Stat 6390
Environmental Statistics

Syllabus

Stat 6390 Environmental Statistics Course Information

Instructor: Dr. Larry P. Ammann
Office hours: Wed, 2:00-3:30 pm, others by appt.
Email: ammann@utdallas.edu
Office: FO 2.410C
Phone: (972) 883-2164
Text: Course notes and web resources

Topics

• One-sample estimation and hypothesis tests
• Two-sample estimation and hypothesis tests
• Regression and ANOVA
• Generalized linear models

Important components of these topics include power, sample size determination, and observable differences for methods based on normality assumptions as well as alternatives to those assumptions. The statistics programming language \textit{R} will be used heavily in this course. An extension to \textit{R} called \textit{Rstudio} also is recommended.

Grading Policy
Course grade will be based on quizzes, homework projects and the final project.
Quizzes: 25%  Homework projects: 50%
Final project: 25%

Note: the complete syllabus is available here:
This course uses the statistical programming language and environment R. This is freely available software with binaries for Linux, MacIntosh, Windows that can be obtained from: http://cran.r-project.org

A very useful extension for R is another freely available open-source package, RStudio. There are versions available for Windows, Mac, and Linux which can be downloaded from: https://www.rstudio.com/products/rstudio/download/

This package provides an intelligent editor for script files, it allows specific projects to be assigned to their own directories to aid in organization of your work, RStudio includes an interactive debugger to help identify and correct errors, and it provides an interactive GUI for easily exporting graphics to image files for incorporation into documents.

Class Notes

Background

Environment regulation had its roots in public health. The U.S. Public Health Service was formed in 1912 from several other organizations. Its initial focus was on water-borne diseases such as typhoid. Later the PHS developed standards for air quality in the industrial workplace in addition to those for water quality. However, it was the tremendous industrial growth after WWII, and all the highly visible pollution associated with that growth, that caused some to call for federal regulatory action. The large scale of factories in the chemical, plastics, petroleum, and automotive industries meant that problems associated with their waste products no longer were confined to a single town, county, or even state.

I was born in and grew up in South Florida after WWII and witnessed first-hand some significant environmental problems. Once home air conditioning became widely available, people began migrating from northern states to Florida. As a result its coastal cities grew rapidly. The attitude among many of those involved with building public infrastructure to support this growth was that the ocean was so vast it never could be polluted, and so cities built large pipes into the ocean to discharge raw sewage. But then under certain weather conditions beaches had to be closed due to bacterial contamination from those sewage outfalls.

Another major problem came from the expansion of farms below Lake Okeechobee. Water from much of the interior of central and south Florida drains into the lake causing it to expand and contract depending on seasonal rainfall. Lake Okeechobee does not discharge into a well-defined river but instead flows slowly through a 50 mile wide area called the Everglades. As a result the soil south of the lake is very rich and crops could be grown without fertilizer. The first large farms there grew sugar cane. Later large vegetable farms were added that supply most of the winter vegetables for the cities of the northeastern and midwestern U.S. But there were two problems that needed to be addressed for these farms to be successful. The periodic flooding of the lake needed to be controlled and so did the
insects that flourished in this semi-tropical environment.

To solve the first problem levees were built around the lake. This had a significant
impact on the Everglades by seriously restricting the amount of water that flowed through
the Everglades into Florida Bay. In the 1950s and early 1960s the glades were viewed as a
worthless swamp fit only for alligators and a few crazy people. It was felt that this water
was better used for farming and as drinking water for the increasingly large numbers of new
residents of south Florida. But Florida is mostly limestone. As the Everglades dried and
more subsurface water along the coast was pumped to water all those new lawns in ever-
expanding coastal cities, these areas began to experience saltwater intrusion which made
that those wells unfit for use. At the same time we began to understand the critical role
of Florida Bay as a nursery for fish and shellfish that inhabit the eastern Gulf of Mexico
and the Atlantic Ocean south and east of Florida. This rich nursery was threatened by the
reduced flow of fresh water into the bay that supported the dense mangrove forests along
the coast.

The insect problem was exacerbated by mosquitoes that feasted on all those new residents.
Not only are mosquitoes a nuisance, but they also carry diseases some of which are deadly.
To combat the insect problem, DDT was applied heavily in farming areas. Anther pesticide,
malathion, was routinely sprayed in residential areas to temporarily reduce adult mosquito
populations, although that relief only lasted about two days at most. Pesticides also were
applied to wetlands to kill mosquito larvae in the water before they could morph into adults.
Although DDT helped eradicate malaria in some areas of the world, not much was known
about its negative impacts until Rachael Carson published her ground-breaking book Silent
Spring in 1962. In this book she argued that DDT and other pesticides were harmful to
wildlife and humans in ways we did not understand. One very influential reader of her book
was President John F. Kennedy who ordered his scientific advisory committee to investigate
those claims. This was an important milestone. JFK asked for science-based decisions
rather than simply listening to other politicians and the opinions of lobbyists who had
significant financial interests in the outcome.

At the same time Floridians began to notice a disturbing trend - birds that depended on
fish were disappearing. These included Osprey, Bald Eagles, herons, egrets, and Florida’s
iconic pelicans. As studies motivated by Silent Spring were undertaken, we learned that
long-lasting compounds in DDT were being absorbed by aquatic insects which then were
consumed by fish. Those compounds remained intact in the fish and accumulated, a pro-
cess referred to as bioaccumulation. Birds that consumed fish were then exposed to those
compounds the effect of which was thinning of their eggshells. As a result the eggs were not
strong enough to support the weight of an adult bird and the eggs ended up getting crushed
in the nest.

Other highly visible environmental disasters such as the Cayahoga River in Cleveland
catching fire brought greater awareness that damage to the environment by human activity
has dangerous consequences. This awareness created strong support for the formation of the
Environmental Protection Agency to oversee development, application, and enforcement of
federal standards for air and water quality. The broad public support for creating such an
agency resulted in President Nixon signing the bill creating the EPA on December 2, 1970. This did not cause the end of environmental disasters in this country and so it became clear to many that the EPA needed to be proactive, not just reactive, to sources of pollution. This led to what are referred to as the Clean Water Act and the Clean Air Act. The first CWA was enacted in 1972 (over Nixon’s veto). This landmark legislation **prohibits the release of toxic substances in toxic amounts** into the waters of the U.S. The CAA first passed in 1970 and is designed to regulate emissions of hazardous air pollutants with a similar goal of prohibiting the release into the atmosphere of toxic substances in toxic amounts.

**Examples**

- **Love Canal, NY.** In the late 1950s about 100 homes and a school were built on property that once was a chemical dumpsite. Then in 1978 it was discovered that carcinogenic compounds from this dumpsite had been leaching into those homes and school producing birth defects, miscarriages, and leukemia among the residents.

- **Hudson River PCB contamination.** PCB is a chemical that was used in electrical transformers until it was banned in 1976 as a highly toxic carcinogen. Tragically over 1 million pounds of PCB were discharged into the Hudson River by G.E. and absorbed by river sediments. Even today an advisory to not eat fish from the Hudson remains in effect due to this PCB contamination. This part of the Hudson River is an EPA Superfund site and delicate dredging is being done to remove PCB from the river sediment. The challenge is to remove contaminated sediment without stirring up the sediment and causing PCB to flow downstream. This will take a long time. Also, hundreds of thousands of pounds of PCB remains in the shale beneath G.E.’s plant next to the river and it continues to leach into the Hudson.

- Environmental disasters have occurred in other countries as well. A Union Carbide plant in **Bhopal, India** released a highly toxic gas, methyl isocyanate, as well as other toxic gases that immediately killed over 2,000 people who lived around the plant. Many others, especially children, experienced long-term health effects.

Incredibly, similar events continue today.

- **West, Texas.** On April 17, 2013 a massive explosion of improperly stored fertilizer, ammonium nitrate, killed 15 people, injured more than 160, and destroyed or damaged over 150 buildings.

- **Beijing, China.** Rapid industrialization and use of coal-fired power plants and heating stoves have created dangerous levels of smog in Beijing and other cities such as Shanghai. At its worst, levels of smog have been recorded that are over seven times greater than safe levels. This smog contains particulates that penetrate deep into lung tissues when inhaled causing permanent damage. These events cause schools and factories to close and residents are advised to stay indoors. China is working now to replace coal with oil and gas for heating and power.
• **Flint, Michigan.** A decision to reduce costs associated with drinking water by using water from nearby Flint River instead of water from Lake Huron. However, higher salinity of the Flint River caused lead in pipes carrying water into homes to be leached into the water. Eventually it was discovered that blood lead levels in children were dangerously high as a result. It has been known for a long time that plumbing in older dwellings contained lead and so were susceptible to leaching. Proactive testing for potential contamination before the switch to the Flint River would have detected this in which case this tragedy would have been avoided.

**Ban of leaded gasoline.** One very important example of the hurdles faced by environmental regulation is the story of how leaded gasoline was banned in the U.S. In the 1920s tetraethyl lead (TEL) was added to gasoline to improve engine performance. Although lead was known to be toxic to human health, instruments sensitive enough to measure atmospheric lead from automobile exhausts did not exist at that time. Scientists employed by the company that produced TEL testified to Congress that low levels of lead had no effect on human health. In spite of the obvious conflict of interest, Congress accepted the opinion of that company’s chief scientist, Robert Kehoe, instead of demanding scientific proof of his assertions like JFK did regarding DDT. So TEL continued to be used as an additive and the producers of TEL continued to profit from it.

In the late 1940s a geology graduate student named Clair Patterson was given a dissertation topic to measure the age of the Earth using the decay of uranium into lead in meteorites. This work required measuring extremely small levels of lead using the most advanced mass spectrometers available at that time. However, a major problem Patterson had to overcome was contamination of his samples by lead in the environment. Patterson eventually realized that he needed to perform his work in a room that was completely isolated from the outside that had been scrubbed thoroughly and he needed to be covered in protective clothing. This was the first *clean room*. This was successful and Patterson determined that the Earth’s age is 4.55 ± .07 billion years (current estimate is 4.54 ± .05 billion years).

After he completed this work, Dr. Patterson wanted to determine the sources of contamination he had found in his initial efforts to measure lead in meteorites. He examined ocean water and found significantly higher lead levels near the surface compared to the ocean floor. He then examined ice cores from Greenland and found that levels of atmospheric lead began to increase significantly at the same time that TEL began to be added to gasoline. To Patterson this was the smoking gun that proved leaded gasoline was a major source of atmospheric lead. Unfortunately, lobbyists and scientists with financial connections to TEL engaged in a campaign to discredit Dr. Patterson and his research. As catalytic converters were introduced as a way to remove smog-producing compounds from engine exhaust, it was found that leaded gasoline corroded them and so these two effects finally led to a ban on leaded gasoline was approved. In the meantime, many more children were exposed to this toxic substance. We know now that there is no safe level of lead and that it causes irreversible damage to the development of children’s brains resulting in behavioral problems and reduced IQs. The story of Dr. Patterson and the ban of leaded gasoline is told in an episode of *Cosmos - A Spacetime Odyssey* that appeared on PBS in 2014.
Sampling

The first step in any research problem is to define the population – the set of all individuals about which you have questions to answer. The next step is to express the main question in terms of observable characteristics of individuals in the population. The types of questions we will discuss in this course involve estimation, model-building, and decision-making (hypothesis testing) when the entire population cannot be measured.

Estimation refers to determining a particular attribute of the population. Model-building refers to determining how multiple attributes are related. For decision-making we must formulate the question into a set of actions that will be taken based on the value of some measure derived from the observable characteristics.

A sample is any subset of the population. If the entire population cannot be measured, then we are forced to answer our questions using a sample. This produces error or uncertainty in our estimates, models, and decisions. These only have meaning if we can quantify this error and that can be done only if the samples are obtained randomly. Suppose for example we want to compare species population densities among different zones of a river where zones are defined in terms of depth, flow, and vegetable type on the banks of the river. Suppose there are 3 zones and a survey of the river gives the following.

<table>
<thead>
<tr>
<th>Distance from start</th>
<th>Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 3</td>
<td>1</td>
</tr>
<tr>
<td>3-5.5</td>
<td>2</td>
</tr>
<tr>
<td>5.5-7</td>
<td>1</td>
</tr>
<tr>
<td>7-10</td>
<td>3</td>
</tr>
<tr>
<td>10-14</td>
<td>2</td>
</tr>
<tr>
<td>14-18</td>
<td>3</td>
</tr>
</tbody>
</table>

Then the total length of zone 1 is 5.5 and so we would generate $n_1$ i.i.d. r.v.’s from $\text{Unif}(0,5.5)$. These r.v.’s would give the locations of our sampling sites within zone 1.

In all cases we must determine the appropriate sample sizes that will result in acceptable levels of error. For estimation we must specify the precision of the estimate along with the confidence expressed as a probability that the actual error is within the stated precision. For hypothesis testing we must specify the level of significance of the test (probability of making Type I error) and the power (probability of rejecting null hypothesis) at a specified distance from the null hypothesis. To accomplish this we must make assumptions about the underlying distribution of the data, or we must apply distribution-free methods. Any assumptions we make must be verified.

Data frames

R had a special data type called a data frame. This structure was introduced to unify modeling functions in R. A matrix in R has the restriction that all elements must be the same type. However, data sets that include a mix of numeric and categorical variables and
so cannot be represented by a matrix in R. A data frame is a structure designed for such cases. It is like a matrix in that a data frame can be considered as a two-dimensional object with columns representing variables and rows representing sampling units. Each column of a data frame must contain the same type of data but different columns may be different types. The column names can be thought of as the names of the variables in the data set. A data frame also is like a list in which the columns are its components. The column names correspond to the names of those components and can be referenced by those names.

The most common way in which a set of tabular data is entered into R is by using one of its functions specifically designed for tabular data. These functions include read.table() and read.csv(). read.table() is the general function and read.csv() is designed for comma-separated-values data. Since data sets often are extracted from spreadsheets or database systems, the easiest way to enter such data into R is to save the data as file type CSV from the spreadsheet or database and then use read.csv() to bring it into R. These functions return the data as a data frame. Typically the first row of the file contains names for the variables and the remaining rows contain the data. In such cases, the optional argument header in read.table() should be set to TRUE. That is the default in read.csv() so that argument does not need to be set in that function.

If a column contains strings instead of numbers then these functions automatically convert those columns to factors. If a categorical variable is coded numerically, then it would need to be explicitly converted to a factor using the factor() function after reading the data. If one of the columns is not data but contains names of the rows, then that column should be specified in the row.names argument. These functions check to see if each line of the file contains the same number of fields and returns an error if not. The function count.fields() can help locate the offending lines. This can happen when extracting data from Excel which often puts in extraneous commas and extra lines into the CSV file it creates. Another problem that may occur is when a column of string data includes some values with apostrophes. By default apostrophes are assumed to be string delimiters in addition to the double quote symbol, so the quote argument must be used.

The file http://www.utdallas.edu/~ammann/stat6390scripts/cars.csv is a data set that gives several measurements for a set of cars. It is a CSV file with column names in the first line and the first row contains vehicle names. Note that rows for diesel cars begin with the # symbol. That is the comment symbol and indicates those rows should be ignored. That is the default for the comment= argument in read.table() but not in read.csv(), so it must be specified for that function. The cars data can be read into R by

```
Cars = read.csv("http://www.utdallas.edu/~ammann/stat6390scripts/cars.csv", comment="#", row.names=1)
```

The plot and modeling functions in R have a formula interface that encourages users to think of the data in terms of dependent variables and independent variables. These formulas
can contain the names of variables in a data frame in which case the `data=` argument must give the name of the data frame object.

```r
plot(mpg ~ weight, data=Cars) # scatterplot since x and y are numeric
plot(mpg ~ origin, data=Cars) # box plots since y is numeric and x is a factor
```

These plots need to be improved visually. For the first plot let’s color points differently for the different levels of `origin`, add an informative title and a legend. In the second plot use the same colors for the boxes as were used for the points in the first plot.

```r
ocols = c("blue","gold","red")
names(ocols) = levels(Cars$origin)
plot(mpg ~ weight, data=Cars, pch=19, col=ocols[Cars$origin])
legend(max(Cars$weight),max(Cars$mpg),legend=levels(Cars$origin),pch=19,col=ocols,xjust=1)
```

```r
title("EPA Mileage vs Weight")
```

```r
plot(mpg ~ origin, data=Cars, col=ocols)
```

```r
title("EPA Mileage by Vehicle Origin")
```

These plots can be saved into graphic files to be included in documents. The graphic file format most often used for \LaTeX{} is `png`. The graphics function `png()` can be given at the beginning of the commands for a graphic or it can be saved interactively if you are using RStudio. If you are preparing graphics for slides, then it usually is more effective to use a black background and white for annotation. This requires some graphical color parameters to be set explicitly to white or some other non-black color.

```r
ocols = c("SkyBlue","gold","red")
names(ocols) = levels(Cars$origin)
png("Cars_wgt_mileage.png",bg="black",width=720,height=720)
plot(mpg ~ weight, data=Cars, pch=19, col=ocols[Cars$origin],
      fg="white",col.axis=white",col.lab=white",col.main="white",col.sub="white")
legend(max(Cars$weight),max(Cars$mpg),legend=levels(Cars$origin),pch=19,
       col=ocols,xjust=1, bty="n", title="Origin",
       text.col="white", title.col="white")
```

```r
title("EPA Mileage vs Weight",col.main="white")
```

```r
graphics.off()
png("Cars_wgt_origin.png",bg="black",width=720,height=720)
plot(mpg ~ origin, data=Cars, col=ocols,
      fg="white",col.axis="white",col.lab="white",col.main="white",col.sub="white",
      medcol="white",whiskcol="white",staplecol="white",outcol="white",outpch=19)
title("EPA Mileage by Vehicle Origin",col.main="white",cex.main=2)
```

```r
graphics.off()
```

### Quantile-Quantile Plots and Probability Distributions

A Quantile-Quantile plot gives a graphical comparison of a data set to a specified probability distribution or another data set. If two distributions are the same, then their quantiles should
be the same. If a r.v. is a linear transformation of another, $y = a + bX$, then their quantiles should exhibit the same linear relationship. Suppose for example that $Y \sim N(\mu, \sigma)$ and we observe a random sample from that distribution. A quantile-quantile plot of the data versus the standard normal distribution should fall on a line with intercept $\mu$ and slope $\sigma$.

Since the mean and standard deviation are easily distorted by an outlier, \texttt{R} uses a robust method for its \texttt{qqline} function to estimate the slope and intercept of a comparison line. By default, \texttt{qqline} estimates the intercept and slope from the 25th and 75th percentiles of the two distributions. Suppose for example $y$ contains the ordered values of the data and we want to compare its distribution to the standard normal distribution. Let $n$ denote the sample size. Then the sample quantile of $y[k]$ is $k/n$. Instead of comparing $y$ to those quantiles of the standard normal distribution, an offset is used, by default .5.

```r
n = 50
mu = 100
sig = 12
y = rnorm(n,mu,sig)
### mimic qqnorm(y)
n = length(y)
a = .5 # offset; qqline uses 3/8 if n <= 10
probs = (seq(n) - a)/n
x = qnorm(probs)
plot(x,sort(y)) # equivalent to qqnorm(y)
# mimic qqline(y)
p = c(.25,.75)
yq = quantile(y,p)
xq = qnorm(p)
slope = diff(yq)/diff(xq)
abline(intercept,slope) # equivalent to qqline(y)
```

An alternative to Q-Q plots is Tukey’s Mean Difference Q-Q plot. For this plot the y-axis contains the difference between the sample and reference quantiles, the x-axis contains the average of those values.

```r
### mimic tmdnorm(y)
n = length(y)
a = .5 # offset; qqline uses 3/8 if n <= 10
probs = (seq(n) - a)/n
x = qnorm(probs)
y = sort(y)
y1 = y-x
x1 = (x+y)/2
plot(x1,y1)
```
A script for TMD plots is here:
http://www.utdallas.edu/~ammann/stat6390scripts/tmdplot.r

This script defines functions for Q-Q and M-D plots. The function \textit{tmdplot}(x,y) is a basic function with arguments \(x,y\) where \(x\) represents the target quantiles and \(y\) represents the sample. The function \textit{tmdnorm}(y) generates a TMD plot with normal distribution target. Other functions are defined that obtain Q-Q and TMD plots for several other common distributions.

A TMD plot is interpreted differently than a standard QQ plot. By default a TMD plot with the normal distribution as target uses the sample mean and s.d. to estimate parameters of the normal distribution from which target quantiles are obtained. In this case if the sample is approximately normal, then the TMD plot will be close to a horizontal line at 0.

```r
source("http://www.utdallas.edu/~ammann/stat6390scripts/tmdplot.r")
n = 100
mu = 50
sig = 8
y = rnorm(n, mu, sig)
par(mfrow=c(1,2),cex.main=.95)
qqnorm(y)
qqline(y)
mtext("Normal data",side=3,line=.5,cex=.9)
tmdnorm(y)
mtext("Normal data",side=3,line=.5,cex=.9)
```

A commonly encountered distribution with environmental data is the lognormal distribution. This is defined to be the distribution of a r.v. \(X\) such that \(Y = \log(X)\) has a normal distribution. It’s density function is given by

\[
f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp \left( -\left[ \log(x) - \mu \right]^2 / (2\sigma^2) \right), \quad x > 0,
\]

where \(\mu, \sigma^2\) are the mean and variance, respectively, of \(\log(X)\). Then

\[
E(X) = \theta = \exp(\mu + \sigma^2/2),
\]

\[
\text{median}(X) = \exp(\mu),
\]

\[
\text{Var}(X) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1] = \theta^2[\exp(\sigma^2) - 1],
\]

\[
CV(X) = \tau = \sqrt{\exp(\sigma^2) - 1}.
\]

It can be seen that

\[
\sigma = \sqrt{\log(\tau^2 + 1)},
\]

\[
\mu = \log \left( \frac{\theta}{\tau^2 + 1} \right),
\]

\[
\log(\text{median}(X)) = \mu,
\]

\[
\log(E(X)) = \mu + \sigma^2/2.
\]
Note that the mean of $\log(X)$ corresponds to $\log(\text{median}(X))$. Parameters of the lognormal distribution can be estimated by the sample mean and s.d. of the log-transformed data.

```r
mu = 5
sig = 2
ly = rlnorm(n, log(mu), log(sig))
qqnorm(ly)
qqline(ly)
mtext("Lognormal data",side=3,line=.5,cex=.9)
tmdnorm(ly)
mtext("Lognormal data",side=3,line=.5,cex=.9)
#
ly = log(ly)
qqnorm(ly)
qqline(ly)
mtext("Lognormal data, log-transformed",side=3,line=.5,cex=.9)
tmdnorm(ly)
mtext("Lognormal data, log-transformed",side=3,line=.5,cex=.9)
```

The lognormal is a skewed distribution. Another commonly encountered skewed distribution is the gamma distribution which has density

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0, \quad \alpha > 0, \quad \beta > 0.$$  

This density is derived from the gamma integral,

$$\beta^\alpha \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx.$$  

If $X \sim \text{Gamma}(\alpha, \beta)$, then

$$E(X) = \alpha \beta$$

$$\text{Var}(X) = \alpha \beta^2.$$  

Basic method of moments estimation of these parameters can be derived from the mean and variance. If $m, s$ represent the sample mean and s.d., respectively, of a random sample, then the method of moments estimators are given by

$$\hat{\beta} = s^2/m$$

$$\hat{\alpha} = m^2/s^2.$$  

A Q-Q plot of gamma distributed data has a similar bowl shape like is seen with skewed data such as the lognormal, so sometimes it can be difficult to distinguish between them. Q-Q and M-D plots using lognormal and gamma targets can help. The file 

http://www.utdallas.edu/~ammann/stat6390scripts/lnormEX.r
gives an example with the standard lognormal distribution and the gamma with same mean and s.d.

The package **EnvStats** includes some data sets taken from EPA guidance documents. One of them is *EPA.94b.tccb.df* which contains measurements of a pollutant TcCB from soil samples in a reference site and a cleanup site. Since these measurements are all positive and their distributions within each site are skewed, we will compare them to the lognormal and gamma distributions.

```r
library(EnvStats)
attach(EPA.94b.tccb.df)
table(Area)
TcCBref = TcCB[Area=="Reference"]
TcCBcleanup = TcCB[Area=="Cleanup"]
source("http://www.utdallas.edu/~ammann/stat6390scripts/tmdplot.r")
par(mfrow=c(2,2))
qqnorm(TcCBref)
qqline(TcCBref)
mtext("Reference Area",line=.25,cex=.8)
tmdnorm(TcCBref)
mtext("Reference Area",line=.25,cex=.8)
#
qqnorm(TcCBcleanup)
qqline(TcCBcleanup)
mtext("Cleanup Area",line=.25,cex=.8)
tmdnorm(TcCBcleanup)
mtext("Cleanup Area",line=.25,cex=.8)
###
lnxy = lnormPlot(TcCBref)
mtext("Reference area",line=.25,cex=.8)
title(sub=paste("Estimated parameters: meanlog = ",round(lnxy$meanlog,2),", sdlog = ",round(lnxy$sdlog,2),sep=""))
gxy = gammaPlot(TcCBref)
mtext("Reference area",line=.25,cex=.8)
title(sub=paste("Estimated parameters: shape = ",round(gxy$shape,2),", scale = ",round(gxy$scale,2),sep=""))
lnormPlot(TcCBref,plot.type="T")
gammaPlot(TcCBref,plot.type="T")
mtext("Reference area",line=.25,cex=.8)
#
lnxyc = lnormPlot(TcCBcleanup)
mtext("Cleanup area",line=.25,cex=.8)
title(sub=paste("Estimated parameters: meanlog = ",round(lnxyc$meanlog,2),", sdlog = ",round(lnxyc$sdlog,2),sep=""))
gxy = gammaPlot(TcCBcleanup)
mtext("Cleanup area",line=.25,cex=.8)
```

12
These plots show that both lognormal and gamma distributions are reasonable models for the reference area, but the gamma distribution gives a better model for the cleanup area. However, note that the estimated shape parameter for the gamma fit is very small. This implies that the cleanup data has mostly very small values with a relatively few larger values.

Now suppose the data represents species counts at some location. A common model for count data is the Poisson distribution. One important property of the Poisson distribution is that it has dispersion equal to 1, where dispersion is the variance divided by the mean. However, if count data is over-dispersed, then the Poisson distribution would not be appropriate. In such cases the negative binomial distribution may be used instead since its dispersion is not 1. The standard definition of the negative binomial is the number of failures before the r-th success, where r is a positive integer. The negative binomial has probability mass function,

\[ P(N = k) = \binom{k + r - 1}{k} p^r (1 - p)^k, \quad k \geq 0. \]

Then \( E(N) = r(1-p)/p \), \( \text{Var}(N) = r(1-p)/p^2 \), and so its dispersion is \( 1/p \). This shows that the negative binomial is over-dispersed. A more general definition of the negative binomial is the distribution of a Poisson r.v. with a random mean that has a gamma distribution. That is,

\[ P(N = k | \Lambda = x) = \frac{x^k}{k!} e^{-x}, \quad k \geq 0, \]

where the density of \( \Lambda \) is

\[ f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta}, \quad x > 0, \quad \alpha > 0, \quad \beta > 0. \]

Then the marginal distribution of \( N \) is

\[
P(N = k) = \frac{1}{k! \beta^\alpha \Gamma(\alpha)} \int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{k+\alpha-1} e^{-x/(1+\beta)/\beta} dx
\]

\[
= \frac{1}{k! \beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{k+\alpha-1} e^{-x(1+\beta)/\beta} dx
\]

\[
= \frac{1}{k! \beta^\alpha \Gamma(\alpha)} \frac{\beta^{k+\alpha} \Gamma(k + \alpha)}{(1 + \beta)^{k+\alpha}}
\]

\[
= \frac{\Gamma(k + \alpha)}{k! \Gamma(\alpha)} \frac{\beta^\alpha (1 + \beta)^{k+\alpha}}{\beta^k (1 - p)^k},
\]
where
\[ p = \frac{1}{1+\beta}. \]

Note that if \( \alpha = r \), then this corresponds to the probability mass function for the negative binomial given above since in that case
\[ \frac{\Gamma(k+r)}{k!\Gamma(r)} = \binom{k+r-1}{k}. \]

The R function \texttt{rnbinom} generates random samples based on this representation of the negative binomial. It has arguments \texttt{size} and \texttt{prob} corresponding to \( \alpha \) and \( p \), respectively. That parameterization is equivalent to
\begin{align*}
\alpha &= \text{size} \\
\beta &= \frac{\mu}{\text{size}} \\
p &= \frac{\text{size}}{\mu + \text{size}}.
\end{align*}

Let \( E(N) = \mu = \alpha(1-p)/p. \) Then
\[ \text{Var}(N) = \frac{\mu}{p}. \]

If \( m, s \) represent the sample mean and s.d. from a random sample of negative binomial observations, then method of moments estimators of \( p, \alpha \) can be derived. Let
\[ d = s^2/m. \]

Then
\begin{align*}
\hat{p} &= 1/d \\
\hat{\alpha} &= m\hat{p}/(1-\hat{p}).
\end{align*}

In addition to these diagnostic plots, the \textit{EnvStats} package includes a function \texttt{gofTest} to perform goodness of fit tests, by default the Shapiro-Wilk test for continuous distributions. Discrete distributions can be tested by a chi-square GOF test, but that test requires additional arguments that make this less useful. In practice one should not rely solely on a GOF test to check assumptions.

The file
\texttt{http://www.utdallas.edu/~ammann/stat6390scripts/negbinEX.r}
gives an example with \( \mu = 45, \text{size}=5 \) which corresponds to \( p = .1 \) for the negative binomial. This script also examines the \textit{quine} data frame from the \textbf{MASS} library.
Estimation

Many questions in environmental studies involve estimation of population characteristics. Problems associated with these questions include identification and verification of appropriate probability models for parameter estimation or application of appropriate nonparametric methods if parametric models are not warranted; determination of sample sizes; precision of estimates; dealing with observations that were below detection thresholds (censored data). We will discuss estimation of proportions and quantiles, location (mean and median), and s.d.’s. Other characteristics such as regression and ANOVA parameters will be discussed later in the course.

Estimation of a population proportion

. Here we are interested in estimating the proportion \( \theta \) of a population that is contained in one particular level of a factor (categorical) variable using a random sample. This problem also provides one method for estimation of quantiles. The basic theory for this problem is the central limit theorem for the binomial distribution. If \( X_1, \ldots, X_n \) are i.i.d. Bernoulli r.v.’s with success probability \( \theta \), then

\[
\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} \xrightarrow{d} N(0,1),
\]

where \( \hat{\theta} \) is the sample proportion,

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i,
\]

The problem with this CLT is that the s.d. is a function of \( \theta \). The Law of Large Numbers implies that

\[
\sqrt{\hat{\theta}(1-\hat{\theta})} \rightarrow \sqrt{\theta(1-\theta)} \text{ wp 1.}
\]

Slutsky’s Theorem then implies that

\[
\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \xrightarrow{d} N(0,1)
\]

We can use this result to obtain a large sample \( 1 - \alpha \) confidence interval for \( \theta \),

\[
\hat{\theta} \pm z\sqrt{\hat{\theta}(1-\hat{\theta})/n},
\]

where \( z \) is the appropriate quantile from the standard normal distribution, \( z_{1-\alpha/2} \).

The performance of this confidence interval degrades for \( \theta \) near 0 or 1. A more accurate approximation can be obtained by solving the quadratic equation

\[
\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} = z.
\]
This was first proposed by Wilson in 1927 and gives the confidence interval
\[
\frac{1}{1 + \frac{z^2}{n}} \left[ \left( \hat{\theta} + \frac{z^2}{2n} \right) \pm z \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n} + \frac{z^2}{4n^2}} \right].
\]

Another confidence interval uses the binomial distribution directly rather than an approximation. For this interval we must find
\[
\begin{align*}
\theta_l &= \max(\theta : P(N_{n,\theta} \leq k) > \alpha/2) \\
\theta_u &= \min(\theta : P(N_{n,\theta} \geq k) > \alpha/2),
\end{align*}
\]
where \( k \) is the number of successes and
\[N_{n,\theta} \sim \text{Binomial}(n, \theta)\].

Although this is straightforward in \( \mathbf{R} \) we also can use the relationship between the Beta distribution and the binomial,
\[P(N_{n,\theta} \geq k) = P(X_{k,n-k-1} \leq \theta),\]
where
\[X_{\alpha,\beta} \sim \text{Beta}(\alpha, \beta)\].

This gives
\[
\begin{align*}
\theta_l &= \text{qbeta}(\alpha/2; k, n-k+1) \\
\theta_u &= \text{qbeta}(1-\alpha/2; k+1, n-k),
\end{align*}
\]
where \text{qbeta} is the quantile function for the Beta distribution. This is referred to as the Clopper-Pearson confidence interval for a population proportion. It can be expressed in \( \mathbf{R} \) as follows. Let \( n, k \) denote the number of trials and number of successes, respectively. Then a \( 1 - \alpha \) C-P confidence interval for \( \theta \) is given by
\[
\begin{align*}
\text{cp.lower} &= \text{qbeta}(\alpha/2, k, n-k+1) \\
\text{cp.upper} &= \text{qbeta}(1-\alpha/2, k+1, n-k)
\end{align*}
\]

We can use the simulation capabilities of \( \mathbf{R} \) to compare these three confidence intervals in terms of their coverage probabilities and widths.

\textbf{Example.} Suppose the sample size is 100 and the observed number of successes is 6, so the sample proportion is 0.06. The following \( \mathbf{R} \) code obtains 95% confidence intervals using these three methods.
n = 400
k = 24
a = .05
phat = k/n
z = qnorm(1-a/2)
shat = phat*(1-phat)/n
p.conf = matrix(0,3,2)
dimnames(p.conf) = list(c("CLT","Wilson","C-P"),c("Lower","Upper"))
p.conf["CLT",] = phat + c(-1,1)*z*sqrt(shat)
b = z^2/n
p.conf["Wilson",] = (phat + b/2 + c(-1,1)*z*sqrt(shat + b/(4*n)))/(1 + b)
p.conf["C-P",] = c(qbeta(a/2, k, n-k+1), qbeta(1-a/2, k+1, n-k))
print(p.conf)
# now generate 4000 samples and compare coverage and width
n = 400
N = 4000
p = .06
a = .05
z = qnorm(1-a/2)
b = z^2/n
X = rbinom(N,n,p)
phat = X/n
shat = phat*(1-phat)/n
CLT.ci = Wilson.ci = CP.ci = matrix(0,2,N)
CLT.ci[1,] = phat - z*sqrt(shat)
CLT.ci[2,] = phat + z*sqrt(shat)
Wilson.ci[1,] = (phat + b/2 - z*sqrt(shat + b/(4*n)))/(1 + b)
Wilson.ci[2,] = (phat + b/2 + z*sqrt(shat + b/(4*n)))/(1 + b)
CP.ci[1,] = qbeta(a/2, X, n-X+1)
CP.ci[2,] = qbeta(1-a/2, X+1, n-X)
CLT.cov = mean(CLT.ci[1,] <= p & p <= CLT.ci[2,])
CLT.w = mean(CLT.ci[2,] - CLT.ci[1,])
Wilson.w = mean(Wilson.ci[2,] - Wilson.ci[1,])
CP.cov = mean(CP.ci[1,] <= p & p <= CP.ci[2,])
CP.w = mean(CP.ci[2,] - CP.ci[1,])

Sample size determination for estimating a population proportion. The half-width of a confidence interval gives the precision of the estimate and is a function of the sample size and level of confidence for a particular method. When possible it is best to plan ahead and determine what sample size would be required to attain specific goals for precision and level of confidence. We can get two estimates of sample size from the CLT confidence
interval,
\[ \hat{\theta} \pm z\sqrt{\theta(1-\theta)/n}. \]

A conservative sample size can be obtained using the inequality
\[ \sqrt{\theta(1-\theta)} \leq 0.5, \ 0 \leq \theta \leq 1. \]

This implies that the half-width is bounded by
\[ \frac{z}{2\sqrt{n}}. \]

Let \( e \) denote the desired precision for the confidence interval and let \( z \) denote the quantile from the standard normal distribution corresponding to the desired level of confidence. Then the required sample size would be
\[ n = \left( \frac{z}{2e} \right)^2. \]

This conservative sample size gives reasonable values when \( \theta \) is moderate, but is too conservative when \( \theta \) is close to 0 or 1. In those cases it is better to use a bound on \( \theta \) if available, or invest in a small preliminary sample to obtain a rough estimate of \( \theta \). If \( \theta_0 \) denotes this prior bound (or preliminary estimate), then we have
\[ \sqrt{\theta(1-\theta)} \leq (\approx)\sqrt{\theta_0(1-\theta_0)} \]

The corresponding sample size is then
\[ n = \left( \frac{z}{e} \right)^2 \theta_0(1-\theta_0). \]

For example, if we wish to estimate a population proportion to within 0.02 with 98% confidence with no prior bound on \( \theta \), then the required sample size would be
\[ n = \left( \frac{qnorm(.99)}{2(.02)} \right)^2 = 3382. \]

If we expect this proportion to be no more than 0.1, then the corresponding sample size would be
\[ n = \left( \frac{qnorm(.99)}{.02} \right)^2 (.1)(.9) = 1218. \]

Of course if the population is much larger than we expect, then this sample size will be too small to attain the stated goals.
Estimation of a population mean

The first case considered is for data that is approximately normally distributed. It is assumed that a quantile-quantile plot of the data has been examined and shows this assumption to be reasonable. Since the population variance never is known in actual applications, it must be treated as an unknown nuisance parameter. A confidence interval for this case is motivated by the linearity property of the normal distribution. If $X_1, \cdots, X_n$ are i.i.d. $N(\mu, \sigma)$ r.v.’s, then

$$
\frac{\bar{X} - \mu}{\sigma/n} \sim N(0, 1).
$$

We cannot construct a confidence interval directly from this result since the s.d. is unknown. William Gosset solved this problem by replacing $\sigma$ with the sample s.d. and then deriving the distribution of

$$
t = \frac{\bar{X} - \mu}{s/n},
$$

where $s^2$ is the sample variance,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.
$$

This distribution is referred to as the $t$-distribution with $n-1$ degrees of freedom. The proof of this result is based on the representation of $t$ as the ratio

$$
\frac{Z}{\sqrt{Q/d}},
$$

where $Z, Q$ are independent r.v.’s, $Z \sim N(0, 1)$ and $Q$ has a chi-square distribution with $d$ degrees of freedom. The t-distribution is symmetric about 0, has infinite mean for 1 df (this is the Cauchy distribution), has finite mean but infinite variance for 2 df, and has finite variance for $df \geq 3$. It has heavier tails than the standard normal distribution. The law of large numbers and Slutsky’s Theorem imply that

$$t_{n-1} \xrightarrow{d} N(0, 1).$$

The Central Limit Theorem implies that this limit also holds for any set of i.i.d. r.v.’s with finite variance.

A confidence interval for $\mu$ is

$$\bar{X} \pm ts/\sqrt{n},
$$

where $t$ is the appropriate quantile from the $t_{n-1}$ distribution. Note that this also can be used as a large sample confidence interval for the mean, but then we must address the question
of how large is large for the coverage probability to be approximately the same as the level of confidence.

Determination of an appropriate sample size for this confidence interval requires a prior bound or preliminary estimate of the population standard deviation. Denote this by $\sigma_0$. An initial estimate for the sample size can be obtained by using $z$ instead of $t$ in the confidence interval, setting the resulting precision equal to the required half-width, and then solving for $n$. This gives

$$n_0 = \left( \frac{z\sigma_0}{e} \right)^2,$$

where $e$ is the goal for precision. Since $t_{n-1} \geq z$ for the same level of confidence, then $n_0$ will be smaller than the required sample size. So we can iterate the following steps,

$$n_1 = (qt(1 - \alpha/2, n_0-1) \cdot s_0/e)^2$$
$$n_0 = \text{ceiling}(n_1)$$

until $n_1 \leq n_0$. Note that the 2nd step rounds up the intermediate sample size. The following defines a function in R that implements this iterative process and returns the final sample size.

```r
normN = function(e, s0, alpha=.05, maxN=5000) {
  prob = 1 - alpha/2
  n0 = (qnorm(prob) * s0/e)^2
  n0 = ceiling(n0)
  n1 = maxN - 1
  checkN = TRUE #ensure loop iterates at least once
  while(checkN & n1 < maxN) {
    n1 = (qt(prob, n0-1) * s0/e)^2
    n1 = ceiling(n1)
    if(n1 <= n0) {
      checkN = FALSE
    }
    n0 = ceiling(n1)
  }
  return(n0)
}
```

The `maxN` argument is a maximum for the sample size to prevent this function from iterating indefinitely under some degenerate conditions.

A weakness of the mean and s.d. is their sensitivity to outliers. This can be managed by using robust alternatives to estimate location and scale. The median is an example of a location estimator that is not sensitive to outliers, but its efficiency relative to the mean is fairly low, so other estimators have been developed. Likewise, the MAD is a robust estimator of scale also with low efficiency. The `MASS` package in R includes a function `hubers` which
provides an alternative to these estimators that is robust and more efficient than the median and MAD. A more robust confidence interval for the mean can be obtained by replacing the sample mean and sample s.d. by the respective estimates returned by \textit{hubers}.

\begin{verbatim}
library(MASS)
alpha = .05
n = 50
tval = qt(1-alpha/2,n-1)
x = rnorm(n)
hx = hubers(x)
m = mean(x)
s = sd(x)
mrob = hx$mu
srob = hx$s
ci = m + c(-1,1)*tval*s/sqrt(n-1)
ci.rob = mrob + c(-1,1)*tval*srob/sqrt(n-1)
# replace 2 values with outliers
xout = c(10,15)
x1 = x
x1[seq(length(xout))] = xout
hx1 = hubers(x1)
m1 = mean(x1)
s1 = sd(x1)
mrob1 = hx1$mu
srob1 = hx1$s
ci1 = m1 + c(-1,1)*tval*s1/sqrt(n-1)
ci1.rob = mrob1 + c(-1,1)*tval*srob1/sqrt(n-1)
\end{verbatim}

Note that the outliers have inflated the sample s.d. which makes the standard confidence interval much wider than it would have been without those outliers.

\textbf{Estimation of a population variance}

. In this section it is again assumed that the assumption of normality is reasonable. The basic theory for this case is given by the following theorem. If \(X_1, \cdots, X_n\) are i.i.d. \(N(\mu, \sigma)\), then \(\bar{X}\) and \(s^2\) are independent r.v.’s and

\[
\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.
\]

We can use this result to derive a confidence interval for \(\sigma^2\). Let \(C_L, C_U\) satisfy

\[
C_L = qchisq(\alpha/2, n-1) \\
C_U = qchisq(1-\alpha/2, n-1).
\]
Then

\[
\alpha/2 = P \left[ \frac{(n - 1)s^2}{\sigma^2} \leq C_L \right] \\
= P \left[ \frac{(n - 1)s^2}{C_L} \leq \sigma^2 \right],
\]

\[
\alpha/2 = P \left[ \frac{(n - 1)s^2}{\sigma^2} > C_U \right] \\
= P \left[ \sigma^2 < \frac{(n - 1)s^2}{C_U} \right],
\]

and so a \(1 - \alpha\) confidence interval for \(\sigma^2\) is

\[
P \left[ \frac{(n - 1)s^2}{C_U} \leq \sigma^2 \leq \frac{(n - 1)s^2}{C_L} \right] = 1 - \alpha.
\]

This interval is referred to as an equal probability interval, but it is not necessarily the interval with minimum width. Derivation of a minimum width confidence interval for the variance requires numerical techniques that often produce only modest reductions in width and so is rarely worth it.

There are situations in which a one-sided confidence interval is required, typically in the form of an upper bound on the variance or s.d. This can be obtained by setting

\[
C_L = qchisq(\alpha, n - 1)
\]

to give

\[
P \left[ \sigma^2 \leq \frac{(n - 1)s^2}{C_L} \right] = 1 - \alpha.
\]

Corresponding confidence intervals for the s.d. are obtained by taking the square root of the intervals for the variance.

**Estimation of parameters for other distributions**

Two other distributions often encountered in environmental data, the lognormal and the gamma will be discussed in this section.

**Lognormal distribution.** Suppose \(Y_1, \cdots, Y_n\) are i.i.d. lognormal with parameters \(\mu, \sigma\), and let \(X_i = \log(Y_i)\). Since the log-transformed data has a normal distribution with mean and s.d. \(\mu, \sigma\), we can use the confidence intervals given above for those parameters. This gives

\[
P(\bar{X} - t_{n-1}s_x/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1}s_x/\sqrt{n}) = 1 - \alpha.
\]

However, if we transform this interval back to the original scale, we have

\[
P(\exp\{\bar{X} - t_{n-1}s_x/\sqrt{n}\} \leq \exp\{\mu\} \leq \exp(\{\bar{X} + t_{n-1}s_x/\sqrt{n}\}) = 1 - \alpha.
\]
Since $\exp\{\mu\} = \text{median}(X)$, then this is a confidence interval for the median of the population, not the mean. If a confidence interval for the mean is required, then we must use a different method.

Let

$$\theta = e^\beta,$$

where

$$\beta = \mu + \sigma^2/2.$$

Then the minimum variance unbiased estimator (MVUE) of $\beta$ is

$$\hat{\beta}_{mvue} = \bar{X} + \frac{s_x^2}{2}$$

and the maximum likelihood estimator (MLE) of $\beta$ is

$$\hat{\beta}_{mle} = \bar{X} + \hat{\sigma}^2_{mle}/2,$$

where

$$\hat{\sigma}^2_{mle} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n-1}{n} s_x^2.$$

The corresponding estimators for $\theta$ are

$$\hat{\theta}_{mvue} = \exp(\hat{\beta}_{mvue})$$
$$\hat{\theta}_{mle} = \exp(\hat{\beta}_{mle}).$$

This shows that estimation of the mean of a lognormal distribution requires simultaneous estimation of the mean and variance of the log-transformed variables. Several authors have proposed methods to obtain confidence intervals for the mean.

A confidence interval for $\beta$ proposed by Land (1971) is

$$\left[ \hat{\beta}_{mvue} + s_x \frac{C_{\alpha/2}}{\sqrt{n-1}}, \hat{\beta}_{mvue} + s_x \frac{C_{1-\alpha/2}}{\sqrt{n-1}} \right]$$

where $C_p$ are critical values derived by Land. The method proposed by Zou, et al (2009) is based on normal approximations to these statistics. Their confidence interval has lower and upper limits given by

$$LL = \hat{\theta}_{mle} \exp\left\{ - \left[ \frac{z^2_{1-\alpha/2} s_x^2}{n} + \frac{s_x^2}{2} - \frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}} \right]^2 \right\}$$

$$UL = \hat{\theta}_{mle} \exp\left\{ - \left[ \frac{z^2_{1-\alpha/2} s_x^2}{n} + \frac{s_x^2}{2} + \frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}} \right]^2 \right\}$$
Cox constructed confidence limits based on the normal approximations for \( \beta \). These confidence intervals are given by

\[
\left[ \exp(\hat{\beta} - t_{1-\alpha/2,n-1}\hat{\sigma}_\beta), \exp(\hat{\beta} + t_{1-\alpha/2,n-1}\hat{\sigma}_\beta) \right]
\]

where

\[
\hat{\sigma}_\beta^2 = \frac{s^2}{n} + \frac{s^4}{2(n+1)}.
\]

Details and additional methods can be found in the help page for function `elnormAlt` in package `EnvStats`.

There are several problems with these confidence intervals. Some data may produce lower confidence limits for \( \theta \) that are negative, but the main problem is due to the sensitivity of these estimators to outlying observations. Since the lognormal distribution can have very heavy tails, some approximations used to derive these confidence limits may have very high upper limits. To address the lack of robustness, Serfling (2002) introduced a family of estimators for \( \theta \) based on generalized-median estimation. These estimators are generated from a kernel function

\[
h(x_1, \cdots, x_k)
\]

which is a median unbiased estimator of a parameter \( \theta \). That is, the median of the distribution of \( h(X_1, \cdots, X_m) \) equals \( \theta \). A generalized median estimator of \( \theta \) is the median of \( h \) over all subsets of size \( m \) of the data. In the case of the lognormal distribution we need estimators for the mean and variance, so we can use the MLE estimators for those kernel functions. The kernel size, \( m \), provides a trade-off between asymptotic relative efficiency (ARE) and break-down point (BP). Higher value for \( m \) increases ARE but decreases BP. An R function to obtain these estimates is defined in

http://www.utdallas.edu/~ammann/stat6390scripts/GMest.r

**GMlnormEst**: Generalized Median estimation of lognormal parameters

**Usage**:

```r
GMlnormEst(y, k=5, m=NULL, alpha=.05, Nsamp=10000)
```

**Arguments**

- `y` vector of data assumed to be lognormally distributed
- `k` size of kernel for the mean, default is 5, minimum value is 2
- `m` size of kernel for variance, default is same as `k`, minimum value is 2
- `alpha` level of confidence interval is 1-`alpha`, default is 0.05
- `N samp` number of samples for kernel functions, default is 10000

**Value**: a list with components

- `ci` vector containing lower and upper 1-`alpha` confidence limits for the mean of a lognormal distribution
- `mu.hat` GM-estimate of the mean of the log-transformed data
- `sig2.hat` GM-estimate of the variance of the log-transformed data
- `theta.hat` GM-estimate of the mean, \( \theta = \exp(\mu + \sigma^2/2) \)
- `cv.hat` GM-estimate of the CV, \( CV = \sqrt{\exp(\sigma^2) - 1} \)
Since resampling is used for the kernel functions instead of all possible subsamples, estimates will vary slightly if the function is applied to the same data more than once.

**Reference**
http://www.utdallas.edu/~serling/papers/naaj02.pdf

**Examples**

```r
source("http://www.utdallas.edu/~ammann/stat6390scripts/GMest.r")
require(EnvStats)
n = 100
mu = 4
sig = 1
theta = exp(mu + sig^2/2)
y = rlnorm(n,mu,sig)
cat(paste("theta =",round(theta,3)),"\n")
elnormAlt(y, ci=TRUE)
GMlnormEst(y)
# add some outliers
ymax = max(y)
cy = c(y,c(5,10,15)*ymax)
elnormAlt(cy, ci=TRUE)
GMlnormEst(cy, k=3)
# repeat with larger sample size
n=250
y = rlnorm(n,mu,sig)
ymax = max(y)
cy = c(y,c(5,10,15)*ymax)
elnormAlt(cy, ci=TRUE)
GMlnormEst(cy, m=3)
```

**Gamma distribution.**

As we saw earlier, simple method of moments estimators exist for the parameters of the gamma distribution,

\[ \hat{\beta} = \frac{s^2}{m} \]
\[ \hat{\alpha} = \frac{m^2}{s^2}. \]

The likelihood equations for MLE give

\[ \beta_{mle} = \frac{1}{\hat{\alpha}} \bar{X}, \]

but the likelihood equation for \( \alpha \) does not have a closed form solution. Therefore numerical optimization algorithms must be used. Confidence intervals can be obtained using a normal approximation for a fractional power of the original data. Functions *egamma* and *egammaAlt*
in package *EnvStats* implement these methods to obtain estimates and confidence intervals for \( \alpha, \beta \) (*egamma*) and the mean, CV (*egammaAlt*). The value returned by these functions is a list that includes the normal transformation power \( p \) used to obtain the confidence intervals. A diagnostic *qqnorm* plot of \( x^p \) can show the appropriateness of this approximation.

**Estimation of quantiles**

Some regulations involve restrictions on quantiles, for example a requirement that concentrations of a chemical can exceed some threshold no more than 5% of the time. Other situations may exist in which parametric models are not appropriate and we must use nonparametric measures such as the median. One problem with quantile estimation is that larger sample sizes are required, especially for more extreme quantiles. We will consider this problem for normal, lognormal, and gamma distributions, and for nonparametric estimates.

Suppose \( X_1, \ldots, X_n \) denote i.i.d. r.v.’s with density \( f \) and let \( X_{(1)}, \ldots, X_{(n)} \) denote their ordered values. For \( 0 < p < 1 \) let \( q_p \) denote the \( p \)th quantile of \( f \),

\[
p = \int_{-\infty}^{q_p} f(x)dx.
\]

The sample quantile \( Q_n(p) \) is given by

\[
Q_n(p) = (1 - \gamma)X_{(\lfloor np \rfloor)} + \gamma X_{(\lceil np \rceil)},
\]

where

\[
\lfloor np \rfloor = \text{floor}(np), \quad \lceil np \rceil = \text{ceiling}(np),
\]

and \( 0 \leq \gamma \leq 1 \). Note that this definition results in a function that is discontinuous in \( p \). This definition can be modified to produce quantiles that are continuous in \( p \) by defining \( \gamma \) to be a linear interpolator between the ordered values.

A CLT exists for sample quantiles,

\[
\frac{\sqrt{n}(Q_n(p) - q_p)}{\sigma_p} \xrightarrow{d} N(0, 1),
\]

where

\[
\sigma_p = \frac{p(1-p)}{f(q_p)}.
\]

Therefore, a large sample confidence interval for \( q_p \) is

\[
Q_n(p) \pm z_{1-\alpha/2} \frac{p(1-p)}{f(q_p)} \sqrt{p(1-p)/n}.
\]

One difficulty here is obtaining an estimate of \( f(q_p) \). The problem is straightforward if \( p = .5 \) and the data are normally distributed. In that case,

\[
f(q_p) = \frac{1}{\sqrt{2\pi\sigma}}.
\]
and so we could use
\[ \hat{f}(q_p) = \frac{1}{\sqrt{2\pi s}} \]
and use t-values rather than z-values,
\[ Q_n(p) \pm t_{1-\alpha/2,n-1} \sqrt{\frac{2\pi s}{n}} \sqrt{\frac{p(1-p)}{n}}. \]

Otherwise, estimation of the density function at the quantile is difficult and other methods have been proposed.

One alternative is based on the equivalence between a confidence interval for a quantile and a tolerance interval for a population. Suppose for example that we wish to find \( C_u \) such that
\[ P(X_{(\lfloor np \rfloor)} \leq C_u) = 1 - \alpha. \]
In this case \((-\infty, C_u)\) is referred to as an upper tolerance interval with coverage probability \( p \) and confidence \( 1 - \alpha \). Note that
\[ X_{(\lfloor np \rfloor)} \]
is the sample \( p^{th} \) quantile and so we are \( 100(1 - \alpha)\% \) confident that \( 100p\% \) of the sample will be no more than \( C_u \). It can be shown that if the data follows a normal distribution then
\[ C_u = \bar{X} + ks, \]
where \( s \) is the sample standard deviation and \( k \) is a function of the sample size given below. Let
\[ \delta = z_p\sqrt{n}, \quad t = qt(1 - \alpha, n - 1, \delta), \]
then \( k = t/\sqrt{n} \). The value of \( k \) for two-sided tolerance intervals is similar but involves the chi-square distribution instead of the t-distribution. The EnvStats library contains a function \texttt{eqnorm} that uses this approach to obtain confidence intervals for percentiles of normal populations. Obviously the assumption of normality for the data must be checked before using this function. These results can be extended immediately to the lognormal distribution since the logarithm of a quantile from the lognormal distribution is the same as the quantile from the normal distribution. The function \texttt{eqlnorm} provides quantile confidence intervals for the lognormal in this way.

**Quantile estimation for gamma distributions.** One approach for gamma distributed data is to follow the same approach used for estimation of parameters by a power transformation to make the data approximately normal. We can use quantile estimates for the normal as above and then transform back to the original scale since the power transformation is monotone. This is the approach taken by the \texttt{eqgamma} function in EnvStats.

**Examples.**
1. Simulate normally distributed data and obtain upper 95% confidence limits for the 0.9 quantile.
source("http://www.utdallas.edu/~ammann/stat6390scripts/tmdplot.r")
require(EnvStats)

n = 80
mu = 10
sig = 2
p = .9
alpha=.05
cat("Actual quantile for normal distribution\n")
qn = qnorm(p,mu,sig)
cat(paste("Actual", p, "quantile =",round(qn,4)),"\n")
X = rnorm(n,mu,sig) # simulate sample from normal distribution
# check assumption of normality
qqnorm(X)
qqline(X)
# obtain upper confidence interval for .9 quantile
qXe = eqnorm(X, p=p, ci=TRUE, ci.type="upper", conf.level=1-alpha)
cat(paste("Upper confidence limit for",p, "quantile =",round(qXe$interval$limits[2],4)),"\n")
# Perform simulation to generate 4000 samples from this normal distribution
# Then find the proportion of samples for which the actual .9 quantile <= upper limits
N = 4000
X = matrix(rnorm(n*N,mu,sig),n,N) # simeulated samples are columns of X
# Define a function to use with apply() that just returns upper conf limit for a sample
upL = function(x,p,alpha) {
  xe = eqnorm(x, p=p, ci=TRUE, ci.type="upper", conf.level=1-alpha)
  return(xe$interval$limits[2])
}
Xu = apply(X,2,upL,p,alpha)
Xp = mean(qn <= Xu)
cat(paste("Proportion of upper confidence limits greater than ",p," quantile: ",round(Xp,3),sep=""))

2. Obtain a 90% confidence interval for the median of TcCB in the reference area of EPA.94b.tccb.df Assume this data has a lognormal distribution.

source("http://www.utdallas.edu/~ammann/stat6390scripts/tmdplot.r")
require(EnvStats)
alpha = .1
p = .5
attach(EPA.94b.tccb.df)
TcCBref = TcCB[Area=="Reference"]
lnxy = lnormPlot(TcCBref)
mtext("Reference area",line=.25,cex=.8)
title(sub=paste("Estimated parameters: meanlog = ",round(lnxy$meanlog,2),", sdlog = ",round(lnxy$sdlog,2),sep=""))

lnormPlot(TcCBref,plot.type="T")
mtex("Reference area",line=.25,cex=.8)

# obtain confidence interval
qXe = eqlnorm(TcCBref, p=p, ci=TRUE, conf.level=1-alpha)
cat(paste("Sample median:",round(qXe$quantiles,3)),"\n")
cat(paste(round(100*(1-alpha)),"% confidence interval for median: ",
          paste(round(qXe$interval$limits,4),collapse=" ",sep=""),"\n")

3. Simulate samples from a gamma distribution and obtain upper confidence limits for the lower quartile. Compare with upper confidence limits using quantile estimate for lognormal.

require(EnvStats)

n = 80 # sample size
N = 4000 # number of samples
p = .25 # lower quartile
alpha = .05
shp = .5 # shape parameter for gamma
sc = 2 # scale parameter for gamma
qn = qgamma(p, shape=shp, scale=sc)
X = matrix(rgamma(n*N, shape=shp, scale=sc),n,N) # matrix of simulated data

# define functions for eqgamma and gamtol.int that return upper confidence limits
upL = function(x,p,alpha) {
  xe = eqgamma(x, p=p, ci=TRUE, ci.type="upper", conf.level=1-alpha)
  return(xe$interval$limits[2])
}
upLln = function(x,p,alpha) {
  xe = eqlnorm(x, p=p, ci=TRUE, ci.type="upper", conf.level=1-alpha)
  return(xe$interval$limits[2])
}
XupL = apply(X,2,upL,p,alpha)
XupLln = apply(X,2,upLln,p,alpha)
qpL = mean(qn <= XupL)
qpLln = mean(qn <= XupLln)
qp = c(qpL,qpLln)
names(qp) = c("eqgamma","eqlnorm")
cat(paste("Proportion of upper confidence limits greater than ",p," quantile:","\n")
print(qp)

# repeat using lognormal simulated data
lmu = 3
lsig = 1
qln = qlnorm(p, lmu, lsig)
Xln = matrix(rlnorm(n*N, lmu, lsig),n,N) # matrix of simulated data
XlnupL = apply(Xln,2,upL,p,alpha)
\[
X_{\text{up}L\text{ln}} = \text{apply}(X_{1,2,\text{up}L\text{ln}},p,\text{alpha})
\]

\[
q_{LP} = \text{mean}(q_{LN} \leq X_{\text{up}L})
\]

\[
q_{LP\text{ln}} = \text{mean}(q_{LN} \leq X_{\text{up}L\text{ln}})
\]

\[
q_{LP} = c(q_{LP},q_{LP\text{ln}})
\]

\[
\text{names}(q_{LP}) = c(\text{"eqgamma"},\text{"eqlnorm"})
\]

\[
\text{cat}(\text{paste}(\text{"Proportion of upper confidence limits greater than ",p," quantile:"},\text{"\n")}
\]

\[
\text{print}(q_{LP})
\]

**Nonparametric quantile estimators.**

Some situations cannot be represented very well by any parametric model. In such situations we can estimate quantiles based solely on the order statistics. Let \(X_{(1)}, \ldots, X_{(n)}\) denote the ordered values of i.i.d. r.v.'s \(X_{1}, \ldots, X_{n}\) with d.f. \(F\) and let \(q_{p}\) denote the \(p^{th}\) quantile of \(F\). Then

\[
Y_{i} = I(X_{i} \leq q_{p}), \ 1 \leq i \leq n
\]

are i.i.d. \(\text{Bernoulli}(p)\) r.v.'s, and so

\[
P(X_{(r)} < q_{p} < X_{(s)}) = \sum_{i=r}^{s} \binom{n}{i} p^{i}(1-p)^{n-i}.
\]

A \(1 - \alpha\) confidence interval for \(q_{p}\) can be obtained by finding

\[
r = \max\{k : \sum_{i=0}^{k} \binom{n}{i} p^{i}(1-p)^{n-i} \leq \alpha/2,\}
\]

\[
s = \min\{k : \sum_{i=k+1}^{n} \binom{n}{i} p^{i}(1-p)^{n-i} \leq \alpha/2.\}
\]

A function to obtain this confidence interval is contained in the file

\href{http://www.utdallas.edu/~ammann/stat6390scripts/nonpar.ci.r}{hturlhttp://www.utdallas.edu/~ammann/stat6390scripts/nonpar.ci.r}

Note that if one of the limits in a two-sided confidence interval is NA, then we can’t obtain a two-sided interval for the specified \(\alpha\). In that case we could switch to a 1-sided confidence interval.

**Example.** First generate a sample from \(N(100,20)\) and estimate the lower quartile. Then replace 10% of the sample with Gamma r.v.'s with mean 200 and s.d. 100. Compare quantile estimates using normal, gamma, and nonparametric confidence intervals for LQ.

\[
\text{library(EnvStats)}
\]

\[
\text{source("http://www.utdallas.edu/~ammann/stat6390scripts/tmdplot.r")}
\]

\[
\text{source("http://www.utdallas.edu/~ammann/stat6390scripts/nonpar.ci.r")}
\]

\[
n = 100
\]

\[
p = .9
\]

\[
\text{alpha} = .05
\]

\[
\text{mu} = 100
\]
sig = 20
X = rnorm(n, mu, sig)
qn = qnorm(p,mu,sig)
norm.ci = eqnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
npar.ci = nonpar.ci(X, p, alpha)
cat(paste("Actual quantile is:", round(qn,3)),"
")
cat(paste("Confidence interval based on normal dist:",
paste(round(norm.ci$interval$limits,3),collapse="", ")),"
")
cat(paste("Nonparametric confidence interval:",
paste(round(npar.ci,3),collapse="", ")),"
")
# now contaminate 10% with gamma
gmu = 200
gsig = 100
pcontam = .1
a = (gmu/gsig)^2
b = gsig^2/gmu
ng = round(pcontam*n)
Xg = rgamma(ng, shape=a, scale=b)
X[1:ng] = Xg
# obtain quantile for contaminated distribution
N = 1000
X0 = seq(min(X),max(X),length=N)
H = pcontam*pgamma(X0, shape=a, scale=b) + (1-pcontam)*pnorm(X0,mu,sig)
qncontam = X0[max(seq(N)[H <= p])]
# show Q-Q plots
qqnorm(X)
qqline(X)
tmdnorm(X)
lnormPlot(X)
lnormPlot(X, plot.type="T")
gammaPlot(X)
gammaPlot(X, plot.type="T")
norm.ci = eqnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
lnorm.ci = eqlnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
gamma.ci = eqgamma(X, p=p, conf.level=1-alpha, ci=TRUE)
npar.ci = nonpar.ci(X, p, alpha)
cat(paste("Actual quantile for contaminated dist is:", round(qncontam,3)),"
")
cat(paste("Confidence interval based on normal dist:",
paste(round(norm.ci$interval$limits,3),collapse="", ")),"
")
cat(paste("Confidence interval based on lognormal dist:",
paste(round(lnorm.ci$interval$limits,3),collapse="", ")),"
")
cat(paste("Confidence interval based on gamma dist:",

Now repeat this for lognormal and gamma distributions.

n = 100
p = .9
alpha = .05
# lognormal data
lmu = 5
lsig = 2
X = rlnorm(n, lmu, lsig)
qn = qlnorm(p, lmu, lsig)
# show Q-Q plots
qqnorm(X)
qqline(X)
tmdnorm(X)
lnormPlot(X)
lnormPlot(X, plot.type="T")
gammaPlot(X)
gammaPlot(X, plot.type="T")
norm.ci = eqnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
lnorm.ci = eqlnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
gamma.ci = eqgamma(X, p=p, conf.level=1-alpha, ci=TRUE)
npar.ci = nonpar.ci(X, p, alpha)
cat(paste("Actual quantile for lognormal dist is: ", round(qn,3)), "\n")
cat(paste("Confidence interval based on normal dist: ",
     paste(round(norm.ci$interval$limits,3),collapse=" ", ",")", "\n")
cat(paste("Confidence interval based on lognormal dist: ",
     paste(round(lnorm.ci$interval$limits,3),collapse=" ", ",")", "\n")
cat(paste("Confidence interval based on gamma dist: ",
     paste(round(gamma.ci$interval$limits,3),collapse=" ", ",")", "\n")
cat(paste("Nonparametric confidence interval: ",
     paste(round(npar.ci,3),collapse=" ", ",")", "\n")
# gamma data
shp = 2
sc = 5
X = rgamma(n, shape=shp, scale=sc)
qn = qgamma(p, shape=shp, scale=sc)
# show Q-Q plots
qqnorm(X)
qqline(X)
tmdnorm(X)
gammaPlot(X)
gammaPlot(X, plot.type="T")
lnormPlot(X)
lnormPlot(X, plot.type="T")
norm.ci = eqnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
lnorm.ci = eqlnorm(X, p=p, conf.level=1-alpha, ci=TRUE)
gamma.ci = eqgamma(X, p=p, conf.level=1-alpha, ci=TRUE)
npar.ci = nonpar.ci(X, p, alpha)
cat(paste("Actual quantile for gamma dist is:", round(qn,3)),"\n")
cat(paste("Confidence interval based on normal dist:",
       paste(round(norm.ci$interval$limits,3),collapse=" ", ")),"\n")
cat(paste("Confidence interval based on lognormal dist:",
       paste(round(lnorm.ci$interval$limits,3),collapse=" ", ")),"\n")
cat(paste("Confidence interval based on gamma dist:",
       paste(round(gamma.ci$interval$limits,3),collapse=" ", ")),"\n")
cat(paste("Nonparametric confidence interval:",
       paste(round(npar.ci,3),collapse=" ", ")),"\n")

A more complete comparison of these three confidence intervals could be made by simulating a large number of samples from various distributions and comparing the coverage probabilities and widths of the intervals.

Homework and Project Assignments

Homework 1

This assignment uses data set from the R package EnvStats. To access them you must enter the following into R.

library(EnvStats)

Answer the following for each of the data sets listed below.
a. Determine what parametric model is most appropriate. Include any diagnostic plots used to support your answer.
b. Construct a 95% confidence interval for the mean of the data. If there is a robust method available for the confidence interval, report it as well and compare it with the standard method.

Details about these data sets can be accessed by the help function in R.

1. EPA.02d.Ex.4.mg.per.kg.vec
2. Gibbons.et.al.09.Vinyl.Chloride.vec
3. Millard.Deverel.88.df
For this data set, extract observations for which Cu.censored is FALSE and answer the questions separately for Cu from Zone Alluvial.Fan and for Cu from Basin.Trough. Note that you also will need to remove any NA observations for Cu. Then do the same for Zn.

4. Benthic.df
For this data set answer the questions for benthic Index in Stratum 104. In addition to those questions, use the Stratum 104 Index values as a preliminary sample to determine the sample size needed to estimate the mean Index with 95% and with precision of ±0.1.

5. For the Benthic.df data, instead of answering questions a and b, use the entire data set to construct a 95% confidence interval for the proportion of sites that have benthic index equal to 1 (the lowest). Do the same for the proportion of sites that have benthic index equal to 5 (the highest).