March 27, 2012

Chapter 1: Introduction

ES/STT 714 is an Environmental Statistics course that covers basic statistical methodologies encountered in collecting and analyzing data from biological surveys and experiments. Students in this class should have had at least one statistics course that covers the basic of probability, statistical thinking, modeling, and inference. The notions of probability distributions (e.g. the normal distribution), hypothesis testing, test statistics, analysis of variance (ANOVA), p-values and regression analysis should be familiar to students in this class.

From this foundation, we shall cover more advanced statistical topics that are commonly encountered in the environmental sciences. The field of statistics is vast and the statistical topics pertinent to environmental applications is also vast. We cannot cover all these topics in great detail but these notes will attempt to give a solid introduction and foundation to many useful topics. The topics covered in this course include:

- Sampling Designs
- Multiple Regression topics
- Logistic Regression and Generalized Linear Models
- Time Series Analysis
- Spatial Statistics

Below we give a brief overview of each of these topics:

Sampling Designs In order to test scientific hypotheses, one needs data. Data can be derived from observational studies where one goes out and observes a population of interest. Alternatively, one can run experiments (e.g. in a lab or in a plot in a field) to collect data. A short introduction to some well-known experimental designs are given in Section 2 below. For observational data, in order to make valid statistical inferences from the data, the data must be representative of the population from which it is obtained. This typically involves randomly sampling the units in the study in some fashion. There are a wide variety of methods available. If one does not obtain a sample in a statistically valid fashion, then the resulting data may fail to be of much use due to inadvertent selection biases. Additionally, a prudent choice of a sampling design may lead to more efficient estimation of population parameters and save time and money in collecting data.

Regression Topics. Regression is perhaps the most used statistical tool in practice. The basic idea of regression is to model how a variable of interest (known as a response variable or the dependent variable), usually denoted by y, depends on one or several "predictor" variables (also known as regressor variables or independent variables). The predictor variables are usually denoted by x's. One useful aspect of regression is that it allows us to model relations between the response and predictor variables. We can perform statistical tests to determine if particular models fit the data well. Sometimes the model is of scientific interest. Models allow us to investigate if particular predictor variables influence the response of interest. In other cases, we may just want to use the regression model to make predictions of future responses and the form of the model may not be of primary interest. In this class, we will briefly review the basics of simple linear regression (only one predictor variable) and cover multiple regression topics involving models with more than one predictor variable. A couple special cases of primary importance are polynomial models and models with indicator variables (analysis of covariance or ANCOVA). In addition, we shall cover nonlinear regression models that are very important in many biological examples.

Generalized Linear Models. Traditional regression models are used when the response variable is continuous. Additionally, a common assumption is that the response variable has a probability distribution that is normal, or approximately normal. This assumption often fails in practice, especially when modeling a response that corresponds to a count or a response that is binary (success or failure). We shall cover the topic of generalized linear models (glm) that can accommodate response variables that are not normal. The most commonly used type of a generalized linear model is *logistic regression* when the response is binary. In logistic regression, the response is a Bernoulli 0 or 1 variable that corresponds to a success or a failure. The estimated regression function gives the probability of success conditional on one or more predictor variables. Another very common example is *Poisson regression* where the response corresponds to a count variable (e.g. number of zooplankton that survive after exposure to a toxin).

Time Series. Time series are very important in environmental studies. Any time data is collected sequentially over time (such as temperatures), the resulting data is a time series. Time series analysis allows us to study seasonal trends and overall trends in the data as well as correlation structure between successive measurements over time.

Spatial Statistics. Environmental studies often involve data collected spatially such as soil samples obtained in a large field. It is of interest in such studies to investigate if clusters exist or if the data is spatially correlated. Spatial models can be used to model how correlations between measurements vary spatially.

1 Statistical Software

After data is collected, it needs to be analyzed and statistical software is usually required to analyze the data, particularly for large data sets and/or complicated models. Sometimes the statistical analysis merely consists of computing some simple summary statistics (means, medians, standard deviations). However, investigators typically want to use their data to perform statistical inferences (e.g. hypothesis testing, estimation and confidence intervals) and to estimate models. These applications often require complex computations involving iterative algorithms where a good statistical software package is essential.

There are many statistical software packages available, most of them commercial. One can use the excel spreadsheet that comes with Microsoft Office to do simple statistics, but this software can be quite limited for statistical analysis. The SAS software is a very popular package that can perform most statistical procedures encountered in practice. We may occasionally use SAS in this course if some students will find that useful. Matlab is a mathematical software package that can be used to program statistical procedures.

The primary software package to be used in this class will be the software R which can be freely downloaded from the Comprehensive R Archive Network (CRAN) at the following cite:

`http://cran.r-project.org`

The syntax for R is very similar to the syntax for the commercial statistical package S-plus. In fact, most code written in R will also work in S-plus and vice-versa.

Learning to use R can be difficult. There exist many manuals and help pages on the internet for using R which can be found via simple Google searches.

There are several reasons for recommending the R software:

- It is free.
- It is very powerful and can perform almost all statistical methods of interest.
- R can produce very nice and high quality graphics fairly easily.
- R is a convenient software to use to run simulations in order to check how certain statistical procedures perform. We shall illustrate this later in the course.
- Scientists and statisticians are producing R software packages that can implement state-of-the-art statistical methodologies. These packages are very easy to load and use on your computer.

In order to install R on a Windows based computer, go to the above web site. Select "R binaries" → "Windows" → "Base". From here, download the most current R-executable file. Once the executable file is downloaded on your computer, simply

double-click on it and follow the instructions. Once the executable file is downloaded on your computer, you can install R in about 30 seconds, it is very fast and easy to do. After installation, an R icon will appear on your screen (desktop). Simply double-click on this icon to start the software. The R software will be illustrated in this course and R programs will be made available for performing the statistical analysis covered in this class.

2 Some Basics of Experimental Design

The next chapter focuses on sampling designs for obtaining observational data. If data is to be collected via an experiment, then the one can consider different experimental designs. In addition, there are some basic principles to follow when conducting experiments which we shall briefly highlight.

First, we introduce some terminology. An **experiment** is when researchers control the allocation of treatments to the **experimental units**. Thus, in an experiment, the researcher has direct control over treatments received by experimental units (e.g. plants, plots, animals, etc.). The *experimental unit* is the smallest unit to which treatment combinations are applied. The *observational unit* is the unit upon which measurements are taken. Experimental units can be people, animals, beakers of liquids, plots of land etc. Sometimes the experimental and observational units are one in the same, and other times they differ. For example, a beaker of water may contain several organisms and the implementation of the experiment is to expose the organisms to a particular chemical. If the chemical is added to the water, then the experimental unit is the beaker. If responses are recorded for individual organisms within the beaker, then the organisms are the observational units.

In observational designs, the researchers do not have control on the allocation of treatments, but instead *observe* a population of interest. The **population** is the collection of all possible sampling units. When identifying the population, the investigator must determine to which group will the conclusions of the statistical inference be applied. One advantage of performing an experiment as opposed to an observational study is that it is easier to make conclusions on *causation* with an experiment via the use of randomization (see below). In an observational study, if we note an association (or correlation) between two variables, it may not be possible to say one effect causes the other. For example, does second-hand smoke cause cancer in humans? There may exist a positive association between these two variables, but there are many other variables that one may not be able to control that could also have an impact on lung cancer rates.

If the goal of an experiment is to study the effect of some treatment, say an environmental toxin, on a population of interest, then good experimental design requires the elimination of all other variables that may effect the outcome to the greatest extent possible. For example, if we want to study the effect of a toxin on a particular organism, then one would want to control to the greatest extent possible the effect of any other variables on the organism. If an effect is discovered when analyzing the data, the experimenter would like to be able to attribute that effect to the toxin. However,

this will not be possible if there are other variables that were not controlled that could have caused the effect. This is sometimes referred to as **local control**. Experimental units should be as similar to each other as possible in order to obtain local control. If there is a high degree of variability among the experimental units, then it may be difficult to detect effects due to the increased variability. On the other hand, if high levels of variability exist among experimental units, then it may not be possible to differentiate any observed effects based on experimental conditions or differences in experimental units. If we want to observe the effect of a toxin on an organism, then we want to factor out **confounding** factors due to differences in temperature, differences in lab conditions and lab technicians, solution preparations, etc.

Blinding: Another aspect of local control is the notion of blinding. There are many examples of studies whose results came out wrong because the experimenter had reason to believe the experiment would come out in a particular way even before the experiment was even conducted. This phenomenon occurs everyday: if someone believes that something is a particular way, they find evidence that supports their view while discarding or ignoring evidence that does not support their view. Often this happens on a subconscious level. Unfortunately, scientists are not exempt from this sort of bias. For this reason, it is important for investigators to protect their research by *blinding* or **masking**. For instance, if one wants to study the effect of an environmental toxin on an organism in an experiment using different levels of toxin exposure, then the investigator should not know which experimental units received which treatment. A single-blind study is one where the subjects do not know which treatment they are receiving. This may not be very relevant in a study on fish say, but it can have a big impact in clinical trials involving humans. A double-blind study is one where neither the subjects nor the experimenter know which subjects are getting which treatments.

Controls: Another fundamental aspect of experimental design is that of a **control group**. If a study of the effect of a toxin on an organism shows an effect, how can the experimenter know the effect is more than what could occur by chance alone? A control group in this situation may be a group of organisms that are as similar as possible to the organisms that are exposed to the toxin except the control group is not subject to exposure. Therefore, if a difference is observed between the treatment group and control group that is too big to have occurred by chance alone, then it may be reasonable to attribute the difference to the toxin. In studies of anti-depressants, the control group is often a group of subjects who take a placebo pill. It is important that these studies be double-blind. It is well-known that there is a high placebo-response rate in studies of depression. In other words, many depressed subjects will report some improvement by simply knowing that they are taking something that is supposed to make them feel better, even if that something does not actually work. If a group of subjects receiving the actual anti-depressant treatment do show an average improvement, the experimenters need to determine if this improvement is greater than what would be observed from the placebo effect alone.

Another way of incorporating controls into an experiment besides a placebo control, is to have the control group receive a well-established standard treatment which will then be compared to the treatment of interest.

When implementing an experiment there will be a variable of interest (or perhaps several variables). We may record something as simple as whether or not an organism survives, or we may record how long an organism survives. The variable of interest may be a quantitative measurement such as the size of a plant or animal (length, width, etc). The experiment will then control **factors** that are thought to effect the variable. For instance, in the toxicity study on fish, the dose of the toxin would be the factor and we may perform the experiment at several different **levels** of the factor, i.e. different doses. Often times an experiment will incorporate other factors as well, for instance water temperature, pH level of the water, etc. A **blocking** factor is one whose effect on the variable is not of interest but is known to effect the outcome and therefore it must be controlled for. A blocking factor then is used to factor out extraneous sources of variability in order that a more focused study of the effect of the treatment can be made. For instance, suppose interest lies in accessing the effect of pesticide runoff on the thickness of turtle eggshells. Suppose also that a turtle's diet has a very big effect on eggshell thickness. If a study is undertaken to study the effect of the pesticide on the eggshell thickness but there is no control over the diet of the turtles, then it could be the case that most of the variability in the observed eggshell thicknesses is due to the difference in diets. If this variability is too great, then it may not be possible to detect a difference in mean eggshell thicknesses due to the pesticide. However, if diet is used as a blocking factor, then this source of variability can be controlled leading to a more powerful statistical analysis, i.e. an analysis that is more likely to find an effect due to pesticide if such an effect actually exists.

Randomization: We have discussed issues related to factoring out variability due to factors that are not of concern in an experiment. It must be noted however that there will always be some intrinsic variability between sampling units. When studying animals, no two animals are exactly the same, even after considering factors such as age, size, etc. In order to balance out uncontrolled systematic effects in an experiment, units are assigned to treatment combinations using **randomization**. Randomization is the process of assigning treatments to experimental units at random. If there exist systematic effects that we have not (or could not) control for, then the use of randomization will hopefully wash out differences between treatments that are not due to the treatments themselves. Suppose in the fish toxicity example that some fish have a genetic predisposition to be immune to the harmful effects of the toxin and the experimenters are unaware of this genetic effect. If the fish are randomized to different treatments based on varying doses of the toxin, then randomization will help make it possible with a high degree of probability that the fish with and without this genetic predisposition will balance out between the treatments. Random number generators can be used to implement randomization in practice. Most statistical software packages, such as R, have random number generators.

Non-random allocation of treatment to experimental units once again puts the experiment at risk of bias, even when it is unintentional.

Finally, remember the most sophisticated statistical analysis available cannot salvage a poorly designed or implemented experiment.

3 Analysis of Variance Review

In this section we give a very brief review of analysis of variance (ANOVA) where generally the goal is to compare the mean outcome over several different experimental treatments.

3.1 Single Factor ANOVA

In some experiments, there may only be a single factor of interest, say dose of a toxin. The statistical analysis for a single factor design is usually based on an ANOVA F-test. The levels of the factor in this example would be the different dose levels used and experimental units (e.g. organisms) would be randomly assigned to the various levels of the factor. If an equal number of units are assigned to each level, then the design is said to be *balanced*. If y represents the variable of interest that is measured, then one may be interested in the mean values of y at each level of the factor. These mean levels can be estimated by the sample means at each level from the experimental data.

Typically, the inference of interest is to see if the mean responses differ among the different levels of the factor. If the factor only has two levels, then a simple two-sample t-test can be conducted (provided the data look reasonably normal – otherwise a nonparametric test would be preferred). If the factor has more than two levels, then the general statistical procedure is to partition the overall variability in the response into two parts: one due to the treatment and one due to random variation within treatments. Hence, the analysis is called analysis of variance. The classical inference procedure is to do an F-test where the F-test statistic is a ratio of the estimated variance between levels of the factor to the estimated variance within each level of the factor. The null hypothesis H_0 is that the means across all levels of the factor are equal. If H_0 if true, the F test statistic follows an F-distribution. If H_0 is true, then F tends to take a value near one on average. If H_0 is false, then F tends to be larger than one and a p-value can be computed to access how likely it is to observe value the obtained value of F when H_0 is true. As with most hypothesis testing procedures, a small p-value (say $p < 0.05$) is evidence against H_0.

3.2 Multifactor Experiments

In some experiments, interest may lie in more than one factor. In a two-factor ANOVA, a treatment consists of a level of one factor and a level of the other factor. For example, the experimental unit may be a plant and the two factors could be fertilizer and water and a particular treatment would correspond to a level of fertilization and a specific amount of water. For a two-factor ANOVA the overall variability in the response can be partitioned into four parts: variability due to the first factor, the second factor, interactions between the two factors and a within treatment variability. Again, the classic inference procedures are based on F-tests. Typically one would first test to see if there is a significant interaction where the F-test statistic is

the ratio of the estimated variability due to the interaction divided by the estimated variance within each treatment which is called the mean squared error (mse). If the interaction is not significant, then it makes sense to test if there are main effects for the two factors.

3.3 Blocking

In ANOVA studies interest may lay primarily in one particular factor but other factors may explain much of the variability in the outcome. We may want to include these other factors to block out variation that is not of primary interest. Often blocks need to be included in the model to appropriately control for certain variables. For example, often an experiment is conducted by taking repeated measurements on the same person, animal or plant etc. Taking repeated measurements on an experimental unit can help factor out variability between units so that a more focused analysis can be performed on the experimental condition of interest, like a particular treatment. In other cases, in an experiment on plants, many plants may exist in the same plot, so the plot can act as a block. An important point here is that outcomes within a block will tend to be correlated with each other and this correlation must be taken into consideration for the model to be valid. In some experiments there is no confusion as to what constitutes a blocking factor, but in other examples, the experimenter can form a blocking variable using their judgement.

3.4 Random Effects

Often the blocking factor is a *random effect* where the levels of the factor may be considered as random instead of fixed. Consider an experiment to study plant growth under a control and an experimental treatment (e.g. a fertilizer). The treatment (control versus fertilizer) is a fixed effect with two levels. The experiment will use several plants, half of which are randomized to the control treatment and the other half to the experimental treatment. Each plant will be repeatedly measured over time. Here we can consider the plants as a random sample from the population of plants and the plant would be considered a random effect.

Here is an illustrative example from the classic book on experimental design by Box, Hunter, and Hunter Box *et al.* (1978):

Example. *open penicillin.r.* Four different processes (which we will call A, B, C, and D) were used to produce penicillin from batches of corn steep liquor. There were 5 batches of liquor for the experiment. The four processes are fixed effects. However, the five batches represent a random sample from a potentially very large collection of batches and hence batch is a random effect here. Also, the production of penicillin can vary quite a bit from batch to batch, so including batch in the model will account for this source of variability. Additionally, penicillin yields from the same batch will be correlated to each other. Introducing a random effect for batch accounts for this correlation. This is an example of a *complete randomized block design.* Note that we are not necessarily interested in the effect of the batch but instead we are primarily

interested if there are differences in yield based on the four treatments. The primary goal of the statistical analysis for this example is to test if there are any differences in mean yield based on the four treatments. Thus the null hypothesis will be that the mean yield across the four treatments are equal. The data from this experiment are below:

Treatment	Batch	yield
A	1	89
B	1	88
C	1	97
D	1	94
A	2	84
B	2	77
C	2	92
D	2	79
A	3	81
B	3	87
C	3	87
D	3	85
A	4	87
B	4	92
C	4	89
D	4	84
A	5	79
B	5	81
C	5	80
D	5	88

Figure 1 shows a boxplot of penicillin yield for the four different treatments. It looks as if treatment C produces a higher yield – we perform a formal test for differences in treatment next.

Letting y denote the yield response, t_j represent the mean effect of the jth treatment and b_i denote the ith batch effect, we can write down a model for this experiment as

$$y_{ij} = \beta_0 + t_j + b_i + \epsilon_{ij},$$

where y_{ij} denotes the yield from the ith batch ($i = 1, 2, 3, 4, 5$) and the jth treatment. The ϵ_{ij} is a random error with mean zero, usually assumed to follow a normal distribution with variance σ^2. The β_0 is usually called the intercept in the contexts of linear models and in particular regression models. Here, β_0 will correspond to an overall mean response. The t_j are considered fixed parameters but the b_i are considered random variables since they correspond to random effects. In particular, we typically assume b_i are normal with mean zero and a random effect variance which we will denote σ_b^2. We also assume the random effects are independent of the random errors. Since we are assuming the distributions of the random effects and the random errors are well-approximated by normal distributions, and they are independent, it follows that

$$y_{ij} \sim N(\beta_0 + t_j, \sigma_b^2 + \sigma^2).$$

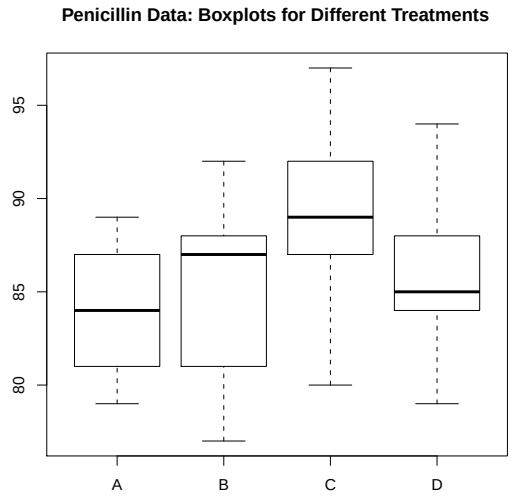

Figure 1: Boxplots for penicillin yield for the four different treatments.

Maximum Likelihood Estimation. The old fashioned way of performing the statistical analysis for this and other related models is to do an ANOVA approach by computing sums of squares, etc. However, for this model and more complicated models, this approach becomes complicated, particularly for unbalanced data (unequal number of observations per treatment). A preferable approach is to use the method of *maximum likelihood*. Since we know the distribution of all the y_{ij}'s we can write down their *joint* density defined in terms of the model parameters (namely, $\beta_0, t_1, t_2, t_3, t_4, \sigma_b^2, \sigma^2$ in this case). The likelihood function is simply this joint density evaluated at the observed data values and regarded as a function of the model parameters. The idea then is to find the values of the parameters that make the likelihood function large as possible, hence the name "maximum likelihood." Intuitively the goal is to find the values of the parameters that make the observed data most likely to have occurred. Typically it is much easier to maximize the logarithm of the likelihood function. Numerical methods are often required to maximize log-likelihoods and most statistical software packages have this capability: in SAS one can use PROC MIXED and in R one can use the "lmer" function from the lme4 package. The name "lmer" stands for "linear mixed effects revised". This function can be applied to linear models (i.e. models where the relation of the coefficient parameters to the response is linear) that have both fixed and/or random effects (hence the name "mixed effects").

Here is the example analyzed using the R software:

```
library(lme4)
dat=read.table("penicillin.dat", header=T)
dat[,2]=as.factor(dat[,2])  # Make the variable batch a factor instead of a numerical varia
names(dat) # print the names of the variables
dat   # print the data
boxplot(yield~Treatment, data=dat,
  main="Penicillin Data: Boxplots for Different Treatments")
```

```
fit=lmer(yield ~ Treatment + (1|Batch), data=dat, REML=F)
summary(fit)
```

Note that the different batches are coded taking values 1–5, but in reality, the different batches are not numerical or quantitative but correspond to different levels of a qualitative factor; therefore, in R you can use the function "as.factor" to change a variable to a factor so that R will recognize it as such. The term "(1|Batch)" tells R that the factor Batch is a random effect. When writing computer programs, regardless of the application, one should always include many comments as to what the program is doing. If this is not done, then it can often be very difficult to recall what the program is supposed to be doing when you come back to look at the program weeks or months later or if someone else is reading the program. In R, anything following the "#" symbol is ignored by R and can be used to add comments to your code. The "REML=F" tells R to fit the usual model using maximum likelihood. The default method is *restricted maximum likelihood* or REML which tends to produce less biased estimates (we will not go into the details of REML here). Note that in order to run this program in your version of R, you must load the library "lme4" (which can be done by clicking "packages" at the top of the R window). Here is part of the output from running this program.

```
Linear mixed model fit by REML
Formula: yield ~ Treatment + (1 | Batch)
   Data: dat
   AIC   BIC logLik deviance REMLdev
 115.8 121.8 -51.91    117.3   103.8
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 11.792   3.4339
 Residual             18.833   4.3397
Number of obs: 20, groups: Batch, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   84.000      2.475   33.94
TreatmentB     1.000      2.745    0.36
TreatmentC     5.000      2.745    1.82
TreatmentD     2.000      2.745    0.73
```

From this output, we see that the maximum value of the log-likelihood is -51.91. The variance of the error term is estimated to be $\hat{\sigma}^2 = 18.833$ and the variance of the Batch random effect is estimated to be $\hat{\sigma}_b^2 = 11.792$. Below this in the output we see the estimated fixed effects for three of the treatments and an associated t-test statistic to see if the effect differs significantly from zero. Note that these t-test statistics are fairly small indicating that there does not appear to be any significant effect for the different treatments.

We can perform a *likelihood ratio test* to test if there is a difference in yield due to the different treatments. We can do this by fitting a model without the treatment

factor and compare it to the original model. Note that the command "REML=F" is required when fitting the model if we wish to use the likelihood ratio test to test significance of fixed effects. We shall call the original model with the treatment factor the *full* model and the model without the treatment effects will be called the *reduced* model which is fit with the following R code:

```
fit2=lmer(yield ~ (1|Batch), data=dat, REML=F) # Drop treatment factor
summary(fit2)
anova(fit2, fit)  # Perform likelihood ratio test
```

The log-likelihood for the "reduced" model without the treatment factor is -60.66. If we take the logarithm of the ratio of the likelihoods from the two models, the logarithm will turn the ratio into a difference. Multiplying this difference by -2 produces a test statistic that has an approximate chi-square distribution on $4-1=3$ degrees of freedom when the null hypothesis is true (i.e. when there is not difference on average between the four production methods). We will discuss the likelihood ratio test in other contexts later in this course.

```
Data: dat
Models:
fit2: yield ~ (1 | Batch)
fit: yield ~ Treatment + (1 | Batch)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit2  3 127.33 130.31 -60.662
fit   6 129.28 135.25 -58.639 4.0474      3     0.2564
```

The chi-square test statistic is computed as

$$-2\{\text{log-likelihood(reduced)}-\text{log-likelihood(full)}\} = -2(-60.662-(-58.639)) = 4.0474$$

and the associated p-value is $p = 0.2564$ which is quite large indicating no significant differences in yield due to the different treatments. The chi-square distribution for the likelihood ratio test is an asymptotic result meaning that the chi-square approximation becomes better as the sample size grows larger. We obtain a more accurate p-value using a method called the *parametric bootstrap* which we will not go into here.

In order to illustrate another feature of R, we note that R has several built-in functions that compute cumulative probability functions, densities, and random number generators for commonly used distributions (e.g. normal, t, F, chi-square, Poisson, etc.). The following R code computes the p-value we see in the above output manually using the "pchisq" function:

```
teststat = as.numeric(-2*(logLik(fit2)-logLik(fit)))
pval=1-pchisq(teststat,3)
teststat
pval
```

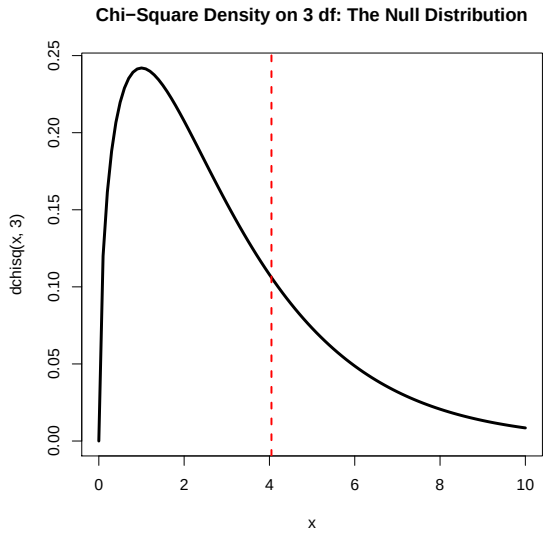

Figure 2: A plot of the chi-square density on 3 degrees of freedom which is the approximate sampling distribution of the likelihood ratio test statistic under the null hypothesis. The dashed vertical line marks the value of the chi-square test statistic that tests for a difference in the four production methods.

Figure 2 shows a plot of the chi-square density function on 3 degrees of freedom which is the approximate distribution of the likelihood ratio test statistic when the null hypothesis is true (that is, when there is no mean difference in penicillin yield for the four treatments). The vertical line marks the value of the observed test statistic which falls roughly in the middle of the distribution. In other words, the observed test statistic is not an unusual value for a chi-square distribution on 3 degrees of freedom which means that the data is not inconsistent with the null hypothesis (i.e. we do not reject the null hypothesis).

4 Simulations

The older and classical statistical inference procedures were developed in the age before high speed computing. Now that high speed computing is easily available, statistical inference procedures have been able to take advantage of this power. Modern statistical methodologies rely very heavily on computationally complex calculations requiring high speed computers. Fortunately most of us have easy access to high speed computing (e.g. our laptop computers). Thanks to the power of computing, we can use simulations to evaluate how certain statistical procedures perform. The idea of simulations is to have the computer generate a set of pseudo data or simulated data. This can easily be done using random number generators which most statistical software packages have available. Actually, the variates generated in computer simulations are not truly random and hence it is more proper to call them "pseudo" random. However, a good random number generator can produce variates that mimic

randomness very well. In this course, we shall use simulations at various times to illustrate certain topics and for statistical inference. Simulations can also be helpful in determining necessary sample sizes needed for proposed surveys or experiments for complicated models where conventional sample size formulas may not be appropriate or available.

In R, one can simulate from most of the well known probability distributions using built-in functions. For non-standard probability distributions, simulated data sets can be obtained using methods like *acceptance-rejection* sampling, details of which can be found in books on simulation.

Let n denote the desired sample size one wants to simulate. Then typing "rnorm(n)" in the R console will produce a set of pseudo-random standard normal variates. For a general normal distribution, suppose the mean is 98.6 and the standard deviation is 0.7, then one can type "rnorm(n, mean=98.6, sd=0.7)" to obtain a random sample of n normal variates. Simulated data sets from other distributions such as the t-distribution, gamma, poisson, etc. can be generated in a similar fashion using the functions "rt", "rgamma", "rpois", etc.

5 Nonparametric Density Estimation

(*open voles.r*). This section gives a very brief introduction to the notion of nonparametric density estimation which will be useful later in the quarter for making graphical comparisons of different estimators and procedures. Figure 3 shows a histogram of 288 skull lengths (mm/100) from a morphometric study of voles (Flury, 1997, page 333). The distribution does not look very normal and there is a good reason for that: the skull lengths were measured on two different species of vole (*microtus multiplex* and *M. subterraneus*). Therefore, the actual data comes from a *finite mixture distribution* with 2 components for the two species. The first column of the data set *voles.dat* is an indicator for the two species of voles and the value of 0 in this indicator variable corresponds to unclassified voles. There were several other variables measured for this study besides skull length.

If x denotes a continuous random variable (e.g. skull length), we would like to know its density function $f(x)$ but in practice the density is unknown. Often we may use a normal density to approximate $f(x)$ but this is not a good idea if the data is clearly non-normal as in the vole example illustrated in Figure 3. The histogram can be thought of as a crude approximation to the density function. The idea of nonparametric density estimation is to obtain a better approximation to the density that is a smooth function. The most popular method of nonparametric density estimation is to use *kernel density estimation*. Denote the data by $x_1, x_2, \ldots, x_n$. Then the kernel density function is of the form

$$\hat{f}_\lambda(x_0) = \frac{1}{n\lambda} \sum_{i=1}^{n} K_\lambda(x_0, x_i), \tag{1}$$

where K is a kernel function. The most popular kernel function is the *Gaussian*

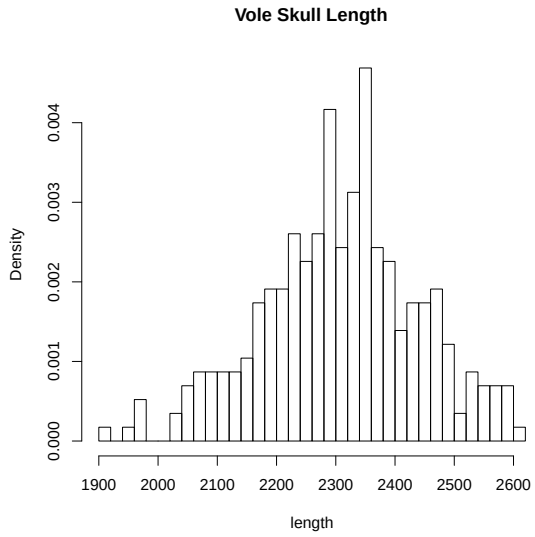

Figure 3: A histogram of vole skull length

kernel given by

$$K_\lambda(x_0, x) = \frac{1}{\sqrt{2\pi\lambda^2}} e^{-(x_0-x)^2/(2\lambda^2)}.$$

This is called the Gaussian kernel because it coincides with the density function for a normal distribution which is also known as the Gaussian distribution after one of the most famous mathematicians ever Johann Carl Friedrich Gauss (1777–1855). The idea of the kernel estimate is that points in the data close to x_0 receive a higher weight than points further from x_0. The nonparametric kernel density estimator (1) produces estimated densities with varying degrees of smoothness. The value of λ (sometimes called the bandwidth) controls the smoothness of the density estimate: as λ gets bigger, one is smoothing the data over a larger window of data points and the density estimator gets smoother. The R function "density" estimates the nonparametric density function with a default Gaussian kernel (other kernel functions could be used) and a default smoothness parameter. Figure 4 shows the histogram of skull lengths again with the nonparametric density estimator overlayed. Note that in the "hist" command in R which draws a histogram, the option "freq=F" is needed here so that a relative frequency histogram is drawn as opposed to a frequency histogram (that gives the counts in each bin instead of the proportion of data in each bin).

In order to compare the two different species of vole (and ignore the unclassified voles), we can draw nonparametric density estimates for the two species on the same plot to make a comparison. Figure 5 shows this plot where it appears that *microtus multiplex* has a longer skull length on average than *subterraneus* voles.

The R program for producing these plots is given below:

```
# Vole data
# (from "A First Course in Multivariate Statistics" by B. Flury)
# Column 1: group indicator variable for species
```

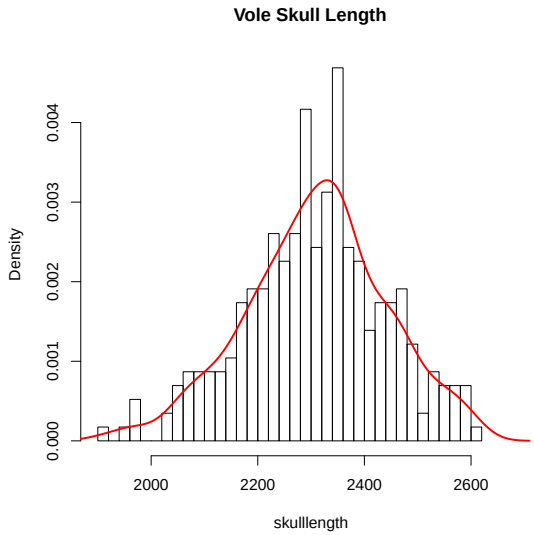

Figure 4: A histogram of vole skull length with the nonparametric density estimator.

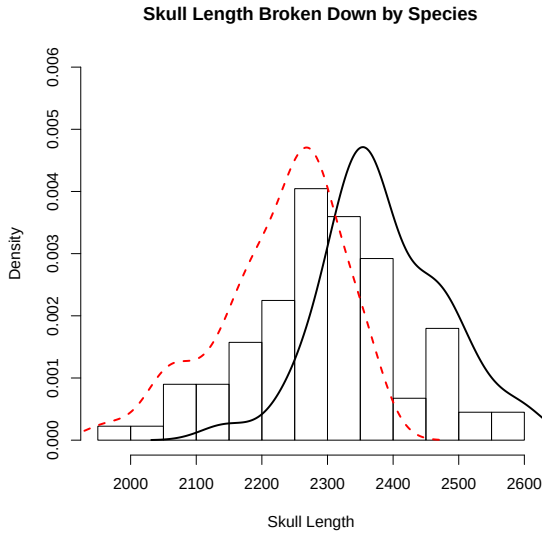

Figure 5: Nonparametric density estimates of vole skull length broken down for the two different species of vole within the data set.

```
#     (1=microtus multiplex, 2=M. subterraneus, 0=not determined)
# Column 2: width of upper left molar 1
# Column 3: width of upper left molar 2
# Column 4: width of upper left molar 3
# Column 5: length of incisive foramen
# Column 6: length of palatal bone
# Column 7: skull length
# Column 8: skull height above bullae
# Column 9: skull width across rostrum
# Columns 2-6 are in units of mm/1000
# Columns 7-9 are in units of mm/100
#
voles=read.table("voles.dat", header=F)
skulllength=voles[,7]
hist(skulllength, 25, freq=F, main="Vole Skull Length", xlim=c(1900,2700))
lines(density(skulllength), lwd=2, col=2)

# Break down the data by species
vole1=voles[voles[,1]==1,]
vole2=voles[voles[,1]==2,]
skulllength1=vole1[,7]
skulllength2=vole2[,7]
# Omit the unclassified voles
hist(voles[voles[,1] != 0,7], 20, ylim=c(0,.006), freq=F,
 main="Skull Length Broken Down by Species", xlab="Skull Length")
lines(density(skulllength1), col=1, lwd=2)
lines(density(skulllength2), col=2, lty=2, lwd=2)
```

References

Box, G. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. Wiley, New York.

Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer, New York.