HANDOUT #6: NUMERICAL SUMMARIES OF DATA

ESTIMATORS FOR PARAMETRIC FAMILIES

1. **Graphical Estimators of Location-Scale Parameters**
2. **Method of Moment Estimators (MOM)**
3. **Maximum Likelihood Estimators (MLE)**
4. **Pdf Based Estimators of Summary Parameters**
5. **Distribution-Free Summaries**
   (a) **Estimators of Measures of Location**
      i. Sample Mean ($\hat{\mu} = \bar{Y}$)
      ii. Sample Median ($\widetilde{Y} = \widehat{Q}_2$)
      iii. Sample Quartiles ($\widehat{Q}_1, \widehat{Q}_3$)
      iv. Trimmed Mean ($\hat{\mu}_{(\alpha)}$)
   (b) **Estimators of Measures of Level of Dispersion**
      i. Range (R)
      ii. Semi-interquartile Range ($SIQR$)
      iii. Sample Standard Deviation ($\hat{\sigma} = S$)
      iv. Mean Absolute Deviation (MAD)
   (c) **Five Number Summary of Data Set:**

      Minimum, $\hat{Q}(.25)$, Median, $\hat{Q}(.75)$, Maximum
   (d) **Estimators of Shape of PDF**
      i. Sample Skewness: $\hat{\beta}_1$
      ii. Sample Kurtosis: $\hat{\beta}_2$
   (e) **Estimators of Measures of Correlation**
      i. Sample Correlation Coefficient: Pearson and Spearman ($\hat{\rho}$, $\hat{\rho}_{sp}$)
      ii. Sample Autocorrelation Coefficient of Lag k ($\hat{\rho}_k$)

Planning a Comparison of Several Populations/Processes

The researcher must carefully design the study by addressing the following questions:

1. What is the specific population or process that is of interest. Carefully specify the particular conditions and limitations associated with the process or population.

- Environmental conditions in Lab, Differences in Technicians, Differences in Equipment

2. What characteristics of the population/process need to be measured?

- Blood pressure, severity of disease, reduction in pollution after new equipment is installed

3. How much data needs to be collected?

- Based on how much accuracy is needed, how much risk of erroneous conclusions is acceptable, how variability in population

4. How will the data be collected?

- Sampling Design: simple random sample, stratified sample, cluster sampling, observational study, historical data

- Experimental Design: completely randomized design, randomized block design, split plot design

5. Is the data independent or correlated temporally/spatially?

- Collected over time
- multiple measurements in close proximity
- observations over a grid on a potential oil field

6. How will the data be summarized?

   Graphically Numerically

7. What comparisons need to be made?
8. To what degree of accuracy do we need to make the comparisons?

- $\widehat{\mu} \pm \Delta$
- How large a difference in $\widehat{\mu}_1 - \widehat{\mu}_2$ is a practical difference?

SUMMARIES FOR PARAMETRIC FAMILIES

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (or iid observations) from a population/process having pdf $f(y)$ which depends on unknown parameters: $\theta_1, \theta_2, \cdots, \theta_k$. Suppose we want to estimate certain population parameters. There are a number of possible methods for obtaining the estimators of the parameters. We will consider three such approaches. Once the estimators of the pdf's parameters are obtained, the population summary parameters are obtained by replacing the unknown parameters involved in these summaries with their sample estimators.

We will illustrate these ideas with an example:

Suppose $Y_1, Y_2, \ldots, Y_n$ be the times to failure of a random sample of n cell phones which are produced in a newly designed production facility. From historical reliability records, failure times for this type of phone have a Weibull cdf $F(y) = 1 - e^{(y/\alpha)^\gamma}$ but both α and γ are unknown due to the changes in the production process.

The mean and standard deviation of Y are given by

$$\mu = \alpha\ \Gamma\left(1 + \frac{1}{\gamma}\right) \qquad \sigma = \sqrt{\alpha^2\left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right]}$$

where the gamma function is defined as follows: $\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}\mathrm{d}y$

First we use the data to obtain estimators of α and γ, $\widehat{\alpha}$ and $\widehat{\gamma}$, respectively, then we substitute these estimators into the formulas for μ and σ to obtain the corresponding estimators of the mean and standard deviation of the Weibull distribution:

$$\widehat{\mu} = \widehat{\alpha}\ \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) \qquad \widehat{\sigma} = \sqrt{\hat{\alpha}^2\left[\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)\right]}$$

We will now discuss several methods for obtaining the point estimators of the unknown parameters in the cdf.

Method 1: Graphical Estimators of Location-Scale Parameters

If the specified cdf is a location-scale family of cdfs, for example,

1. Normal(μ, σ^2)
2. Logistic(μ, β)
3. Exponential(β)
4. Weibull(γ, α)

then we can use a graphical procedure which use reference distribution plots to obtain *rough* estimators of the location and scale parameters.

Let $Q_o(u)$ be the quantile function of the standard member of the family and $\widehat{Q}(u)$ be the sample quantile for the n data values $Y_1, Y_2, \ldots, Y_n$.

Plot $\widehat{Q}(u_i)$ versus $Q_o(u_i)$ for $u_i = \frac{i-.5}{n}; \quad i = 1, 2, \ldots, n.$

If the n plotted points are reasonably close to a straight-line then **graphical estimators** of θ_1 and θ_2 are given by

$\widehat{\theta_1}$ = Y-intercept of fitted line

$\widehat{\theta_2}$ = Slope of fitted line

We will discuss in detail **Reference Distribution Plots** in Handout 8.

EXAMPLE The time to failure, in 100 hours, of a random sample of 25 newly designed fuel pumps are recorded as follows:

15.321	9.008	20.104	7.729	45.154	8.404	5.332	0.577	4.305
4.517	12.594	6.829	3.291	37.175	0.841	1.317	7.613	20.582
2.030	10.001	4.666	12.933	0.591	39.454	8.875		

The researcher states that from previous studies that the Weibull distribution was a good approximation to cdf of the r.v. Y, Time to Failure. The cdf of Y is given by

$$F_Y(y) = 1 - e^{-(y/\alpha)^\gamma}.$$

From the form of the cdf of Y, it can be observed that α is a scale parameter but γ is a shape parameter. Therefore, the Weibull family of cdf's is not a location-scale family.

However, a transformation of the data to $W = log(Y)$ yields the following results:

$W = log(Y)$ has cdf given by

$$\begin{aligned} F_W(w) = P[W \le w] = P[log(Y) \le w] = P[Y \le e^w] &= 1 - e^{-(e^w/\alpha)^\gamma} \\ &= 1 - e^{-(e^{\gamma w}/\alpha^\gamma)} \\ &= 1 - e^{-e^{\gamma w - log(\alpha^\gamma)}} \\ &= 1 - e^{-e^{\gamma(w - log(\alpha))}} \\ &= 1 - e^{-e^{(w-log(\alpha))/\frac{1}{\gamma}}} \end{aligned}$$

From the above we can conclude that the family of cdf's for W is a location-scale family with

$$\theta_1 = log(\alpha) \quad \text{and} \quad \theta_2 = \tfrac{1}{\gamma}$$

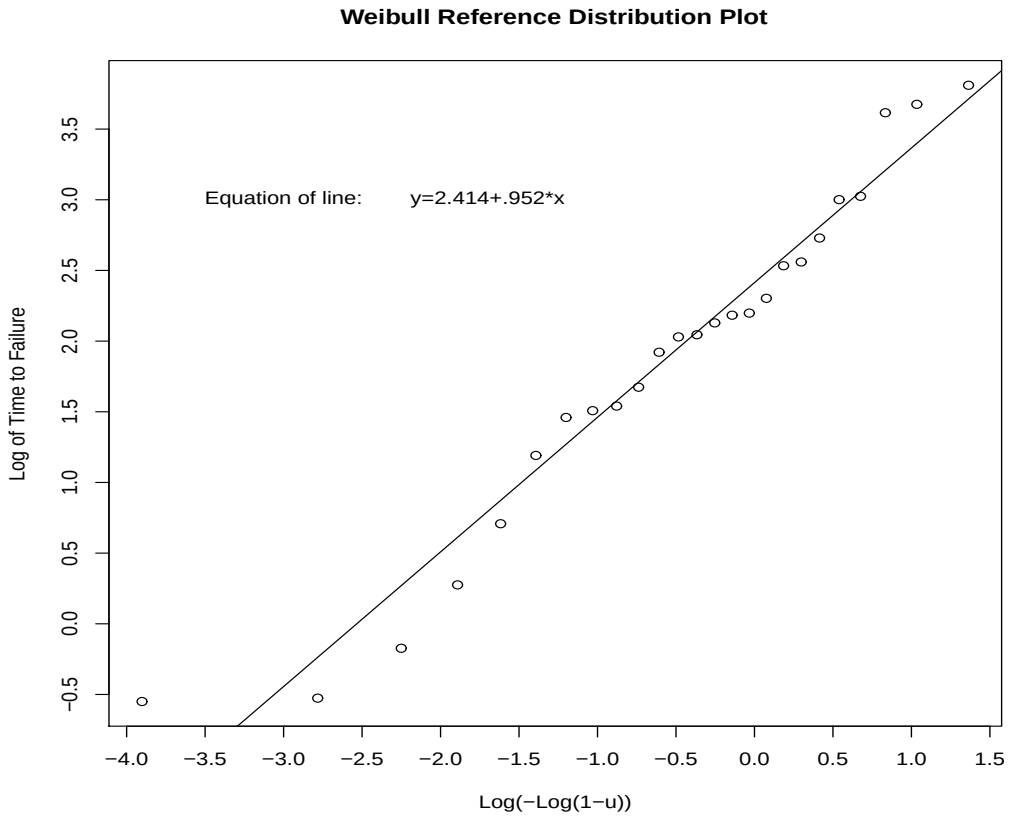
Plot the data on a Weibull Reference Distribution Plot with the sample quantile function for $W_i = log(Y_i)$, $\widehat{Q}_W(u_i)$ on the vertical axis and the standard member,

$\theta_1 = 0,\ \theta_2 = 1$, of the W family of quantiles, $Q_o(u_i)$, on the horizontal axis.

$\theta_1 = 0 \ \Rightarrow\ log(\alpha) = 0 \ \Rightarrow\ \alpha = 1$ and $\theta_2 = 1 \ \Rightarrow \gamma = 1$

Next, determine $Q_o(u) = F_o^{-1}(u)$, in the following manner:

$u = F_o(w_u) = 1 - e^{-e^{(w_u - 0)/1}} \ \Rightarrow\ Q_o(u) = w_u = log(-log(1-u))$



The 25 data values are relatively close to the fitted line:

$$\widehat{Q}_W(u_i) = 2.414 + .952 Q_o(u_i)$$

Thus, we can conclude that graphical estimators of θ_1 and θ_2 are given by

$$\widehat{\theta}_1 = 2.414 \quad \text{and} \quad \widehat{\theta}_2 = .952$$

From these estimators we can then compute:

$$\widehat{\gamma} = \frac{1}{\widehat{\theta}_2} = \frac{1}{.952} = 1.05042 \approx 1 \text{ (which implies Exponential cdf) and } \widehat{\alpha} = e^{\widehat{\theta}_1} = e^{2.414} = 11.178586$$

From these values, we can then estimate μ and σ:

$$\widehat{\mu} = \widehat{\alpha}\ \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) = (11.178586)\ \Gamma\left(1 + \frac{1}{1.05042}\right) = 10.96$$

$$\widehat{\sigma} = \sqrt{\widehat{\alpha}^2 \left[\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)\right]}$$

$$\widehat{\sigma} = \sqrt{(11.178586)^2 \left[\Gamma\left(1 + \frac{2}{1.05042}\right) - \Gamma^2\left(1 + \frac{1}{1.05042}\right)\right]} = 10.44$$

From the data we have the distribution-free estimators:

$$\widehat{\mu} = \bar{Y} = 11.57 \qquad\qquad \widehat{\sigma} = S = 12.29$$

Both of these methods are very crude estimators of the population mean and standard deviation. When we know the underlying population distributions, there are much more accurate estimators.

Method of Moments (MOM) Estimators:

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a population or n iid realizations from a process having cdf which depends on k unknown parameters: $\theta_1, \theta_2, \ldots, \theta_k$. The population moments $m_i = E(Y^i)$ depend on the k θs:

$$m_i = E(Y^i) = \int_{-\infty}^{\infty} y^i dF(y) = g_i(\theta_1, \theta_2, \ldots, \theta_k) \quad \text{for} \quad i = 1, 2, 3, 4$$

We obtain sample estimators of these $m_i s$ by replacing F with the edf $\widehat{F}$:

$$\widehat{m}_i = \int_{-\infty}^{\infty} y^i d\widehat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} Y_j^i$$

To obtain estimators of the θs, we just equate the sample moments to the population moments (with θ_is replaced with $\widehat{\theta}_i$s and solve for the $\widehat{\theta}_i$s:

$$\widehat{m}_i = g_i(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k) \quad \text{for} \quad i = 1, 2, \ldots, k$$

We have k equations in k unknowns.

EXAMPLE #1 Suppose F is a $N(\theta_1, \theta_2^2)$ cdf and we have a random sample:

$Y_1, Y_2, \ldots, Y_n$ iid $N(\theta_1, \theta_2^2)$. Find $\widehat{\theta}_1$ and $\widehat{\theta}_2$

$$m_1 = E[Y] = \theta_1 = g_1(\theta_1, \theta_2) \quad m_2 = E[Y^2] = Var(Y) + (E[Y])^2 = \theta_2^2 + \theta_1^2 = g_2(\theta_1, \theta_2)$$

Next, we equate sample moments to population moments:

$$(1) \quad \frac{1}{n} \sum_{1}^{n} Y_i = \bar{Y} = \widehat{m}_1 = g_1(\widehat{\theta}_1, \widehat{\theta}_2) = \widehat{\theta}_1$$

$$(2) \quad \frac{1}{n} \sum_{1}^{n} Y_i^2 = \widehat{m}_2 = g_2(\widehat{\theta}_1, \widehat{\theta}_2) = \widehat{\theta}_2^2 + \widehat{\theta}_1^2$$

Solving equations (1) and (2) we obtain

$$(1) \Rightarrow \widehat{\theta}_1 = \bar{Y} \qquad (2) \Rightarrow \widehat{\theta}_2^2 = \widehat{m}_2 - \widehat{\theta}_1^2 = \frac{1}{n} \sum_{1}^{n} Y_i^2 - \bar{Y}^2 = \frac{1}{n} \sum_{1}^{n} (Y_i - \bar{Y})^2$$

Therefore, we obtain:

$$\widehat{\theta}_1 = \bar{Y} \qquad \widehat{\theta}_2 = \sqrt{\frac{1}{n} \sum_{1}^{n} (Y_i - \bar{Y})^2}$$

EXAMPLE #2 Suppose F is a $Gamma(\alpha, \beta)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $Gamma(\alpha, \beta)$. Find $\widehat{\alpha}$ and $\widehat{\beta}$.

$m_1 = E[Y] = \alpha\beta = g_1(\alpha, \beta) \quad m_2 = E[Y^2] = Var(Y) + (E[Y])^2 = \alpha\beta^2 + (\alpha\beta)^2 = g_2(\alpha, \beta)$

Next, we equate sample moments to population moments:

$$(1) \quad \frac{1}{n}\sum_{1}^{n} Y_i = \bar{Y} = \widehat{m}_1 = g_1(\widehat{\alpha}, \widehat{\beta}) = \widehat{\alpha}\widehat{\beta}$$

$$(2) \quad \frac{1}{n}\sum_{1}^{n} Y_i^2 = \widehat{m}_2 = g_2(\widehat{\alpha}, \widehat{\beta}) = \widehat{\alpha}\widehat{\beta}^2 + \widehat{\alpha}^2\widehat{\beta}^2$$

Solving equations (1) and (2) we obtain

$$(1) \Rightarrow \quad \widehat{\alpha} = \bar{Y}/\widehat{\beta} \qquad\qquad (2) \Rightarrow \quad \widehat{m}_2 = \bar{Y}\widehat{\beta} + \bar{Y}^2 \quad \Rightarrow \quad \widehat{\beta} = (\widehat{m}_2 - \bar{Y}^2)/\bar{Y}$$

Therefore, we obtain:

$$\widehat{\alpha} = \bar{Y}^2/(\widehat{m}_2 - \bar{Y}^2) \qquad\qquad \widehat{\beta} = (\widehat{m}_2 - \bar{Y}^2)/\bar{Y}$$

EXAMPLE #3 Suppose F is a $Weibull(\gamma, \alpha)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $Weibull(\gamma, \alpha)$. Find $\widehat{\gamma}$ and $\widehat{\alpha}$.

$$m_1 = \alpha\ \Gamma\left(1 + \frac{1}{\gamma}\right) = g_1(\gamma, \alpha), \quad \text{where} \ \ \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

$$m_2 = \alpha^2\ \Gamma\left(1 + \frac{2}{\gamma}\right) = g_2(\gamma, \alpha)$$

Next, we equate sample moments to population moments:

$$(1) \quad \frac{1}{n}\sum_{1}^{n} Y_i = \bar{Y} = \widehat{m}_1 = g_1(\widehat{\gamma}, \widehat{\alpha}) = \widehat{\alpha}\ \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right)$$

$$(2) \quad \frac{1}{n}\sum_{1}^{n} Y_i^2 = \widehat{m}_2 = g_2(\widehat{\gamma}, \widehat{\alpha}) = \widehat{\alpha}\ \Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right)$$

The two equations are then solved numerically, closed form solutions are not possible.

Maximum Likelihood Estimation (MLE)

MOM's only used the first few moments of the distribution and hence do not use the full knowledge of the structure of the population distribution. The MLE's will directly use the pdf in obtaining the estimates of the unknown parameters.

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (or iid observations) from a population/process having pdf $f(y)$ which depends on unknown parameters: $\theta_1, \theta_2, \cdots, \theta_k$, where θs are elements of a parameter space Θ.

Define the likelihood function as

$$L(\theta_1, \theta_2, \cdots, \theta_k; y) = f(y_1, y_2, \ldots, y_n; \theta) \quad \text{joint pdf of} \quad Y_1, Y_2, \ldots, Y_n$$

Because the $Y_i s$ are iid we have

$$L(\theta_1, \theta_2, \cdots, \theta_k; y) = f(y_1; \theta) f(y_2; \theta) \cdots f(y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

The MLEs of the θs is that vector $(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k)$ which maximizes the likelihood function:

$$L(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k) = \max_{\theta \epsilon \Theta} L(\theta_1, \theta_2, \cdots, \theta_k)$$

Example #1 Exponential pdf:

Suppose F is an Exponential (β) cdf and $T_1, T_2, \ldots, T_n$ is a random sample from F. Find the MLE of β : $\widehat{\beta}$.

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f(t_i; \beta) \\ &= \prod_{i=1}^{n} \frac{1}{\beta} e^{-t_i/\beta} \\ &= \left(\frac{1}{\beta}\right)^n e^{-\frac{1}{\beta}\sum_{i=1}^{n} t_i} \end{aligned}$$

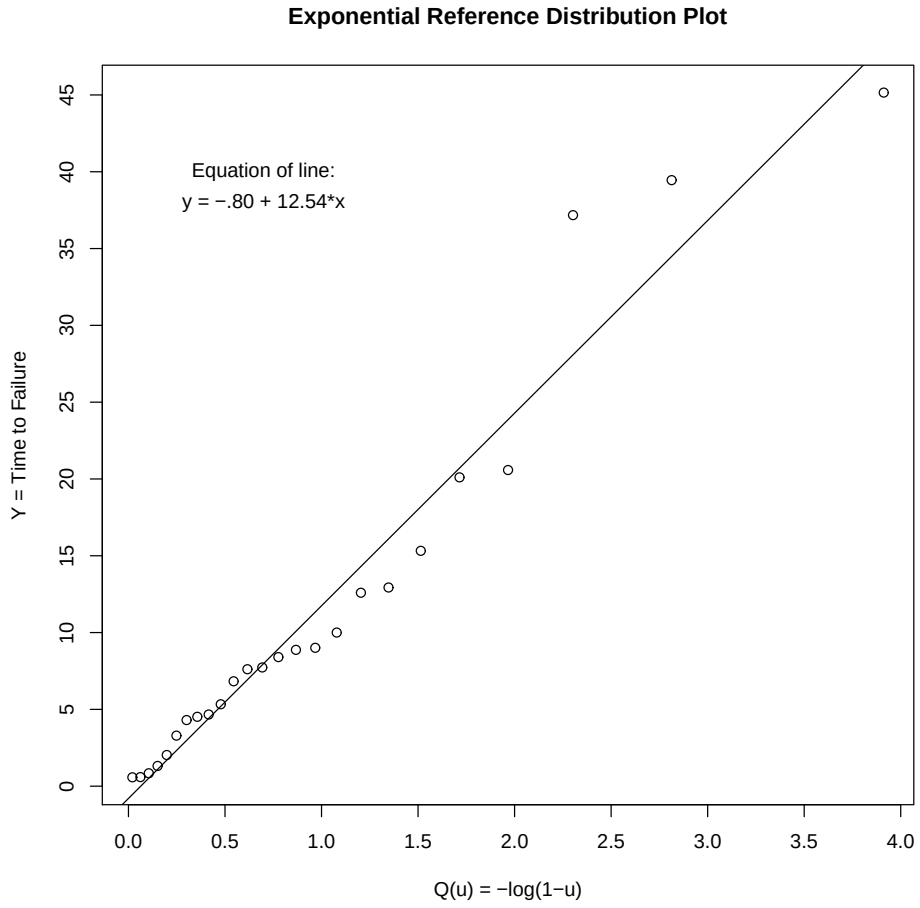
We will demonstrate the derivation of MLE's for the time to failure example.

The time to failure, in 100 hours, of a random sample of 25 newly designed fuel pumps are recorded as follows:

15.321	9.008	20.104	7.729	45.154	8.404	5.332	0.577	4.305
4.517	12.594	6.829	3.291	37.175	0.841	1.317	7.613	20.582
2.030	10.001	4.666	12.933	0.591	39.454	8.875		

We compute $\sum_{i=1}^{25} T_i = 289.243$ and $\bar{T} = 11.570$

From the Exponential Distribution Plot given on the next page, we conclude that the times to failure are adequately modeled by an Exponential Distribution.



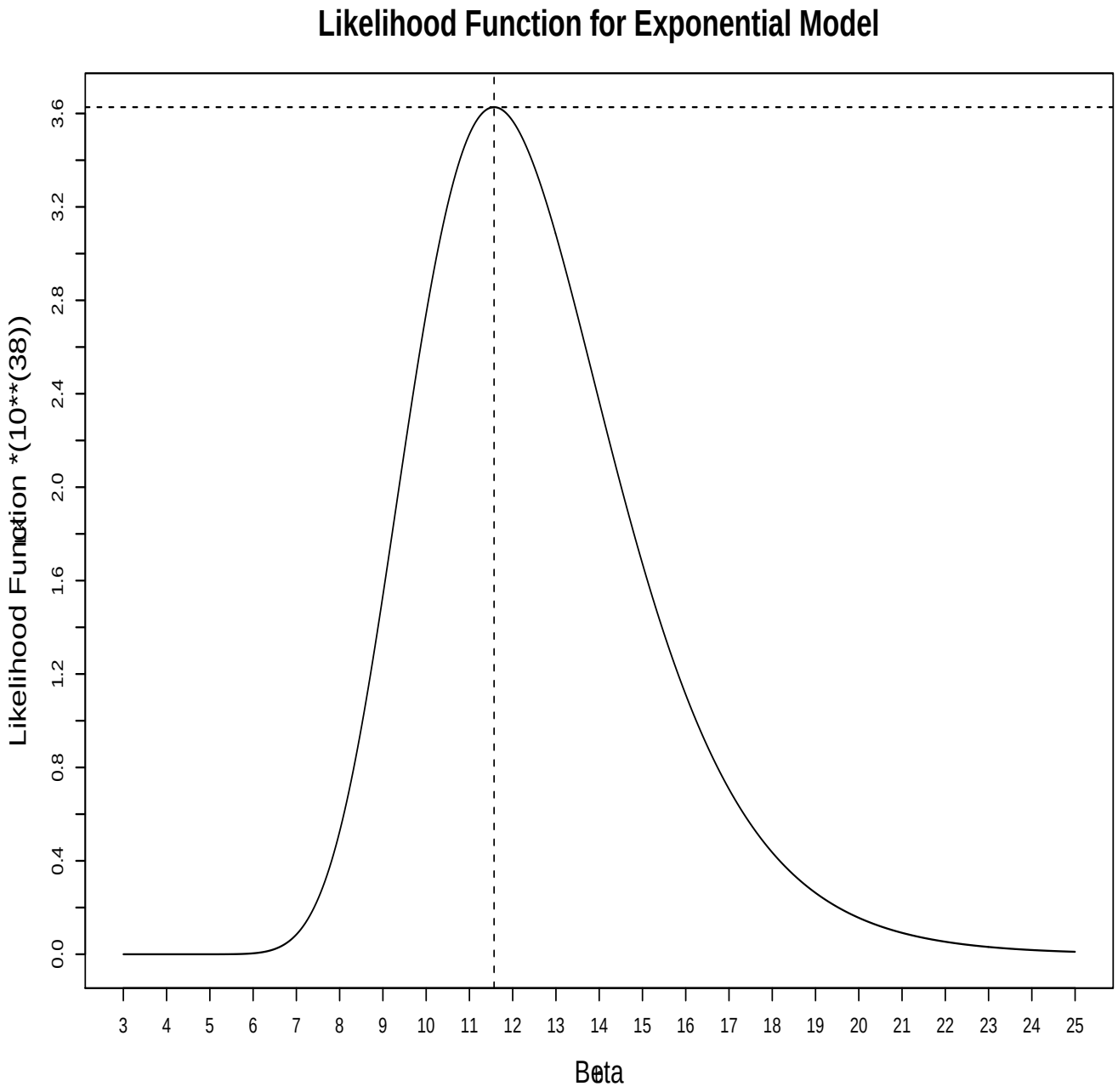
We next need to estimate the parameter β in the exponential distribution:

$$f(t) = \frac{1}{\beta} e^{-t/\beta}$$

We will use Maximum Likelihood techniques, that is, find the value of β which maximizes the likelihood function:

$$L(\beta) = \prod_{i=1}^{n} f(T_i) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-T_i/\beta} = \beta^{-n} e^{-\sum_{i=1}^{n} T_i/\beta} = \beta^{-25} e^{-289.243/\beta}.$$

Next, we plot the likelihood function, $L(\beta)$ and determine the value of β which maximizes this function.



From the plot of the likelihood function, we can observe that the maximum occurs at $\beta \approx 11.57$.

R code used to produce Exponential Reference plot and plot of the likelihood function

```
# program name: mle_exp.R

t = c(15.321,9.008,20.104,7.729,45.154,8.404,5.332,0.577,4.305,4.517,12.594,
     6.829,3.291,37.175,0.841,1.317,7.613,20.582,2.030,10.001,4.666,12.933,
     0.591,39.454,8.875)
t = sort(t)
n = 25
i = seq(1,25,1)
u = (i-.5)/n
x =  -log(1-u)

postscript("u:/meth1/Rfiles/exp_refplot.ps",height=8,horizontal=F)

plot(x,t, xlab="-Log(1-u)",ylab="Time to Failure",lab=c(13,11,7),
main="Exponential Reference Distribution Plot")
abline(lm(t~x))
text(.7,30,"Equation of line:")
text(.7,28,"y=2.414+.952*x")

postscript("u:/meth1/Rfiles/mle_exp.ps",height=8,horizontal=F)

b = seq(3,25,.01)
LK = (b^(-25))*exp(-289.243/b)*(10)^38
out =  cbind(b,LK)
LKmax  =  max(LK)
bmax  =  which(LK==max(LK))
MLE  =  b[bmax]
par(cex=.65)
plot(b,LK, type="l",lab=c(30,16,7))
par(cex=.99)
title("Likelihood Function for Exponential Model",xlab="Beta",
ylab="Likelihood Function *(10**(38))")
abline("v"=b[bmax],lty=2)
abline("h"=max(LK),lty=2)
graphics.off()



library(MASS)
mle_exp=fitdistr(t,"exponential")
mle_weibull=fitdistr(t,"weibull")
```

Using the last 3 lines of code in the R program, given on the previous page:

```
library(MASS)
mle_exp=fitdistr(t,"exponential")
```

we can obtain the MLE estimates of the parameters in the Exponential model:
Output from R code:

```
> fitdistr(t,"exponential")
      rate
  0.08643252
 (0.01728650)
```

Note that the estimate in the exponential model is given in terms of $1/\beta$, called the "rate" in R: $\hat{\beta}$ =1/rate = 1/.08643252 = 11.5697 which is the value we determined.

In fact, we can show that the MLE of β in the exponential distribution occurs at $\widehat{\beta} = \bar{T}$:

Suppose F is a Exponential (β) cdf and we have a random sample: $T_1, T_2, \ldots, T_n$ from F, i.e., iid Exponential (β). Find the MLE's : $\widehat{\beta}$.

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f(t_i;\beta) \\ &= \prod_{i=1}^{n} \beta^{-1} e^{-t_i/\beta} \\ &= \beta^{-n} e^{-\frac{1}{\beta}\sum_{i=1}^{n} t_i} \end{aligned}$$

Taking logarithms of both sides of the equation yields:

$$l(\beta; y) = -nlog(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} t_i$$

Take the partial derivate of the log-likelihood wrt β, set derivate equal to 0, and then solve for $\widehat{\beta}$:

$$\frac{\partial l(\beta; y)}{\partial \beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} t_i \quad \Rightarrow \quad \widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Note that the second partial of the log-likelihood is negative at $\widehat{\beta}$:

$$\frac{\partial^2 l(\beta; y)}{\partial^2 \beta} \text{ at } (\beta = \widehat{\beta}) \quad = \frac{n}{\widehat{\beta}^2} - \frac{2}{\widehat{\beta}^3}\sum_{i=1}^{n} t_i \quad = -\frac{n}{\widehat{\beta}^2} < 0$$

Thus, the likelihood function is a maximum at $\widehat{\beta}$.

In our data set, we had $\bar{T} = 11.5692$ which to roundoff error is the value we obtained for the MLE of β using 1/rate = = 1/.08643252 = 11.5697

EXAMPLE #2 Suppose F is a $N(\theta_1, \theta_2^2)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $N(\theta_1, \theta_2^2)$. Find the MLE's : $\widehat{\theta_1}$ and $\widehat{\theta_2}$

$$\begin{aligned} L(\theta_1, \theta_2) &= \prod_{i=1}^{n} f(y_i; \theta_1, \theta_2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}(y_i - \theta_1)^2} \\ &= \frac{1}{(2\pi)^{n/2}(\theta_2)^n} e^{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(y_i - \theta_1)^2} \end{aligned}$$

Taking the natural log of both sides of the equation yields with $l(\theta_1, \theta_2; y) = log(L(\theta_1, \theta_2); y)$:

$$l(\theta_1, \theta_2; y) = -\frac{n}{2} log(2\pi) - n log(\theta_2) - \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (y_i - \theta_1)^2$$

Take the partial derivates wrt θ_1 and θ_2, set derivates equal to 0, and then solve for $\widehat{\theta_1}$ and $\widehat{\theta_2}$.

$$\frac{\partial l(\theta_1, \theta_2; y)}{\partial \theta_1} = \frac{1}{\theta_2^2} \sum_{i=1}^{n} (y_i - \theta_1) \quad \Rightarrow \widehat{\theta_1} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\frac{\partial l(\theta_1, \theta_2; y)}{\partial \theta_2} = \frac{-n}{\theta_2} + \frac{1}{\theta_2^3} \sum_{i=1}^{n} (y_i - \theta_1)^2 \quad \Rightarrow \widehat{\theta_2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\theta_1})^2}$$

The same answers that we obtained using MOM.

To be complete, we would need to verify that these solutions in fact yield the maximums in the likelihood function and not the minimums. Thus, we would need to examine the second derivatives and the mixed derivatives of the likelihood function to determine if a relative maximum, relative minimum, or an undetermined point has been achieved.

EXAMPLE #3 Suppose F is a $Gamma(\alpha, \beta)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $Gamma(\alpha, \beta)$. Find the MLE's : $\widehat{\alpha}$ and $\widehat{\beta}$.

$$\begin{aligned} L(\alpha, \beta; y) &= \prod_{i=1}^{n} f(y_i; \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} e^{-y_i/\beta} \\ &= \frac{1}{\Gamma^n(\alpha)\beta^{n\alpha}} (\prod_{i=1}^{n} y_i)^{\alpha-1} e^{-\frac{1}{\beta}\sum_{i=1}^{n} y_i} \end{aligned}$$

Taking logarithms of both sides of the equation yields:

$$l(\alpha, \beta; y) = -nlog(\Gamma(\alpha)) - n\alpha log(\beta) + (\alpha - 1)\sum_{i=1}^{n} log(y_i) - \frac{1}{\beta}\sum_{i=1}^{n} y_i$$

Take the partial derivates wrt α and β, set derivates equal to 0, and then solve for $\widehat{\alpha}$ and $\widehat{\beta}$.

$$\frac{\partial l(\alpha, \beta; y)}{\partial \alpha} = -\frac{n\Gamma^{'}(\alpha)}{\Gamma(\alpha)} - nlog(\beta) + \sum_{i=1}^{n} log(y_i) \quad \Rightarrow \quad -\frac{n\Gamma^{'}(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} - nlog(\widehat{\beta}) + \sum_{i=1}^{n} log(y_i) = 0$$

$$\frac{\partial l(\alpha, \beta; y)}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} y_i \quad \Rightarrow \quad -\frac{n\widehat{\alpha}}{\widehat{\beta}} + \frac{1}{\widehat{\beta}^2}\sum_{i=1}^{n} y_i = 0$$

Need to verify that the solutions to these two equations $\widehat{\alpha}$ and $\widehat{\beta}$ are maximums of the log-likelihood and not minimums.

For the gamma family of pdfs, there is no closed form solution to the equations because of the terms involving :

$$\Gamma(c) = \int_0^{\infty} y^{c-1} e^{-y} dy \quad \text{and its derivative} \quad \Gamma^{'}(c)$$

.

Thus, we would need to obtain a numerical solution to the equations.

Using the following R code we obtain the MLE's:

```
y = data

library(MASS)
fitdistr(y,"gamma")
```

The estimators will be outputted as "Shape" and "Rate"

Then we obtain

$$\widehat{\beta} = \frac{1}{\text{Rate}} \quad \text{and} \quad \widehat{\alpha} = \text{Shape}$$

EXAMPLE #4 Suppose F is a $Weibull(\gamma, \alpha)$ cdf and we have a random sample: $T_1, T_2, \ldots, T_n$ iid $Weibull(\gamma, \alpha)$. Find the MLE's : $\widehat{\gamma}$ and $\widehat{\alpha}$.

$$\begin{aligned} L(\gamma, \alpha) &= \prod_{i=1}^{n} f(t_i; \gamma, \alpha) \\ &= \prod_{i=1}^{n} \frac{\gamma}{\alpha} \left(\frac{t_i}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{t_i}{\alpha}\right)^{\gamma}} \\ &= \left(\frac{\gamma}{\alpha}\right)^{n} \left(\prod_{i=1}^{n} \left(\frac{t_i}{\alpha}\right)^{\gamma-1}\right) e^{-\frac{1}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma}} \end{aligned}$$

Taking logarithms of both sides of the equation yields:

$$l(\gamma, \beta; y) = nlog(\gamma) - n\gamma log(\alpha) + (\gamma - 1) \sum_{i=1}^{n} log(t_i) - \frac{1}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma}$$

Take the partial derivates wrt γ and α, set derivates equal to 0, and then solve for $\widehat{\gamma}$ and $\widehat{\alpha}$:

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \alpha} = -\frac{n\gamma}{\alpha} + \frac{\gamma}{\alpha^{\gamma+1}} \sum_{i=1}^{n} t_i^{\gamma}$$

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \alpha} = 0 \quad \Rightarrow \quad \widehat{\alpha} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i^{\widehat{\gamma}}\right)^{1/\widehat{\gamma}}$$

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \gamma} = \frac{n}{\gamma} - nlog(\alpha) + \sum_{i=1}^{n} log(t_i) + \frac{log(\alpha)}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma} - \frac{1}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma} log(t_i)$$

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \gamma} = 0 \quad \Rightarrow \quad \frac{\sum_{i=1}^{n} t_i^{\widehat{\gamma}} log(t_i)}{\sum_{i=1}^{n} t_i^{\widehat{\gamma}}} - \frac{1}{\widehat{\gamma}} = \frac{1}{n} \sum_{i=1}^{n} log(t_i)$$

Numerical techniques would be required to find the solutions to these two equations. R and SAS have a routines for estimating the parameters in the Weibull distribution.

We will illustrate the code using the data from the Weibull example used earlier in this handout to illustrate the graphical estimation of parameters.

Recall that the Weibull Reference distribution plot indicated that a Weibull distribution would be an appropriate model for the data. From the reference distribution plot we obtained the following "rough" estimates of the scale and shape parameters:

$$\widehat{\gamma} = 1.05 \qquad \text{and} \qquad \widehat{\alpha} = 11.18$$

Using the following R code we obtain the MLE's:

```
y = c(15.321, 9.008, 20.104, 7.729, 45.154, 8.404, 5.332, 0.577, 4.305, 4.517,
12.594, 6.829, 3.291, 37.175, 0.841, 1.317, 7.613, 20.582, 2.030, 10.001,
4.666, 12.933, 0.591, 39.454,  8.875)

library(MASS)
fitdistr(y,"weibull")


OUTPUT from R:

     shape         scale
  0.9839245    11.4852981
( 0.1512936) ( 2.4660607)
```

The values in parentheses are the standard errors of the estimators.

Recall, that we also used the exponential model for this data set and obtained $\hat{\beta} = 11.57$.

Furthermore, $E(Y) = \beta = \sqrt{Var(Y)}$.

Thus, we have $\hat{\mu} = 11.57 = \hat{\sigma}$.

How would these values change if modeled the data with the Weibull distribution?

$$\widehat{\mu} = \widehat{\alpha}\ \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) = (11.4852981)\Gamma\left(1 + \frac{1}{.9839245}\right) = 11.5659$$

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\widehat{\alpha}^2\left[\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)\right]} \\ &= \sqrt{(11.4852981)^2\left[\Gamma\left(1 + \frac{2}{0.9839245}\right) - \Gamma^2\left(1 + \frac{1}{0.9839245}\right)\right]} = 11.75532\end{aligned}$$

Note: $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ which can be obtain using the R function **gamma(x)**

From the data we compute $\bar{Y} = 11.57 \qquad S = 12.284.$

Thus, the two estimates of the mean are closely matched but there is some difference in the two estimates of the standard deviation.

The next page contains the SAS code needed to obtain the MLS's for the parameters in the Weibull model.

SAS code for estimating parameters from a Weibull Distribution:

```
option ls=75 ps=55 nocenter nodate;
title 'Weibull MLE Estimation of Fuel Pump Data Data';
data cords;
input F @@;
label F  = 'Time to Failure of Pumps';
cards;
15.321   9.008    20.104    7.729    45.154    8.404    5.332    0.577     4.305
 4.517  12.594     6.829    3.291    37.175    0.841    1.317    7.613    20.582
 2.030  10.001     4.666   12.933     0.591   39.454    8.875
run;
proc print;
proc lifereg data=cords;
 model  F = /dist=weibull covb;
run;
```

OUTPUT

```
Weibull MLE Estimation of Fuel Pump Data Data                            11

The LIFEREG Procedure

                          Model Information

Data Set                      WORK.CORDS
Dependent Variable                Log(F)    Time to Failure of Pumps
Number of Observations                25
Noncensored Values                    25
Right Censored Values                  0
Left Censored Values                   0
Interval Censored Values               0
Name of Distribution             Weibull
Log Likelihood              -39.33978867


Number of Observations Read          25
Number of Observations Used          25

                   Analysis of Parameter Estimates

                           Standard   95% Confidence     Chi-
Parameter     DF Estimate    Error       Limits       Square Pr > ChiSq

Intercept      1   2.4411   0.2147   2.0202   2.8619  129.25     <.0001
Scale          1   1.0163   0.1563   0.7519   1.3738
Weibull Scale  1  11.4853   2.4661   7.5401  17.4947
Weibull Shape  1   0.9839   0.1513   0.7279   1.3300


     Estimated Covariance Matrix

              Intercept         Scale

Intercept      0.046102     -0.010810
Scale         -0.010810      0.024423
```

Note the estimates, Weibull Scale, $\hat{\alpha}$ and Weibull Shape, $\hat{\gamma}$ are identical to the values obtained from R on the previous page.

DISTRIBUTION-FREE SUMMARIES

Let $Y_1, Y_2, \ldots, Y_n$ a random sample (or iid observations) from a population in which the pdf $f(y)$ is not specified. Suppose estimators of population summaries are desired. There are a number of possible methods to obtain the estimators. In general, we can simply replace the the population distribution function (cdf) with the sample distribution function (edf) in the definition of the summary parameter.

Estimators Based on Population Moments

Suppose we have a random sample (iid r.v.s) $Y_1, Y_2, \ldots, Y_n$ with cdf F which is unknown. We want to estimate the various population summaries of location:

In the definition of $\mu_i = \int_{-\infty}^{\infty}(y-\mu)^i dF(y)$ replace $F(y)$ with the edf

$$\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y)$$

We then obtain

$$\widehat{\mu} = \int_{-\infty}^{\infty} y d\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} Y_{(i)} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

and

$$\widehat{\mu}_k = \int_{-\infty}^{\infty}(y-\widehat{\mu})^k d\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n}(Y_{(i)}-\widehat{\mu})^k = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^k$$

Thus, we have the following estimators of the population standard deviation σ, population skewness β_1, population kurtosis β_2, and trimmed mean $\mu_{(\alpha)}$:

1. **Sample Standard Deviation**

$$\widehat{\sigma} = \sqrt{\widehat{\mu}_2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

2. **Sample Skewness**

$$\widehat{\beta}_1 = \frac{\widehat{\mu}_3}{(\widehat{\mu}_2)^{3/2}} = \frac{\widehat{\mu}_3}{(\widehat{\sigma})^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^3}{(\widehat{\sigma})^3}$$

3. **Sample Kurtosis**

$$\widehat{\beta}_2 = \frac{\widehat{\mu}_4}{(\widehat{\mu}_2)^2} = \frac{\widehat{\mu}_4}{(\widehat{\sigma})^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^4}{(\widehat{\sigma})^4}$$

   Some software packages, e.g., SAS, report the excess kurtosis defined as $\hat{\beta}_2 - 3$. The word excess referring to the difference from the kurtosis for the normal distribution.

4. **Sample α-Trimmed Mean**

Recall,

$$\mu_{(\alpha)} = \frac{1}{1-2\alpha}\int_{Q(\alpha)}^{Q(1-\alpha)} y\ dF(y)$$

thus using our idea of replacing the quantile function with its sample estimator we have

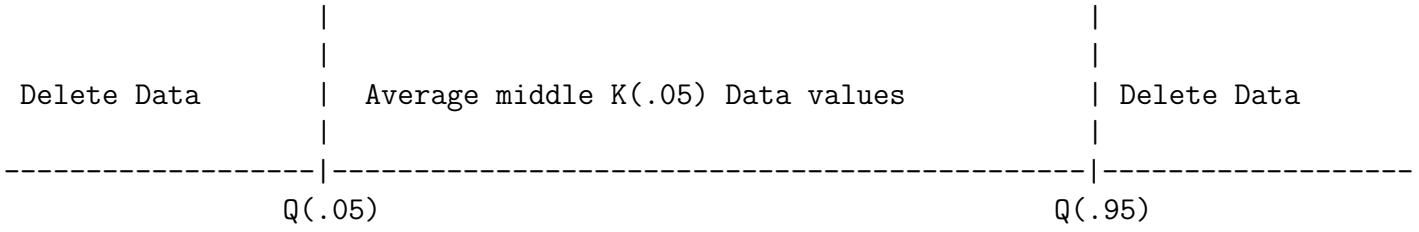
$$\widehat{\mu}_{(\alpha)} = \frac{1}{1-2\alpha}\int_{\widehat{Q}(\alpha)}^{\widehat{Q}(1-\alpha)} y\ d\widehat{F}(y) = \frac{1}{K(\alpha)/n}\sum_{i=[n\alpha+1]}^{n-[n\alpha]} Y_{(i)}\frac{1}{n} = \frac{1}{K(\alpha)}\sum_{i=[n\alpha+1]}^{n-[n\alpha]} Y_{(i)},$$

where $[A]$ is the largest integer less than or equal to the real number A and

$$K(\alpha) = n - [n\alpha] - [n\alpha + 1] + 1$$

is the number of data values between $\widehat{Q}(\alpha)$ and $\widehat{Q}(1-\alpha)$.

Thus, $\widehat{\mu}_{(\alpha)}$ is the average of the data values that remain after removing the $[n\alpha]$ smallest and $[n\alpha]$ largest values in the data set. We are trimming exactly $100\alpha\%$ of the data from both tails when $n\alpha$ is an integer and slightly less than $100\alpha\%$ when $n\alpha$ is not an integer.



For example, suppose $n = 30$, $\alpha = .05$ then $n\alpha = (30)(.05) = 1.5$

$$K(\alpha) = n - [n\alpha] - [n\alpha + 1] + 1 = 30 - [1.5] - [2.5] + 1 = 28$$

Thus, we would trim $(1 - \frac{28}{30})/2 = .033$ from the left and right tails, not $\alpha = .05$.

If $n\alpha$ is an integer, then exactly $100\alpha\%$ is trimmed from both tails.

For example, suppose $n = 20$, $\alpha = .10$ then $n\alpha = (20)(.1) = 2$

$$K(\alpha) = n - [n\alpha] - [n\alpha + 1] + 1 = 20 - [2] - [3] + 1 = 16$$

Thus, we would trim $(1 - \frac{16}{20})/2 = .10$ from the left and right tails and this would be exactly $\alpha = .1$.

Modified Estimators

In practice, the previously defined estimators are often modified so that the estimators are **unbiased** estimators of the specified parameter when the data are from a normal distribution. In particular, many computer packages use the following quantities.

1. **Unbiased Estimator of μ:**

$$\text{Let} \quad \hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{then} \quad E[\bar{Y}] = \mu$$

- That is, $\bar{Y}$ is an unbiased estimator of $\hat{\mu}$ for any distribution for which $|\mu| < \infty$

2. **Unbiased Estimator of σ^2:**

$$\text{Let} \quad \widehat{\sigma}^{2*} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \text{then} \quad E[S^2] = \sigma^2$$

- That is, S^2 is an unbiased estimator of σ^2 for any distribution for which $\sigma^2 < \infty$

- However, S is a biased estimator of σ, that is, $E[S] \neq \sigma$.

  In Handout 10, we will prove the results in 1. and 2.

3. **Modified Estimator of Skewness Parameter β_1:**

$$\text{Let} \quad \widehat{\beta}_1^* = \left(\frac{n^2}{(n-1)(n-2)}\right) \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^3}{(\widehat{\sigma}^*)^3} = \left(\frac{n^2}{(n-1)(n-2)}\right) \frac{\hat{\mu}_3}{(S^2)^{3/2}}$$

- If $Y_1, Y_2, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, then $\beta_1 = 0$ and $E[\widehat{\beta}_1^*] = 0$, therefore, $\widehat{\beta}_1^*$ is an unbiased estimator of β_1. This result does not necessarily hold when the data is from a non-normal distribution.

4. **Modified Estimator of Excess Kurtosis Parameter $\beta_2 - 3$:**

$$\text{Let} \quad \widehat{\beta}_2^* - 3 = \frac{(n+1)n^2}{(n-1)(n-2)(n-3)} \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^4}{(\widehat{\sigma}^*)^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

- If $Y_1, Y_2, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, then $\beta_2 = 3$ and $E[\widehat{\beta}_2^*] = 3$ so $\widehat{\beta}_2^*$ is an unbiased estimator of β_2. This result does not hold necessarily hold when the data is from a non-normal distribution.

These are the forms of the estimators used in most software packages. However, not in all. You must check the definitions if you want to know what exactly the software program is computing.

Estimators Based on Quantiles

To obtain estimators of population parameters which are defined in terms of the population quantile $Q(u)$, we will just replace the population quantile with the sample quantile in the definition of the population parameter:

1. **Median** The population median $\tilde{\mu} = Q(.5)$ is estimated by

$$\widehat{Q}(.5) = \begin{cases} Y_{((n+1)/2)} & if \quad n \text{ is odd} \\ \left(Y_{(n/2)} + Y_{(n/2+1)}\right)/2 & if \quad n \text{ is even} \end{cases}$$

2. **Quartiles** The lower and upper quartiles $Q_1 = Q(.25)$ and $Q_3 = Q(.75)$ are estimated by their corresponding sample quartiles in most instances: $\widehat{Q}_1 = \widehat{Q}(.25)$ and $\widehat{Q}_3 = \widehat{Q}(.75)$.

   However, in some software packages and textbooks an alternative definition is given:

   Divide the data set into two equal halves:

   If n is even:

   Set 1: $Y_{(1)}, \cdots, Y_{(n/2)}$ and

   Set 2: $Y_{(n/2+1)}, \cdots, Y_{(n)}$

   Then $\widehat{Q}_1$ is the median of Set 1 and $\widehat{Q}_3$ is the median of Set 2

   If n is odd:

   Set 1: $Y_{(1)}, \cdots, Y_{((n+1)/2)}$ and

   Set 2: $Y_{((n+1)/2)}, \cdots, Y_{(n)}$

   Then $\widehat{Q}_1$ is the median of Set 1 and $\widehat{Q}_3$ is the median of Set 2

   For small n, these values for $\widehat{Q}_1$ and $\widehat{Q}_3$ may differ from the values obtained from $\widehat{Q}(.25)$ and $\widehat{Q}(.75)$

3. **Five Number Summary of Data:** Produced in R using the function **quantile(y)**

$$\left[Min = Y_{(1)}; \qquad Q_1; \qquad Q_2; \qquad Q_3; \qquad Max = Y_{(n)}\right]$$

   In R, the function **summary(y)** provides the 5 number summary of data in y but also provides the sample mean $\bar{y}$.

4. **Interquartile Range (IQR)** The sample estimator of the population

   $IQR = Q(.75) - Q(.25)$ is just $\widehat{IQR} = \widehat{Q}(.75) - \widehat{Q}(.25)$

5. **Range**: The population range is $Q(1) - Q(0) = R$. The sample estimator is taken to be

   $\widehat{R} = Y_{(n)} - Y_{(1)}$

6. **MAD**: $MAD = Median\{|Y - Median\{Y\}|\}/.6745 = Median\{|Y - Q_Y(.5)|\}/.6745$

   Therefore, $\widehat{MAD} = Median\{|Y - \widehat{Q}_Y(.5)|\}/.6745$

   To compute $\widehat{MAD}$

   (a) Find $\widehat{Q}_Y(.5)$ from $Y_1, Y_2, \ldots, Y_n$

   (b) Compute $W_i = |Y_i - \widehat{Q}_Y(.5)|$ for $i = 1, \ldots, n$

   (c) Find $\widehat{Q}_W(.5)$ from $W_1, W_2, \ldots, W_n$

   (d) $\widehat{MAD} = \widehat{Q}_W(.5)/.6745$

7. **m-estimator of Location** A modification of the $\alpha-$trimmed mean is the m-estimator. In place of deleting observations the m-estimator $\hat{m}$ assigns weights w_i to each data value such that the more extreme a data value is from the "center" of the data, the smaller the weight. Thus, in place of deleting extreme data values as we did with the $\alpha-$trimmed mean, we just reduce their influence.

   The m-estimator is defined as

$$\widehat{m} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i Y_i$$

   The m-estimator involves three parameters:

   (a) v a robust measure of dispersion, $\widehat{MAD}$, for example.

   (b) t a "tuning constant"

   The tuning constant is generally taken to be a value in (1.345, 1.5). Smaller values of t place a larger penalty on data values for being extreme to the center of the data.

   (c) w_i a weight assigned to each data value Y_i with

$$w_j = \begin{cases} -\frac{tv}{Y_j - \hat{m}} & if \quad Y_j < \hat{m} - tv \\ 1 & if \quad \hat{m} - tv \leq Y_j \leq \widehat{m} + tv \\ \frac{tv}{Y_j - \hat{m}} & if \quad Y_j > \hat{m} + tv \end{cases}$$

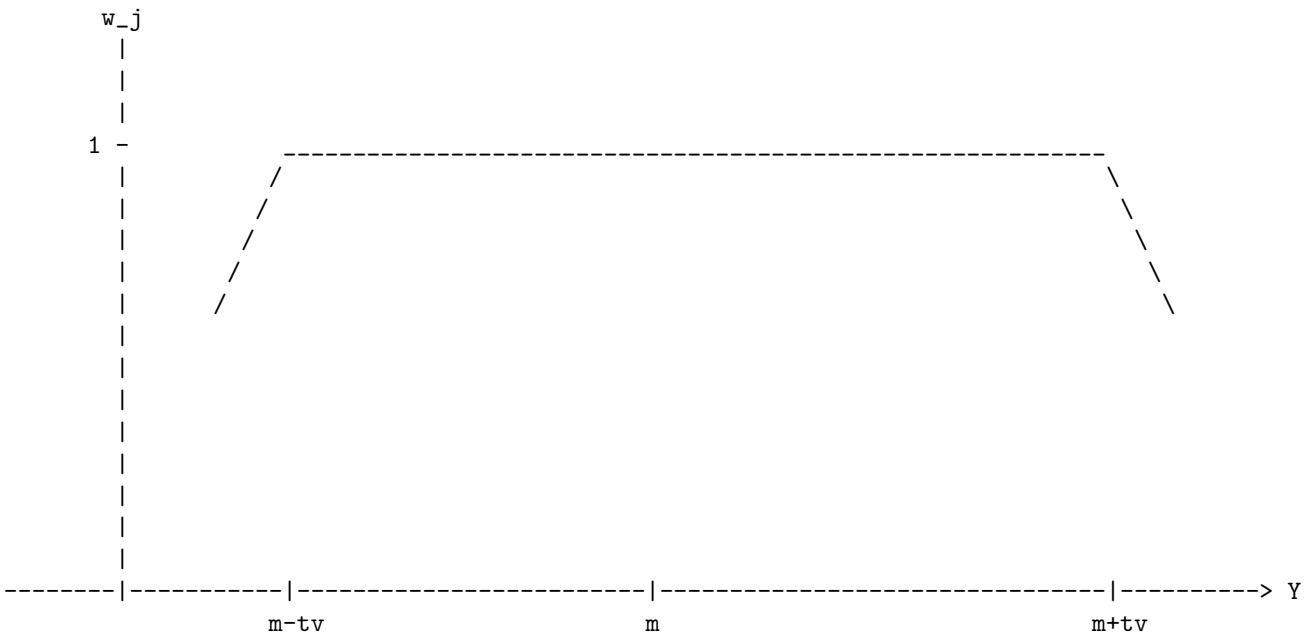
Note that $0 \leq w_j \leq 1$

For $\widehat{m} - tv \leq Y_j \leq \widehat{m} + tv$ we have $w_j = 1$, thus there is no downweighting of the data value Y_j:

For $Y_j < \widehat{m} - tv$ or $Y_j > \widehat{m} + tv$, then $w_j < 1$ and the data value Y_j receives a greater downweighting as it moves further from the center, i.e., w_j decreases as $|Y_j - \widehat{m}|$ increases

The computation of $\hat{m}$ is iterative:

1. Select an initial value for $\widehat{m}$, for example, $\widehat{m}_1 = \widehat{Q(.5)}$.
2. Select value for t (e.g. $t = 1.345$)
3. Select number of iterations OR
4. Select a level of relative accuracy RA = $|\frac{\widehat{m}_{k+1} - \widehat{m}_k}{\widehat{m}_k}| < \epsilon$ (a stopping point)
5. Select a robust estimator of dispersion ($v = \widehat{MAD}$)
6. Calculate $w_j s$ using $\widehat{m}_1$
7. Calculate $\widehat{m}_2$
8. Calculate RA: If RA$< \epsilon$ STOP calculations and output $\hat{m} = \hat{m}_2$
9. If RA$\geq \epsilon$, return to step 6
10. continue the above loop until either RA is achieved or the specified number of iterations occurs

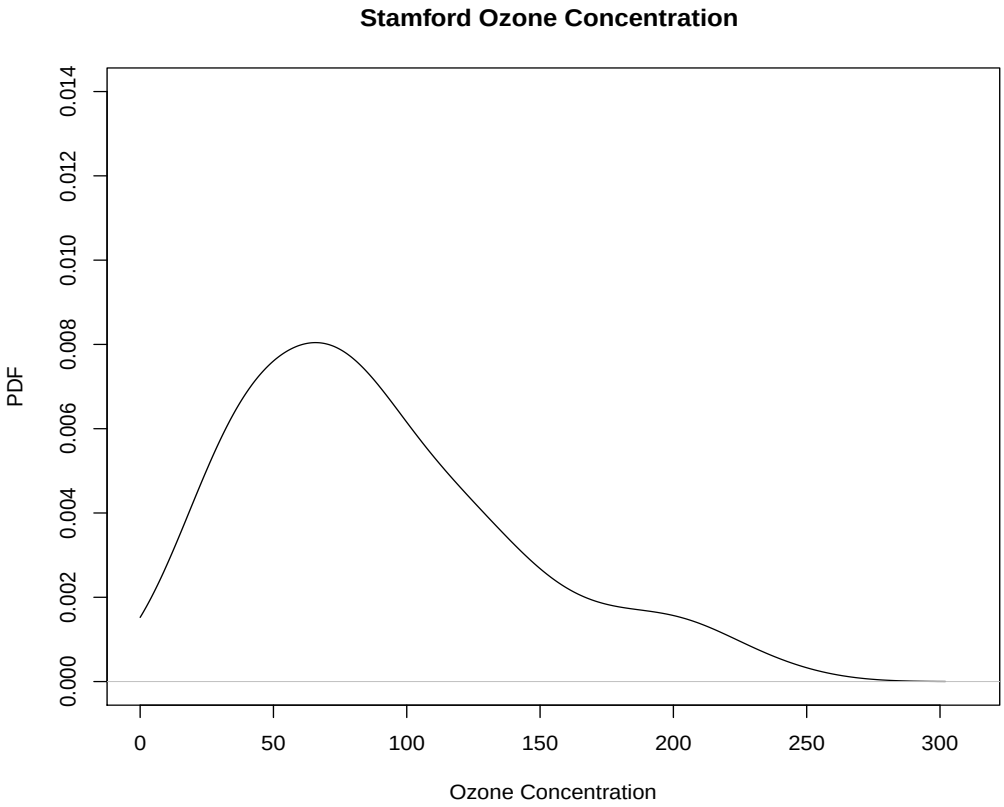


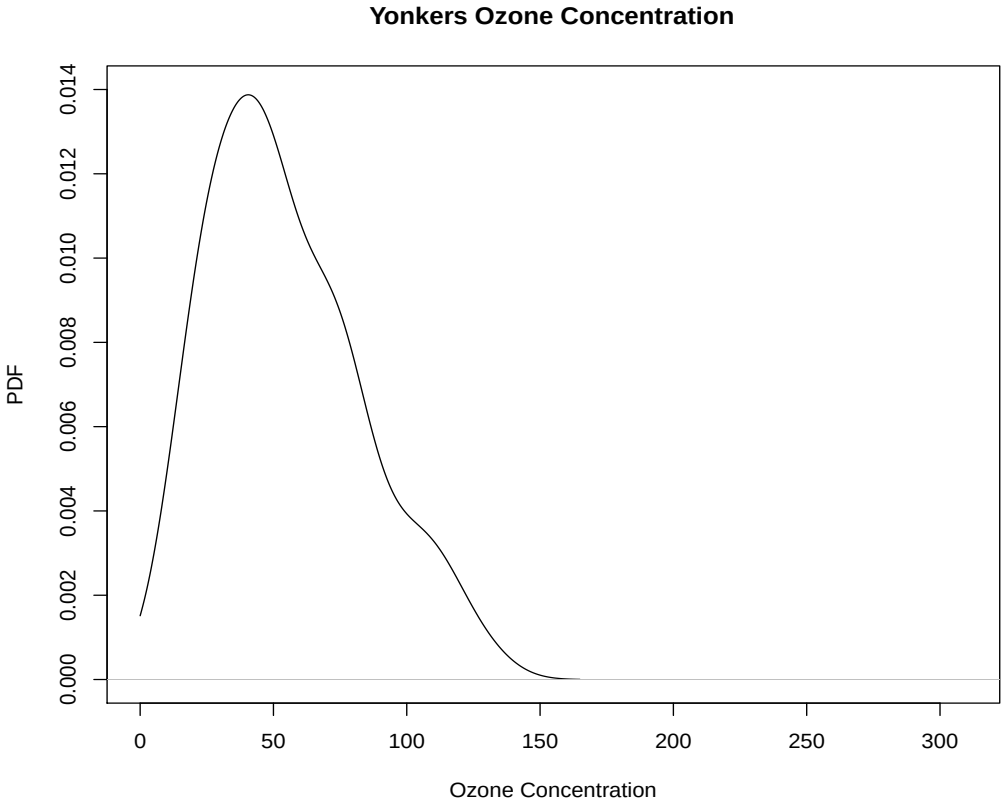
The computation of the sample statistics will be illustrated using the Ozone Data:

Maximum Daily Ozone Concentrations

Daily maximum ozone concentrations in parts per billion (ppb) at ground level recorded between May 1 and September 30, 1974 at sites in Stamford,Connecticut and Yonkers, New York are given below. (There are 17 missing days ofdata at Stamford and 5 at Yonkers due to equipment malfunction.) The current federal standard for ozone states that the concentration should not exceed 120 ppb more than one day per year at any particular location. A day with ozone concentration above 220 ppb is regarded as heavily polluted.

May		June		July		August		September	
Stmf	Ykrs	Stmf	Ykrs	Stmf	Ykrs	Stmf	Ykrs	Stmf	Ykrs
66	47	61	36	152	76	80	66	113	66
52	37	47	24	201	108	68	82	38	18
—	27	—	52	134	85	24	47	38	25
—	37	196	88	206	96	24	28	28	14
—	38	131	111	92	48	82	44	52	27
—	—	173	117	101	60	100	55	14	9
49	45	37	31	119	54	55	34	38	16
64	52	47	37	124	71	91	60	94	67
68	51	215	93	133	—	87	70	89	74
26	22	230	106	83	50	64	41	99	74
86	27	—	49	—	27	—	67	150	75
52	25	69	64	60	37	—	127	146	74
43	—	98	83	124	47	170	96	113	42
75	55	125	97	142	71	—	56	38	—
87	72	94	79	124	46	86	54	66	38
188	132	72	36	64	41	202	100	38	23
118	—	72	51	75	49	71	44	80	50
103	106	125	75	103	59	85	44	80	34
82	42	143	104	—	53	122	75	99	58
71	45	192	107	46	25	155	86	71	35
103	80	—	56	68	45	80	70	42	24
240	107	122	68	—	78	71	53	52	27
31	21	32	19	87	40	28	36	33	17
40	50	114	67	27	13	212	117	38	21
47	31	32	20	—	25	80	43	24	14
51	37	23	35	73	46	24	27	61	32
31	19	71	30	59	62	80	77	108	51
47	33	38	31	119	80	169	75	38	15
14	22	136	81	64	39	174	87	28	21
—	67	169	119	—	70	141	47	—	18
71	45			111	74	202	114		


Stamford Ozone Concentration
PDF
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0
50
100
150
200
250
300
Ozone Concentration


Yonkers Ozone Concentration
PDF
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0
50
100
150
200
250
300
Ozone Concentration

We will now analyze the ozone data using the following SAS and R code:

```
#The following R code generates various summary statistics for the
#Ozone data. The ozone data is in the files ozone1.DAT and ozone2.DAT
#The following file can be found in ~longneck/meth1/Rfiles/ozonesum.R
#------------------------------------------------------------------------------

#input the data from data files:

          y1  =  scan("u:/meth1/Rfiles/ozone1.DAT")
          y2  =  scan("u:/meth1/Rfiles/ozone2.DAT")
          y1p  =  scan("u:/meth1/Rfiles/ozone1+.DAT")
          y2p  =  scan("u:/meth1/Rfiles/ozone2+.DAT")

#compute summary statistics for Ozone data:

    MeanYkrs  =  mean(y1)
    MeanStmf  =  mean(y2)
    VarYkrs   =  var(y1)
    VarStmf   =  var(y2)
    StDevYkrs  =  sd(y1)
    StDevStmf  =  sd(y2)
    StErrMeanYkrs  =  sqrt(mean(y1)/length(y1))
    StErrMeanStmf  =  sqrt(mean(y2)/length(y2))
    MedianYkrs  =  median(y1)
    MedianStmf  =  median(y2)
    MinYkrs  =  min(y1)
    MinStmf  =  min(y2)
    MaxYkrs  =  max(y1)
    MaxStmf  =  max(y2)
    RangeYkrs  =  max(y1) - min(y1)
    RangeStmf  =  max(y2) - min(y2)
    Q.25Ykrs  =  quantile(y1,.25)
    Q.25Stmf  =  quantile(y2,.25)
    Q.75Ykrs  =  quantile(y1,.75)
    Q.75Stmf  =  quantile(y2,.75)
    IQRYkrs  =  Q.75Ykrs-Q.25Ykrs
    IQRStmf  =  Q.75Stmf-Q.25Stmf
    MadStmf  =  mad(y2)
    MadYkrs  =  mad(y1)

#compute summary statistics for Ozone data with 5 outliers added (1000, 1200, 1500, 2000, 2500):

    MeanYkrsp  =  mean(y1p)
    MeanStmfp  =  mean(y2p)
    VarYkrsp   =  var(y1p)
    VarStmfp   =  var(y2p)
    StDevYkrsp  =  sqrt(var(y1p))
    StDevStmfp  =  sqrt(var(y2p))
    StErrMeanYkrsp  =  sqrt(mean(y1p)/length(y1p))
    StErrMeanStmfp  =  sqrt(mean(y2p)/length(y2p))
    MedianYkrsp  =  median(y1p)
    MedianStmfp  =  median(y2p)
    MinYkrsp  =  min(y2p)
    MinStmfp  =  min(y1p)
    MaxYkrsp  =  max(y1p)
    MaxStmfp  =  max(y2p)
    RangeYkrsp  =  max(y1p) - min(y1p)
    RangeStmfp  =  max(y2p) - min(y2p)
    Q.25Ykrsp  =  quantile(y1p,.25)
    Q.25Stmfp  =  quantile(y2p,.25)
    Q.75Ykrsp  =  quantile(y1p,.75)
    Q.75Stmfp  =  quantile(y2p,.75)
    IQRYkrsp  =  Q.75Ykrsp-Q.25Ykrsp
    IQRStmfp  =  Q.75Stmfp-Q.25Stmfp
    MadStmfp  =  mad(y1p)
    MadYkrsp  =  mad(y2p)
```

```
  SumStat  =  c(MeanYkrs,MeanStmf,StDevYkrs,StDevStmf,MedianYkrs,MedianStmf,
                MinYkrs,MinStmf,MaxYkrs,MaxStmf,Q.25Ykrs,Q.25Stmf,Q.75Ykrs,
                Q.75Stmf,IQRYkrs,IQRStmf,MadYkrs,MadStmf)
  SumStatName  =  c("MeanYkrs","MeanStmf","StDevYkrs","StDevStmf","MedianYkrs",
                   "MedianStmf","MinYkrs","MinStmf","MaxYkrs","MaxStmf",
                   "Q.25Ykrs","Q.25Stmf","Q.75Ykrs","Q.75Stmf","IQRYkrs",
                   "IQRStmf","MadYkrs","MadStmf")
   SumStatp  =  c(MeanYkrsp,MeanStmfp,StDevYkrsp,StDevStmfp,MedianYkrsp,
                MedianStmfp,MinYkrsp,MinStmfp,MaxYkrsp,MaxStmfp,Q.25Ykrsp,
                Q.25Stmfp,Q.75Ykrsp,Q.75Stmfp,IQRYkrsp,IQRStmfp,MadYkrsp,
                MadStmfp)

Round summary statistics to 2 decimal points:

  SumStat  =  round(SumStat,2)

  SumStatp  =  round(SumStatp,2)

  SumStat  =  cbind(SumStatName,SumStat,SumStatp)


 #Output summary statistics to file named SumOzone:

  sink("SumOzone")

  SumStat

  sink()

#     SumStatName  SumStat SumStatp
# ---------------------------------
#     "MeanYkrs"   "89.67"  "144.65"
#     "MeanStmf"   "54.69"  "106.5"
#     "StDevYkrs"  "52.11"  "309.95"
#     "StDevStmf"  "28.11"  "300.92"
#     "MedianYkrs" "80"     "80"
#     "MedianStmf" "49.5"   "50"
#     "MinYkrs"    "14"     "14"
#     "MinStmf"    "9"      "9"
#     "MaxYkrs"    "240"    "2500"
#     "MaxStmf"    "132"    "2500"
#     "Q.25Ykrs"   "48.5"   "51"
#     "Q.25Stmf"   "33.75"  "34"
#     "Q.75Ykrs"   "119.75" "124"
#     "Q.75Stmf"   "74"     "75"
#     "IQRYkrs"    "71.25"  "73"
#     "IQRStmf"    "40.25"  "41"
#     "MadYkrs"    "30.39"  "31.13"
#     "MadStmf"    "49.67"  "54.86"
```

```
* SAS program to obtain summary statistics for ozone data. ;


option ls=75 ps=55 nocenter nodate;
title 'Ozone Concentration';
data OZONE1;
   infile 'C:\sasdata\ozone1.DAT';           * input data;
      input ozone1 @@;
data OZONE2;
   infile 'C:\sasdata\ozone2.DAT';           * input data;
      input ozone2 @@;
data ozone;
      merge ozone1 ozone2;                              * combine data sets;

   label OZONE1='Ozone Level in Stamford'
         OZONE2='Ozone Level in Yonkers';

run;
proc print;
run;

* generates various summary statistics and plots;

proc univariate plot normal def=5;
   var ozone1 ozone2;

* create data set with ozone values in one column
  and city name in second column;

data ozonebox1;
set ozone1;
ozone=ozone1;drop ozone1;
run;
data ozonebox2;
set ozone2;
ozone=ozone2;drop ozone2;
run;
data ozonebox;
set ozonebox1 ozonebox2;
if _n_<137 then city="Stamford";
if _n_>=137 then city="Yonkers";drop obs;
run;

* creates side-by-side box plots;

proc boxplot;
   plot ozone*city/boxstyle=schematic;
run;
```

```
The UNIVARIATE Procedure
Variable:  ozone1  (Ozone Level in Stamford)

                          Moments

N                         136    Sum Weights               136
Mean                89.6691176    Sum Observations        12195
Std Deviation       52.1074467    Variance              2715.186
Skewness            0.88432102    Kurtosis            0.18080786
Uncorrected SS         1460065    Corrected SS         366550.11
Coeff Variation     58.1108057    Std Error Mean      4.46817669

    Location                  Variability              Quantile      Estimate
                         100% Max         240.0
Mean      89.66912    Std Deviation           52.10745   75% Q3           120.5
Median    80.00000    Variance                    2715   50% Median        80.0
Mode      38.00000    Range                  226.00000   25% Q1            48.0
                      Interquartile Range     72.50000   0% Min            14.0

                  Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.928464    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.116014    Pr > D     <0.0100
Cramer-von Mises      W-Sq  0.422695    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  2.754894    Pr > A-Sq  <0.0050


   Stem Leaf                     #  Boxplot
     24 0                        1     0
     23 0                        1     0
     22
     21 25                       2     |
     20 1226                     4     |
     19 26                       2     |
     18 8                        1     |
     17 034                      3     |has has
     16 99                       2     |
     15 025                      3     |
     14 1236                     4     |
     13 1346                     4     |
     12 2244455                  7  +-----+
     11 1334899                  7  |     |
     10 013338                   6  |     |
      9 1244899                  7  |  +  |
      8 0000002235667779        16  *-----*
      7 11111122355             11  |     |
      6 0114444668889           13  |     |
      5 1222259                  7  |     |
      4 023677779                9  +-----+
      3 11223788888888          14     |
      2 3444467888              10     |
      1 44                       2     |
        ----+----+----+----+
    Multiply Stem.Leaf by 10**+1
```

```
The UNIVARIATE Procedure
Variable:  ozone2  (Ozone Level in Yonkers)

                          Moments

N                         148    Sum Weights               148
Mean                54.6891892    Sum Observations         8094
Std Deviation       28.1148885    Variance            790.446957
Skewness            0.65601648    Kurtosis            -0.2458541
Uncorrected SS          558850    Corrected SS        116195.703
Coeff Variation      51.408494    Std Error Mean       2.3110296

    Location                    Variability              Quantile      Estimate

Mean      54.68919     Std Deviation 28.11489            100% Max         132.0
Median    49.50000     Variance               790.44696   75% Q3            74.0
Mode      27.00000     Range                  123.00000   50% Median        49.5
                       Interquartile Range     40.50000   25% Q1            33.5
                              0% Min             9.0

                  Tests for Normality

Test                 --Statistic---    -----p Value------

Shapiro-Wilk         W     0.952488    Pr < W     <0.0001
Kolmogorov-Smirnov   D     0.092739    Pr > D     <0.0100
Cramer-von Mises     W-Sq  0.296356    Pr > W-Sq  <0.0050
Anderson-Darling     A-Sq  1.927373    Pr > A-Sq  <0.0050

   Stem Leaf                     #  Boxplot
     13 2                        1     |
     12 7                        1     |
     12                          0     |
     11 779                      3     |
     11 14                       2     |
     10 66778                    5     |
     10 04                       2     |
      9 667                      3     |
      9 3                        1     |
      8 5678                     4     |
      8 00123                    5     |
      7 55556789                 8     |
      7 0001124444              10  +-----+
      6 6677778                  7  |     |
      6 0024                     4  |     |
      5 556689                   6  |  +  |
      5 000111223344            12  *-----*
      4 5555667777899           13  |     |
      4 011223444                9  |     |
      3 5566677777889           13  |     |
      3 01112344                 8  +-----+
      2 55557777778             11     |
      2 011122344                9     |
      1 5678899                  7     |
      1 344                      3     |
      0 9                        1     |
        ----+----+----+----+
    Multiply Stem.Leaf by 10**+1
```

```
#The following R code will be used to
#calculate the m-estimator of a location parameter using 30 iterations

# mest1.R:
x  =  scan("u:/meth1/ozone1.DAT")
t  =  1.345
mx  =  median(x)
y  =  abs(x-mx)
my  =  median(y)
v  =  my/.6745
n  =  length(x)
k  =  30
r  =  k+1
mest1  =  matrix(0,r,1)
s =  matrix(0,r,1)
p =  matrix(0,r,1)
mest1[1]  =  median(x)
w  =  matrix(0,n,1)
for(j in 1:k)
{
m  =  mest1[j]
for(i in 1:n)
{
if               (x[i] < m-t*v)  w[i]  =  -t*v/(x[i]-m)
if (m-t*v <= x[i] && x[i] <= m+t*v)      w[i]  =  1
if               (x[i] > m+t*v)  w[i]  =  t*v/(x[i]-m)
}
s[j]  =  sum(w)
p[j]  =  t(x)%*%w
mest1[j+1]  =  p[j]/s[j]
}



#-----------------------------------------------------------
#
#The following R code will be used to
#calculate the m-estimator of a location parameter using iterations
#until a specified (1.e-8) degree of precision is achieved.
#-------------------------------------------------------------------------
#mest2.s:

x  =  scan("u:/meth1/ozone1.DAT")
t  =  1.345
mx  =  median(x)
y  =  abs(x-mx)
mad  =  median(y)
v  =  mad/.6745
n  =  length(x)
mest2  =  median(x)
m  =  mest2
lastm  =  2*m
w  =  matrix(0,n,1)
nit  =  1
while(abs((mest2-lastm)/lastm)>1.e-8)
{
lastm  =  mest2
m  =  mest2
for(i in 1:n)
{
if               (x[i] < m-t*v)  w[i]  =  -t*v/(x[i]-m)
if (m-t*v <= x[i] && x[i] <= m+t*v)      w[i]  =  1
if               (x[i] > m+t*v)  w[i]  =  t*v/(x[i]-m)
}
s  =  sum(w)
p  =  t(x)%*%w
mest2  =  p/s
```

```
nit  =  nit+1

}
```

OUTPUT FROM mest1.s and mest2.s

```
         Stamford Ozone                          Yonkers Ozone

        mest1       mest2                   mest1           mest2
 [1,] 80.00000     84.39501           [1,] 49.50000      52.88504
 [2,] 83.93129      nit=9             [2,] 52.56577        nit=9
 [3,] 84.34605                        [3,] 52.85504
 [4,] 84.38983                        [4,] 52.88221
 [5,] 84.39446                        [5,] 52.88477
 [6,] 84.39495                        [6,] 52.88501
 [7,] 84.39501                        [7,] 52.88503
 [8,] 84.39501                        [8,] 52.88504
 [9,] 84.39501                        [9,] 52.88504
[10,] 84.39501                       [10,] 52.88504
...............                      [29,] ........
[30,] 84.39501                       [30,] 52.88504
```

With the outliers(1000, 1200, 1500, 2000, 2500) added to the data,
we obtain:

```
        Stamford     Yonkers
------------------------------
mest1   88.3884      54.61086
mest2   88.3884      54.61086
nit       9            9
```

mest1 was specified to quit after 30 iterations
mest2 took 9 iterations to achieve convergence

Comparison of Summary Statistics for Ozone Data

Estimators of Center of Data Set				
	Yonkers		Stamford	
Data Set	Without	With	Without	With
Sample Size: n	148	153	136	141
Mean: $\bar{Y}$	54.69	106.50	89.67	144.65
$\%Y_{i's} \leq \bar{Y}$	58.1%	90.2%	67.6%	82.3%
Median: M	49.0	50.0	80.0	80.0
$\%Y_{i's} \leq M$	50.0%	50.3%	52.9%	51.1%
5% Trimmed Mean: $T_{.05}$	53.75	55.66	86.4	91.17
$\%Y_{i's} \leq T_{.05}$	56.7%	58.8%	57.3%	58.9%
M-EST: $MEST$	52.89	54.61	84.33	88.40
$\%Y_{i's} \leq MEST$	55.4%	56.2%	55.1%	57.4%
Estimators of Level of Dispersion in Data Set				
	Yonkers		Stamford	
Data Set	Without	With	Without	With
Sample Size: n	148	153	136	141
Range: R	123	2491	226	2486
Semi-Interquarile Range: $SIQR$	20.38	20.50	36.75	36.50
Standard Deviation: S	28.11	300.90	52.11	310.00
MAD_3	30.39	31.13	49.67	54.86
$Skewness$	0.66	6.16	0.88	5.76
$Kurtosis$	-0.25	39.94	0.18	35.53
Five Number Summaries of Data				
	Yonkers		Stamford	
Data Set	Without	With	Without	With
Minimum	9	9	14	14
Q(.25)	33.75	34	48.5	51
Q(.5)	49.5	50	80	80
Q(.75)	74	75	119.75	124
Maximum	132	2500	240	2500

The Without Column refers to the original data set

The With Column data set with 5 large values added: 1000, 1200, 1500, 2000, 2500

Selecting a Measure of Center and Dispersion about Center

Why not just use the pair $(\widehat{Q(.5)}, \widehat{MAD})$ for all data sets?

When the data is not heavily skewed or has only a few outliers (near normal in shape), $(\widehat{\mu}, \widehat{\sigma})$ provide a more complete picture of the distribution. Furthermore, $(\widehat{\mu}, \widehat{\sigma})$ are more efficient estimators than are $(\widehat{Q(.5)}, \widehat{MAD})$. We will discuss these comments in a later handout and in STAT 611. We will now consider some recommendations on how to select an estimator for a broad class of distributions.

Robust Estimation of Location and Scale

A robust estimator should be relatively unaffected by two types of anomalies that are often encountered in data sets:

1. A few outliers - values that are large relative to the other data values
2. Many relatively small deviations in the data which may occur due to rounding or grouping of the data

There are three broad classes of robust estimators of location and scale:

1. R-Estimators: Estimators based on linear combinations of the ranks of the data values - Rank based regressions
2. M-Estimators: Estimators which minimize an objective function - Least Squares Estimators
3. L-Estimators: Estimators which are linear combinations of the order statistics

A comparison of a number of L-Estimators is given in the book, *Understanding Robust and Exploratory Data Analysis*, by D. Hoaglin, F. Mosteller, and J. Tukey. All of the following estimators have symmetric coefficients and hence are unbiased estimators of a location parameter.

1. α-Trimmed Mean:

$$T(\alpha) = \frac{1}{n(1-2\alpha)} \left(pY_{([n\alpha]+1)} + pY_{(n-[n\alpha])} + \sum_{i=[n\alpha+2]}^{n-[n\alpha]-1} Y_{(i)} \right)$$

   where $p = 1 + [n\alpha] - n\alpha$. This definition is slightly modified from our definition in order to better approximate trimming exactly $100\alpha\%$ from each tail of the data. When $n\alpha$ is an integer, the definitions agree. When $n\alpha$ is not an integer, we are only using $100p\%$ of the smallest and largest untrimmed data values in the average.

   Example Suppose $n = 30,\ \alpha = .05,\ n\alpha = 1.5,\ p = 1 + [1.5] - 1.5 = .5$. Thus, trim $Y_{(1)}$ and $Y_{(30)}$ and partially trim $Y_{(2)}$ and $Y_{(29)}$ yielding

$$T(.05) = \frac{1}{27} \left(.5Y_{(2)} + .5Y_{(29)} + \sum_{i=3}^{28} Y_{(i)} \right)$$

2. Mean: The average of all n data values, a 0% trimmed mean.

3. MidMean: The average of the central half of the order statistics. This is just $T(.25)$.

4. Median (M): A variably trimmed mean with trimming proportion equal to $\alpha_n = \frac{1}{2} - \frac{1}{2n}$

5. Trimean (TRI): $TRI = \frac{1}{4}\left(\widehat{Q}(.25) + 2M + \widehat{Q}(.75)\right)$

6. Best Linear Unbiased Estimator (BLUE): The linear combination of the order statistics (L-estimator) which is unbiased and has smallest variance among all L-estimators.

The goal is to evaluate the above estimators of location over a broad class of symmetric distributions. In order to compare the mass in the tails of the distributions relative to the normal distribution, we will define the following index of tail weight in a distribution:

Let F be a symmetric distribution with quantile function Q. Let Q_o be the standard normal quantile function. The Tail Weight Index of a distribution is defined as

$$\tau(F) = \frac{Q(.99) - Q(.5)}{Q(.75) - Q(.5)} \Big/ \frac{Q_o(.99) - Q_o(.5)}{Q_o(.75) - Q_o(.5)}$$

The value of $\tau(F)$ for various distributions is given in the following table:

Dist.	Uniform	Triangular	Normal	CN(.05;3)	Logistic	D-Exp	CN(.05,10)	Slash	Cauchy
$\tau(F)$	.57	.86	1.00	1.20	1.21	1.63	3.42	7.85	9.22

1. CN(α,k) is a contaminated normal distribution: $F(x) = (1-\alpha)\Phi(x) + \alpha F_2(x)$ where Φ is $N(0,1)$ and F_2 is $N(0,k^2)$

2. D-Exp is the Double Exponential Distribution

3. Slash: $F(x) = \Phi(x) - xf(x)$ with $f(x) = \frac{1-e^{-x^2/2}}{x^2\sqrt{2\pi}}$ for $x \neq 0$ and $f(0) = \frac{1}{2\sqrt{2\pi}}$

Two other sampling schemes will be considered in comparing the estimators.

1. One-Out is a sample with one observation from a $N(0,3^2)$ distribution and $n-1$ observations from a $N(0,1)$

2. One-Wild is a sample with one observation from a $N(0,10^2)$ distribution and $n-1$ observations from a $N(0,1)$

The following tables display the variance of the estimators for the above distributions. For each of the following distributions, the preferred estimator is the one having smallest variance.

Table 1: Variance of Estimators based on a sample of size 10

	Distributions and Sampling Situations						
Estimator	Normal	One-Out	Logistic	One-Wild	D-Exp	Slash	Cauchy
Best Trim	.1000	.1295	.0940	.1416	.1403	.6995	.3362
(Trim %)	(0%)	(11%)	(13%)	(16%)	(34%)	(38%)	(40%)
Mean(0%)	.1000	.1800	.1000	1.090	.2000	∞	∞
T(10%)	.1053	.1296	.0943	.1432	.1617	∞	∞
T(20%)	.1133	.1339	.0962	.1433	.1463	.9649	.5377
TriMean	.1136	.1348	.0971	.1448	.1503	1.114	.6348
MidMean(25%)	.1164	.1366	.0975	.1454	.1424	.8389	.4498
T(30%)	.1238	.1438	.1019	.1521	.1408	.7252	.3672
BMED(35%)	.1270	.1472	.1039	.1554	.1404	.7063	.3500
Median(40%)	.1383	.1596	.1120	.1679	.1452	.7048	.3362
BLUE	.1000	*	.0935	*	.1399	*	.3263

Table 2: Variance of Estimators based on a sample of size 20

	Distributions and Sampling Situations						
Estimator	Normal	One-Out	One-Wild	Logistic	Slash	D-Exp	Cauchy
Best Trim	.0500	.0578	.0610	.0464	.3039	.0638	.1357
(Trim %)	(0%)	(6%)	(9%)	(12%)	(34%)	(37%)	(39%)
Mean(0%)	.0500	.0700	.2975	.0500	∞	.1000	∞
T(5%)	.0512	.0578	.0619	.0472	∞	.0854	∞
T(10%)	.0528	.0583	.0611	.0465	.6964	.0778	.4141
T(20%)	.0568	.0616	.0637	.0474	.3579	.0690	.1874
TriMean	.0576	.0625	.0646	.0482	.3849	.0702	.2045
MidMean(25%)	.0593	.0640	.0659	.0486	.3221	.0664	.1591
T(30%)	.0621	.0668	.0686	.0502	.3071	.0647	.1444
BMED(37.5%)	.0664	.0713	.0731	.0530	.3048	.0638	.1361
T(40%)	.0689	.0739	.0757	.0547	.3092	.0644	.1358
Median(45%)	.0734	.0787	.0806	.0581	.3229	.0666	.1395
BLUE	.0500	*	*	.0462	*	*	.1257

For the group of distributions Normal, One-Out, Logistic, One-Wild, Slash, the best overall estimator is the MidMean for n=10 and $T(20\%)$ for n=20.

For the group of distributions One-Out, Logistic, One-Wild, Slash, the best overall estimator is the $T(30\%)$ for n=10 and MidMean(25%) for n=20.

For the group of distributions Normal, One-Out, Logistic, One-Wild, the best overall estimator is the $T(10\%)$ for n=10 and T(5%) for n=20.

Measures of Association Amongst Vectors of R.V.s

We will now define the sample estimators of the correlation coefficient and autocorrelation function:

Pearson Correlation Coefficient

Definition: The Pearson Correlation Coefficient is based on having n independent pairs (Y_i, W_i) of observations on possibly correlated random variables Y and W.

$$r_{Y,W} = \widehat{Corr}(Y,W) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(W_i - \bar{W})}{(s_Y)(s_W)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(W_i - \bar{W})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(W_i - \bar{W})^2}}$$

where s_Y and s_W are the sample standard deviations

1. The Pearson correlation coefficient is a unit-free measure of the linear relationship between the two variables.
2. $-1 \leq r_{Y,W} \leq 1$
3. $r_{Y,W} = \pm 1$ implies $Y_i = \beta_o + \beta_1 W_i$ where the sign of β_1 is the same as the sign of $r_{Y,W}$
4. $r_{Y,W}$ has the limitation of only measuring linear relationship. Thus, higher order relationships may not be detected
5. $r_{Y,W}$ only measures linear relationships between two of the many variables under study. Thus, may fail to detect linear/nonlinear relationships that exist between several of the variables simultaneously.

FIGURE 11.20
Samples of size 1,000 from the bivariate normal distribution

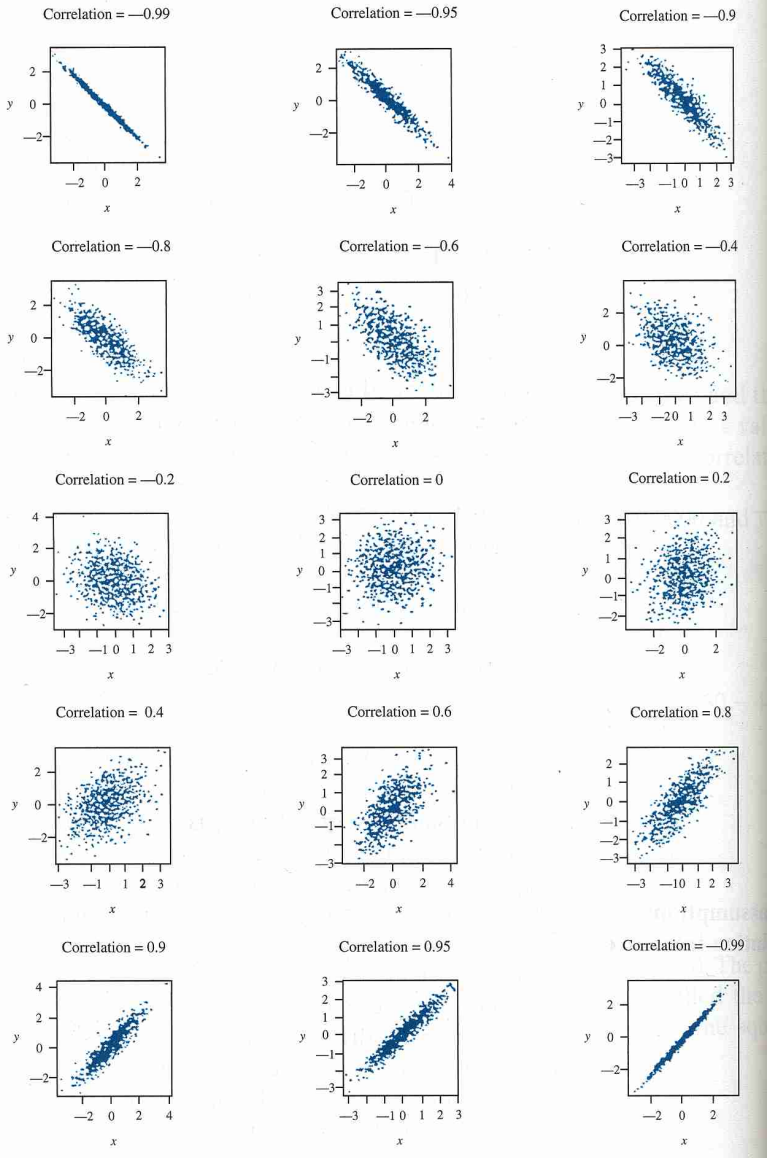


TABLE 3.10 Data for epilepsy study: successive 2-week seizure counts for 59 epileptics. Covariates are adjuvant treatment (0 = placebo, 1 = Progabide), 8-week baseline seizure counts, and age (in years)

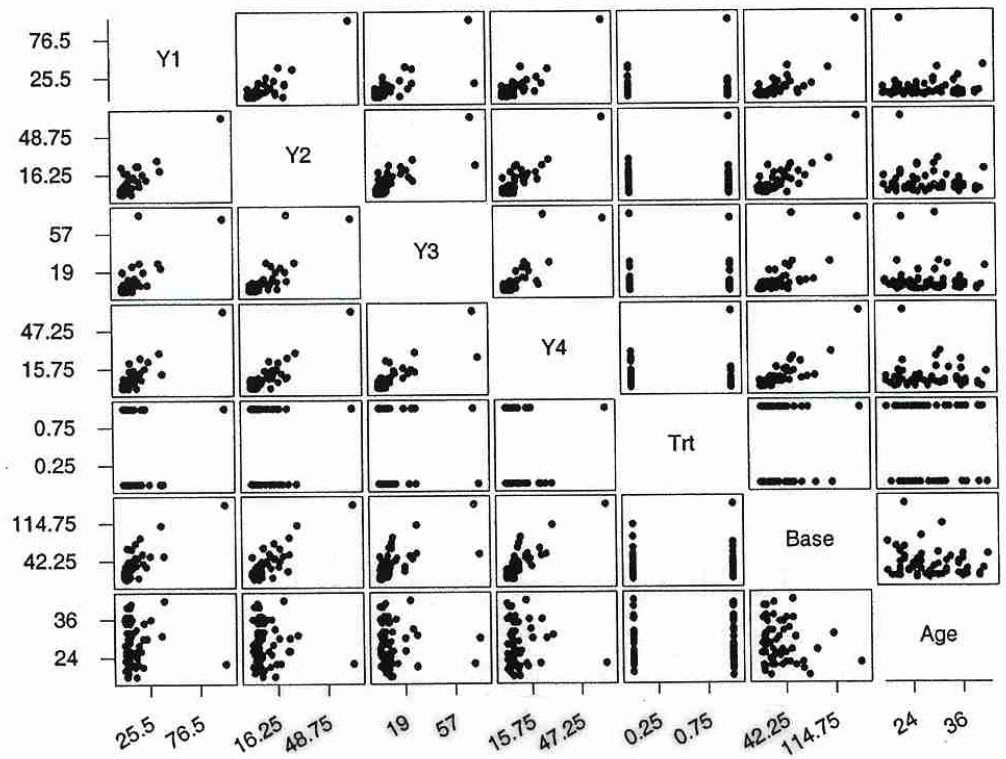
ID	y_1	y_2	y_3	y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

Correlations: Y1, Y2, Y3, Y4, Base, Age

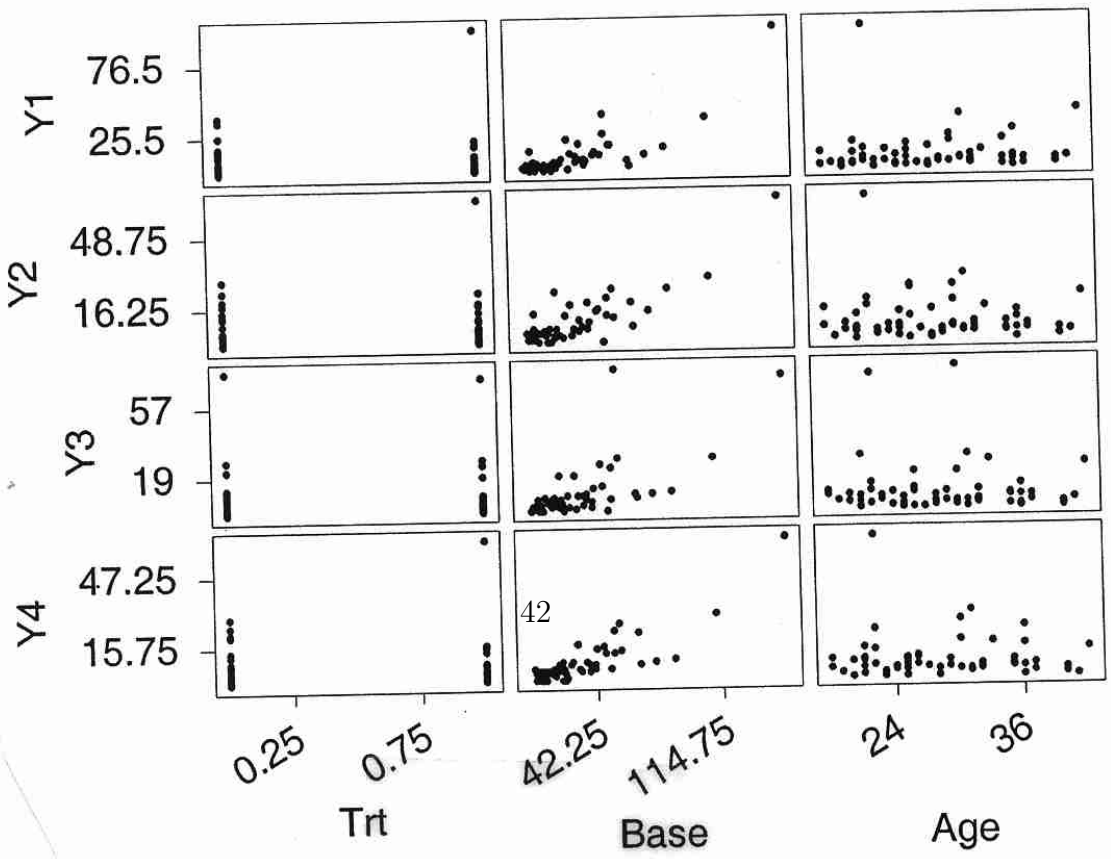
	Y1	Y2	Y3	Y4	Base
Y2	0.871				
	0.000				
Y3	0.738	0.802			
	0.000	0.000			
Y4	0.892	0.895	0.824		
	0.000	0.000	0.000		
Base	0.796	0.831	0.672	0.843	
	0.000	0.000	0.000	0.000	
Age	0.008	-0.116	-0.049	-0.077	-0.181
	0.955	0.384	0.714	0.563	0.171

Cell Contents: Pearson correlation
P-Value

Matrix Plot of Epilpsy Data



Draftsman Plot of Epilpsy Data



Spearman Rank Correlation

When the relationship between the observations on two variables Y and X has a monotone relationship which is not linear, the Pearson correlation coefficient may not reflect the strength of the relationship between Y and X. Also, when there are extreme values in the data set with respect to the x-variable, called influential observations, the value of the correlation coefficient can be inflated. These two cases are illustrated in the following two graphs.

Figure 1: Nonlinear Relationship Y vs X

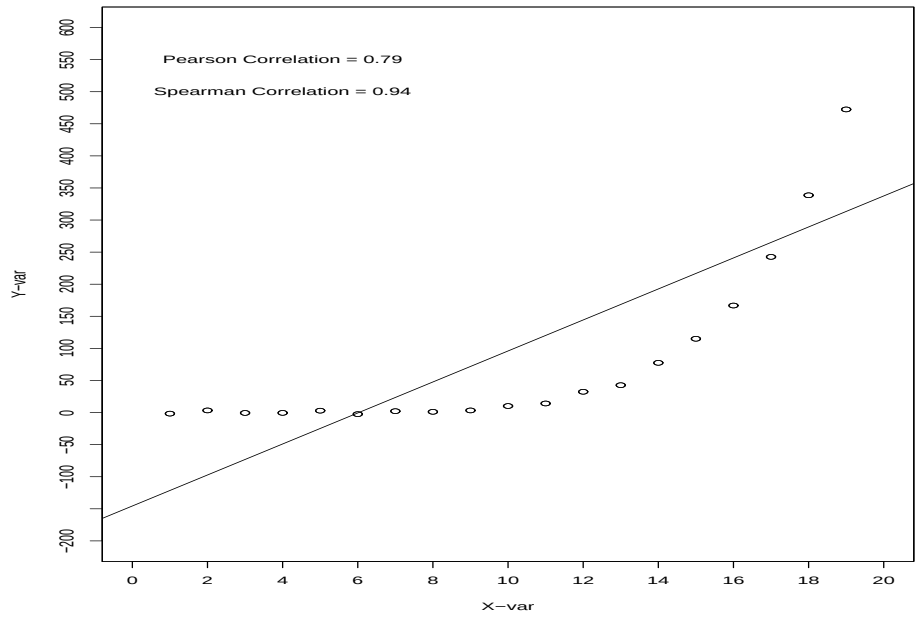
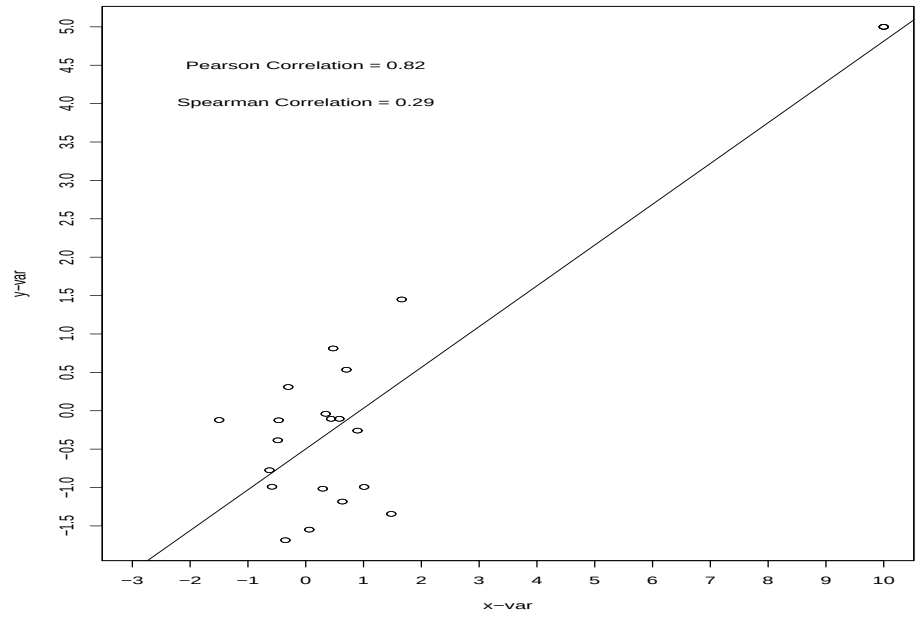


Figure 2: One Influential X Value



Spearman Correlation Coefficient An alternative to the Pearson correlation coefficient is the Spearman Rank Correlation coefficient.

Suppose we have n pairs of observation $(X_i, Y_i), \quad i = 1, \ldots, n$. Let R_i denote the rank of X_i among the Xs and S_i denote the rank of Y_i among the Ys. If there are tied X values, the average rank is assigned to the tied observations. A similar assignment is done for the Ys. The Spearman rank correlation, denoted by r_{sp} is obtained by applying the formula for the Pearson correlation coefficent to the pairs of ranks $(R_i, S_i), \quad i = 1, \ldots, n$.

Definition The Spearman Correlation Coefficient: let $(X_i, Y_i) \;\; i = 1, \ldots, n$ be independent pairs of random variables. Let $(R_i, S_i), \quad i = 1, \ldots, n$ be the corresponding ranks of (X_i, Y_i). The Spearman Correlation Coefficient is given by

$$r_{sp} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{s_R s_S} = \frac{12\sum_1^n R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

If there are no ties in the data or if all the values of R_i and S_i are integers, then we have the following simplification to r_{sp}:

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad \text{where} \quad d_i = R_i - S_i$$

The values of r_{sp} is now compared to r for two data sets given above.

Note that in the case of a nonlinear relationship between X and Y, Figure 1, the Pearson Correlation Coefficient underestimates the strength of the relationship, $r = .79$. The Spearman Correlation Coefficient identifies the strong non-linear relationship between X and Y, $r_{sp} = .94$.

In Figure 2, there is no relationship between X and Y but there is one data value which is very extreme to the other data values in the X-direction. This results in a relatively large value for the Pearson Correlation Coefficient, $r = .82$ whereas the Spearman Correlation Coefficient is small, $r_{sp} = .29$. Thus, reflecting the lack of a relationship between X and Y.

Sample Autocorrelation Function

When we are observing a physical characteristic over time, we are interested in the degree to which these measurements are associated. One measure of this association is the autocorrelation:

Definition: The AutoCorrelation of Order k in a series of stationary random variables: $X_t : t = 1, 2, 3, \ldots$ having the same mean μ and standard deviation σ is given by

$$\rho_k = \frac{E[(X_t - \mu)(X_{t-k} - \mu)]}{\sigma^2} \quad \text{for} \quad k = 1, 2, \ldots$$

ρ_k measures the degree of linear relationship in the random variable X over time or space. ρ_1 the 1st order autocorrelation is the most widely used of these correlations.

A very simple, but widely used model for correlated over time or space observations is the $AR(1)$ model:

$$X_t = \mu + \rho X_{t-1} + e_t,$$

where $e_t s$ are iid with $E[e_t] = 0, Var(e_t) = \sigma^2$, $e_t s$ are independent of the $X_t s$ and $|\rho| < 1$.

Under this model, we can show that the $X_t s$ are not independent and
$\rho_k = \rho^k \to 0$ as $k \to \infty$

The computation of the sample autocorrelation function of lag k is given by

$$\hat{\rho_k} = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

A plot of the time series and autocorrelation functions is given on the next pages. These plots and the computation of the sample autocorrelation function are obtained using the following R code:

```
#The following R code produces plots which take into account the
#time element in the ozone data. The file is named ozone_ts.R
#---------------------------------------------------------------------
#input data:

          y1 = scan("u:/meth1/Rfiles/ozone1.DAT")
          y2 = scan("u:/meth1/Rfiles/ozone2.DAT")

          y1na = scan("u:/meth1/Rfiles/ozone1,na.DAT")
          y2na = scan("u:/meth1/Rfiles/ozone2,na.DAT")

          t1 = c(1:136)
          t2 = c(1:148)

          postscript("u:/meth1/psfiles/ozonetime_Stamford.ps",height=8,horizontal=F)

         plot.ts(y1na,type="b",ylab="Ozone Conc-Stamford (ppb)",xlab="DAY",
          main="Time Series Plot of Stamford Data",cex=.9)
```

```
         abline(h=90)

     postscript("u:/meth1/psfiles/ozonetime_Yonkers.ps",height=8,horizontal=F)

         plot.ts(y2na,type="b",ylab="Ozone Conc-Yonkers (ppb)",xlab="DAY",
          main="Time Series Plot of Yonkers Data",cex=.9)
          abline(h=55)

postscript("u:/meth1/psfiles/ozoneacf_Stamford.ps",height=8,horizontal=F)

         acf_S=acf(y1,main="ACF for Stamford Ozone Concentration")
postscript("u:/meth1/psfiles/ozoneacf_Yonkers.ps",height=8,horizontal=F)

         acf_Y=acf(y2,main="ACF for Yonkers Ozone Concentration")
graphics.off()

         sink("u:/meth1/psfiles/autocorr")
         acf_S = acf(y1,plot=F)
         acf_S
         acf_Y = acf(y2,plot=F)
         acf_Y
        sink()
```

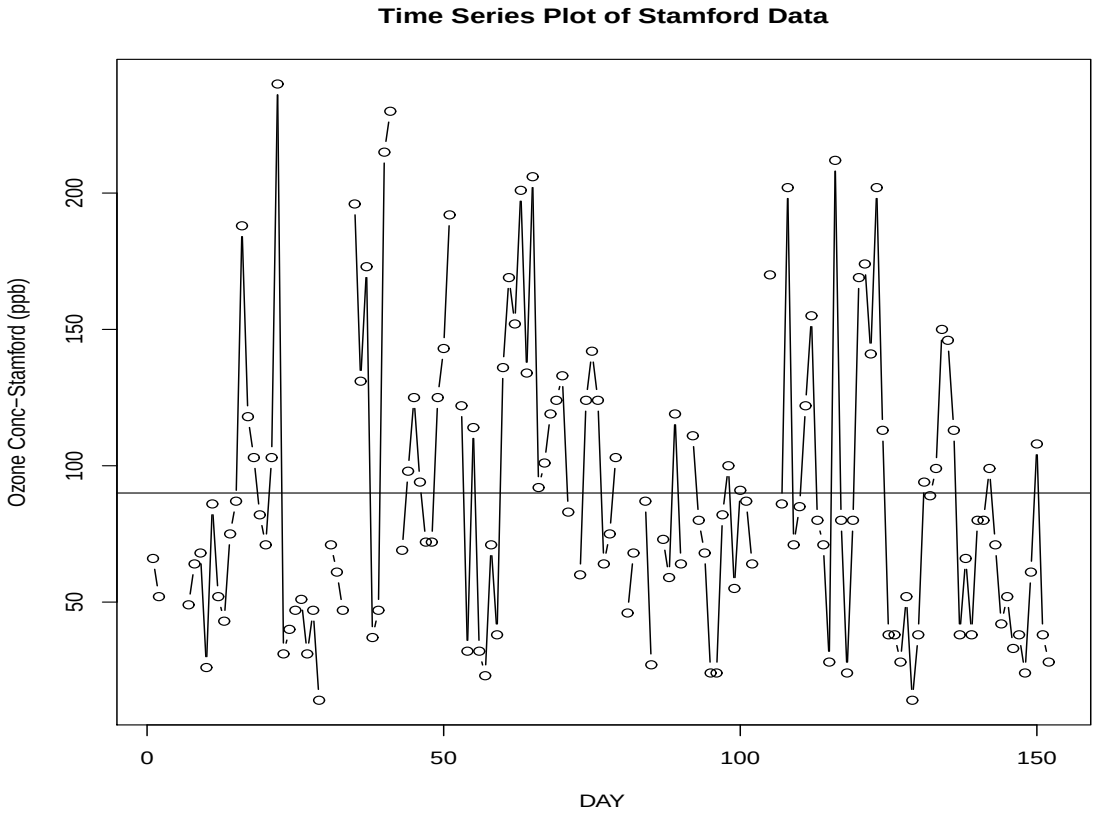
The sample estimators of ρ_k are given in the following table:

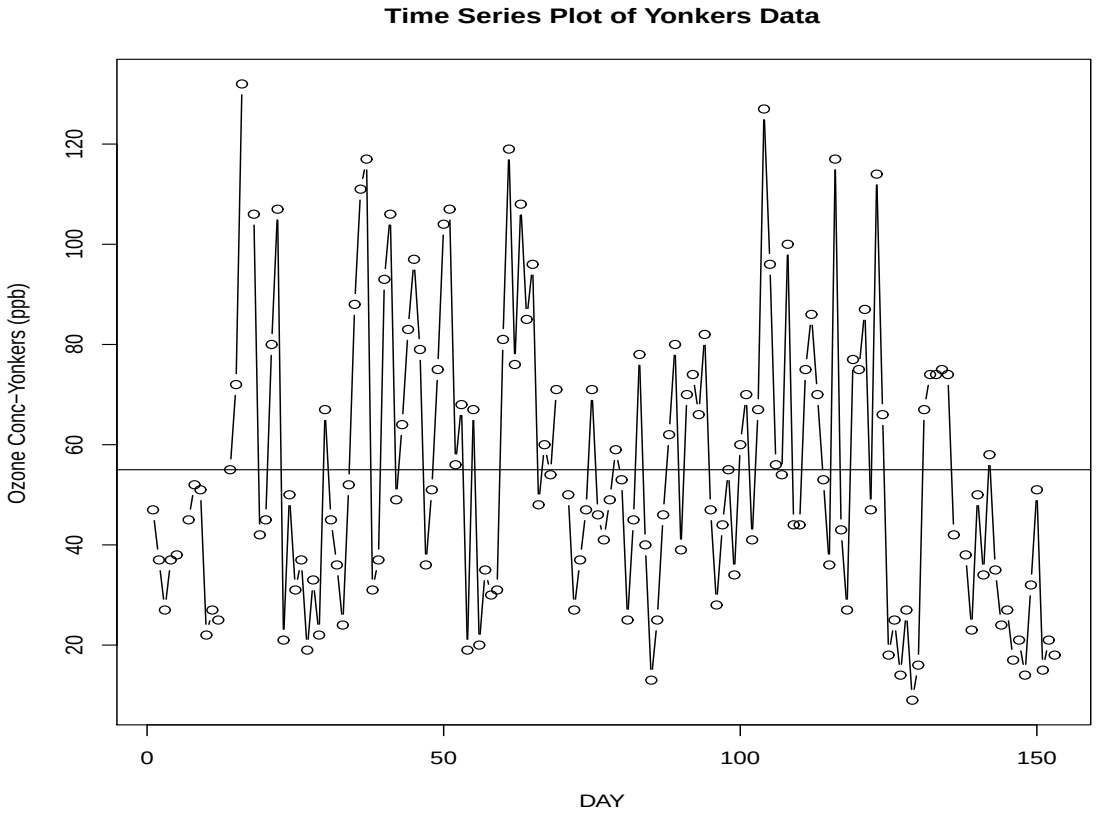
AUTOCORRELATION MATRIX:

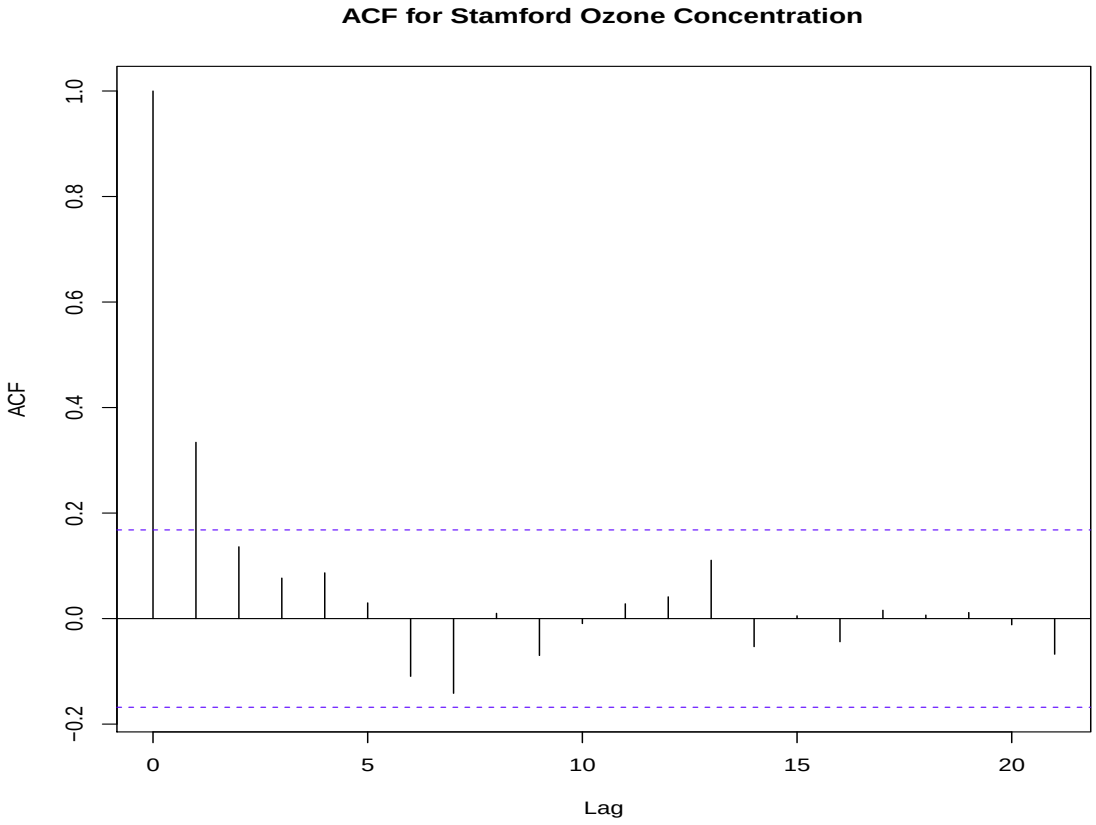
LAG k	YONKERS OZONE rho_k
0	1.0000
1	0.4342
2	0.1352
3	0.0805
4	0.1828
5	0.0621
6	-0.0993
7	-0.0694
8	-0.0130
9	-0.0237
10	0.0008
11	-0.0138
12	0.0385
13	0.0251
14	0.0651
15	0.1280
16	0.0144
17	-0.0690
18	0.0583
19	0.1566
20	0.0553
21	-0.0617

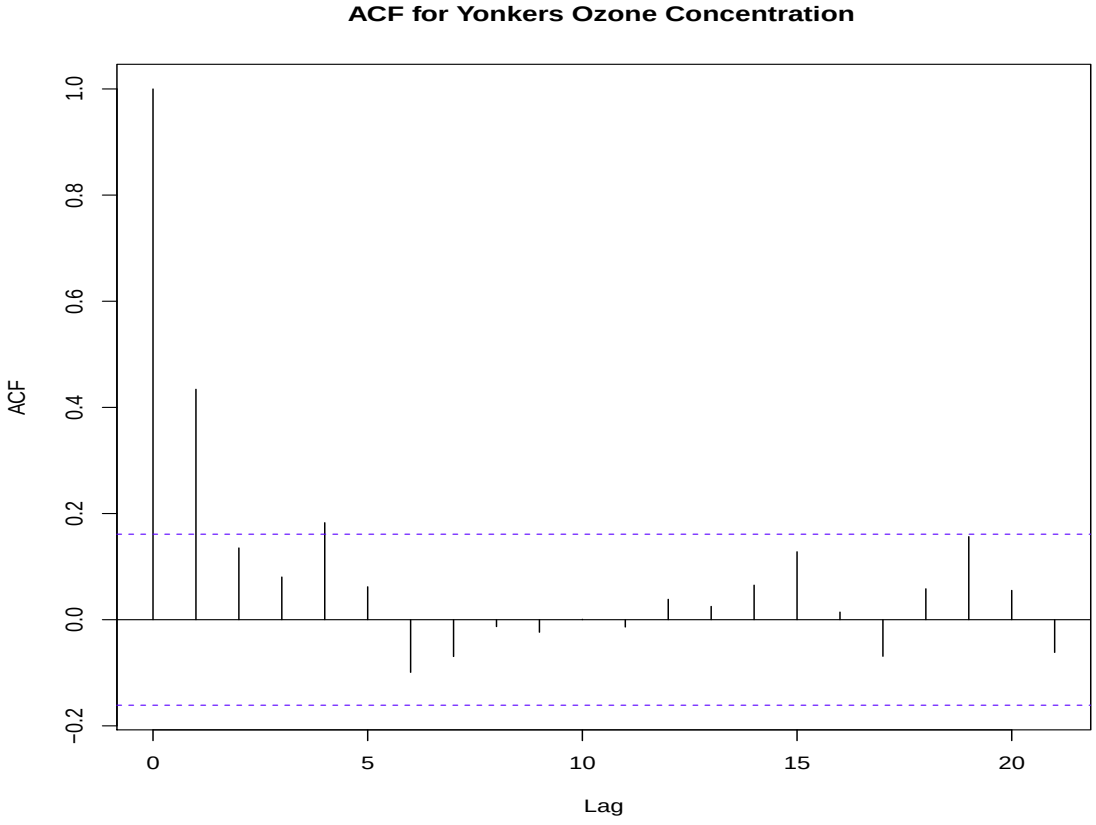
AUTOCORRELATION MATRIX:

LAG k	STAMFORD OZONE rho_k
0	1.0000
1	0.3342
2	0.1361
3	0.0768
4	0.0868
5	0.0298
6	-0.1095
7	-0.1417
8	0.0101
9	-0.0700
10	-0.0095
11	0.0281
12	0.0413
13	0.1106
14	-0.0532
15	0.0054
16	-0.0440
17	0.0159
18	0.0067
19	0.0116
20	-0.0118
21	-0.0676


Time Series Plot of Stamford Data
Ozone Conc−Stamford (ppb)
50
100
150
200
0
50
100
150
DAY


Time Series Plot of Yonkers Data
Ozone Conc−Yonkers (ppb)
20
40
60
80
100
120
0
50
100
150
DAY


ACF for Stamford Ozone Concentration
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
Lag


ACF for Yonkers Ozone Concentration
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
Lag

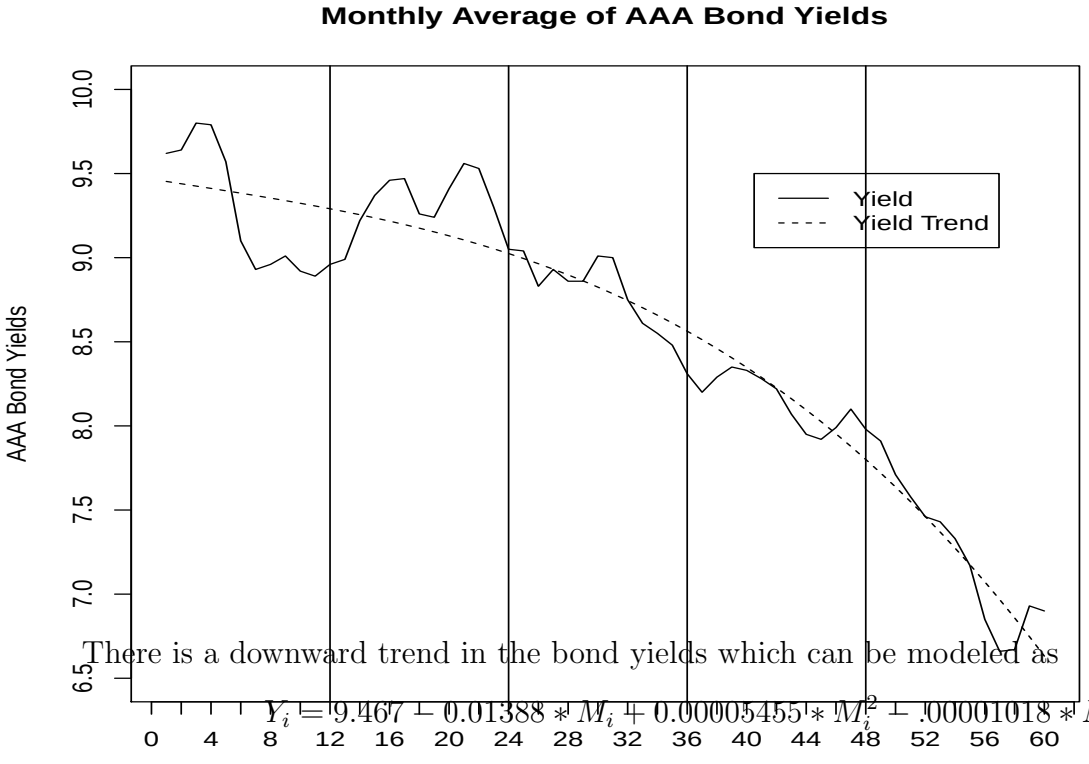
The time series plots for the Stamford and Yonkers ozone data appears to be relatively stationary with no trends over the 5 month summer period.

The following data is the monthly average of daily yields of Moody's AAA bonds for the years 1989 to 1993.

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1989	9.62	9.64	9.80	9.79	9.57	9.10	8.93	8.96	9.01	8.92	8.89	8.96
1990	8.99	9.22	9.37	9.46	9.47	9.26	9.24	9.41	9.56	9.53	9.30	9.05
1991	9.04	8.83	8.93	8.86	8.86	9.01	9.00	8.75	8.61	8.55	8.48	8.31
1992	8.20	8.29	8.35	8.33	8.28	8.22	8.07	7.95	7.92	7.99	8.10	7.98
1993	7.91	7.71	7.58	7.46	7.43	7.33	7.17	6.85	6.66	6.67	6.93	6.90

The researcher was interested in determining if there was a trend and how the mean and variance changed over time.

- Create a time series plot of the data.



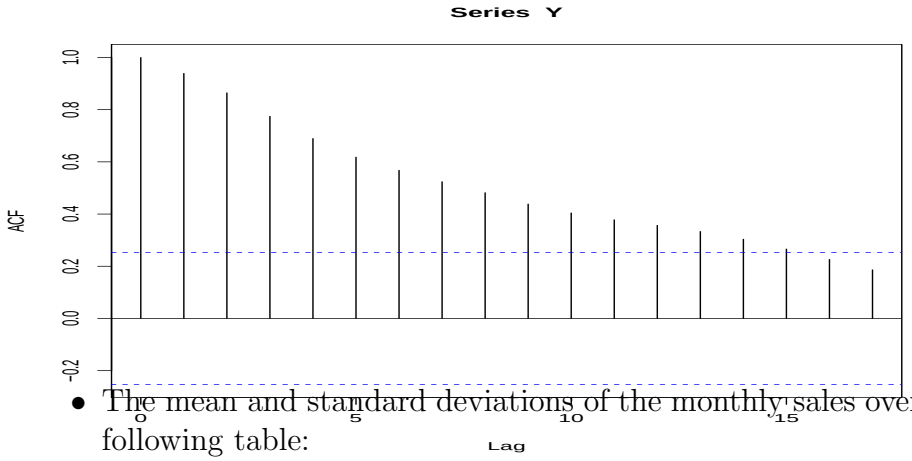
There is a downward trend in the bond yields which can be modeled as

$$Y_i = 9.467 - 0.01388 * M_i + 0.00005455 * M_i^2 - .00001018 * M_i^3$$

where Y_i is the bond yield for Month M_i, $\quad i = 1, 2, \ldots, 60$.

- To evaluate if there is correlation in the data, calculate the values of ρ_k, the autocorrelation coefficients. The lag k autocorrelations are given below. Based on these correlations and the plot it would appear that the adjacent monthly sales have a strong positive correlation. There appears to be a very slow decline in the autocorrelations with a pattern such as $\rho_k = (\rho_1)^k = (.939)^k$ for $k = 1, 2, \cdots, 17$, as would be seen in an AR(1) model. This would indicate that the monthly average yields of the AAA bonds are strongly correlated.

i	0	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}_i$	1.000	0.939	0.864	0.775	0.690	0.618	0.568	0.524	0.482	0.438	0.404
i	11	12	13	14	15	16	17				
$\hat{\rho}_i$	0.378	0.357	0.333	0.304	0.266	0.226	.187				



- The mean and standard deviations of the monthly sales over the five years are given in the following table:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Mean	8.752	8.738	8.806	8.780	8.722	8.584	8.482	8.384	8.352	8.332	8.340	8.240
St.Dev.	0.690	0.760	0.871	0.927	0.889	0.808	0.857	1.007	1.119	1.085	0.907	0.872

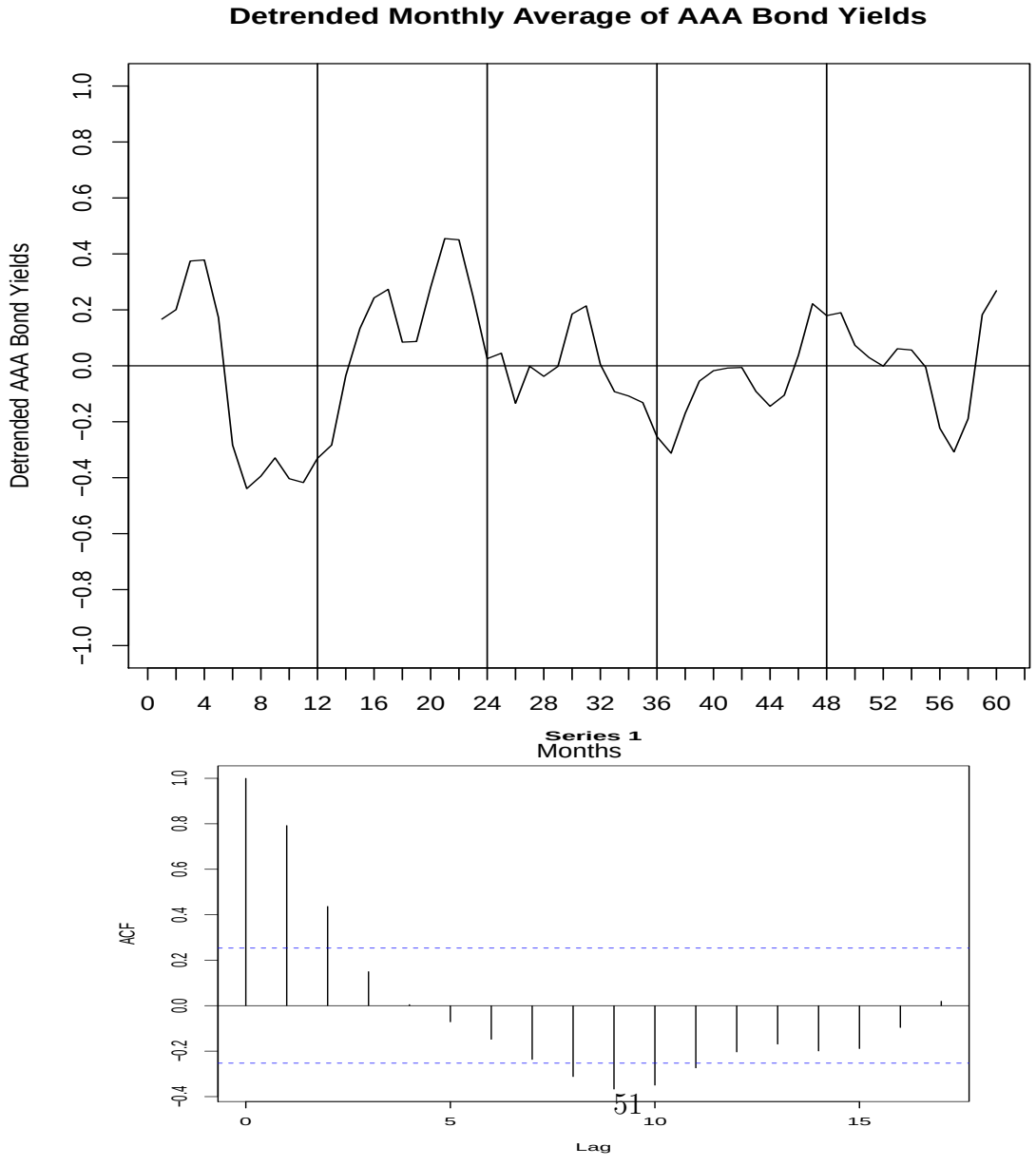
The sales appear non-stationary with an overall decline in yields along with a somewhat cyclic behavior over the five years. However, the monthly means and standard deviations over the five years are somewhat stable with a pattern of higher values for January through May and then lower values through the remaining months.

The analysis of the monthly yields is somewhat misleading if the trend is not taken into account.

If the data is detrended by defining

$$Z_i = Y_i - (9.467 - 0.01388 * M_i + 0.00005455 * M_i^2 - .00001018 * M_i^3)$$

The plot of Z_i and the autocorrelation function for the Z_i are given below:



i	0	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}_i$	1.000	0.792	0.437	0.150	0.005	-0.071	-0.148	-0.237	-0.312	-0.367	-0.350
i	11	12	13	14	15	16	17				
$\hat{\rho}_i$	-0.274	-0.204	-0.169	-0.199	-0.189	-0.096	0.020				

Not surprising, that there is still a strong correlation in the detrended data with $\rho_1 = 0.792$ and significant values for $\rho_1, \rho_2, \rho_8, \rho_9, \rho_{10}, \rho_{11}$.

- The mean and standard deviations of the detrended monthly sales over the five years are given in the following table:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Mean	-0.039	-.013	0.096	0.113	0.099	0.007	-0.047	-0.095	-0.076	-0.043	-0.021	-0.022
St.Dev.	0.243	0.152	0.170	0.187	0.121	0.177	0.247	0.255	0.316	0.318	0.289	0.263

The detrended monthly sales now appear to be relatively stationary over the five years. The regression line relating Z_i to M_i is $Z_i = .00024 + .000005 * M_i$ that is essentially correlated noise about a horizontal line through 0. Also, the monthly means and standard deviations over the five years are relatively stable considering that the values are based on only 5 data values each.

The difficult phase of the analysis of time series data is the modeling of the correlation. There are many different types of models autoregressive (AR), moving averages (MA), a combination of AR and MA (ARMA), and many more. I refer you to STAT 626 which is taught every summer.