely among those who lived near (and drank from) the Broad Street water pump. He had the handle of the contaminated pump removed, ending the neighborhood epidemic which had taken more than 500 lives. [Source: Edward R. Tufte, The Visual Display of Quantative Information, Cheshire, 1983, reprint 1995, p. 24.]
One more Example: As part of my dissertation research (a long, long time ago), I developed methods of statistical inference for autologistic models. Autologistic models are spatial models used for relating spatially correlated binary data to any number of other variables (covariates). By “spatially correlated,” we mean that the value at one location is influenced by the value at nearby locations in a quantifiable way.

The motivation for my research stemmed from an application in plant pathology where interest was in studying mechanisms of spread for the pathogen *Phytophthora capsici* in bell pepper fields. These fields consisted of plants in a lattice structure so that the entire field could be viewed as a grid. The primary questions of interest were to what degree the pathogen was spread through root-to-root contact in the soil versus excess surface water from over-irrigation and to what degree soil moisture played a role.

We looked at several plots through the growing season; however, two interesting plots at a single time are shown below. The first displays a greyscale map of the underlying soil moisture and the second the disease status of the plants combined with moisture levels on the 20x20 grid.

What patterns do you see?