HANDOUT #8: GRAPHICAL SUMMARIES OF DATA AND COMPARISON GRAPHS

TYPES OF GRAPHS

1. Quantile-Reference Distribution Plots
2. Quantile-Quantile Plots
3. Quantile Plots for Mixture Distributions
4. Mixtures of Normal Distributions
5. Comparison Plots: Q-Q plots, Side-by-Side Box Plots, etc.
6. Times Series Plots
7. Stacked Bar Plots: Relationship between Two Categorical Variables
8. Side-by-Side Box Plots: Relationship between Categorical and Continuous variables
9. Matrix and Draftsmans Plots: Relationship between Many Variables
10. Fitted lines to Scatter Plots

Reference Distribution Plots:

Given a random sample from a population: $Y_1, Y_2, \cdots, Y_n$ with cdf F, we often want to evaluate whether or not the cdf F has some specified form, F_o. For example, is F from a normal family or Weibull family. We will consider several cases. Case 1 will have F_o completely specified with no unspecified parameters, Case 2 will have F_o stated to be of a particular family but the parameters are unspecified, and Case 3 will have F_o from a location family, or scale family, or location-scale family. In all three cases, we will use the following graphical representation of the data to evaluate whether or not F is well represented by F_o.

Let Q_o be the quantile function associated with F_o and $\widehat{Q}$ be the sample quantile.

A Quantile-Quantile $(Q-Q)$ Plot has Q_o on the horizontal axis and $\widehat{Q}$ on the vertical axis. The n plotted points are

$$\left(Q_o(u_i), \widehat{Q}(u_i)\right), \quad \text{for} \quad u_i = \frac{i-.5}{n}, \quad i = 1, \ldots, n$$

(Note: $\widehat{Q}(u_i) = Y_{(i)}$ for our version of the sample quantile function.)

Case 1: F_o Completely Specified

Suppose F_o is completely specified with no unknown parameters. For example, F_o is a Gamma distribution with $\alpha = 2.3$ and $\beta = 4.5$ or F_o is Poisson with $\lambda = 5$. If the n plotted points in the $Q-Q$ plot are very close to a 45° line through the origin, then we would infer that F is equal to (or is very well approximated by) F_o.

Example: Suppose we have $n = 100$ observations on a process having cdf F and want to evaluate whether or not F has a Gamma Distribution with $\alpha = 2.3$ and $\beta = 4.5$.

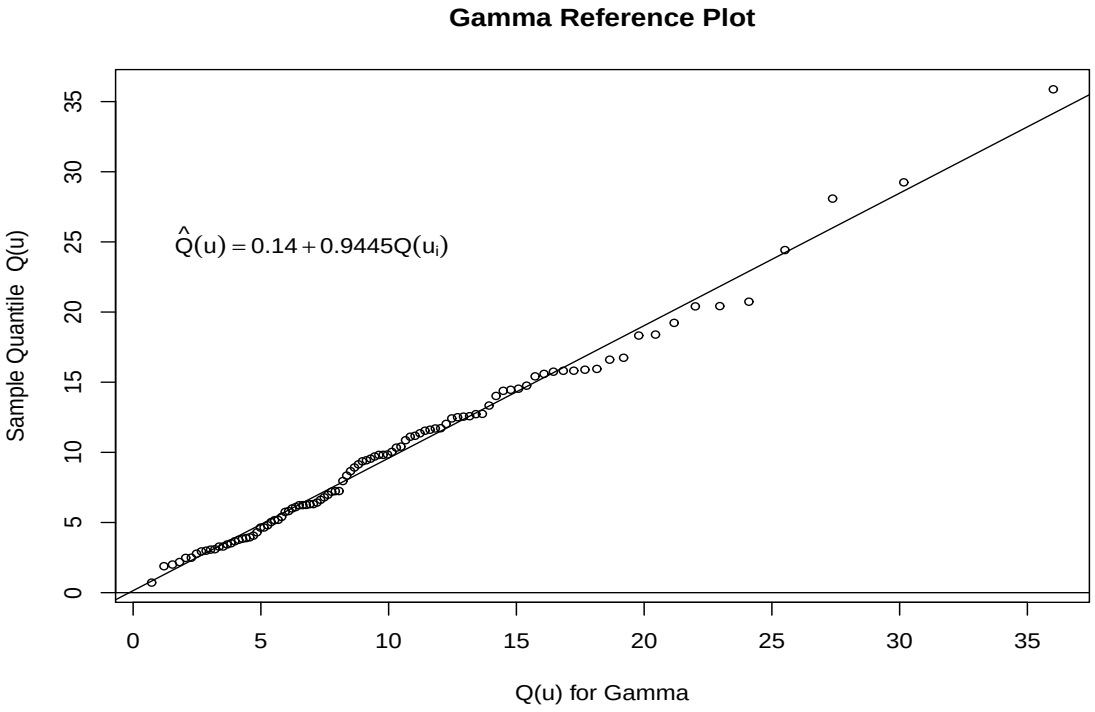
1. plot $\left(Q_o(u_i), Y_{(i)}\right)$

   Note: $Y_{(i)} = \widehat{Q}(u_i)$, where $\widehat{Q}(u_i)$ is the sample quantile evaluated at

   $u_i = (i-.5)/100$ for $i = 1, 2, \ldots, 100$

2. $Q_o(u_i)$ is the Gamma$(\alpha = 2.3, \beta = 4.5)$ quantile function, obtain from R by

3. $u = seq(1/2n, 1-1/2n, 1/n) = seq(.005, .995, .01)$ and $Q_o(u)$=qgamma$(u, 2.3, 1/4.5)$



Case 2: F_o Not Completely Specified

Let Q_o be the quantile function associated with F_o and $\widehat{Q}$ be the sample quantile. Suppose F_o is stated to be from a specific parametric family but the parameters are unspecified. For example, F_o is from a Weibull family but β and λ are not specified. We could just estimate the parameters using the observed data and use these values in our $Q-Q$ plot. However, this could result in varying degrees of inaccuracy in specifying Q_o and could result in a misstatement of the match between the Q_o and the sample quantile $\widehat{Q}$. This approach should be used only if one of the following approaches is not feasible and it should be emphasized in the documentation that you are using estimated parameters in the procedure, not the true parameters.

Case 3: F_o Member of Location-Scale Family

Suppose F_o is stated to be from a specific location family, scale family, or location-scale parametric family but the parameters are unspecified. Furthermore, suppose that the family has no unspecified parameters other than location and/or scale parameters.

For example,

F_o member of Gaussian family with the values of μ and σ not specified.

F_o member of Cauchy family with the values of θ_1 and θ_2 not specified.

In this situation, we will alter our plotting technique by placing the quantile function from the standard member of the family ($\theta_1 = 0,\ \theta_2 = 1$) on the horizontal axis. That is, if Q_Z is the quantile function of the standard member of the family specified for F_o then the plotted points will be

$$\left(Q_Z(u_i), \widehat{Q}(u_i),\right), \quad \text{for} \quad u_i = \frac{i-.5}{n}, \quad i = 1, \ldots, n$$

If the n plotted points in the $Q-Q$ plot are very close to any straight line, then we would infer that F is equal to (or is very well approximated by) F_o.

Why does a plot of $\widehat{Q}(u_i)$ versus $Q_Z(u_i)$ provide us information of the appropriateness of using F_o to model F?

Claim: Let F is a member of a location/scale family with standard member having cdf F_Z and let Q and Q_Z be the corresponding quantiles.

Then

$$Q(u) = \theta_1 + \theta_2 Q_Z(u) \quad \text{for all} \quad 0 \le u \le 1,$$

where θ_1 and θ_2 are the location and scale parameters in the family of distributions.

Proof of Claim: Let Y have cdf F_Y and Z have cdf F_Z, where Y is a member of the location-scale family for which Z is the standard member. For all $0 < u < 1$

$$u = F_Y(Q_Y(u)) = P[Y \le Q_Y(u)] = P\left[\frac{Y-\theta_1}{\theta_2} \le \frac{Q_Y(u)-\theta_1}{\theta_2}\right] = P\left[Z \le \frac{Q_Y(u)-\theta_1}{\theta_2}\right]$$

Therefore,

$$u = P\left[Z \le \frac{Q_Y(u)-\theta_1}{\theta_2}\right] \quad \Rightarrow \quad Q_Z(u) = \frac{Q_Y(u)-\theta_1}{\theta_2} \quad \Rightarrow \quad Q_Y(u) = \theta_1 + \theta_2 Q_Z(u)$$

The Standard Member of any location-scale family retains all information about the family once the values of θ_1 and θ_2 are specified.

Suppose we have $Y_1, Y_2, \ldots, Y_n$ iid with continuous cdf F. We want to evaluate whether or not $F = F_o$, where F_o is a member of a location-scale family of cdfs.

Example # 1: F_o is a member of $N(\mu, \sigma^2)$ with μ and σ both unknown:

1. For $i = 1, 2, \ldots, n$, compute $Q_Z(u_i)$ using standard normal tables or R function
2. Plot the n points, $\left(Q_Z(u_i), Y_{(i)}\right)$ for $u_i = (i - .5)/n$

Using R, $Q_Z(u) = qnorm(u, 0, 1)$ with $u = seq(1/(2*n), 1 - 1/(2*n), 1/n)$

Example # 2: F_o is an Exponential cdf with parameter β

The parameter β is a scale parameter thus use the quantile function for the standard exponential distribution ($\beta = 1$):

The cdf for the standard exponential is $F_Z(z) = 1 - e^{-z} \Rightarrow Q_Z(u) = -log(1-u)$

Plot the n points, $\left(Q_Z(u_i), Y_{(i)}\right) = \left(-log(1-u_i), Y_{(i)}\right)$ for $u_i = (i - .5)/n$

Example # 3: F_o is a Uniform cdf on $[\theta, \theta+1]$

$F_o(x) = x - \theta \qquad$ for $\quad \theta < x < \theta + 1.$

This demonstrates that θ is a location parameter with standard member of the family having cdf

$F_Z(x) = x$ for $0 < x < 1$ and quantile function $Q_Z(u) = u$,

Therefore, we plot the n points, $\left(u_i, Y_{(i)}\right) \quad$ for $\quad u_i = (i - .5)/n$

Sometimes we can reparametrize a non-location-scale families into location-scale families. The next example will illustrate this idea.

Example # 4: F_o is a Weibull cdf with parameters α and γ

$F_o(y) = 1 - e^{-(y/\alpha)^\gamma}$ for $y \geq 0$

From the pdf of the Weibull it is observed that α is a scale parameter but γ is a shape parameter. However, using the following transformation:

$X = log(Y)$ has a location-scale distribution with location/scale parameters:

$\theta_1 = log(\alpha)$ and $\theta_2 = \frac{1}{\gamma}$.

Thus, if we wanted to evaluate if the observed data is from a Weibull distribution, then we could just evaluate whether or not $X = log(Y)$ had a log(Weibull)-distribution.

Determine the quantile function of X, the log-Weibull distribution, and the standard member, Z of the log-Weibull distribution. Finally, plot $\left(Q_Z(u_i), W_{(i)}\right) \quad$ for $\quad u_i = (i - .5)/n$

The cdf and quantile function of the log-Weibull distribution is obtained as follows:

$$F_X(x) = P[X \leq x] = P[log(Y) \leq x] = P[Y \leq e^x] = 1 - e^{-(e^x/\alpha)^\gamma} \text{ for } e^x \geq 0 \Rightarrow \ -\infty < x < \infty$$

Rearranging terms we obtain:

$$F_X(x) = 1 - e^{-e^{\gamma x} e^{log(\alpha^{-\gamma})}} = 1 - e^{-e^{(x - log(\alpha))/\frac{1}{\gamma}}} = 1 - e^{-e^{(x-\theta_1)/\theta_2}}$$

$F_X(x)$ is now in a location/scale form with location parameter $\theta_1 = log(\alpha)$ and scale parameter $\theta_2 = \frac{1}{\gamma}$

$$u = F_X(y_u) \ \Rightarrow \ y_u = \theta_1 + \theta_2 log(-log(1-u)) \ \Rightarrow \ Q_X(u) = \theta_1 + \theta_2 Q_Z(u)$$

Thus, we have that the standard member of the log-Weibull distribution has quantile function:

$$Q_Z(u) = log(-log(1-u))$$

General method of obtaining quantile function for the r.v. Y when $Y = h(X)$:

We will derive a more general formulation of how to find the quantile function of a random variable which is defined in terms of a second random variable.

Let $W = h(Y)$. Find the quantile function of W in terms of the quantile function of Y:

Case 1: h is an increasing function.

$$u = P[W \leq Q_W(u)] \;=\; P[h(Y) \leq Q_W(u)] \;=\; P[Y \leq h^{-1}(Q_W(u))] \Rightarrow$$

$$Q_Y(u) \;=\; h^{-1}(Q_W(u)) \Rightarrow \quad Q_W(u) \;=\; h(Q_Y(u))$$

Case 2: h is a decreasing function.

$$u = P[W \leq Q_W(u)] = P[h(Y) \leq Q_W(u)] = P[Y \geq h^{-1}(Q_W(u))] = 1 - P[Y \leq h^{-1}(Q_W(u))] \Rightarrow$$

$$P[Y \leq h^{-1}(Q_W(u))] = 1 - u \Rightarrow$$

$$Q_Y(1-u) = h^{-1}(Q_W(u)) \Rightarrow$$

$$Q_W(u) = h(Q_Y(1-u))$$

Example: Weibull Distribution:

For the transformation of the Weibull distribution, $W = log(Y)$ where Y has a Weibull distribution, we have that

$h(y) = log(y)$ is an increasing function.

Thus we have that the quantile function of W is obtained as follows:

Recall, that for Y having a Weibull distribution:

$$Q_Y(u) = \alpha[-log(1-u)]^{1/\gamma}$$

therefore, the quantile for W=log(Y) is given by

$$Q_W(u) = log(Q_Y(u)) = log(\alpha) + \frac{1}{\gamma}[log(-log(1-u))] = \theta_1 + \theta_2 log(-log(1-u))$$

Thus, the standard member of the family ($\theta_1 = 0, \theta_2 = 1$) has quantile function:

$$Q_Z(u) = log(-log(1-u)).$$

Thus, we plot the n points $[Q_Z(u_i), W_{(i)}]$, that is,

$$\big[log(-log(1-u_i)), log(Y_{(i)})\big] \quad \text{for} \quad u_i = (i - .5)/n.$$

Assessing the Fit of Line in a Reference Distribution Plot

In order to assess if the data values are **close to** a straight line, we obtain the least squares line through the points

$(Q_Z(u_i), \widehat{Q}(u))$, that is,

obtain the least squares line

$\widehat{Q}(u) = b_1 + b_2 Q_Z(u)$.

Then, examine the R^2 value from this fit. If the R^2 is fairly close to 1.0, then we have a good fit. In this case, we can obtain *rough* estimates of the location/scale parameters by equating b_1 to the location parameter and b_2 to the scale parameter.

The least squares line may be highly affected by a few outliers and hence the straight line through the two points $(Q_Z(.25), \widehat{Q}(.25))$ and $(Q_Z(.75), \widehat{Q}(.75))$

is often placed on the plot to assess the fit.

Note: $Q_Z(.25) = -0.675$, and $Q_Z(.75) = 0.675$.

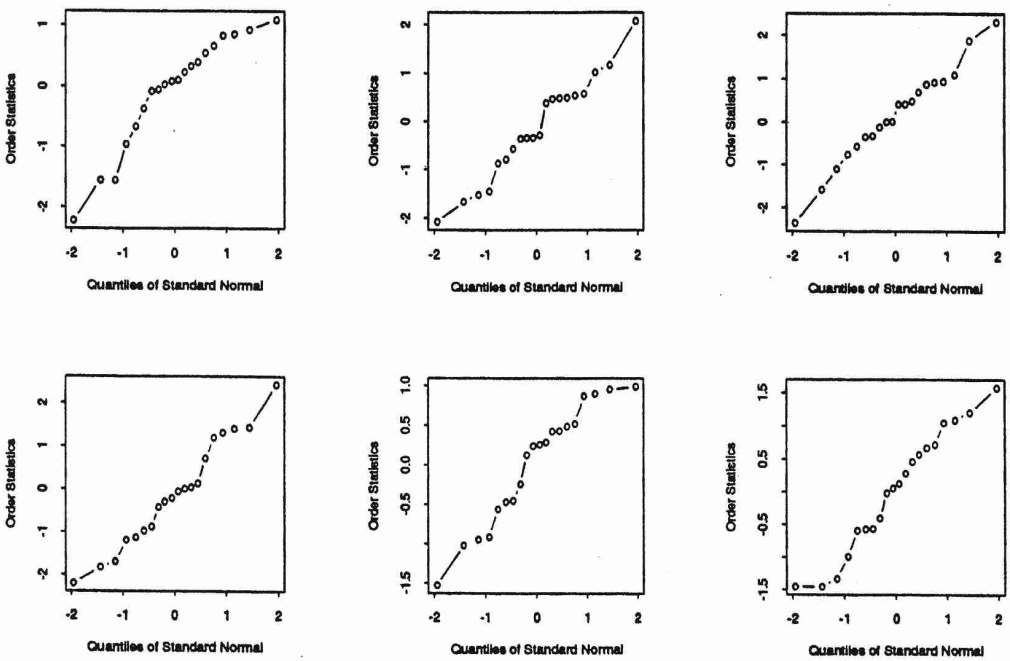
This line measures the fit of the model to the observed data in the middle portion of the values. Depending on the size of n, the points may or not fit very close to a straight line as will be demonstrated with the series of plots given on the next four pages. A short description of the graphs will now be given:

Six random samples of size 20 and 200 were obtained from a Normal Distribution, a Chi-squared (df=4) Distribution, a t (df=2) Distribution, and a Uniform on (0,1) Distribution. The sample quantile from each of these 48 samples were plotted versus a standard normal quantile. The plots indicate the types of deviations from a straight line we may expect to see in a normal reference plot when sampling from the four distributions:

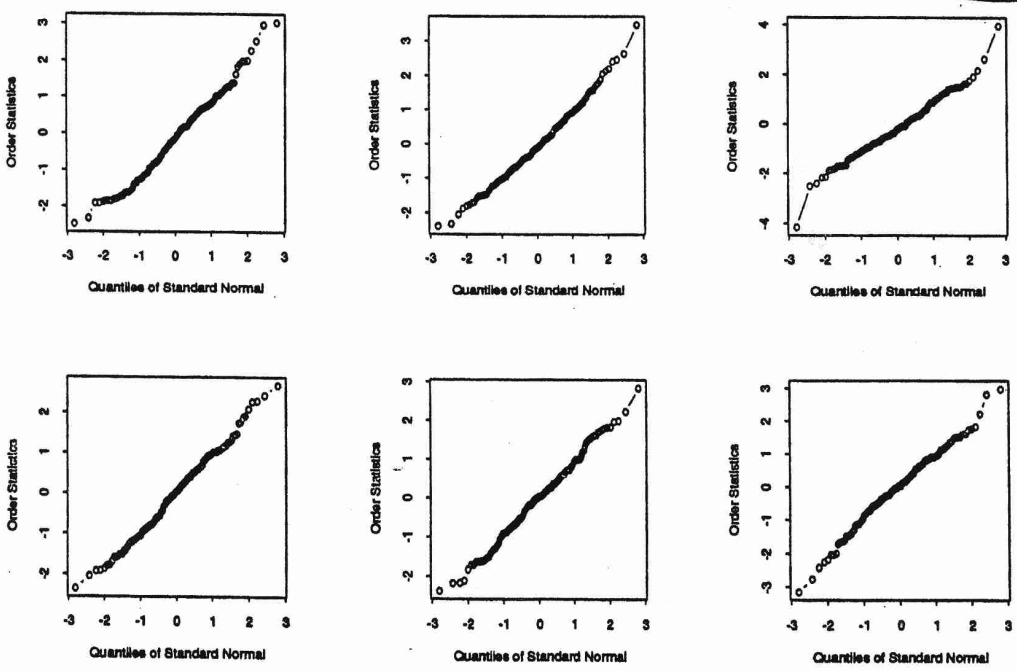
1. Normal Distribution - all 12 plots should be relatively close to line
2. Chi-squared Distribution(df=4) - Right skewed distribution with short tail on the left end of distribution - 12 plots should have plotted points curving away from line with most points above the line on the left and curving away from line with most points above the line on right
3. t distribution(df=2) - symmetric heavy-tailed distribution - 12 plots should have plotted points curving away from line with most points below the line on the left and curving away from line with most points above the line on right
4. Uniform - symmetric short-tailed distribution - 12 plots should have plotted points curving away from line with most points above the line on the left and curving away from line with most points below the line on right

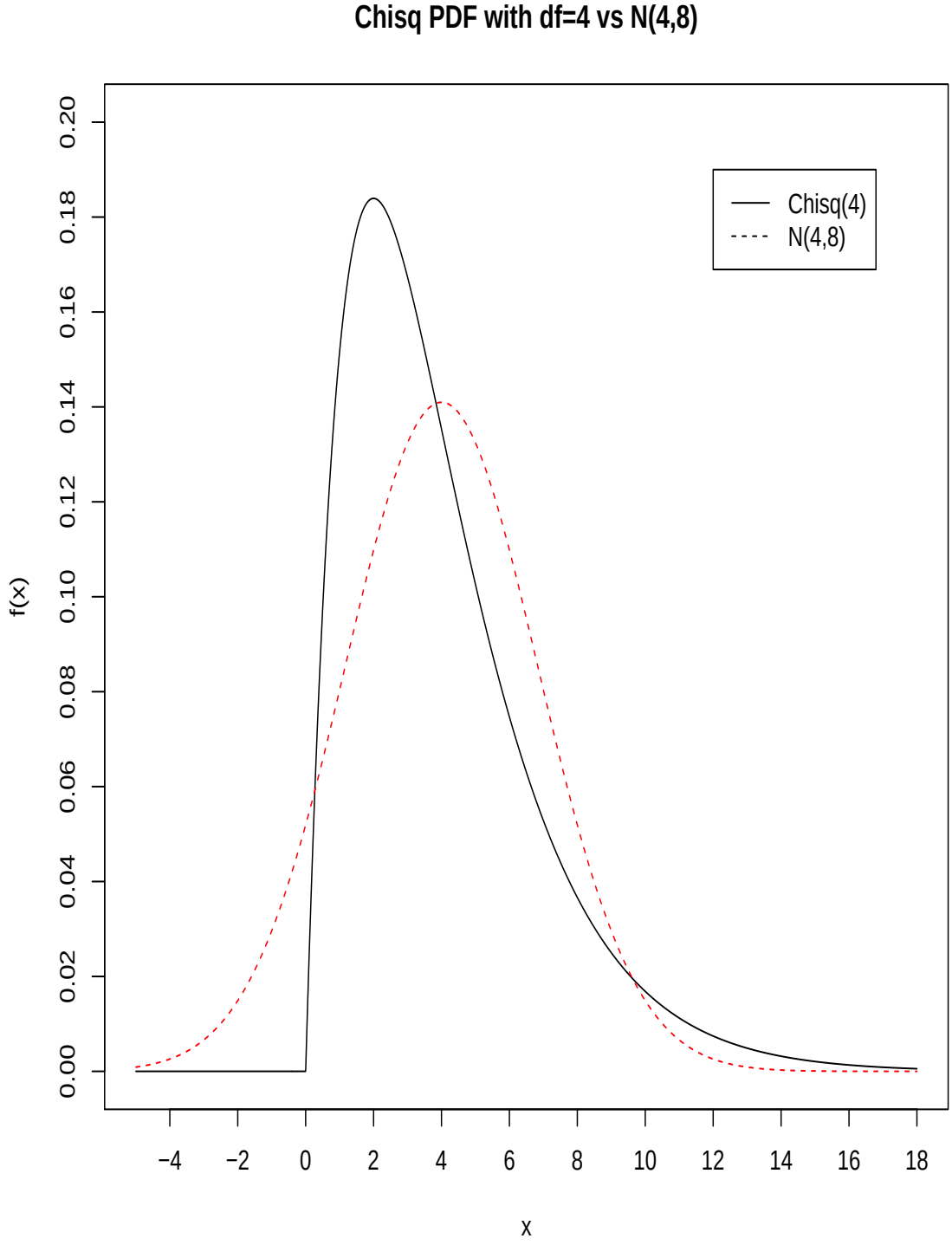
Normal Probability Plots, Normal Data, n=20/200

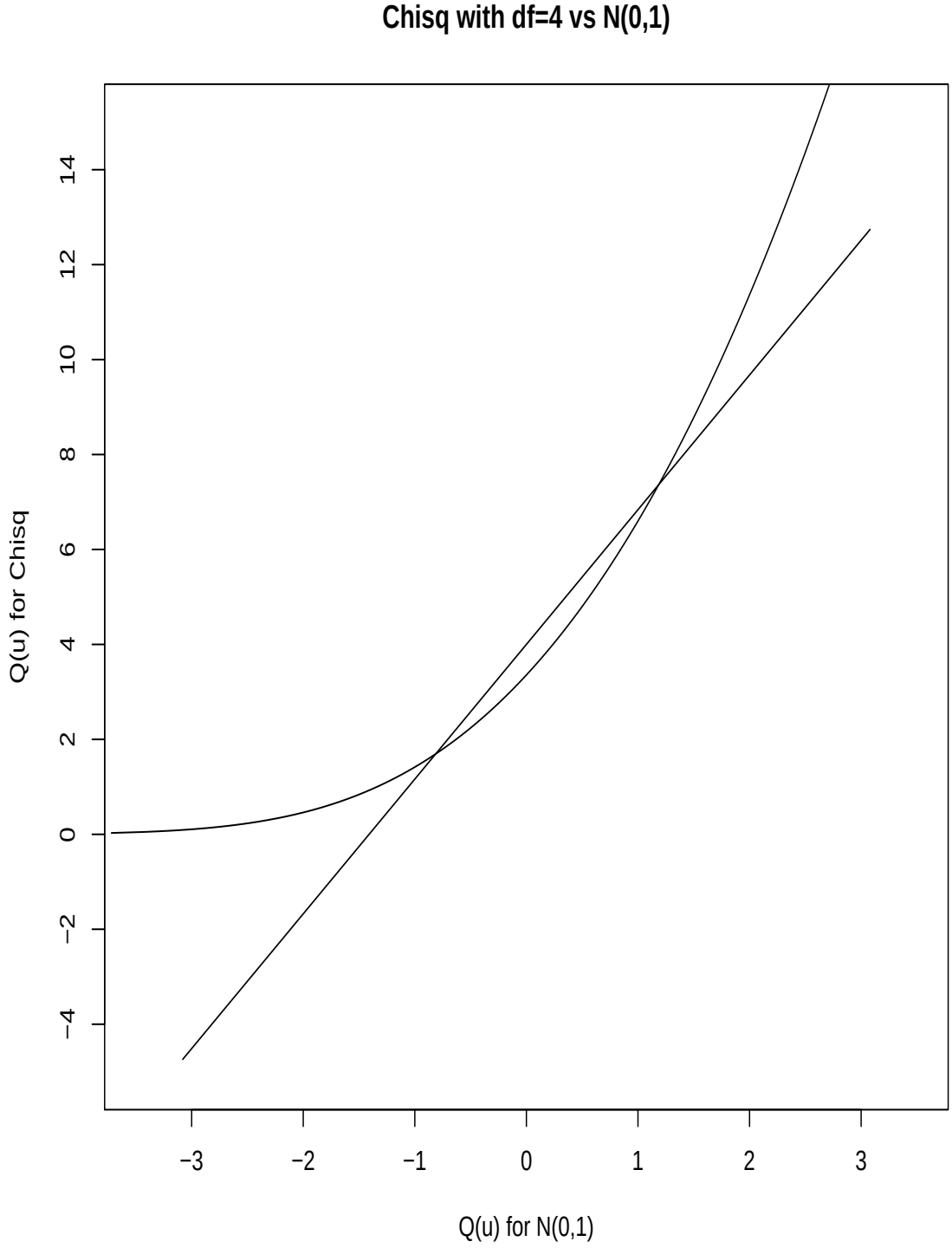
$N = 20$



$N = 200$

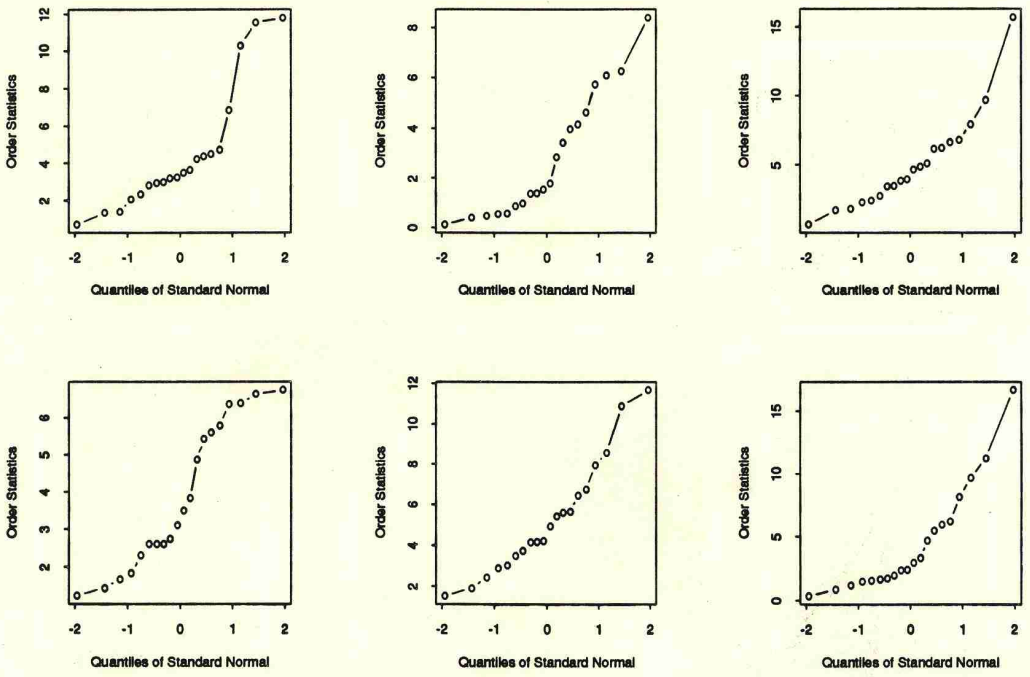



Chisq PDF with df=4 vs N(4,8)
Chisq(4)
N(4,8)
f(x)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
−4
−2
0
2
4
6
8
10
12
14
16
18
x

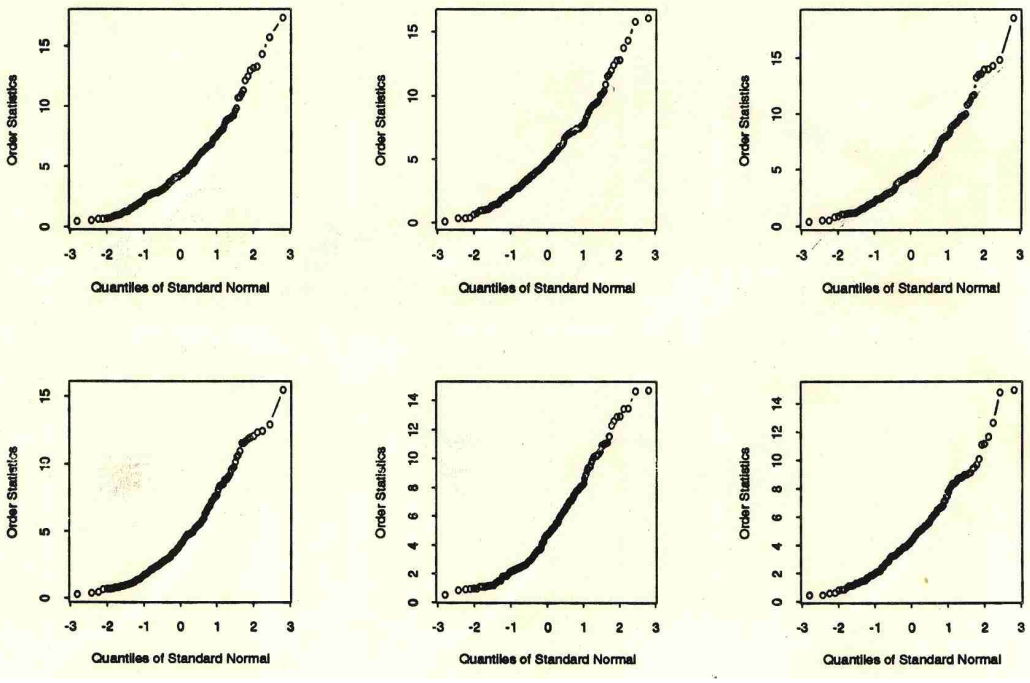

Chisq with df=4 vs N(0,1)
Q(u) for Chisq
Q(u) for N(0,1)
−4
−2
0
2
4
6
8
10
12
14
−3
−2
−1
0
1
2
3

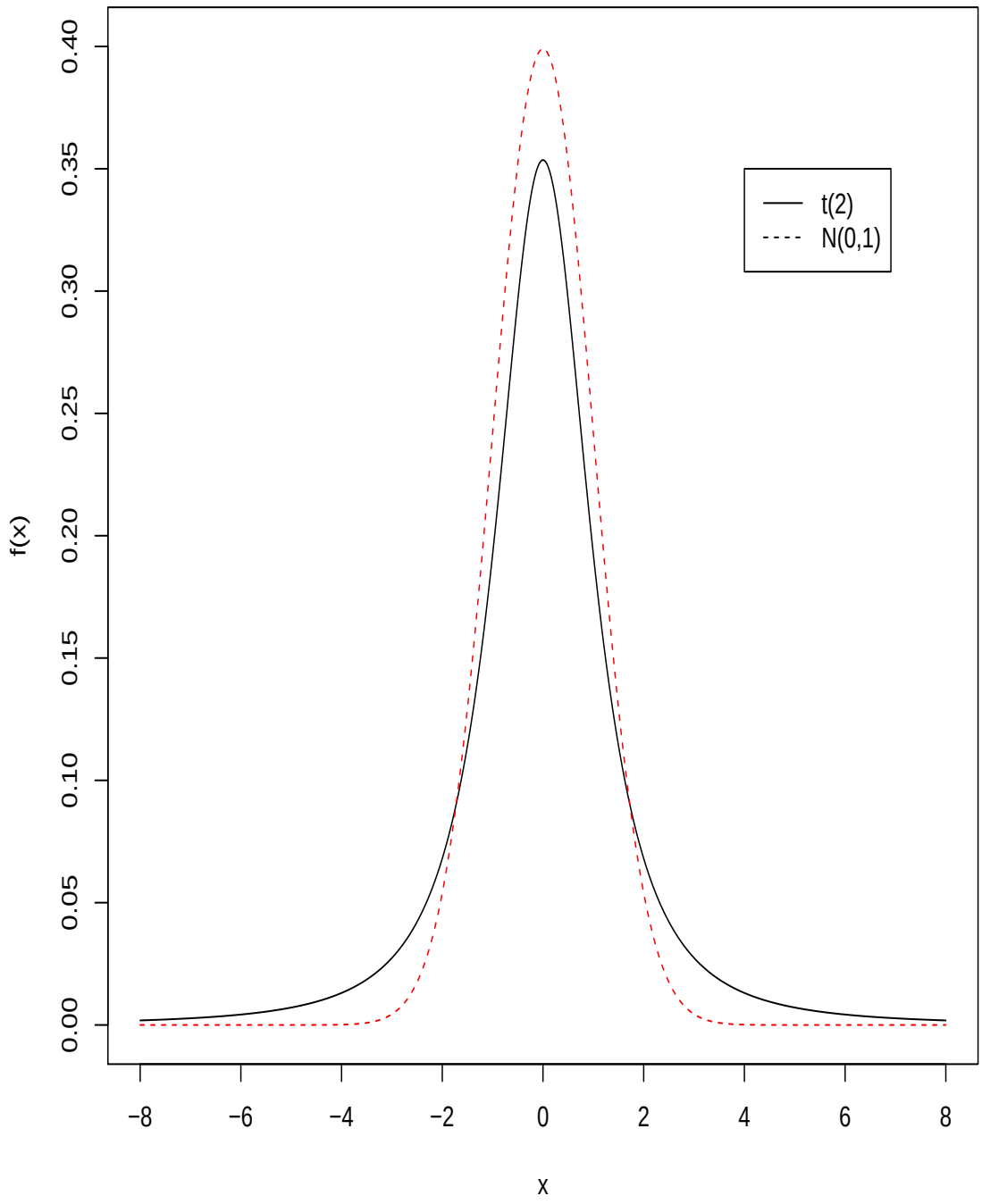
Normal Probability Plots, Chi-Squared (4 df) Data, n=20/200

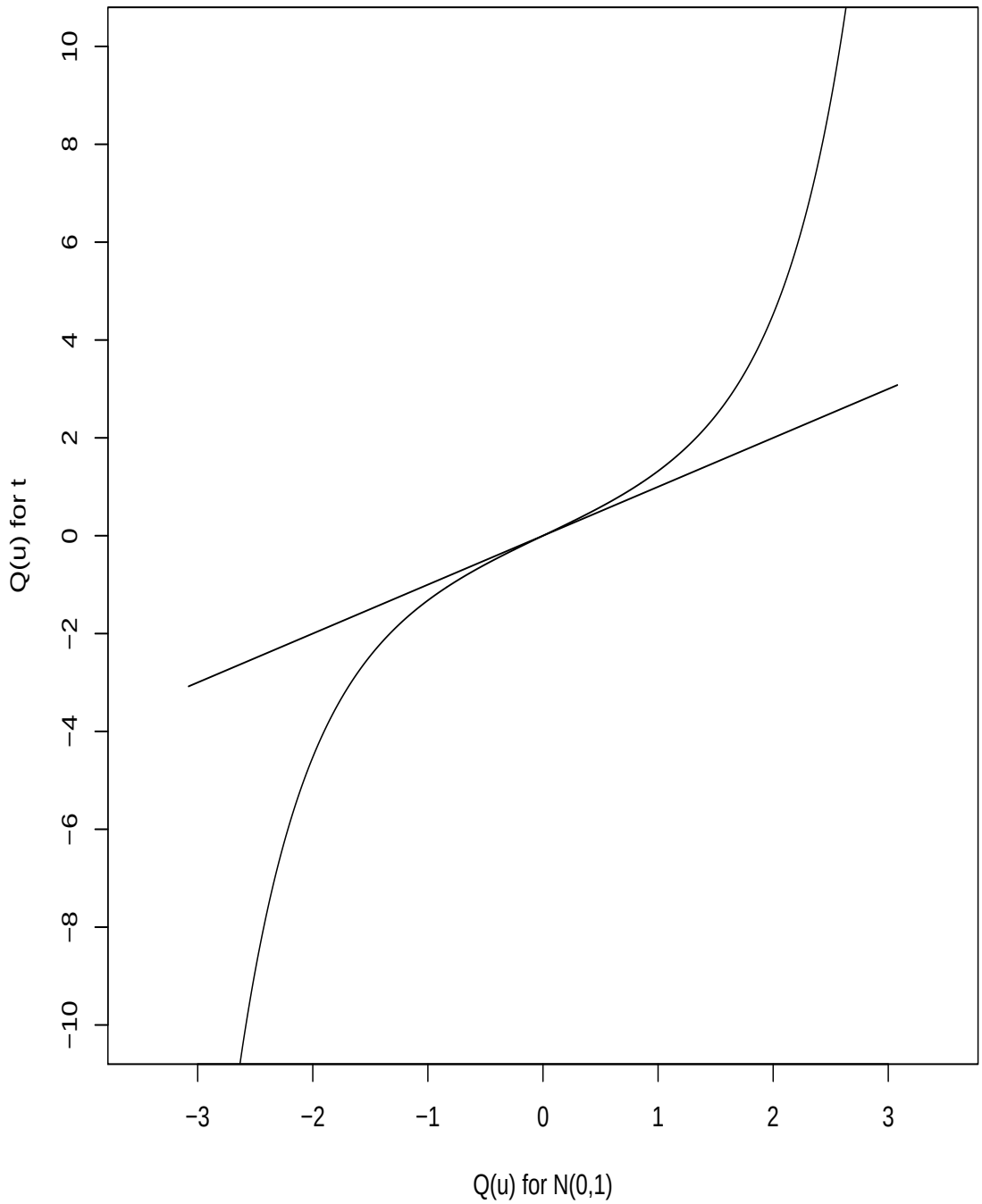
$N = 20$



$N = 200$

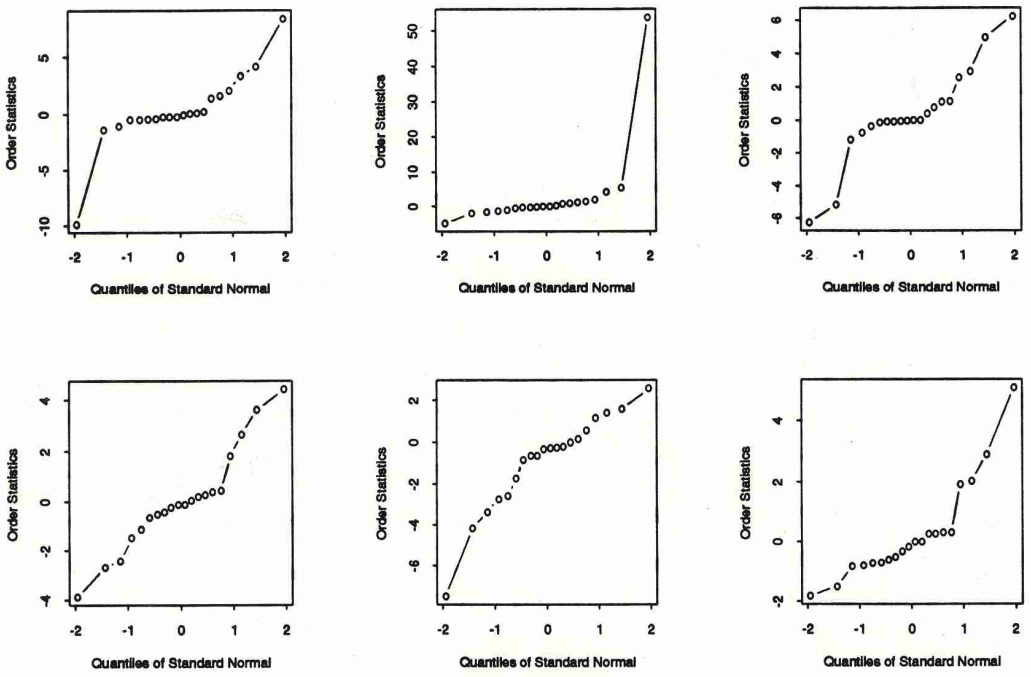



t PDF with df=2 vs N(0,1)
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−8
−6
−4
−2
0
2
4
6
8
x
t(2)
N(0,1)

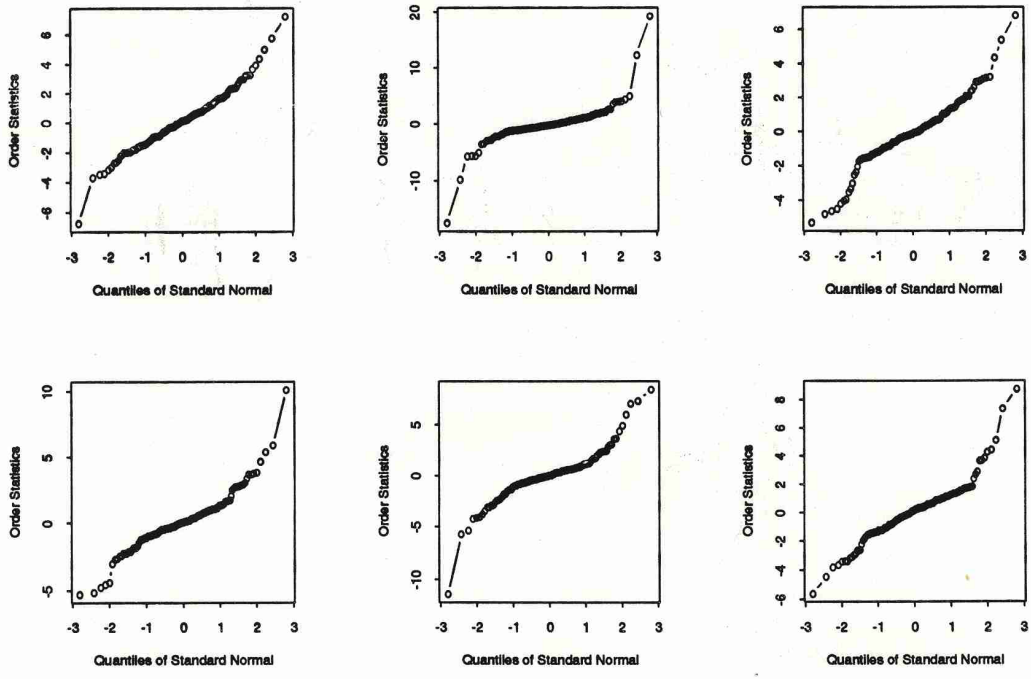

t Quantile with df=2 vs N(0,1)
Q(u) for t
10
8
6
4
2
0
−2
−4
−6
−8
−10
−3
−2
−1
0
1
2
3
Q(u) for N(0,1)

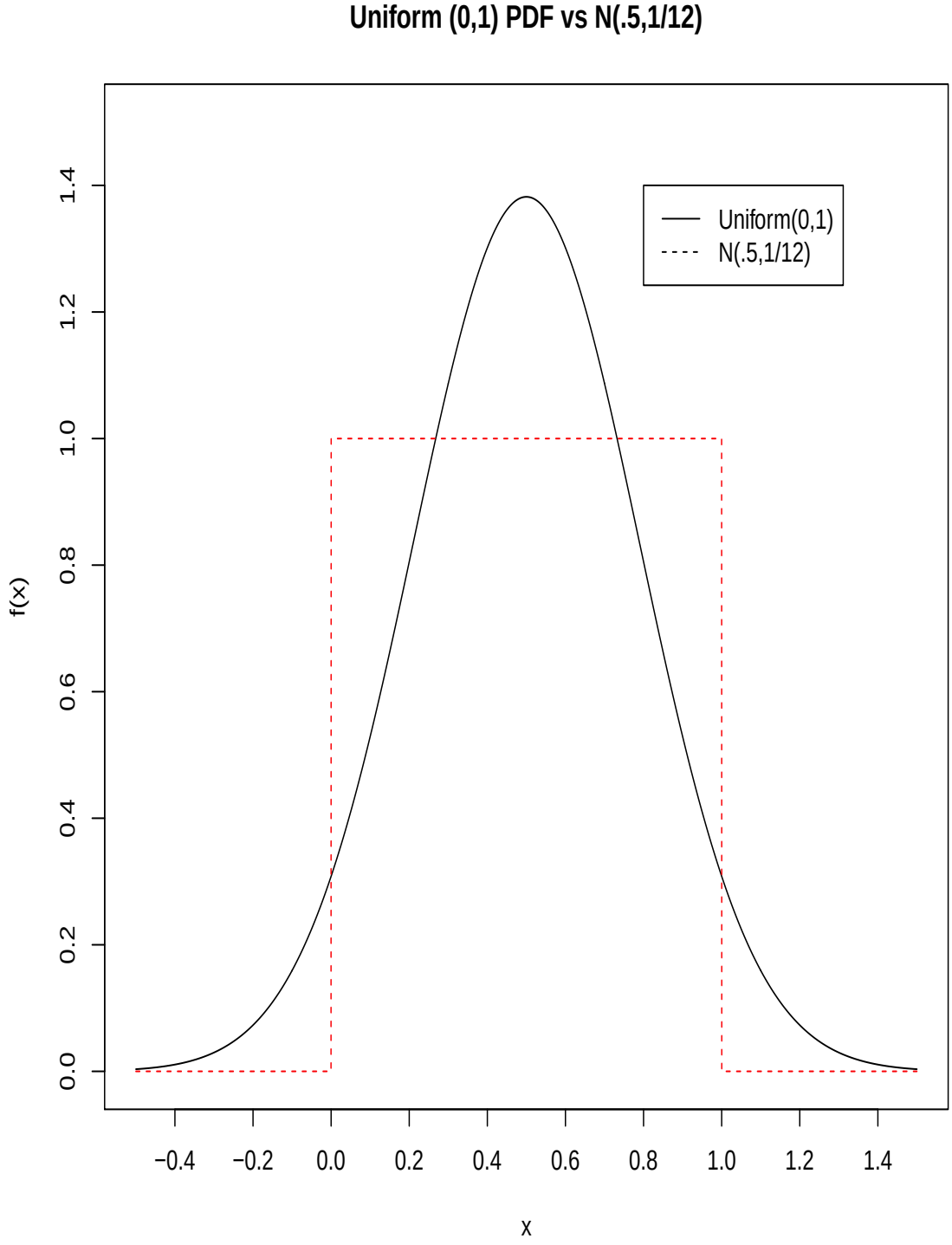
Normal Probability Plots, Student's t (2 df) Data, n=20/200

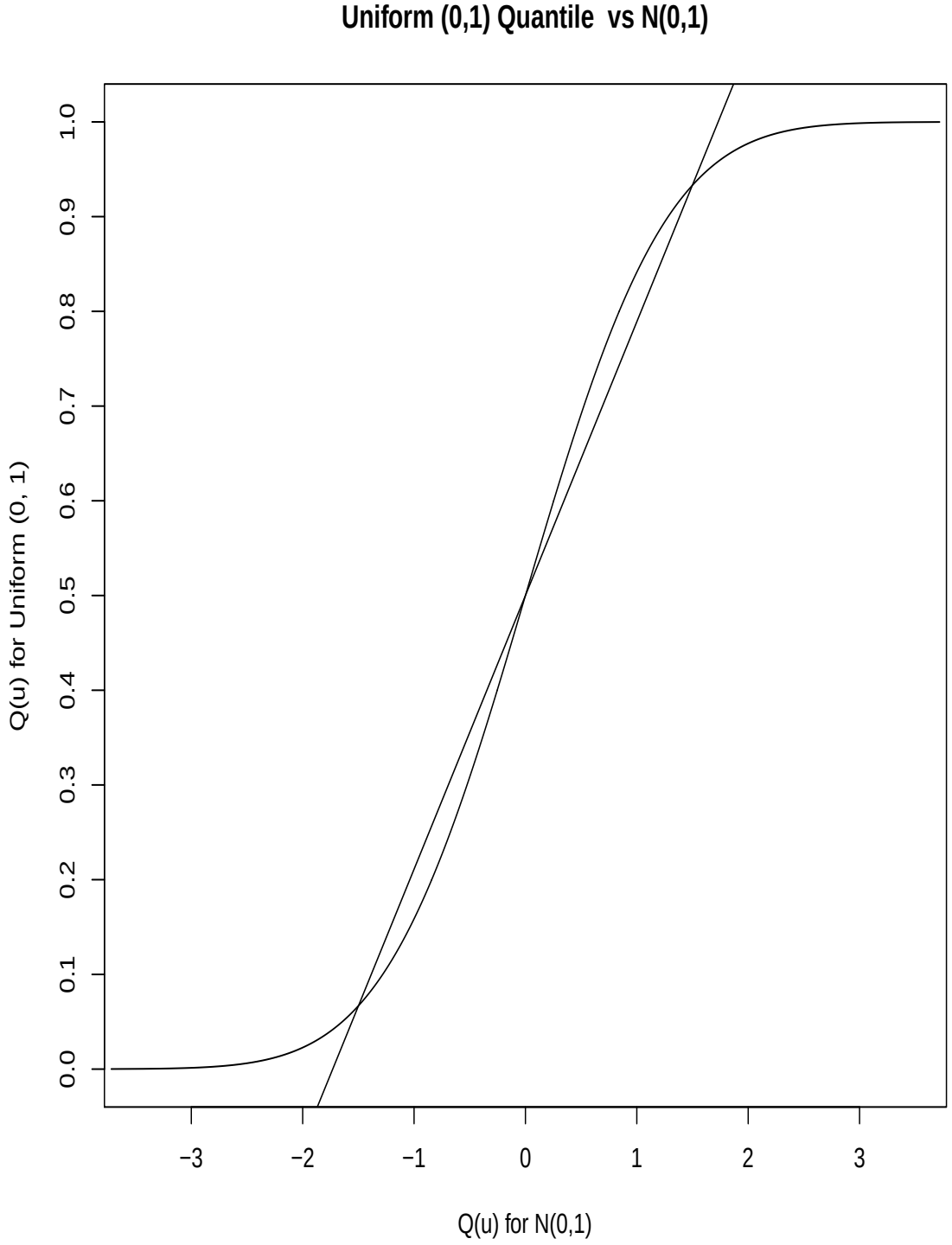
$N = 20$



$N = 200$

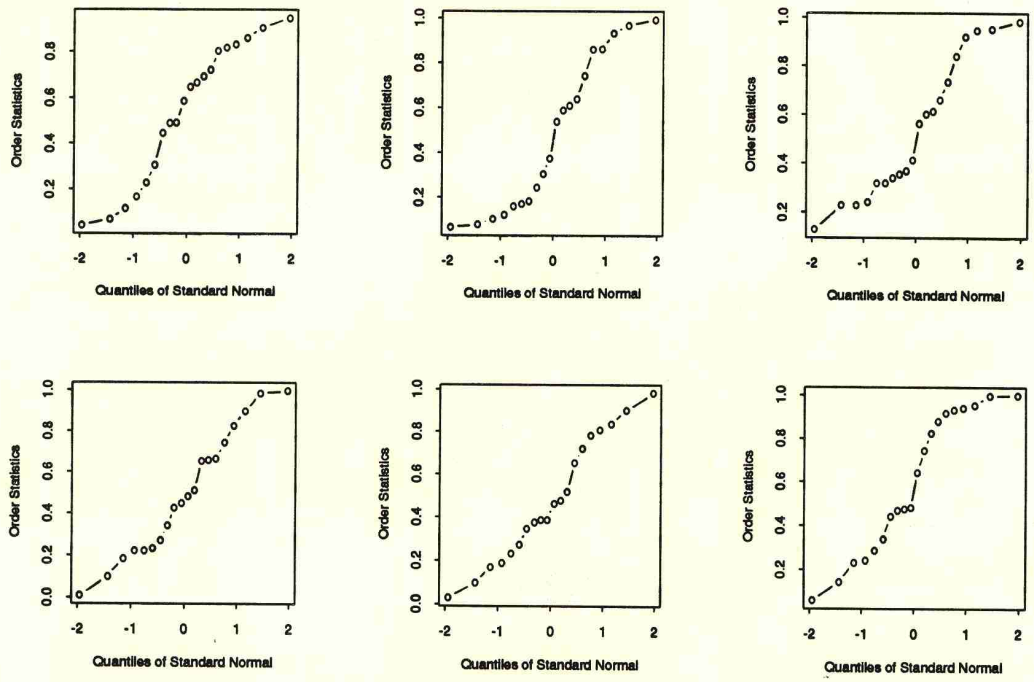



Uniform (0,1) PDF vs N(.5,1/12)
Uniform(0,1)
N(.5,1/12)
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x

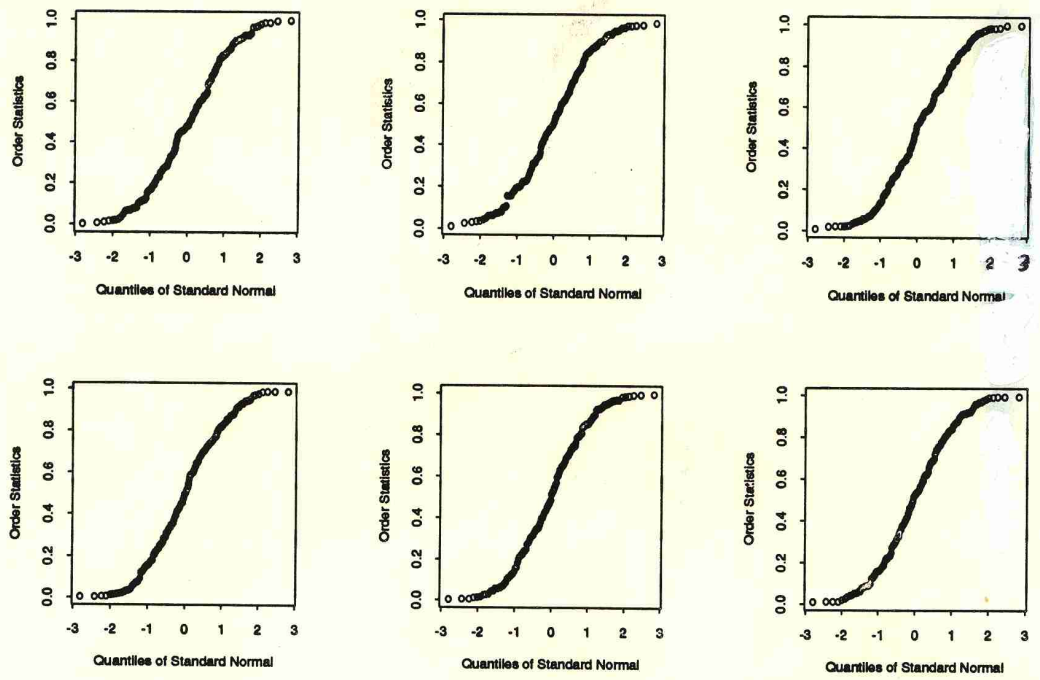

Uniform (0,1) Quantile vs N(0,1)
Q(u) for Uniform (0, 1)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
−3
−2
−1
0
1
2
3
Q(u) for N(0,1)

Normal Probability Plots, Uniform Data, n=20/200

$N = 20$

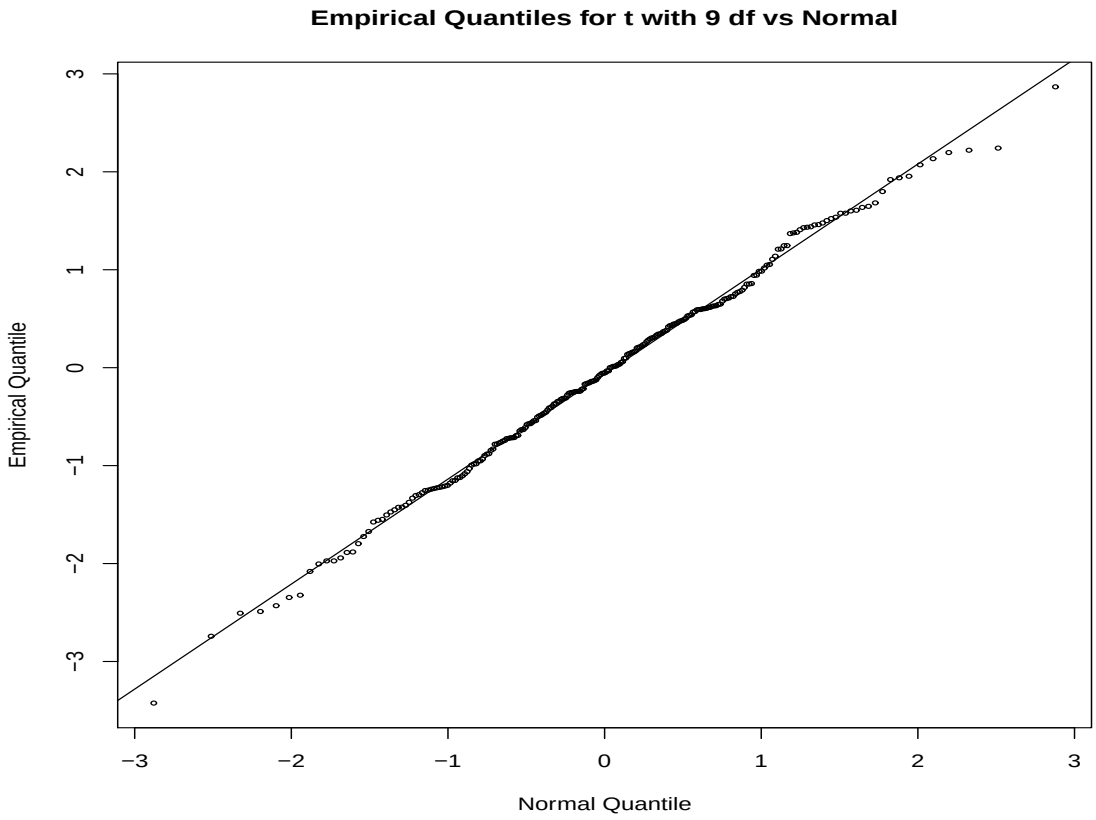


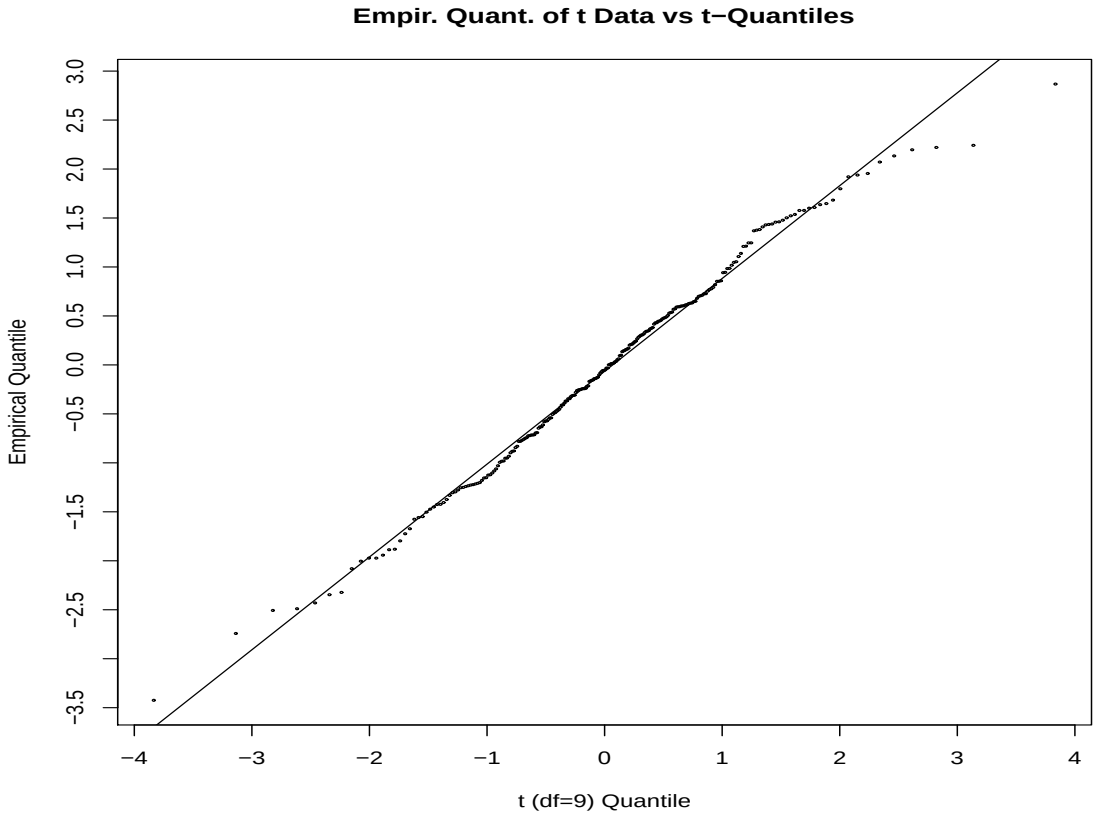
$N = 200$

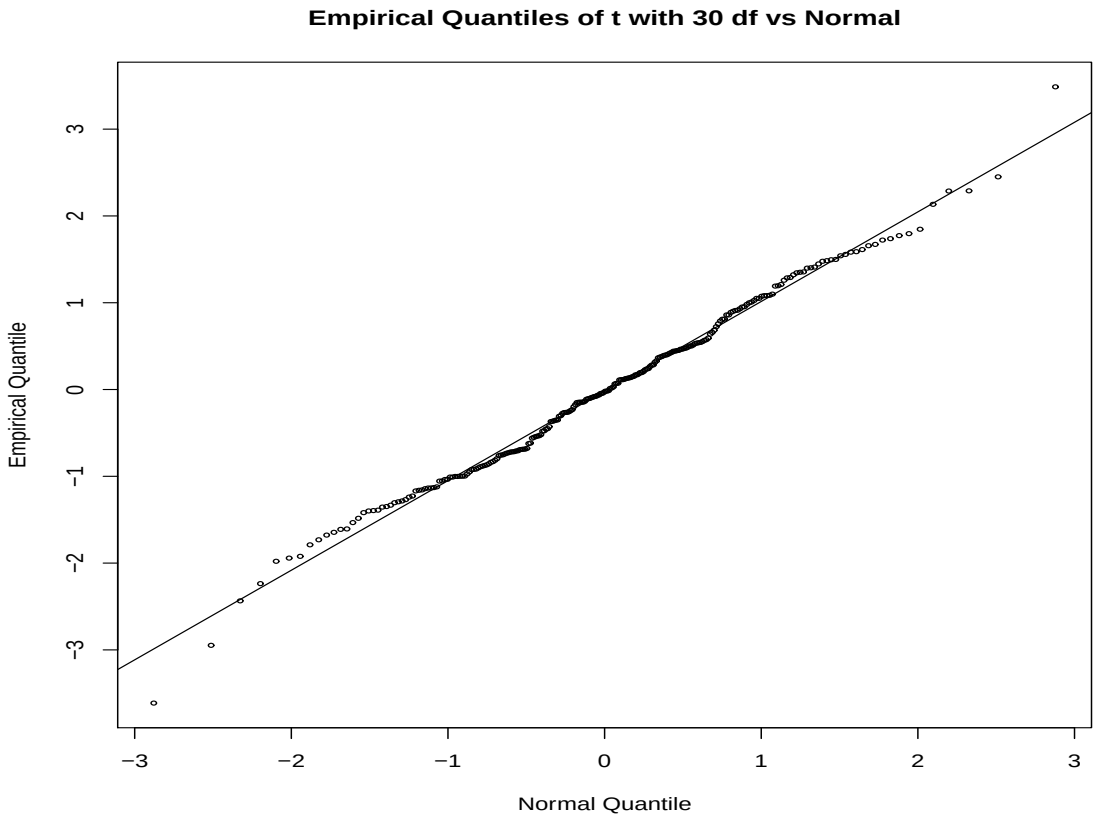


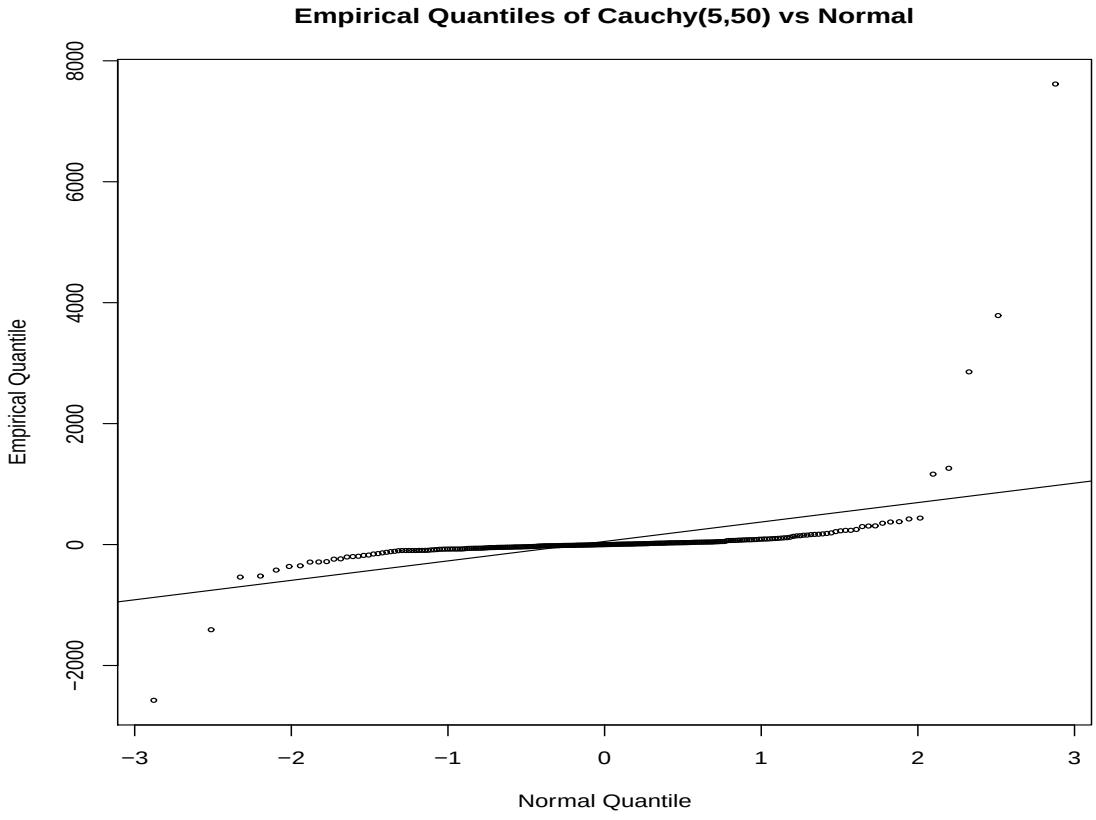
```
#The following R program generates data from various specified distributions
#and the plots the generated data various various reference distributions.
# This program is refdist.R in Files/Rcode on Dostat
#-------------------------------------------------------------------------------

#generates 250 observations from t with df=9 and 30, Cauchy, Gamma with
#shape=2 and scale=1/3, weibull with scale=16 and shape=2, uniform on (-2,5):
#Note in the gamma function, gamma(a,b): a=shape, b=1/scale

n=250

i= seq(1:n)
u= (i-.5)/n
z   = sort(qnorm(u))

t9  = sort(rt(n,9))
t30 = sort(rt(n,30))
cau = sort(rcauchy(n,5,50))
wei = sort(rweibull(n,2,16))
gam = sort(rgamma(n,2,3))
uni = sort(runif(n,-2,5))

#The following commands will generate various normal probability plots:

postscript("u:/meth1/psfiles/refdistp1.ps",height=8,horizontal=F)

par(mfrow=c(2,1))

# Empirical Quantile of t with df=9  vs Normal Quantiles:

plot(z,t9,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
       lab=c(7,8,7),main="Empirical Quantiles for t with 9 df vs Normal",
cex=.5)
abline(lm(t9~z))

# Empirical t with df= 9-Quantile vs t with df=9 Quantiles:

t=sort(qt(u,9))
plot(t,t9, xlab="t (df=9) Quantile",
     ylab="Empirical Quantile",lab=c(6,9,7), main=
"Empir. Quant. of t Data vs t-Quantiles",cex=.25)
abline(lm(t9~t))
graphics.off()
postscript("u:/meth1/psfiles/refdistp2.ps",height=8,horizontal=F)
par(mfrow=c(2,1))

# empirical t with df=30-Quantile vs normal quantiles:

plot(z,t30,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
       lab=c(7,8,7),main="Empirical Quantiles of t with 30 df vs Normal",cex=.5)
abline(lm(t30~z))
```

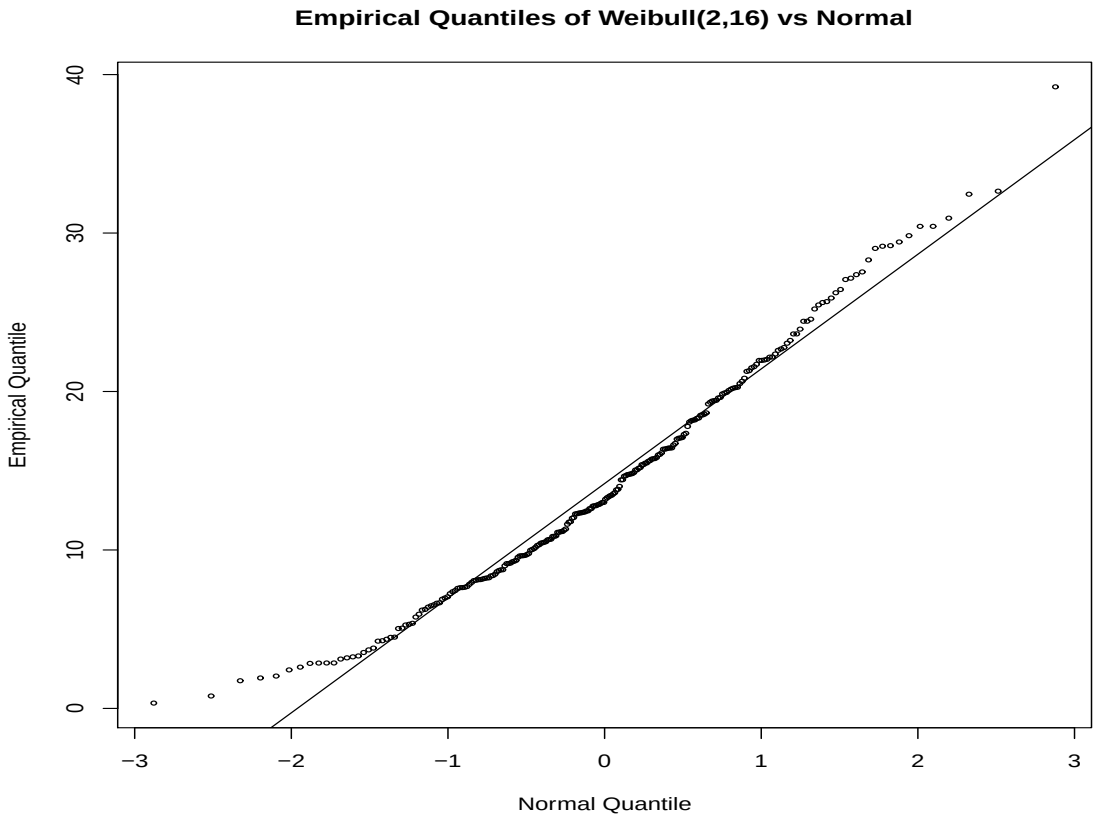
```
# empirical Cauchy-Quantile vs Normal Quantiles:

plot(z,cau,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
main="Empirical Quantiles of Cauchy(5,50) vs Normal",cex=.5)
abline(lm(cau~z))
graphics.off()


postscript("u:/meth1/psfiles/refdistp3.ps",height=8,horizontal=F)
par(mfrow=c(2,1))



# empirical Weibull-Quantile vs Normal Quantiles:

plot(z,wei,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
main="Empirical Quantiles of Weibull(2,256) vs Normal",cex=.5)
abline(lm(wei~z))

# empirical Weibull-Quantile vs Weibull-Quantiles:

x= sort(qweibull(u,2,16))
plot(x,wei, xlab="Weibull Quantile",ylab="Empirical Quantile",main=
"Empir. Quant. of Weibull Data vs Weibull-Quantiles",cex=.5)
abline(lm(wei~x))
graphics.off()


postscript("u:/meth1/psfiles/refdistp4.ps",height=8,horizontal=F)
par(mfrow=c(2,1))



# empirical Gamma-Quantile vs Normal Quantiles:

plot(z,gam,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
main="Empirical Quantiles of Gamma(2,1/3) vs Normal",cex=.5)
abline(lm(gam~z))


# empirical Uniform-Quantile vs Normal Quantiles:

plot(z,uni,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
lab=c(7,7,7), main="Empirical Quantiles of Uniform(-2,5) vs Normal",cex=.5)
abline(lm(uni~z))

graphics.off()
```

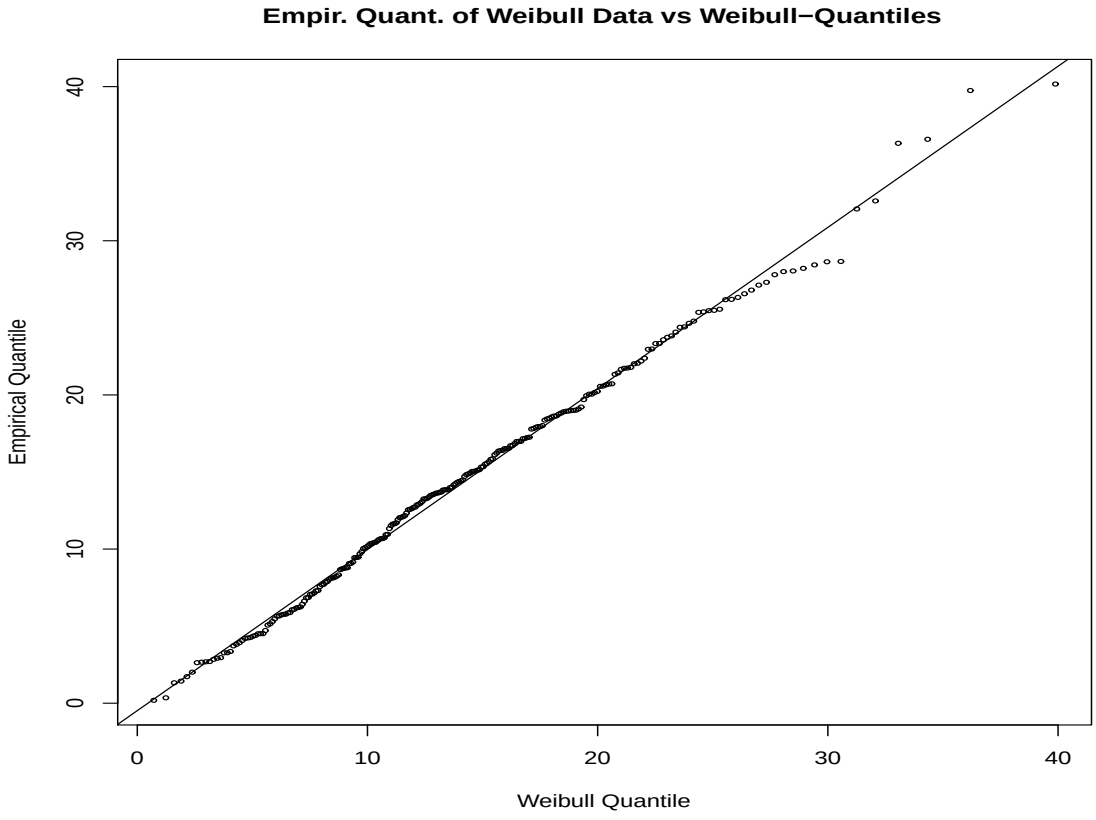

Empirical Quantiles for t with 9 df vs Normal
Empirical Quantile
Normal Quantile
−3
−2
−1
0
1
2
3

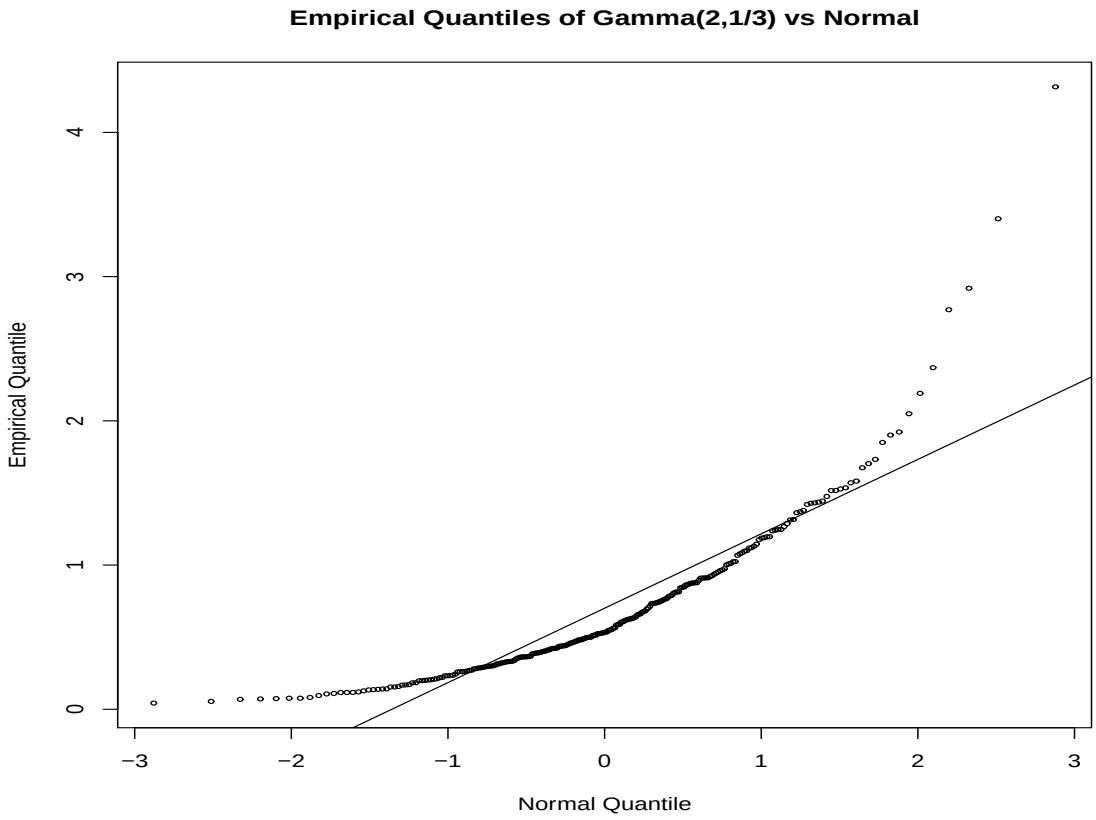

Empir. Quant. of t Data vs t−Quantiles
Empirical Quantile
t (df=9) Quantile
−3.5
−2.5
−1.5
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−4
−3
−2
−1
0
1
2
3
4

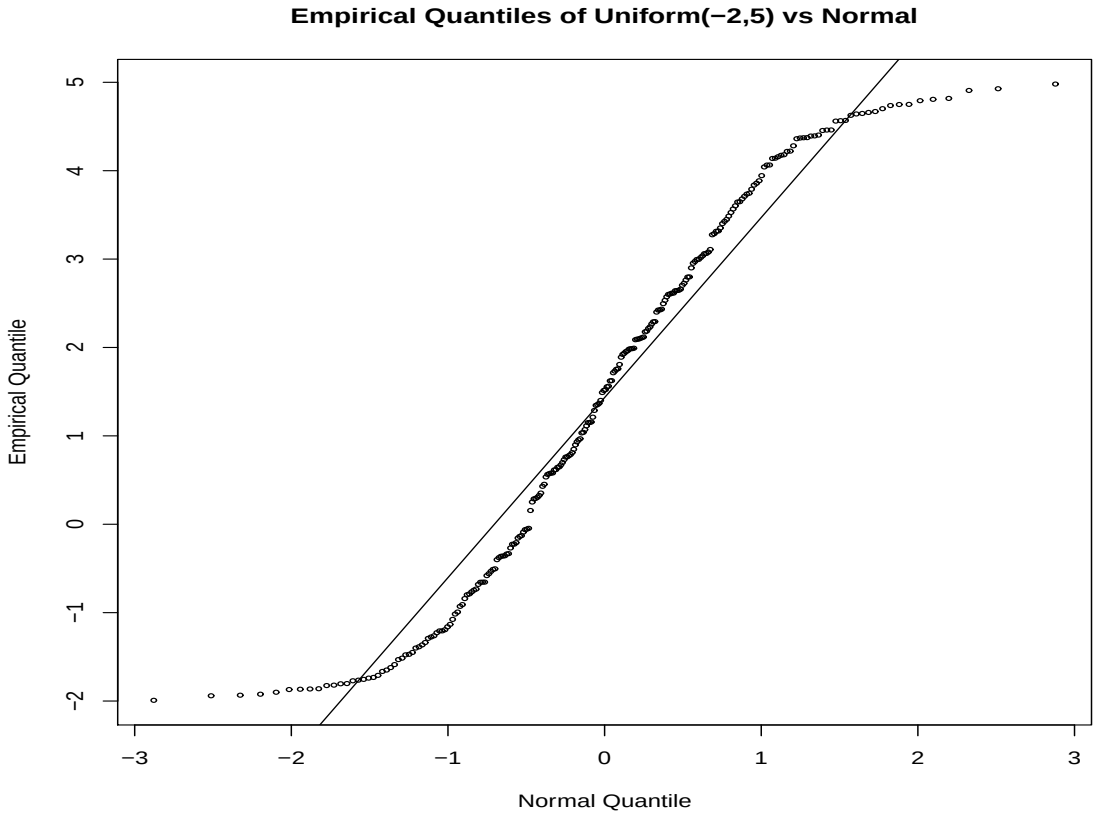

Empirical Quantiles of t with 30 df vs Normal
Empirical Quantile
Normal Quantile
−3
−2
−1
0
1
2
3


Empirical Quantiles of Cauchy(5,50) vs Normal
Empirical Quantile
Normal Quantile
−2000
0
2000
4000
6000
8000
−3
−2
−1
0
1
2
3


Empirical Quantiles of Weibull(2,16) vs Normal
Empirical Quantile
Normal Quantile
0
10
20
30
40
−3
−2
−1
0
1
2
3


Empir. Quant. of Weibull Data vs Weibull−Quantiles
Empirical Quantile
Weibull Quantile
0
10
20
30
40


Empirical Quantiles of Gamma(2,1/3) vs Normal
Empirical Quantile
Normal Quantile
0
1
2
3
4
−3
−2
−1
0
1
2
3


Empirical Quantiles of Uniform(−2,5) vs Normal
Empirical Quantile
Normal Quantile
−2
−1
0
1
2
3
4
5
−3
−2
−1
0
1
2
3

Sample Quantile-Quantile (q-q) Plots: Comparing Two Distributions

Suppose we have two populations/processes that we would like to compare with respect to their distributions.

Suppose we have a random sample from a population/process with cdf F_Y and

and a second random sample from a population/process with cdf F_X.

The Question of interest is "How is F_Y related to F_X?"

One way to answer this question is to plot the two sample quantiles:

$(\widehat{Q}_X(u), \widehat{Q}_Y(u))$

and see how close the plotted points are to a line.

That is, suppose we have

$X_1, X_2, \cdots, X_{n_1}$ from the population/process having cdf F_X and

$Y_1, Y_2, \cdots, Y_{n_2}$ from the population/process having cdf F_Y,

with $n = min(n_1, n_2)$.

Plot the n points

$$\left(\widehat{Q}_X(u_i),\ \widehat{Q}_Y(u_i)\right), \quad \text{for} \quad u_i = (i - .5)/n, \quad i = 1, \cdots, n = min(n_1, n_2).$$

If the n points fall close to a **line**, (not necessarily a straight line), the conclusions that we make depend on the nature of the line:

Case 1 If the plotted points, $\left(\widehat{Q}_X(u_i), \widehat{Q}_Y(u_i)\right)$, are close to a **45° line** through the origin then we can conclude that there is evidence that F_X and F_Y are the same cdf.

$$Q_X(u) = Q_Y(u) \text{ for all u if and only if } F_X(y) = F_Y(y) \text{ for all y}$$

Case 2 If the plotted points $\left(\widehat{Q}_X(u_i), \widehat{Q}_Y(u_i)\right)$ are close to a **straight line**, then we can conclude that there is evidence that X and Y are linearly related. Furthermore, **if** F_X is a member of a location-scale family, then there is evidence that F_Y is a member of the same family.

If $Y = \beta_o + \beta_1 X$ then $Q_Y(u) = \beta_o + \beta_1 Q_X(u)$.

If the distribution of X is a location/scale family of distributions, then

$Q_X(u) = \theta_1 + \theta_2 Q_Z(u)$, where $Q_Z(u)$ is the quantile of standard member.

Thus, $Q_Y(u) = \beta_o + \beta_1(\theta_1 + \theta_2 Q_Z(u)) = (\beta_o + \beta_1\theta_1) + \beta_1\theta_2 Q_Z(u)$

We can then conclude that the distribution of Y is location/scale with

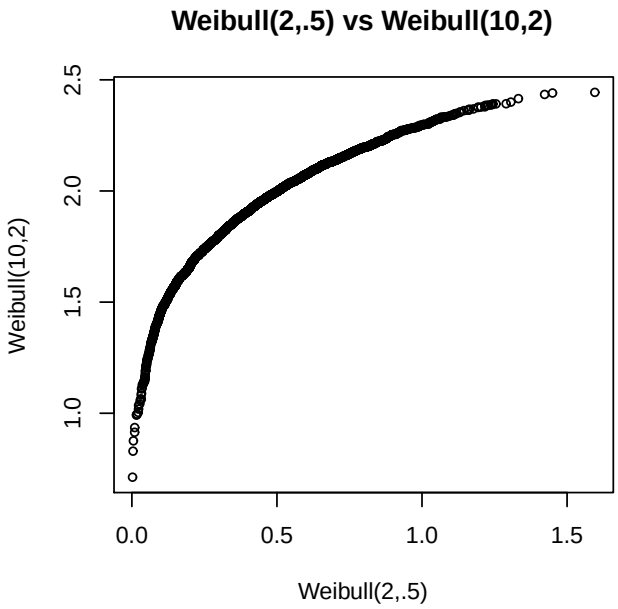
location parameter,$(\beta_o + \beta_1\theta_1)$ and scale parameter, $\beta_1\theta_2$

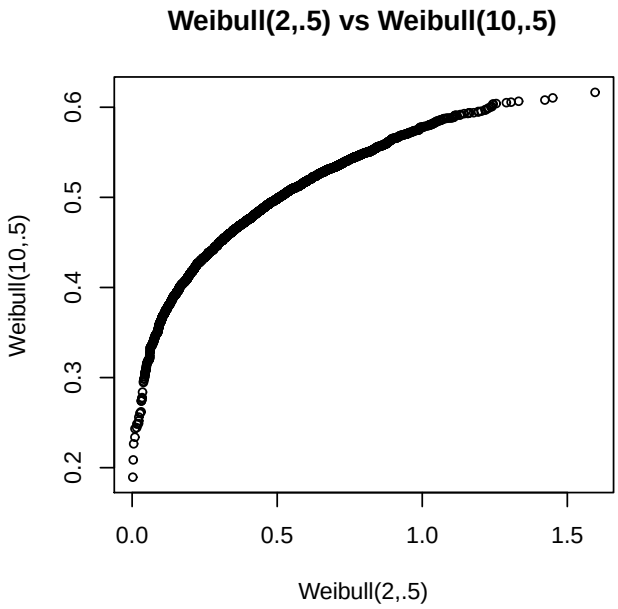
Case 3 If the plotted points, $\left(\widehat{Q}_X(u_i), \widehat{Q}_Y(u_i)\right)$, are close to a **non-linear line,** $\widehat{Q}_Y(u_i) = h(\widehat{Q}_X(u_i))$, for example, $h(x) = \beta_o e^{\beta_1 x}$, then we can conclude that there is evidence that X and Y are related by the same function, that is, $Y = h(X)$.

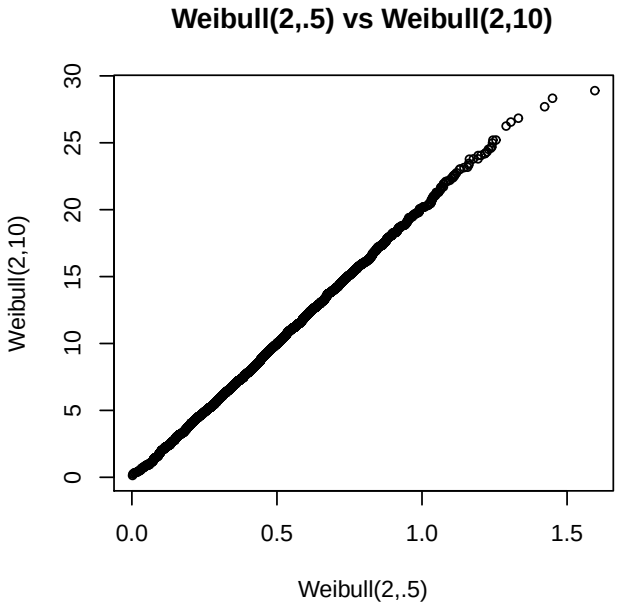
If $Y = h(X)$, then $Q_Y(u) = h(Q_X(u))$

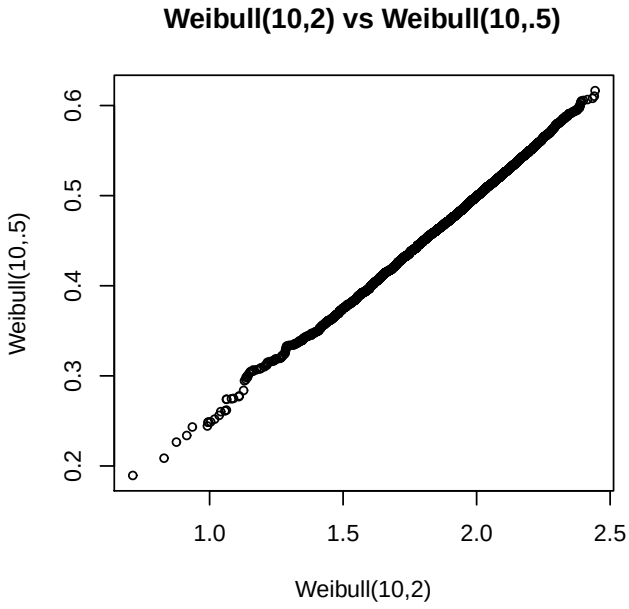
Case 4 If the plotted points, $\left(\widehat{Q}_X(u_i), \widehat{Q}_Y(u_i)\right)$, are **NOT** close to a **straight line**, then we **CANNOT** conclude that the distributions of X and Y are **NOT** members of the same family. For example, X and Y may be members of the Weibull family but have different shape parameters.

See the Q-Q plot on the following page.


Weibull(2,.5) vs Weibull(10,2)
Weibull(10,2)
Weibull(2,.5)
0.0
0.5
1.0
1.5
1.0
1.5
2.0
2.5


Weibull(2,.5) vs Weibull(10,.5)
Weibull(10,.5)
Weibull(2,.5)
0.0
0.5
1.0
1.5
0.2
0.3
0.4
0.5
0.6


Weibull(2,.5) vs Weibull(2,10)
Weibull(2,10)
Weibull(2,.5)
0.0
0.5
1.0
1.5
0
5
10
15
20
25
30


Weibull(10,2) vs Weibull(10,.5)
Weibull(10,.5)
Weibull(10,2)
1.0
1.5
2.0
2.5
0.2
0.3
0.4
0.5
0.6

Plots Associated with Mixture Distributions

Suppose that the population being studied consists of k subpopulations with pdf's $f_1, f_2, \cdots, f_k$ in proportions $p_1, p_2, \cdots, p_k$.

Let Y represent a randomly selected unit from the population having pdf f_Y.

The pdf of Y, f_Y can be represented as follows:

$$f_Y(y) = p_1 f_1(y) + p_2 f_2(y) + \cdots + p_k f_k(y) = \sum_{i=1}^{k} p_i f_i(y).$$

The mean and variance of Y satisfy the following relationships with the subpopulation means and variances: (Recall: $E[Y^2] = \sigma_Y^2 + \mu_Y^2$)

$$\mu_Y = E[Y] = \sum_{i=1}^{k} p_i E[Y_i] = \sum_{i=1}^{k} p_i \mu_i \qquad E[Y^2] = \sum_{i=1}^{k} p_i E[Y_i^2] = \sum_{i=1}^{k} p_i (\sigma_i^2 + \mu_i^2)$$

$$\sigma_Y^2 = E[Y^2] - \mu_Y^2 = \sum_{i=1}^{k} p_i \left(\sigma_i^2 + \mu_i^2\right) - \left(\sum_{i=1}^{k} p_i \mu_i\right)^2 \neq \sum_{i=1}^{k} p_i \sigma_i^2$$

Note that $\sigma_Y^2 = \sum_{i=1}^{k} p_i \sigma_i^2$ only if $\mu_i = \mu$ for all $i = 1, \ldots, n$.

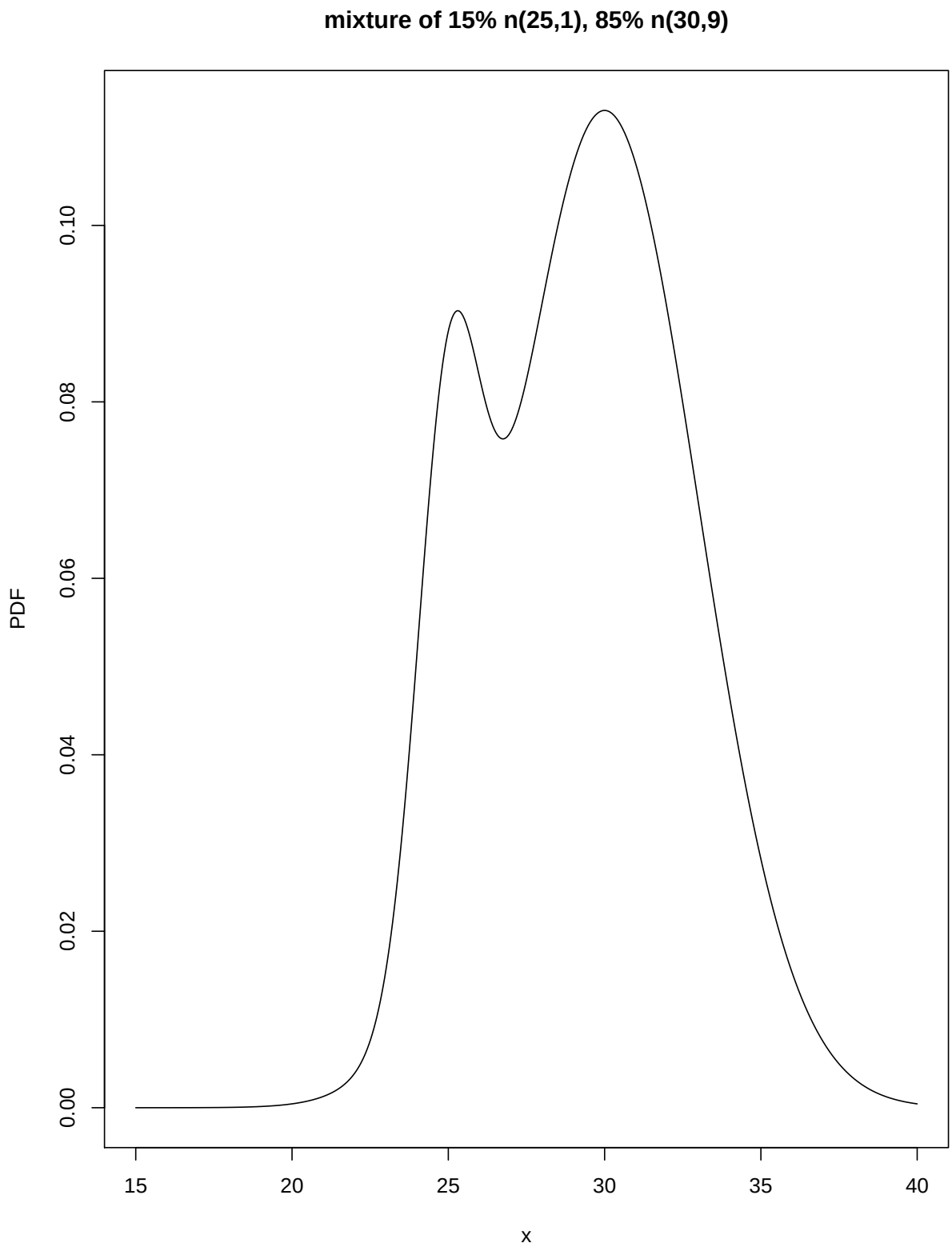
If we have a random sample from a population $Y_1, Y_2, \cdots, Y_n$, how can we tell whether or not the population has distinct subpopulations?

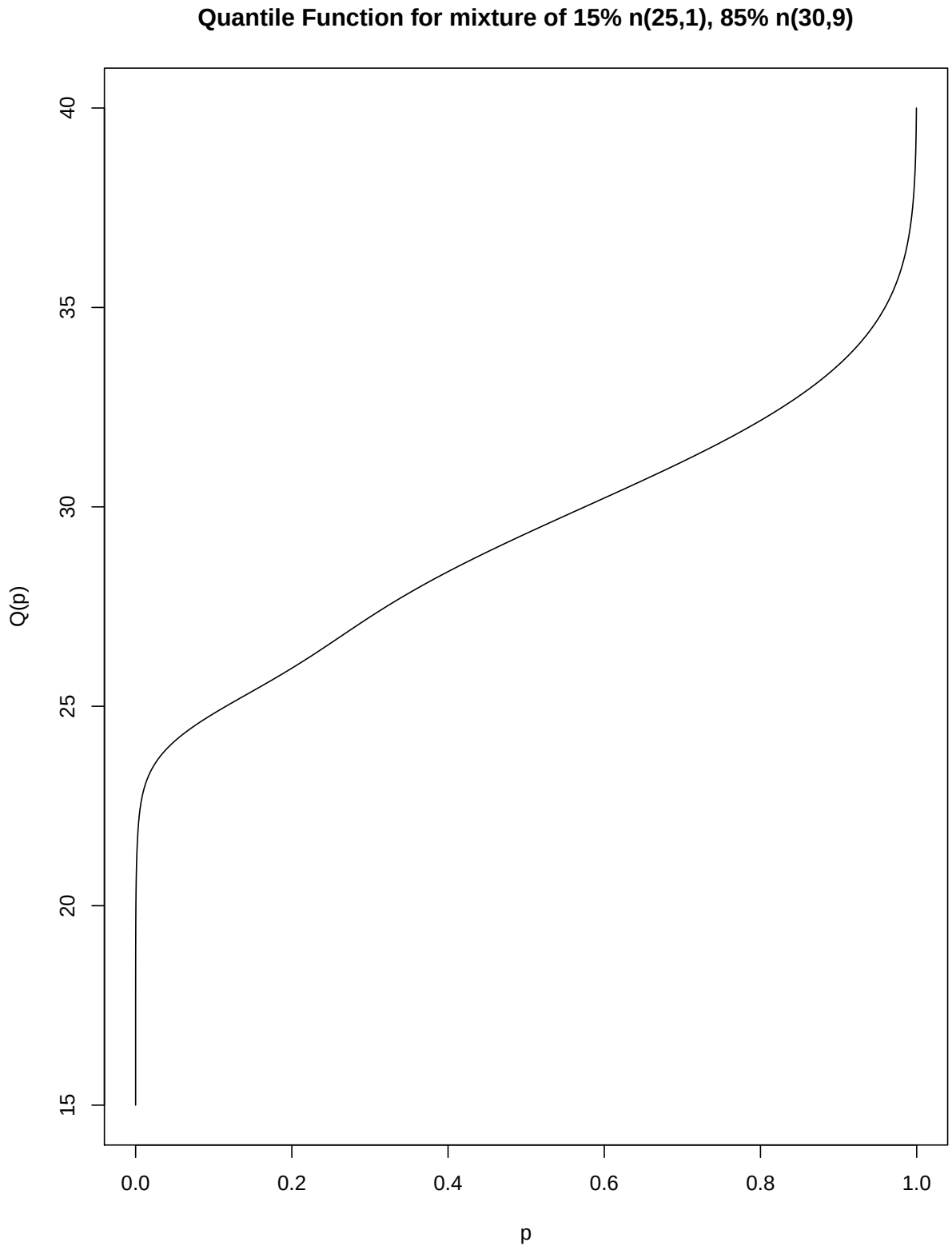
1. We can plot the kernel density estimator $\widehat{f}_Y(\cdot)$ and look for distinct modes in the plot. A problem with this method is that modes can appear and disappear as we vary the bandwidth. Thus it is very crucial to select an optimal bandwidth when using the kernel density estimator in such situations.

2. We can plot $\widehat{Q}(y)$ versus y and look for jumps in the plot. Once again, the question arises *when is a jump in the sample quantile plot a true jump in the population quantile.*

The plots on the next two pages will attempt to illustrate these problems in the simple case of mixing two normal distributions. We will consider four cases:

1. Case 1: 5% N(20,1) and 95% N(30,9)
2. Case 2: 15% N(25,1) and 85% N(30,9)
3. Case 3: 90% N(20,1) and 10% N(30,1)
4. Case 4: 15% N(28,1) and 85% N(30,9)


mixture of 15% n(25,1), 85% n(30,9)
PDF
0.00
0.02
0.04
0.06
0.08
0.10
15
20
25
30
35
40
x


Quantile Function for mixture of 15% n(25,1), 85% n(30,9)
Q(p)
40
35
30
25
20
15
0.0
0.2
0.4
0.6
0.8
1.0
p

Does the Mixture of Normal pdfs Result in a Normal pdf?

A population consists of four subspecies of a particular animal mixed together in varying proportions. A biologist samples the population and measures a physical characteristic, Y, which is distinct to the subspecies.

Let $f_i, \quad i = 1, 2, 3, 4$ be the pdf associated with the characteristic for each of the four subspecies with mixing proportions $p_i, \quad i = 1, 2, 3, 4$.

Suppose the four pdfs each have a normal distribution with possibly different parameters:

$N(\mu_i, \sigma_i) \quad i = 1, 2, 3, 4.$

Let f_Y be the pdf for the characteristic in the overall population.

Does this characteristic have a normal distribution, i.e., is f_Y a member of the normal family?

We will consider several cases:

1. **Case 1:** No restrictions on the parameters: p_i, μ_i, σ_i for $i = 1, 2, \ldots, k$
2. **Case 2:** No restrictions on the parameters: $p_1, \ldots, p_k$ and $\sigma_1, \ldots, \sigma_k$ but have $\mu_1 = \mu_2 = \cdots = \mu_k$
3. **Case 3:** No restrictions on the parameters: $\sigma_1, \ldots, \sigma_k$ but have $p_1 = p_2 = \cdots = p_k$ and $\mu_1 = \mu_2 = \cdots = \mu_k$

In case 1, the unequal μ_i obviously produce a nonnormal distribution.

In case 2 and case 3, the pdf f_Y is more peaked than a normal pdf having the same mean and variance.

The graphs on the next page will illustrate these ideas.

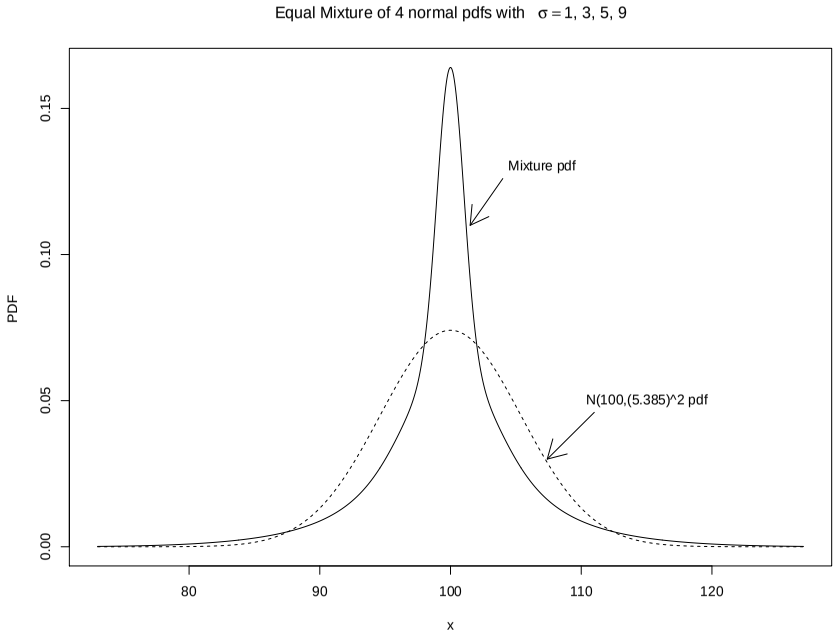
We will consider the situations depicted in the following table with $\mu_1 = \mu_2 = \mu_3 = \mu_4$. In this case, recall

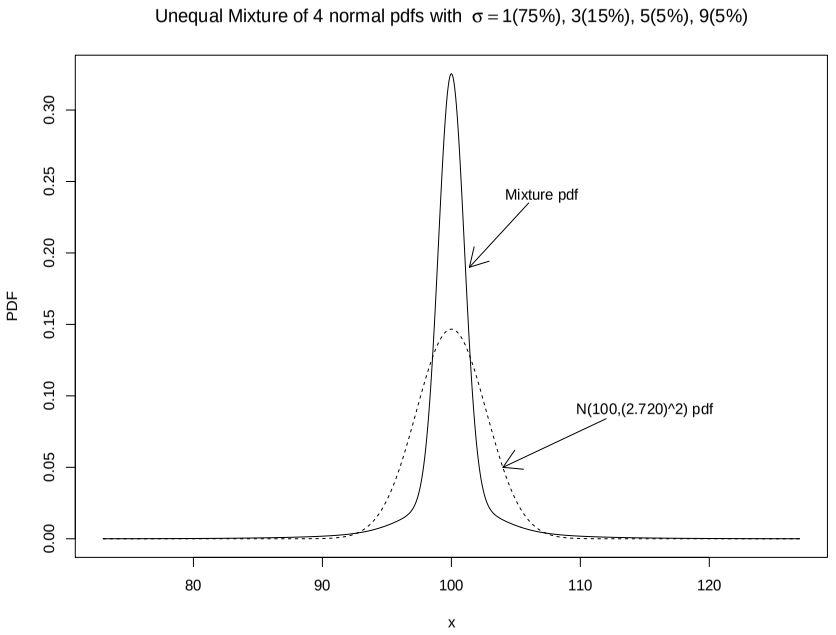
$$\sigma_Y^2 = \sum_{i=1}^{k} p_i \left(\sigma_i^2 + \mu_i^2\right) - \left(\sum_{i=1}^{k} p_i \mu_i\right)^2 = \sum_{i=1}^{k} p_i \sigma_i^2$$

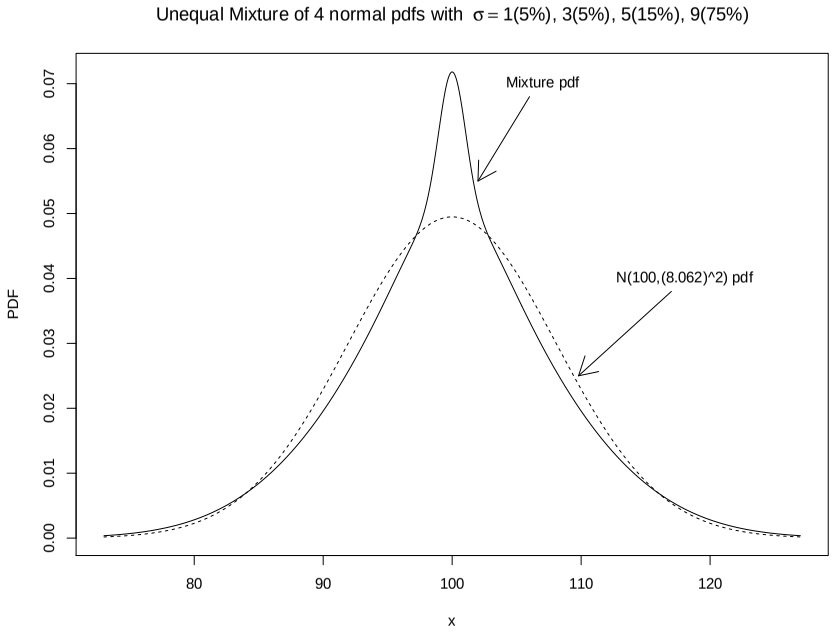
The following table contains four sets of values for the $p_i s$:

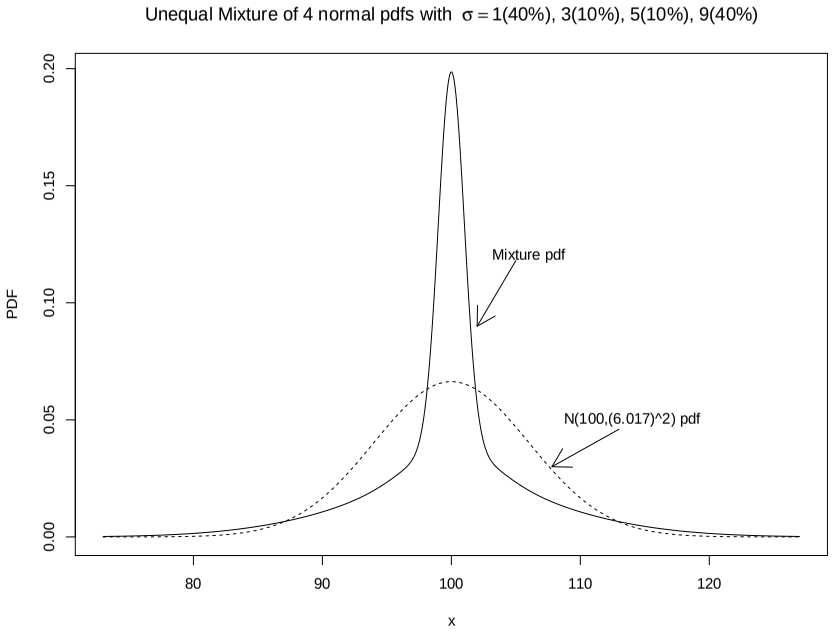
Four sets of values for $p_i' s$

Set	$\sigma_1 = 1$	$\sigma_2 = 3$	$\sigma_3 = 5$	$\sigma_4 = 9$	σ_Y
1	.25	.25	.25	.25	5.385
2	.75	.15	.05	.05	2.720
3	.05	.05	.15	.75	8.062
4	.40	.10	.10	.40	6.017


Equal Mixture of 4 normal pdfs with σ = 1, 3, 5, 9
Mixture pdf
N(100,(5.385)^2 pdf
PDF
x
0.00
0.05
0.10
0.15
80
90
100
110
120


Unequal Mixture of 4 normal pdfs with σ = 1(75%), 3(15%), 5(5%), 9(5%)
Mixture pdf
N(100,(2.720)^2) pdf
PDF
x
0.00
0.05
0.10
0.15
0.20
0.25
0.30
80
90
100
110
120


Unequal Mixture of 4 normal pdfs with σ = 1(5%), 3(5%), 5(15%), 9(75%)
Mixture pdf
N(100,(8.062)^2) pdf
PDF
x
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
80
90
100
110
120


Unequal Mixture of 4 normal pdfs with σ = 1(40%), 3(10%), 5(10%), 9(40%)
Mixture pdf
N(100,(6.017)^2) pdf
PDF
x
0.00
0.05
0.10
0.15
0.20
80
90
100
110
120

Box Plots and Their Shapes

A box and whiskers plot of the data depicts the data using the quartiles from the data:

$$\widehat{Q}_1 = \widehat{Q}(.25) \qquad \widehat{Q}_2 = \widehat{Q}(.5) \qquad \widehat{Q}_3 = \widehat{Q}(.75)$$

.

The box plot has the following features:

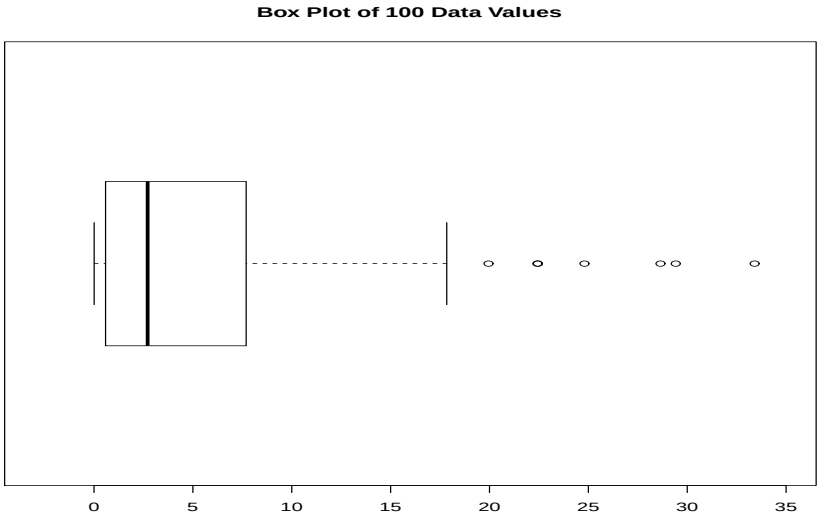
1. A rectangular box is drawn extending from $\widehat{Q}_1$ to $\widehat{Q}_3$
2. A line is drawn across the box at $\widehat{Q}_2$
3. Often the sample mean is depicted in the plot with a "+".
4. $\widehat{Q}_1$ and $\widehat{Q}_3$ are referred to as the **hinges** of the box plot.
5. Data values are classified as possible **outliers** if they fall beyond the fences of the data:

   $$\text{Lower Fence:} \quad \widehat{Q}_1 - 1.5(IQR)$$

   and

   $$\text{Upper Fence:} \quad \widehat{Q}_3 + 1.5(IQR),$$

6. $IQR = \widehat{Q}_3 - \widehat{Q}_1$ is referred to as the InterQuartile Range.
7. Two lines denoted as the **Whiskers** of the plot are drawn from $\widehat{Q}_1$ and $\widehat{Q}_3$ to the most extreme data values that are still *inside* the fences.
8. Data values having values outside the fences are regarded as *outliers*, that is, values which are unusual relative to the remaining data values. They are often denoted by asterisks.



9. In some instances, outliers are further classified as moderate outliers or extreme outliers by defining another set of fences:

$$\text{Lower Outer Fence:} \quad \widehat{Q}_1 - 3(IQR)$$

and

$$\text{Upper Outer Fence:} \quad \widehat{Q}_3 + 3(IQR).$$

If an observation falls between the two Lower Fences or two Upper Fences, then it is referred to as a **mild outlier**.

10. If an observation falls beyond the Lower Outer Fence or Upper Outer Fence, then it is referred to as an **extreme outlier**.

Some comments on the expected shape of box plots:

1. If the data is from a symmetric distribution, we would expect to obtain a box plot with whiskers of equal length and with median line in the center of the box.

   If there are any outliers then we would expect approximately an equal number of outliers beyond both the upper and lower fences.

2. If the data is from a distribution with normal distribution type tails then we would expect a very small proportion of outliers.

3. If the data is from a skewed to the right distribution, we would expect to obtain a box plot with longer whiskers emitting from $\widehat{Q}_3$ than from $\widehat{Q}_1$ and with a median line closer to $\widehat{Q}_1$ than to $\widehat{Q}_3$. If there are any outliers then we would expect more outliers beyond the upper fence than beyond the lower fence.

4. If the data is from a symmetric distribution having tails much heavier than the normal distribution, we would expect to obtain a box plot with whiskers of equal length and with median line in the center of the box. However, we would expect to have a significant proportion of outliers.

5. A box plot of data from a distribution having multiple modes **will not** reveal the multiple modes.

Outlier Detection

We will now illustrate the calculation of the probability of obtaining an outlier.

We will define an observation, Y, to be an **outlier** if

$$Y < Q_1 - 1.5(IQR) \quad \text{or} \quad Y > Q_3 + 1.5(IQR).$$

Case 1: The cdf of Y, F_Y is Known

The probability can be directly calculated:

$$P[\text{outlier}] = F_Y\left[Q_1 - 1.5(IQR)\right] + 1 - F_Y\left[Q_3 + 1.5(IQR)\right]$$

Case 2: Population distribution member of a location-scale family

It is possible to calculate the probabilities without any information about the location or scale parameters.

Suppose that the cdf of Y is a member of a location-scale family with parameters, θ_1 and θ_2.

Let Z be the standard member of the family with cdf, F_Z and quantile function, Q_Z

$$Q_Y(u) = \theta_1 + \theta_2 Q_Z(u)$$

$$IQR_Y = Q_Y(.75) - Q_Y(.25) = (\theta_1 + \theta_2 Q_Z(.75)) - (\theta_1 + \theta_2 Q_Z(.25)) = \theta_2(IQR_Z)$$

The probability of an outlier for the r.v. Y can be computed as follows.

$$P\left[Y < Q_Y(.25) - 1.5(IQR_Y)\right] + P[Y > Q_Y(.75) + 1.5(IQR_Y)]$$

$$= P\left[\frac{Y-\theta_1}{\theta_2} < \frac{(\theta_1+\theta_2 Q_Z(.25))-1.5(\theta_2 IQR_Z)-\theta_1}{\theta_2}\right] + P\left[\frac{Y-\theta_1}{\theta_2} > \frac{(\theta_1+\theta_2 Q_Z(.75))+1.5(\theta_2 IQR_Z)-\theta_1}{\theta_2}\right]$$

$$= P\left[Z < Q_Z(.25) - 1.5(IQR_Z)\right] + P\left[Z > Q_Z(.25) + 1.5(IQR_Z)\right]$$

$$= F_Z\left[Q_Z(.25) - 1.5(IQR_Z)\right] + 1 - F_Z\left[Q_Z(.75) + 1.5(IQR_Z)\right]$$

We can then use the cdf, F_Z and quantile function, Q_Z of the standard member of the family to complete the above calculation.

Thus, for location-scale families of distributions, the probability of an outlier is the same for every member of the family no matter the values of the location and scale parameters. However, this is not true for non-location-scales families. For example, the probability of an outlier for one member of the Weibull family of distributions will be much higher than for other members of the family.

Example: Let Y have a Weibull(γ, α) distribution, then

$$Q(u) = \alpha\,[-log(1-u)]^{1/\gamma} \;\Rightarrow\; Q_1 = \alpha\,[-log(.75)]^{1/\gamma}; \quad Q_3 = \alpha\,[-log(.25)]^{1/\gamma}; \quad IQR = Q_3 - Q_1$$

$$P[\text{Y is an outlier}] = F_Y\,[Q_1 - 1.5(IQR)] + 1 - F_Y\,[Q_3 + 1.5(IQR)]$$

Consider the following four cases:

- Case 1: Y is Weibull$(\gamma = 2, \alpha = 5)$ distribution

  $Q_1 = 2.6818 \quad Q_3 = 5.8871 \quad IQR = 3.2053$

$$P[\text{Y is an outlier}] = pweibull(2.6818 - 1.5(3.2053), 2, 5) + 1 - pweibull(5.8871 + 1.5(3.2053), 2, 5) = .0103$$

- Case 2: Y is Weibull$(\gamma = 2, \alpha = 1.18)$ distribution

  $Q_1 = 0.6329 \quad Q_3 = 1.3893 \quad IQR = 0.7564$

$$P[\text{Y is an outlier}] = pweibull(0.6329 - 1.5(0.7564), 2, 1.18) + 1 - pweibull(1.3893 + 1.5(0.7564), 2, 1.18) = .0103$$

- Case 3: Y is Weibull$(\gamma = .2, \alpha = 1.18)$ distribution

  $Q_1 = 0.002325 \quad Q_3 = 6.0417 \quad IQR = 6.0394$

$$P[\text{Y is an outlier}] = pweibull(0.002325 - 1.5(6.0394), .2, 1.18) + 1 - pweibull(6.0417 + 1.5(6.0394), .2, 1.18) = .1892$$

- Case 4: Y is Weibull$(\gamma = .2, \alpha = 5)$ distribution

  $Q_1 = .009852 \quad Q_3 = 25.6004 \quad IQR = 25.5906$

$$P[\text{Y is an outlier}] = pweibull(.009852 - 1.5(25.5906), .2, 5) + 1 - pweibull(25.6004 + 1.5(25.5906), .2, 5) = .1892$$

What can we conclude from the above calculations?

The probability of an outlier remains the same if the scale parameter changes provided the shape parameter stays constant.

The probability of an outlier changes if the shape parameter changes.

Expected Number of Outliers in n Data Values

The **expected number** of outliers in a random sample of n observations, would be

$$E_n = nP[Y \text{ is an Outlier}].$$

This follows from the expected value of number of successes in n iid Bernoulli trials.

Thus, the expected number of outliers would depend on the population distribution.

For the following distributions, the probability of obtaining an outlier and the expected number of outliers in a random sample of $n = 100$ or $n = 1000$ observations from the specified distribution are given in the following table:

Distribution	p=P[Outlier]	$E_n = 100p$	$E_n = 1000p$
Normal	0.007	.7	7
Logistic	0.02439	2.44	24.4
Double Exponential	.0625	6.25	62.5
Cauchy	0.156	15.6	156

As can seen in the above table, the number of outliers observed in a box plot will depend heavily on the type of distribution from which the sample was taken and the size of the sample.

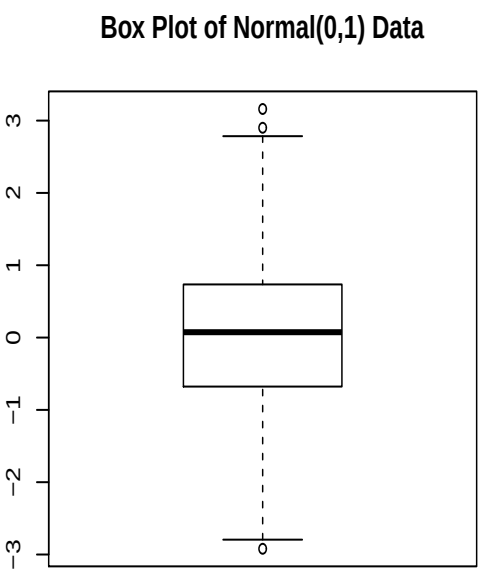
Note: For a symmetric distribution with location parameter = 0,

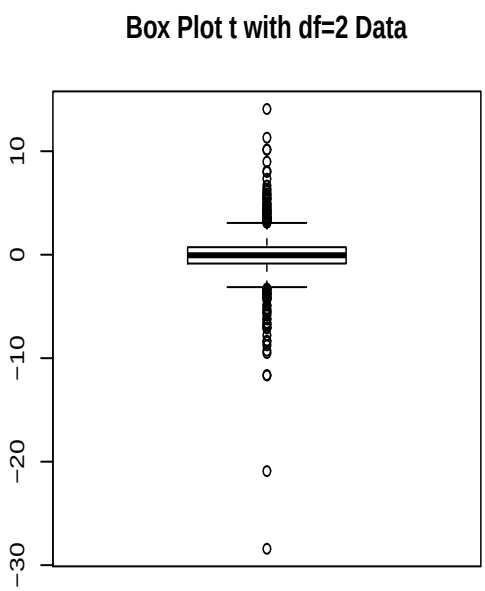
$$Q(.75) = -Q(.25) \;\Rightarrow\; IQR = -2Q(.25) \;\Rightarrow$$

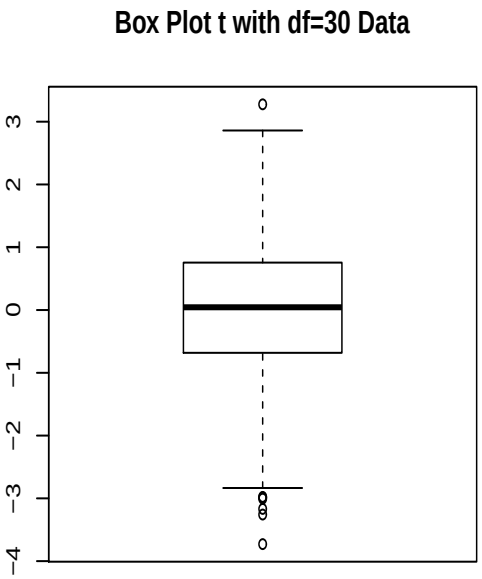
$$\begin{aligned} P[Y \text{ is an Outlier}] &= P\left[Y < Q_Y(.25) - 1.5(IQR_Y)\right] + P[Y > Q_Y(.75) + 1.5(IQR_Y)] \\ &= 2P\left[Y < 4Q(.25)\right] = 2F_Y(4Q(.25)) \end{aligned}$$

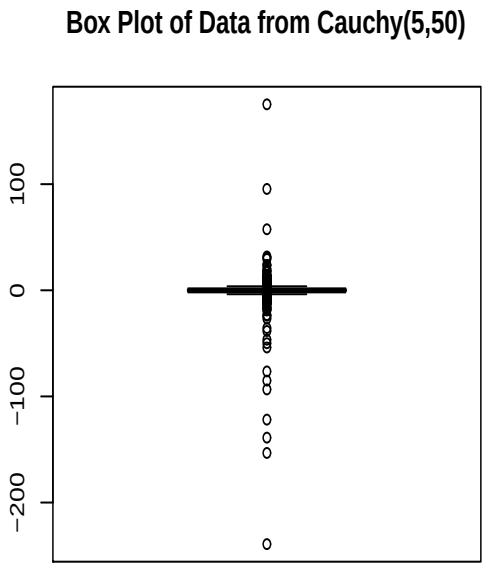
$$Q_{Normal}(.25) = -.6745 \quad Q_{logistic}(.25) = -1.0986 \quad Q_{DouExp}(.25) = -.9093 \quad Q_{Cauchy}(.25) = -1.0$$

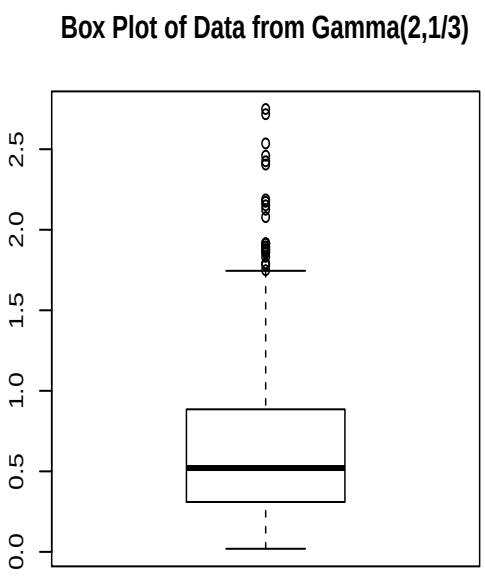
Box plots for six distributions were generated based on 1000 observations from each of the six distributions. The plots are displayed on the following page.

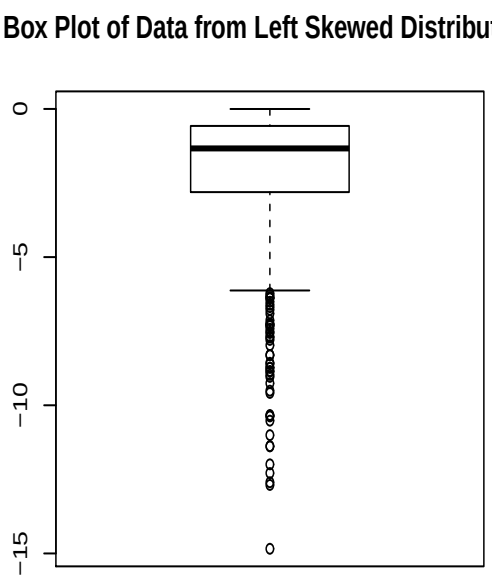

Box Plot of Normal(0,1) Data
3
2
1
0
−1
−2
−3

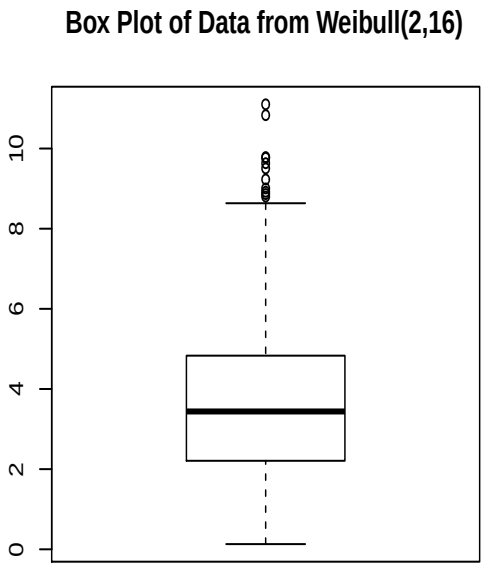

Box Plot t with df=2 Data
10
0
−10
−20
−30

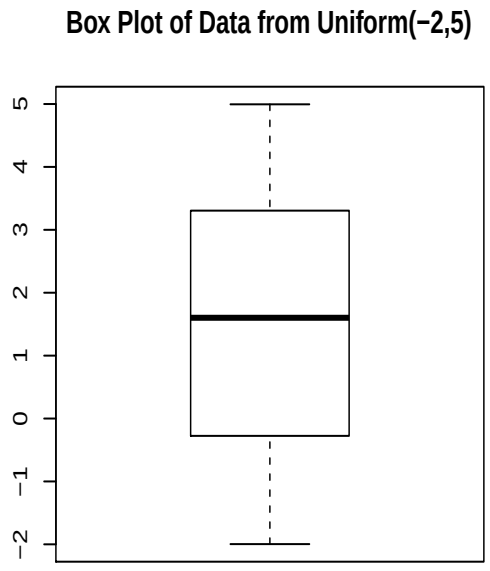

Box Plot t with df=30 Data
3
2
1
0
−1
−2
−3
−4


Box Plot of Data from Cauchy(5,50)
100
0
−100
−200


Box Plot of Data from Gamma(2,1/3)
0.0
0.5
1.0
1.5
2.0
2.5


Box Plot of Data from Left Skewed Distribution
0
−5
−10
−15


Box Plot of Data from Weibull(2,16)
0
2
4
6
8
10

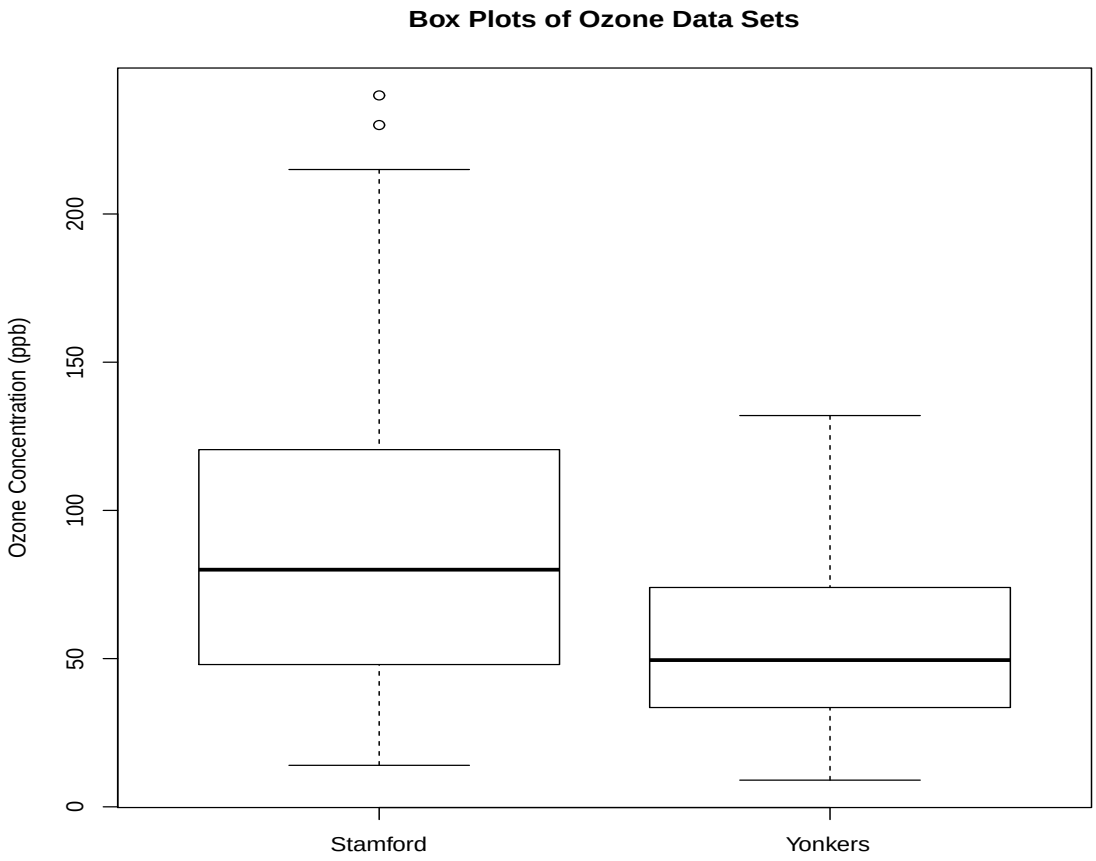

Box Plot of Data from Uniform(−2,5)
−2
−1
0
1
2
3
4
5

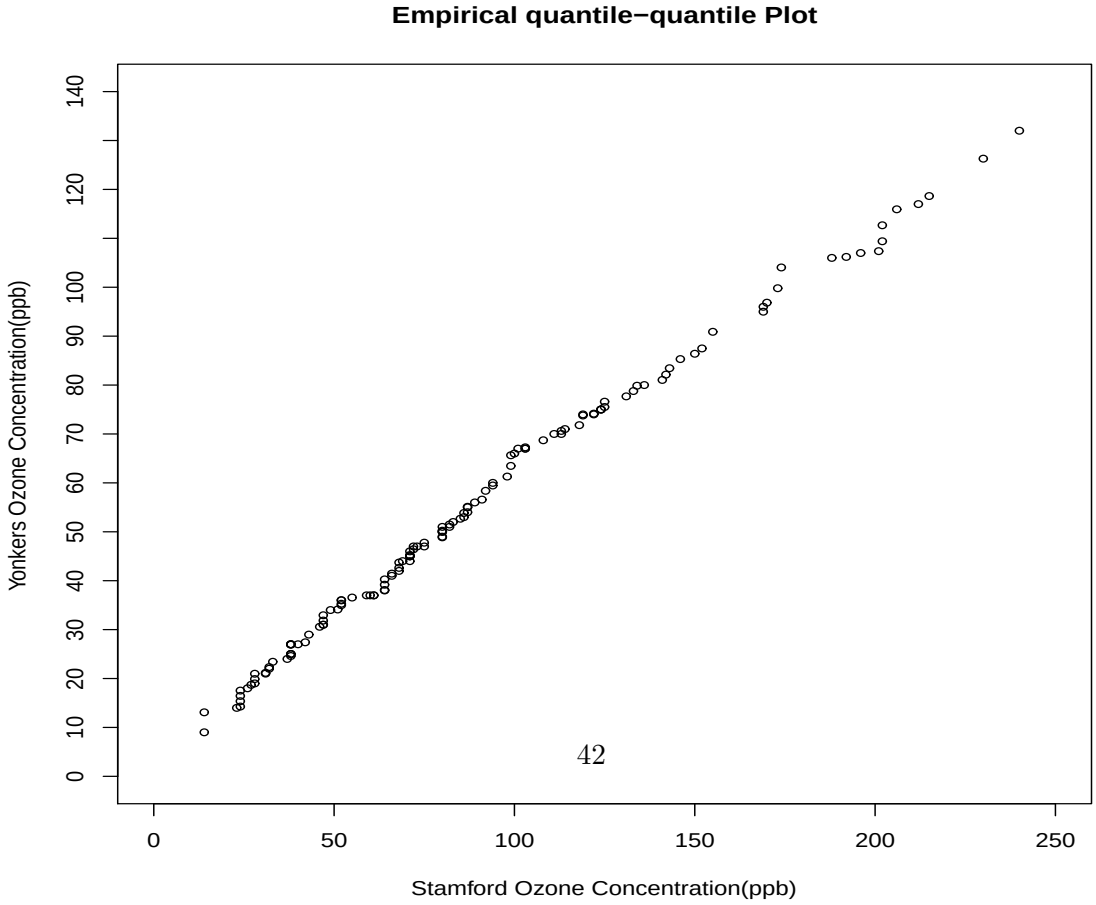
Box Plots, Quantile Plots, and Time Series Plots for Ozone Data

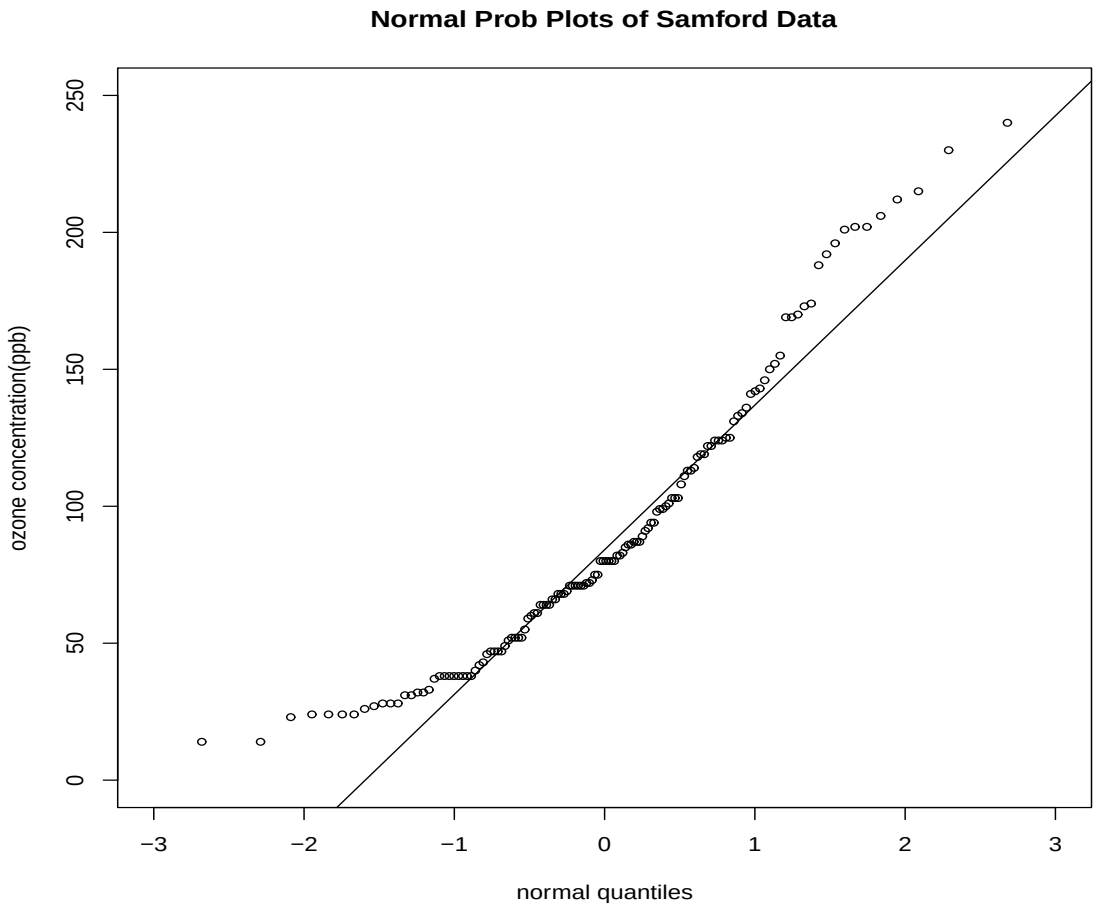
The following R code, generates various graphical comparisons of the Ozone data.

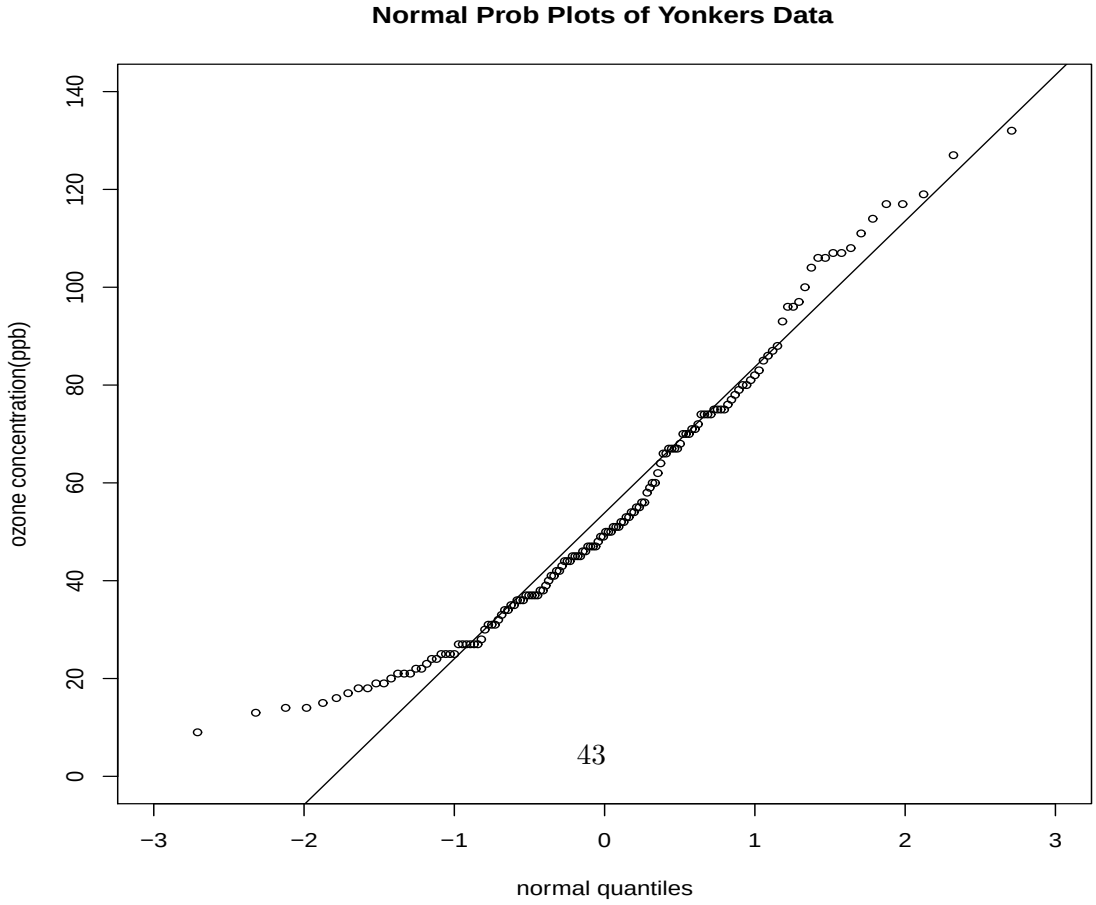
```
#The following R program generates various graphical comparisons of the
#Ozone data. The ozone data is in the files ozone1.DAT and ozone2.DAT
#The code is in Dostat in the folders: Files/Rcode/ozonecompare.R
#-------------------------------------------------------------------------
#input data:

          y1 = scan("u:/meth1/Rfiles/ozone1.DAT")
          y2 = scan("u:/meth1/Rfiles/ozone2.DAT")
          y1 = sort(y1)
          y2 = sort(y2)
          n1=length(y1)
          i=seq(1,n1,1)
          n2=length(y2)
          j=seq(1,n2,1)
          u1=(i-.5)/n1
          u2=(j-.5)/n2
          w1 = sort(log(y1))
          w2 = sort(log(y2))
          QZ1 = log(-log(1-u1))

          QZ2 = log(-log(1-u2))

#creates side-by-side box plots:

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

          boxplot(y1,y2,
                  names=c("Stamford","Yonkers"),
                  main="Box Plots of Ozone Data Sets",
                  ylab="Ozone Concentration (ppb)",plot=TRUE  )


#creates a quantile by quantile (q-q) plot:

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

          qqplot(y1,y2,
                 main="Empirical quantile-quantile Plot",cex=.75,
                 ylab="Yonkers Ozone Concentration(ppb)",
                 xlab="Stamford Ozone Concentration(ppb)",ylim=c(0,140),
                 xlim=c(0,250),lab=c(7,10,7))
```

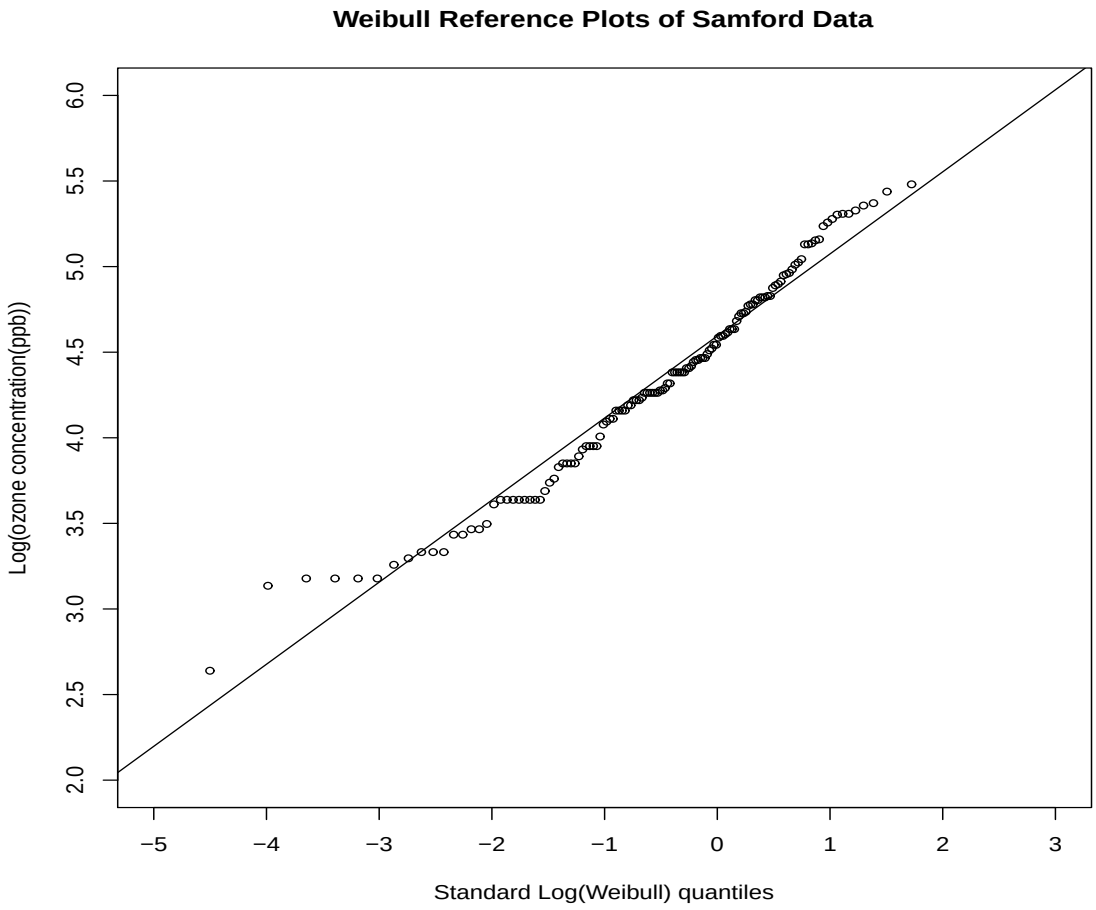
```
#creates a normal Reference Distribution plot:

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


          qqnorm(y1,main="Normal Prob Plots of Samford Data",
                 xlab="normal quantiles",ylab="ozone concentration(ppb)",
                 ylim=c(0,250),xlim=c(-3,3),lab=c(7,7,7),cex=.75)
          qqline(y1)

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


          qqnorm(y2,main="Normal Prob Plots of Yonkers Data",
                 xlab="normal quantiles",ylab="ozone concentration(ppb)",
                 ylim=c(0,140),xlim=c(-3,3),lab=c(7,7,7),cex=.75)
          qqline(y2)

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


#creates Weibull Reference Distribution Plots:

          plot(QZ1,w1,main="Weibull Reference Plots of Samford Data",
                 xlab="Standard Log(Weibull) quantiles",
                 ylab="Log(ozone concentration(ppb))",
                 ylim=c(2,6),xlim=c(-5,3),lab=c(7,10,7),cex=.75)
           abline(lm(w1~QZ1))
       postscript("u:meth1/psfiles/ozonecompareWeibullYonkers.ps",height=8,horizontal=F)


          plot(QZ2,w2,main="Weibull Reference Plots of Yonkers Data",
                 xlab="Standard Log(Weibull) quantiles",
                 ylab="Log(ozone concentration(ppb))",
                 ylim=c(2,6),xlim=c(-5,3),lab=c(7,10,7),cex=.75)
           abline(lm(w2~QZ2))

          graphics.off()
```

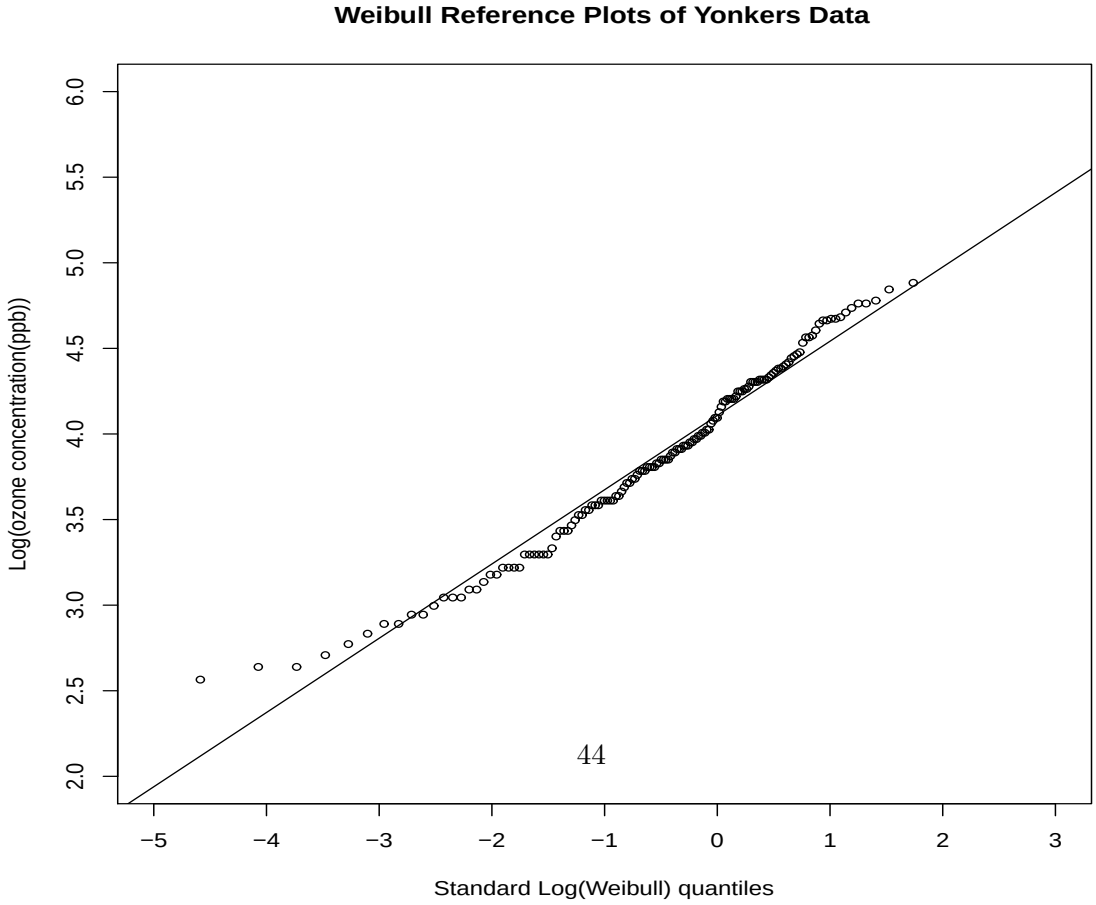

Box Plots of Ozone Data Sets
Ozone Concentration (ppb)
0
50
100
150
200
Stamford
Yonkers


Empirical quantile−quantile Plot
Yonkers Ozone Concentration(ppb)
0
10
20
30
40
50
60
70
80
90
100
120
140
0
50
100
150
200
250
Stamford Ozone Concentration(ppb)


Normal Prob Plots of Samford Data
ozone concentration(ppb)
0
50
100
150
200
250
−3
−2
−1
0
1
2
3
normal quantiles


Normal Prob Plots of Yonkers Data
ozone concentration(ppb)
0
20
40
60
80
100
120
140
−3
−2
−1
0
1
2
3
normal quantiles


Weibull Reference Plots of Samford Data
Log(ozone concentration(ppb))
Standard Log(Weibull) quantiles
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
−5
−4
−3
−2
−1
0
1
2
3


Weibull Reference Plots of Yonkers Data
Log(ozone concentration(ppb))
Standard Log(Weibull) quantiles
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
−5
−4
−3
−2
−1
0
1
2
3

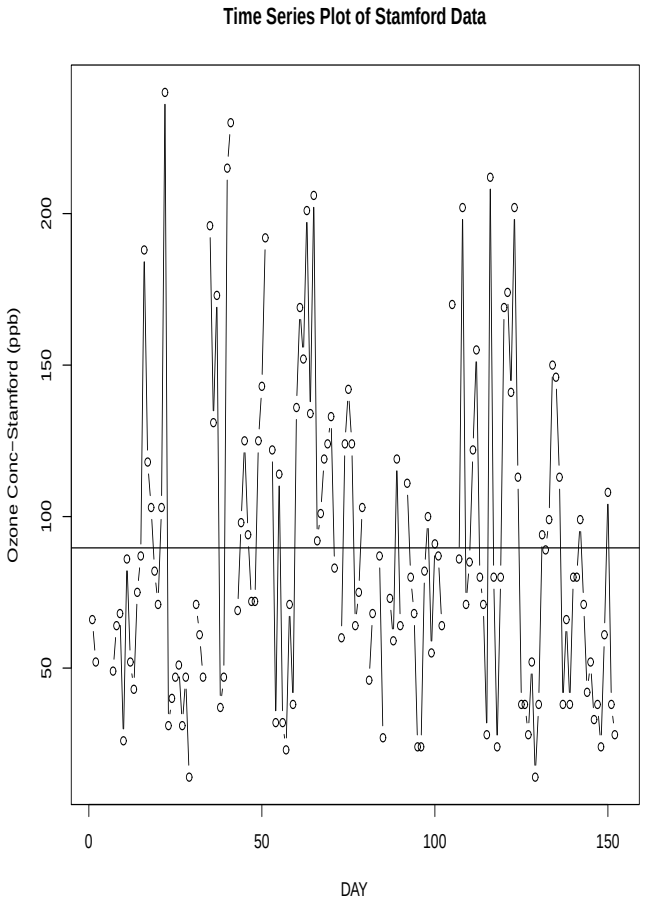
Time Series Plots

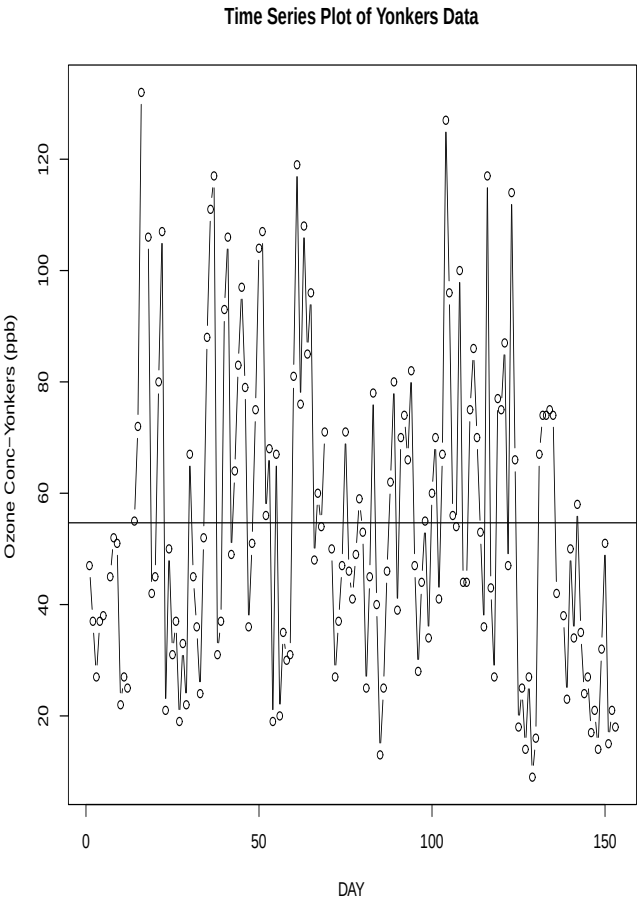
The following R code in the Dostat folders: Files/Rcode/ozonecompare,time produces plots which take into account the time element in the ozone data.

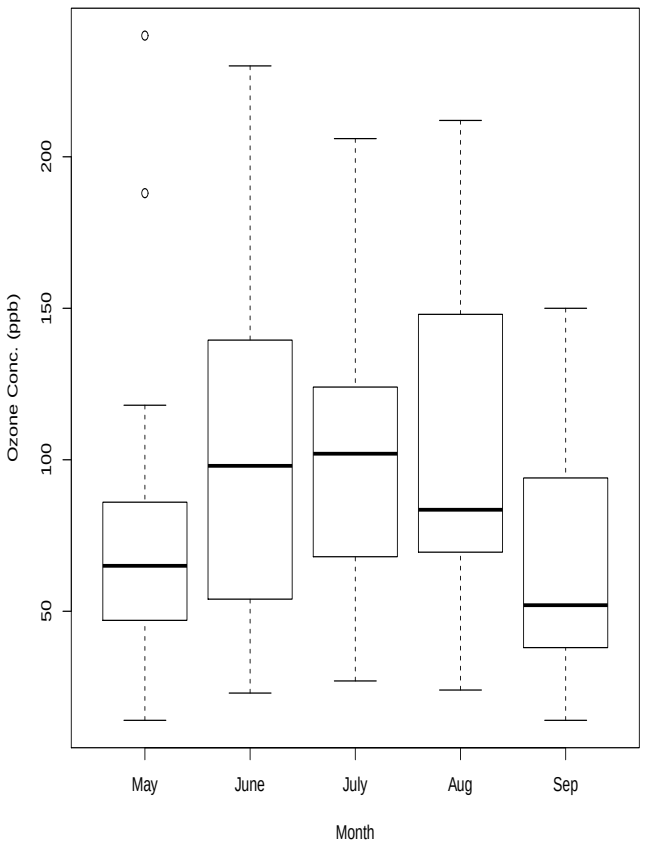
```
#input data:
         y1 = scan("u:/meth1/Rfiles/ozone1.DAT")
          y2 = scan("u:/meth1/Rfiles/ozone2.DAT")

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

#create time index

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

#classify data by month

          y1m = rep(0,26)
          y1jn= rep(0,27)
          y1jl= rep(0,26)
          y1a = rep(0,,28)
          y1s = rep(0,29)
          y2m = rep(0,28)
          y2jn= rep(0,30)
          y2jl= rep(0,30)
          y2a = rep(0,31)
          y2s = rep(0,29)
for (i in 1:26) {  y1m[i] = y1[i]}
for (i in 27:53) {  y1jn[i-26] = y1[i]}
for (i in 54:79) { y1jl[i-53] = y1[i]}
for (i in 80:107) {  y1a[i-79] = y1[i]}
for (i in 108:136) {  y1s[i-107] = y1[i]}
for (i in 1:28) { y2m[i] = y2[i]}
for (i in 29:58) { y2jn[i-28] = y2[i]}
for (i in 59:88) { y2jl[i-58] = y2[i]}
for (i in 89:119) { y2a[i-88] = y2[i]}
for (i in 120:148) { y2s[i-119] = y2[i]}
#ozone vs time plot:

 postscript("u:/meth1/psfiles/ozonetime1.ps",height=8,horizontal=F)

       plot(y1na,type="b",ylab="Ozone Conc-Stamford (ppb)",xlab="DAY",
          main="Time Series Plot of Stamford Data",cex=.9)
          abline(h=mean(y1))
graphics.off()

 postscript("u:/meth1/psfiles/ozonetime2.ps",height=8,horizontal=F)

         plot(y2na,type="b",ylab="Ozone Conc-Yonkers (ppb)",xlab="DAY",
          main="Time Series Plot of Yonkers Data",cex=.9)
         abline(h=mean(y2))
graphics.off()
```

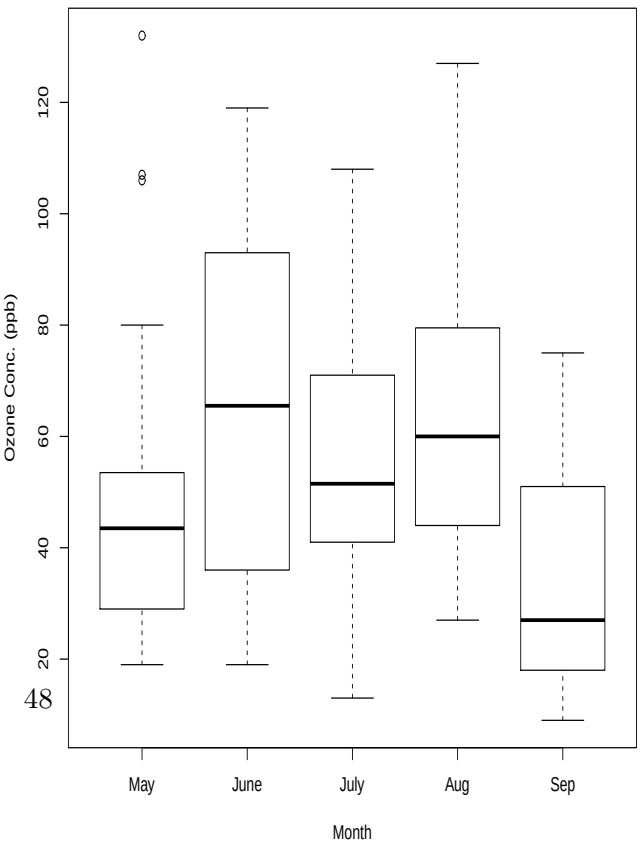
```
 postscript("u:/meth1/psfiles/ozonetime1a.ps",height=8,horizontal=F)

       plot(y1na,type="b",ylim=c(0,250),
          ylab="Ozone Conc-Stamford (ppb)",xlab="DAY",
          main="Time Series Plot of Stamford Data",cex=.9)
          abline(h=mean(y1))
graphics.off()


 postscript("u:/meth1/psfiles/ozonetime2a.ps",height=8,horizontal=F)

         plot(y2na,type="b",ylim=c(0,250),
          ylab="Ozone Conc-Yonkers (ppb)",xlab="DAY",
          main="Time Series Plot of Yonkers Data",cex=.9)
         abline(h=mean(y2))
graphics.off()


#side by side boxplots for the various months:

 postscript("u:/meth1/psfiles/ozonetimebox1.ps",height=8,horizontal=F)

          boxplot(y1m,y1jn,y1jl,y1a,y1s,xlab="Month",ylab="Ozone Conc. (ppb)",
          main="Boxplots of Ozone Conc. for Stamford by Month",
               names=c("May","June","July","Aug","Sep"))
 graphics.off()


 postscript("u:/meth1/psfiles/ozonetimebox2.ps",height=8,horizontal=F)

          boxplot(y2m,y2jn,y2jl,y2a,y2s,xlab="Month",ylab="Ozone Conc. (ppb)",
          main="Boxplots of Ozone Conc. for Yonkers by Month",
               names=c("May","June","July","Aug","Sep"))
graphics.off()

   postscript("u:/meth1/psfiles/ozonetimebox1a.ps",height=8,horizontal=F)

          boxplot(y1m,y1jn,y1jl,y1a,y1s,xlab="Month",ylab="Ozone Conc. (ppb)",
          main="Ozone Conc. for Stamford",ylim=c(0,250),
               names=c("May","June","July","Aug","Sep"),cex=.75)
 graphics.off()


    postscript("u:/meth1/psfiles/ozonetimebox2a.ps",height=8,horizontal=F)

          boxplot(y2m,y2jn,y2jl,y2a,y2s,xlab="Month",ylab="Ozone Conc. (ppb)",
          main="Ozone Conc. for Yonkers", ylim=c(0,250),
               names=c("May","June","July","Aug","Sep"))

         graphics.off()
```

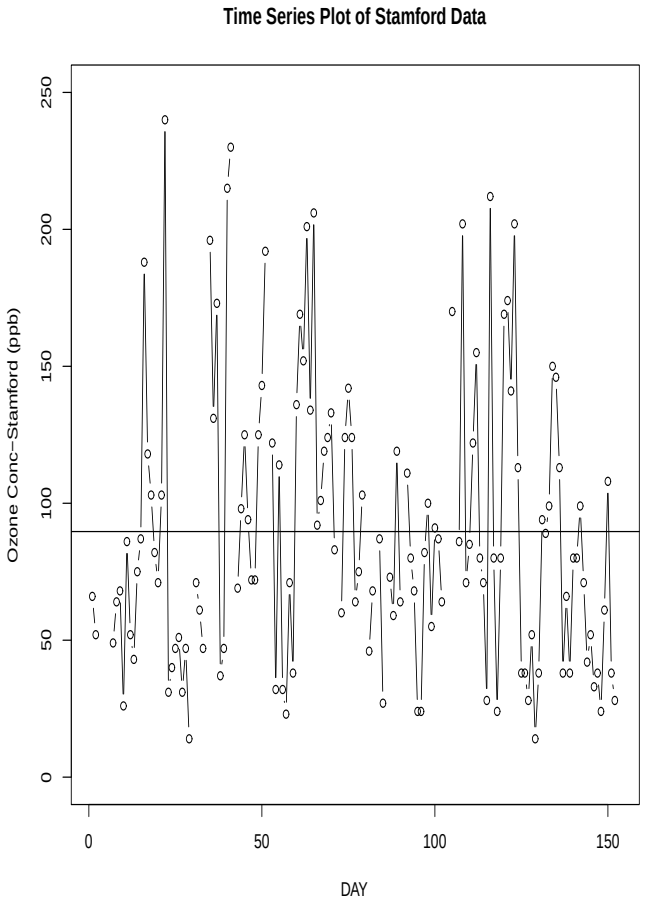
```
Call: acf(x = StamfordOzone)


AUTOCORRELATION MATRIX:

   LAG    YONKERS OZONE
   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    STAMFORD OZONE
   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
```

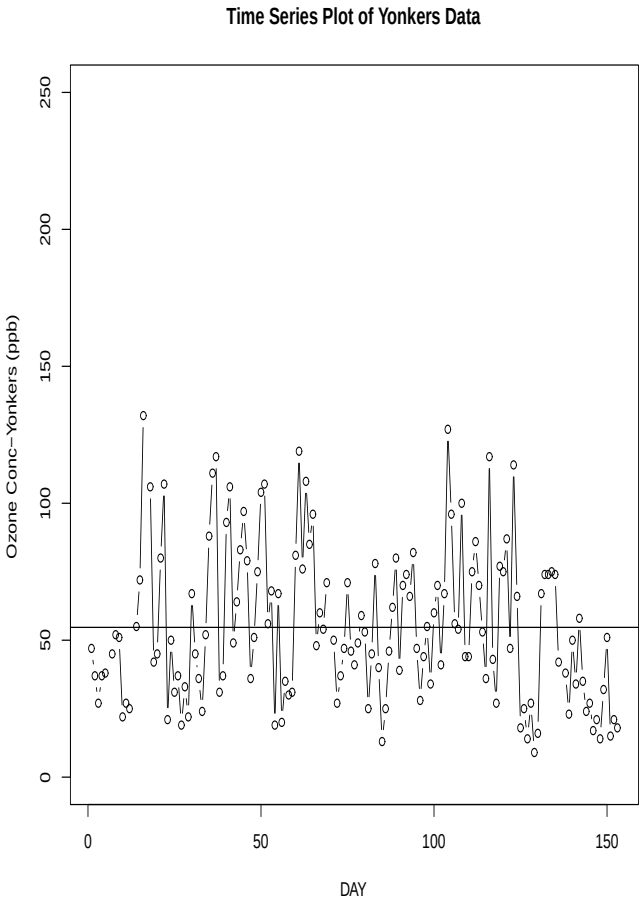

Time Series Plot of Stamford Data
Ozone Conc−Stamford (ppb)
50
100
150
200
0
50
100
150
DAY

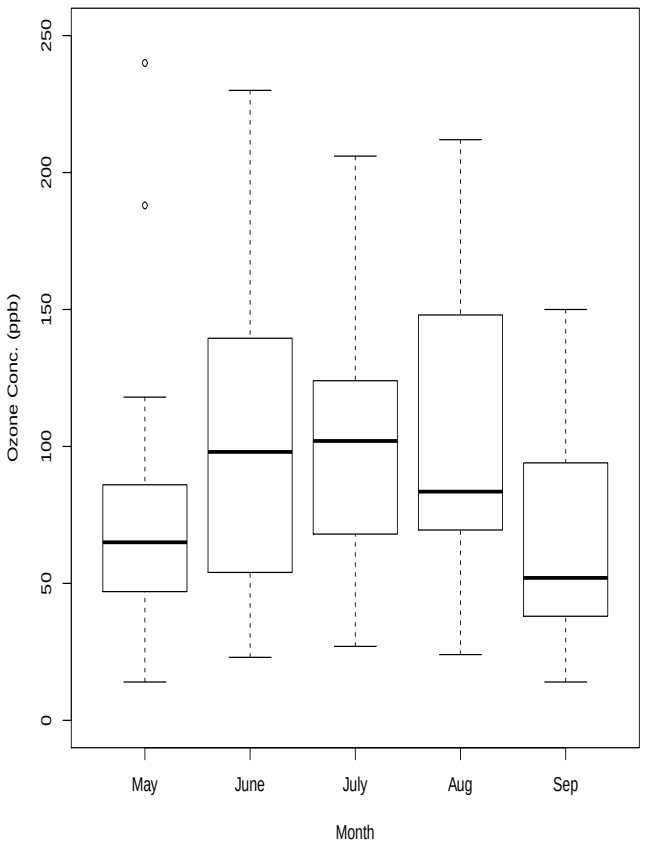

Time Series Plot of Yonkers Data
Ozone Conc−Yonkers (ppb)
20
40
60
80
100
120
0
50
100
150
DAY

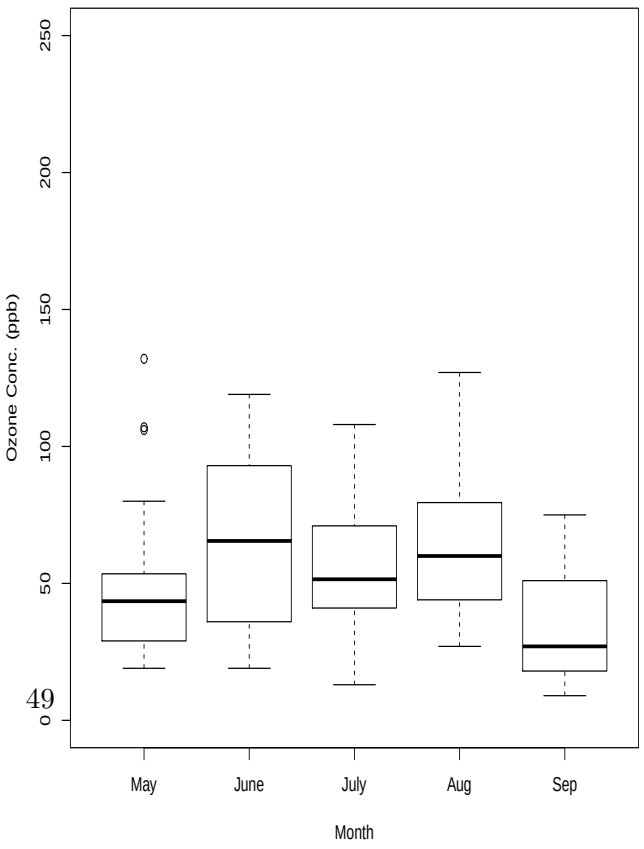

Boxplots of Ozone Conc. for Stamford by Month
Ozone Conc. (ppb)
50
100
150
200
May
June
July
Aug
Sep
Month


Boxplots of Ozone Conc. for Yonkers by Month
Ozone Conc. (ppb)
20
40
60
80
100
120
May
June
July
Aug
Sep
Month


Time Series Plot of Stamford Data
Ozone Conc−Stamford (ppb)
0
50
100
150
200
250
0
50
100
150
DAY


Time Series Plot of Yonkers Data
Ozone Conc−Yonkers (ppb)
0
50
100
150
200
250
0
50
100
150
DAY


Ozone Conc. for Stamford
Ozone Conc. (ppb)
0
50
100
150
200
250
May
June
July
Aug
Sep
Month


Ozone Conc. for Yonkers
Ozone Conc. (ppb)
0
50
100
150
200
250
May
June
July
Aug
Sep
Month

Scatterplots of Multiple Response Data

The following set of graphs will examine various situations when we have more than one response from the sampled experimental unit. We will consider the following situations:

1. **Case 1:** Two Qualitative Variables - Figure 3.29

2. **Case 2:** A Qualitative Variable and A Quantitative Variable - Figure 3.30

3. **Case 3:** Several Quantitative Variables - Figures 3.32

4. **Case 4:** Several Quantitative Variables - Matrix Plot

5. **Case 5:** Several Quantitative Variables - Draftsman Plot

6. **Case 6:** Several Quantitative Variables - Regression Plot

Consider first the problem of summarizing data from two qualitative variables. Cross-tabulations can be constructed to form a **contingency table.** The rows of the table identify the categories of one variable, and the columns identify the categories of the other variable. The entries in the table are the number of times each value of one variable occurs with each possible value of the other. For example, a television viewing survey was conducted on 1,500 individuals. Each individual surveyed was asked to state his or her place of residence and network preference for national news. The results of the survey are shown in Table 3.7. As you can see, 144 urban residents preferred ABC, 135 urban residents preferred CBS, and so on.

TABLE 3.7

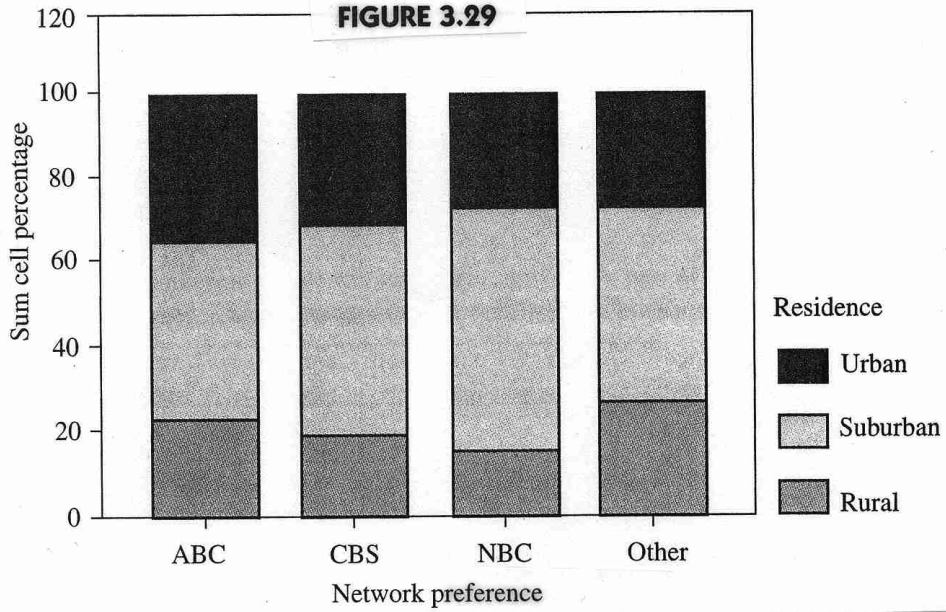
	Residence			
Network Preference	**Urban**	**Suburban**	**Rural**	**Total**
ABC	144	180	90	414
CBS	135	240	96	471
NBC	108	225	54	387
Other	63	105	60	228
Total	450	750	300	1,500

TABLE 3.8

	Residence			
Network Preference	**Urban**	**Suburban**	**Rural**	**Total**
ABC	34.8	43.5	21.7	100 (n = 414)
CBS	28.7	50.9	20.4	100 (n = 471)
NBC	27.9	58.1	14.0	100 (n = 387)
Other	27.6	46.1	26.3	100 (n = 228)

An extension of the bar graph provides a convenient method for displaying data from a pair of qualitative variables. Figure 3.29 is a **stacked bar graph,** which displays the data in Table 3.8.

The graph represents the distribution of television viewers of each of the major network's news programs based on the location of the viewer's residence. This type of information is often used by advertisers to determine on which networks' programs they will place their commercials.



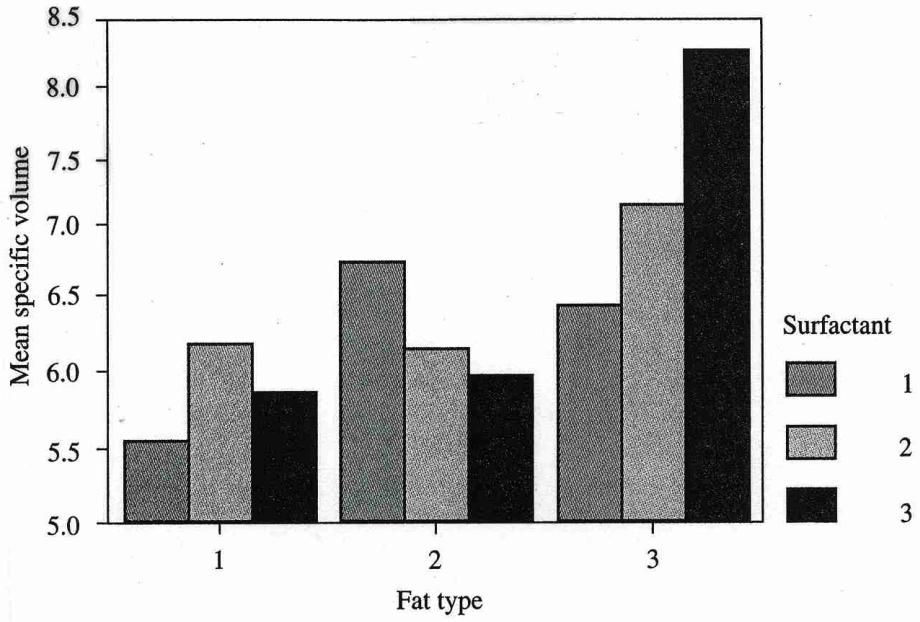
A second extension of the bar graph provides a convenient method for displaying the relationship between a single quantitative and a qualitative variable. A food scientist is studying the effects of combining different types of fats with different surfactants on the specific volume of baked bread loaves. The experiment is designed with three levels of surfactant and three levels of fat, a 3×3 factorial experiment with a varying number of loaves baked from each of the nine treatments. She bakes bread from dough mixed from the nine different combinations of the types of fat and types of surfactants and then measures the specific volume of the bread. The data and summary statistics are displayed in Table 3.9.

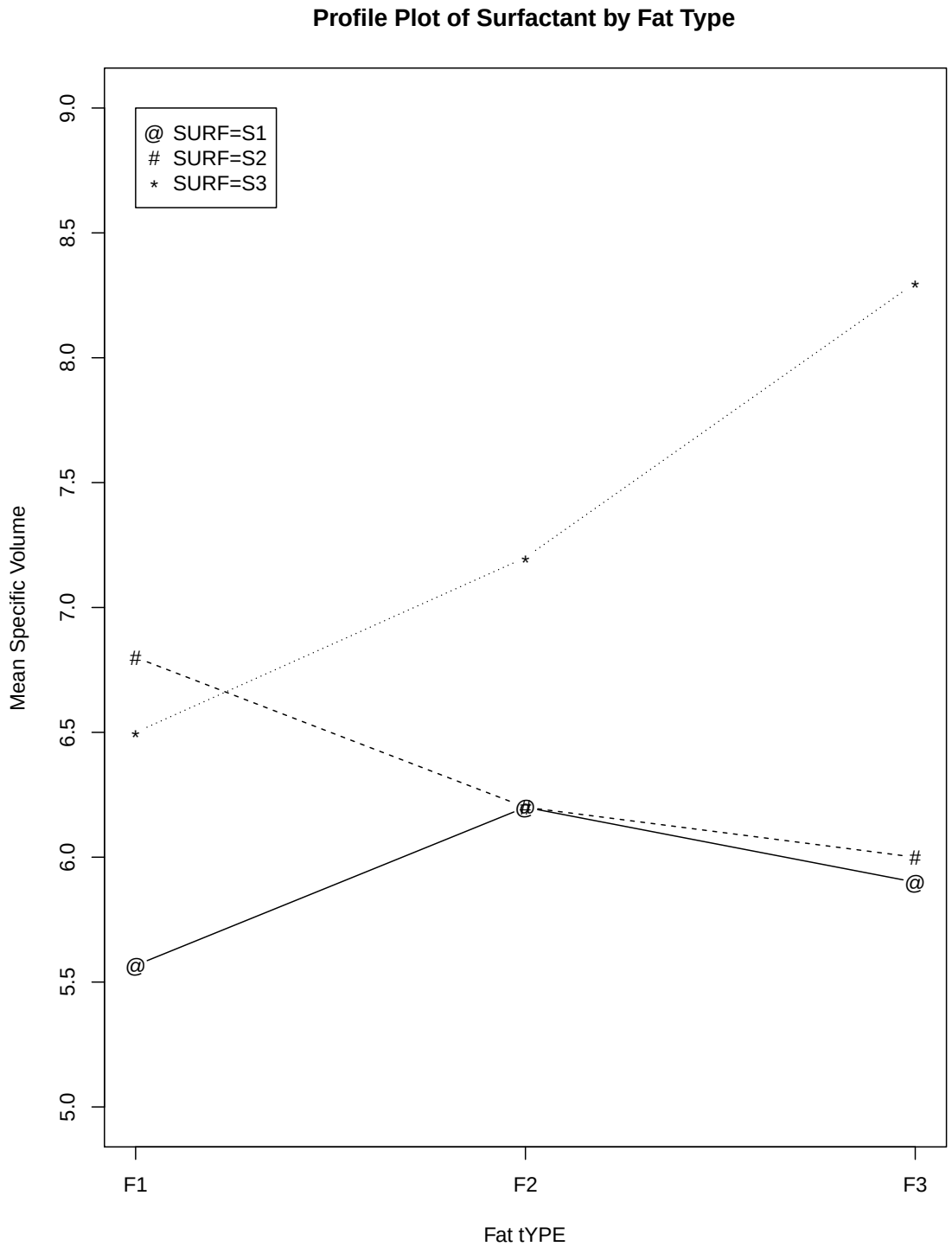
In this experiment, the scientist wants to make inferences from the results of the experiment to the commercial production process. Figure 3.30 is a **cluster bar graph** from the baking experiment. This type of graph allows the experimenter to examine the simultaneous effects of two factors, type of fat and type of surfactant, on the specific volume of the bread. Thus, the researcher can examine the differences in the specific volumes of the nine different ways in which the bread was formulated.

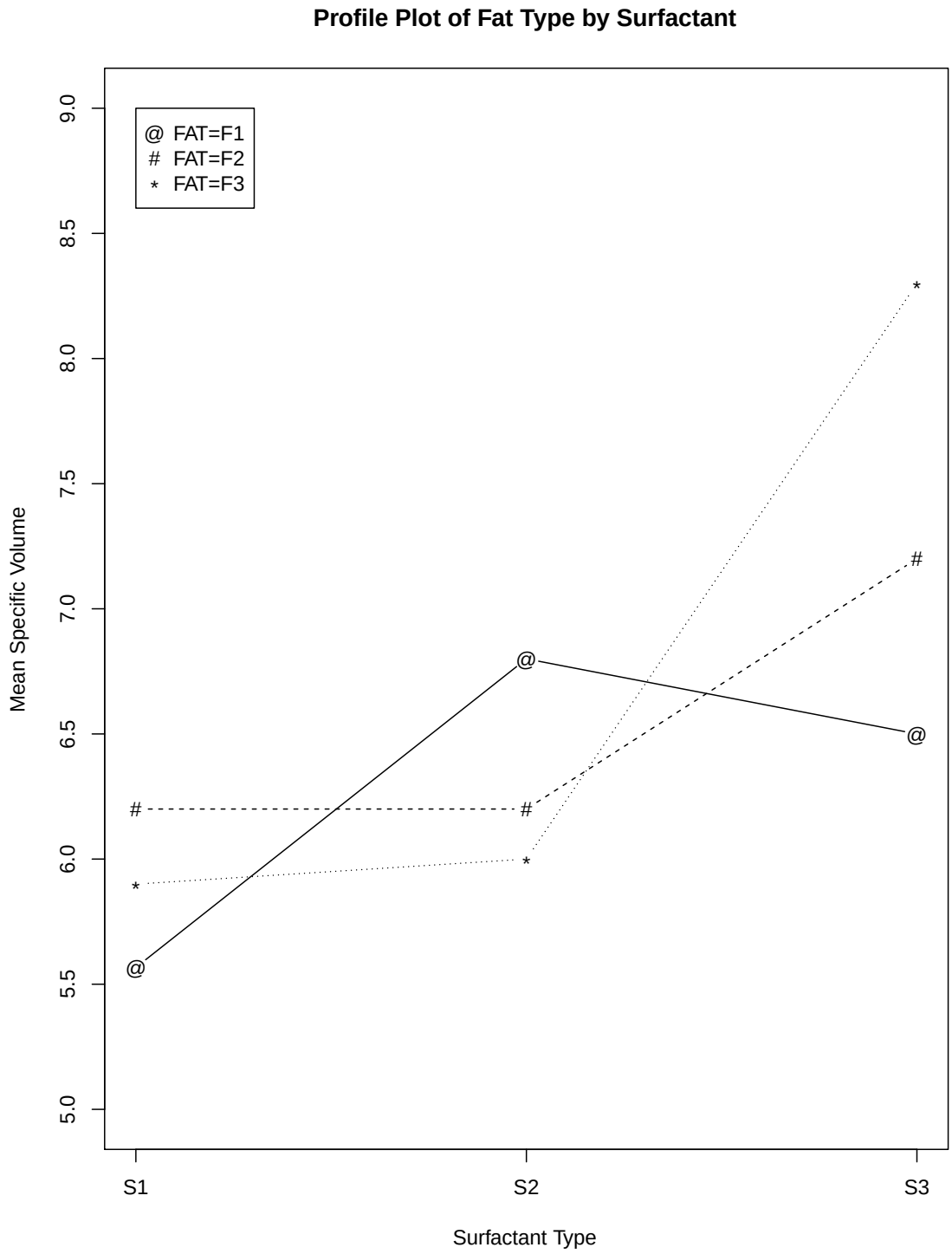
TABLE 3.9

Fat	Surfactant	Mean	Standard Deviation	*N*
1	1	5.567	1.206	3
	2	6.200	.794	3
	3	5.900	.458	3
	Total	5.889	.805	9
2	1	6.800	.794	3
	2	6.200	.849	2
	3	6.000	.606	4
	Total	6.311	.725	9
3	1	6.500	.849	2
	2	7.200	.668	4
	3	8.300	1.131	2
	Total	7.300	.975	8
Total	1	6.263	1.023	8
	2	6.644	.832	9
	3	6.478	1.191	9
	Total	6.469	.997	26

FIGURE 3.30



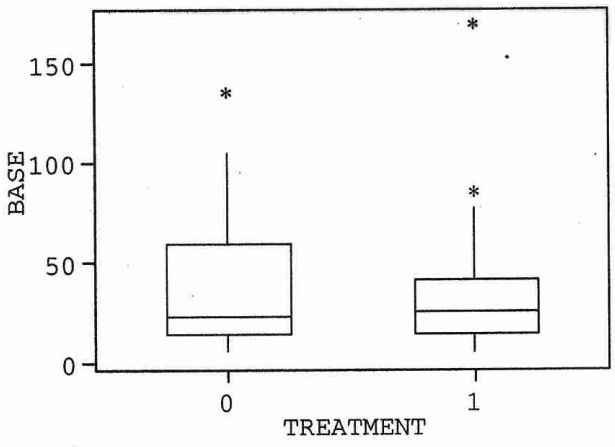

Profile Plot of Surfactant by Fat Type
@ SURF=S1
# SURF=S2
* SURF=S3
Mean Specific Volume
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
F1
F2
F3
Fat tYPE


Profile Plot of Fat Type by Surfactant
@ FAT=F1
# FAT=F2
* FAT=F3
Mean Specific Volume
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
S1
S2
S3
Surfactant Type

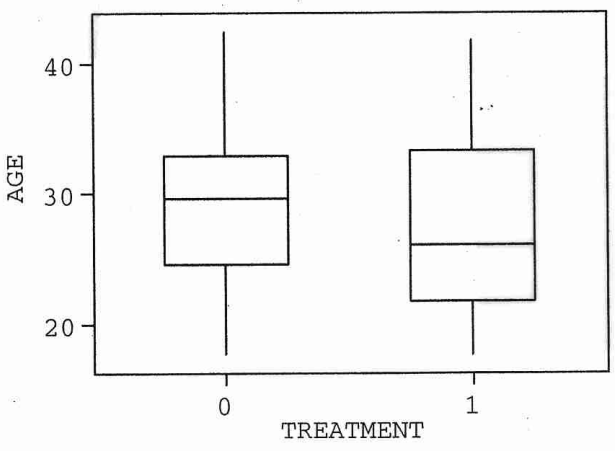
Finally, we can construct data plots for summarizing the relation between several quantitative variables. Consider the following example. Thall and Vail (1990) described a study to evaluate the effectiveness of the anti-epileptic drug progabide as an adjuvant to standard chemotherapy. A group of 59 epileptics was selected to be used in the clinical trial. The patients suffering from simple or complex partial seizures were randomly assigned to receive either the anti-epileptic drug progabide or a placebo. At each of four successive postrandomization clinic visits, the number of seizures occurring over the previous 2 weeks was reported. The measured variables were $y_i (i = 1, 2, 3, 4$—the seizure counts recorded at the four clinic visits); Trt (x_1) (0 is the placebo, 1 is progabide); Base (x_2), the baseline seizure rate; Age (x_3), the patient's age in years. The data and summary statistics are given in Tables 3.10 and 3.11.

The first plots are **side-by-side boxplots** that compare the base number of seizures and ages of the treatment patients to the patients assigned to the placebo. These plots provide a visual assessment of whether the treatment patients and placebo patients had similar distributions of age and base seizure counts prior to the start of the clinical trials. An examination of Figure 3.32(a) reveals that the number of seizures prior to the beginning of the clinical trials has similar patterns for the two groups of patients. There is a single patient with a base seizure count greater than 100 in both groups. The base seizure count for the placebo group is somewhat more variable than for the treatment group—its box is wider than the box for the treatment group.

FIGURE 3.32



(a) Boxplot of base by treatment



(b) Boxplot of age by treatment

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

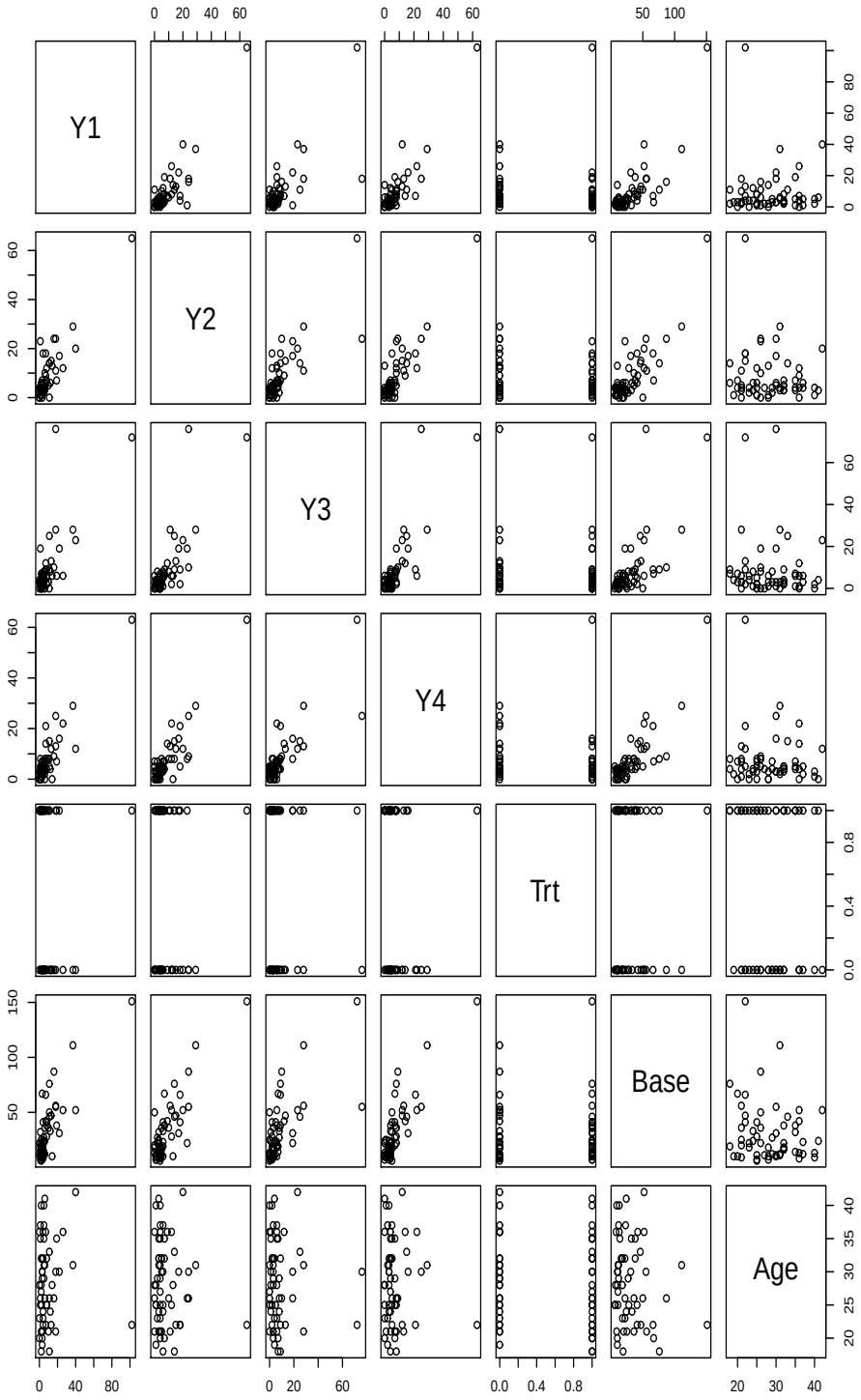
Cell Contents: Pearson correlation
P-Value

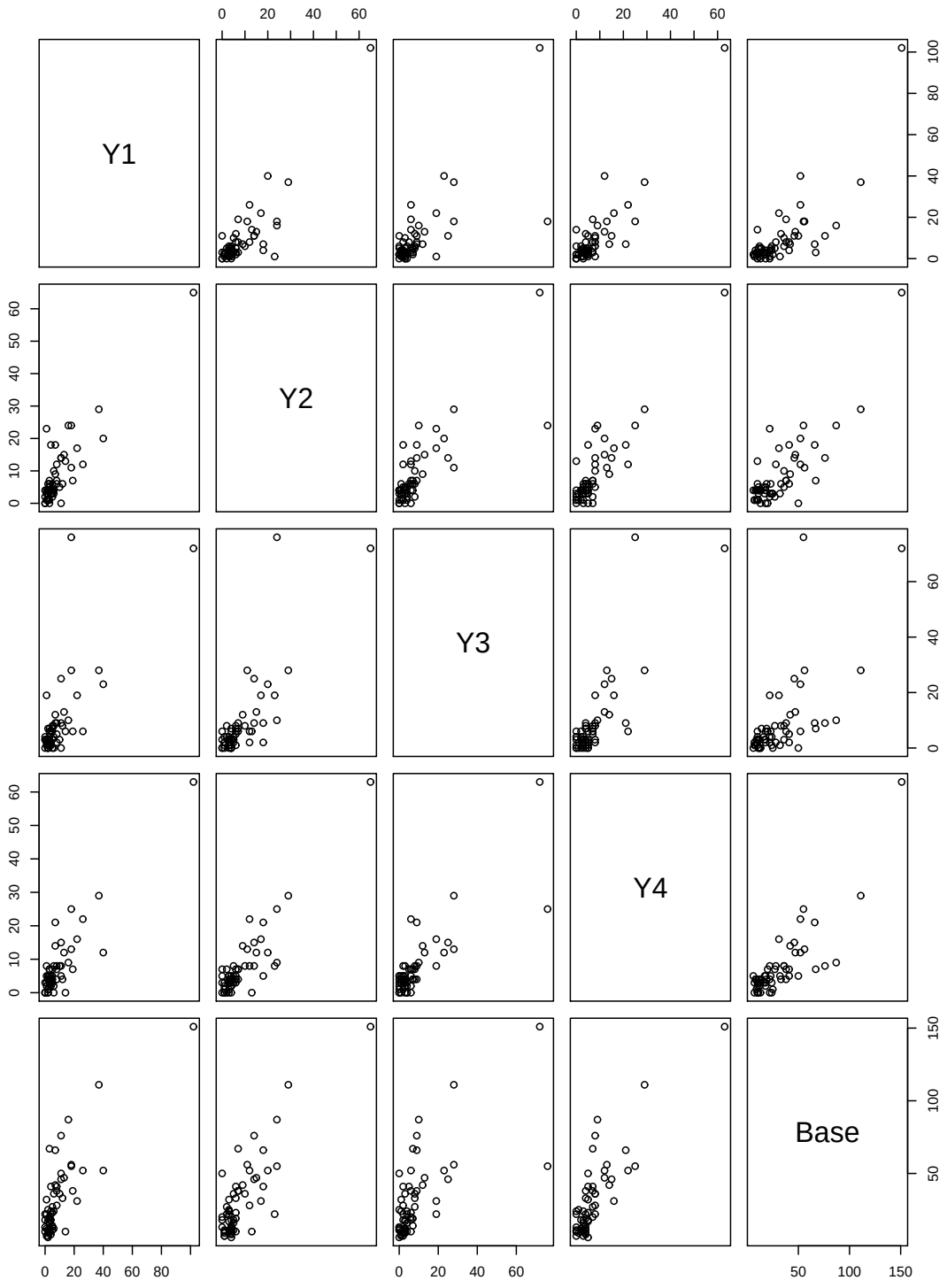
The following R code will produce various plots of the Epilepsy data:

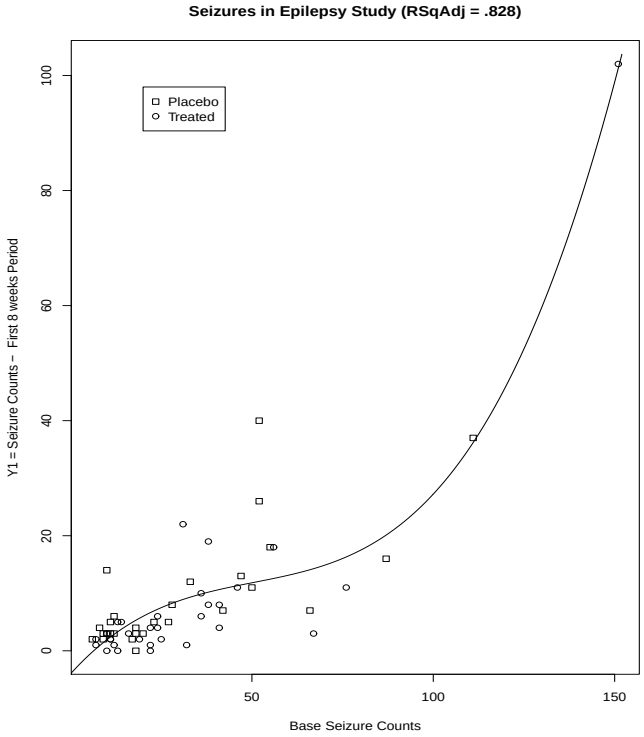
```
par(ask=TRUE)

x = read.csv("u:\\meth1\\Rfiles\\epilpsy.csv",header=TRUE)
attach(x)

postscript("u:/meth1/psfiles/epsBoxBase.ps")

boxplot(split(Base,Trt),ylab="Base Seizure Count",xlab="Treatment",
main="Boxplots for Epilepsy Study")

postscript("u:/meth1/psfiles/epsBoxAge.ps")

boxplot(split(Age,Trt),ylab="Age of Patient",xlab="Treatment",
main="Boxplots for Epilepsy Study")

Y <- cbind(Y1,Y2,Y3,Y4,Base,Age)
c = cor(Y)
c = round(c,3)

postscript("u:/meth1/psfiles/epsMatrix1.ps")

pairs(~Y1+Y2+Y3+Y4+Trt+Base+Age)

postscript("u:/meth1/psfiles/epsMatrix2.ps")

pairs(~Y1+Y2+Y3+Y4+Base)

postscript("u:/meth1/psfiles/RegY1Base.ps")

#Regression Plot of Y1 vs Base
m1 = lm(Y1~Base+I(Base^2)+I(Base^3))
summary(m1)
plot(Base,Y1,main="Seizures in Epilepsy Study (RSqAdj = .828)",xlab="Base Seizure Counts",
ylab="Y1 = Seizure Counts -  First 8 weeks Period",pch=as.numeric(Trt))
legend(20,98,c("Placebo","Treated"),pch=0:1)
av=seq(0,152,.1)
bv = predict(m1,list(Base=av))
lines(av,bv)

postscript("u:/meth1/psfiles/RegY1Age.ps")

#Regression Plot of Y1 vs Age
m2 = lm(Y1~Age+I(Age^2)+I(Age^3))
summary(m2)
plot(Age,Y1,main="Seizures in Epilepsy Study  (RSqAdj = 0)",xlab="Age of Patient",
ylab="Y1 = Seizure Counts - First 8 weeks Period",pch=as.numeric(Trt))
legend(36,98,c("Placebo","Treated"),pch=0:1)
av=seq(0,42,.1)
bv = predict(m2,list(Age=av))
lines(av,bv)

postscript("u:/meth1/psfiles/RegY1BaseWO.ps")

#Regression Plot of Y1 vs Base with outlier removed
m3 = update(m1,subset=(1:59)[-c(49)])
summary(m3)
```

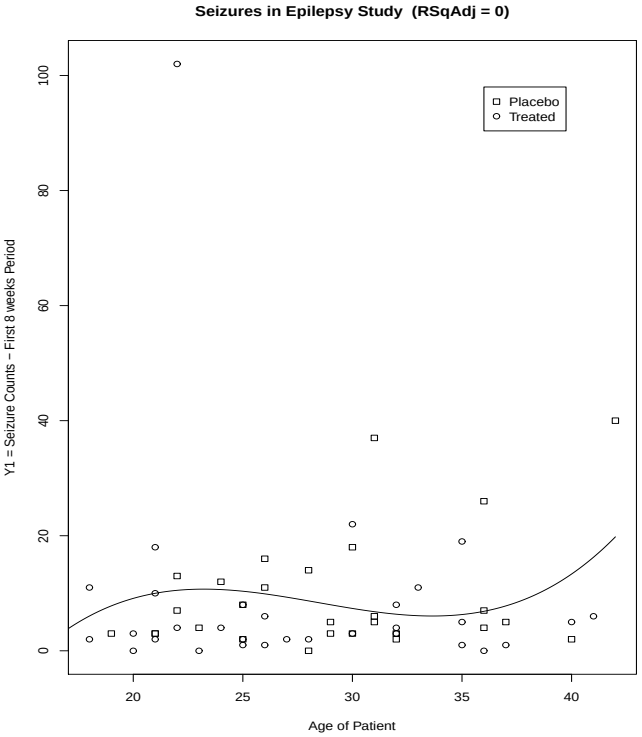
```
plot(Base[-c(49)],Y1[-c(49)],main="Seizures in Epilepsy Study Without Outlier
(RSqAdj = .449)",xlab="Base Seizure Counts",
ylab="Y1 = Seizure Counts - First 8 weeks Period",pch=as.numeric(Trt))
legend(13,38,c("Placebo","Treated"),pch=0:1)
av=seq(0,112,.1)
bv = predict(m1,list(Base=av))
lines(av,bv)

postscript("u:/meth1/psfiles/RegY1AgeWO.ps")

#Regression Plot of Y1 vs Age with outlier removed
m4 = update(m2,subset=(1:59)[-c(49)])
summary(m4)
plot(Age[-c(49)],Y1[-c(49)],main="Seizures in Epilepsy Study Without Outlier
(RSqAdj = .026)",xlab="Age of Patient",
ylab="Seizure Counts - First 8 wks",pch=as.numeric(Trt))
legend(18,38,c("Placebo","Treated"),pch=0:1)
av=seq(0,42,.1)
bv = predict(m4,list(Age=av))
lines(av,bv)

postscript("u:/meth1/psfiles/RegY4Base.ps")

#Regression Plot of Y4 vs Base
m5 = lm(Y4~Base+I(Base^2)+I(Base^3))
summary(m5)
plot(Base,Y4,main="Seizures in Epilepsy Study  (RSqAdj = .812)",xlab="Base Seizure Counts",
ylab="Y4 = Seizure Counts - Fourth 8 weeks Period",pch=as.numeric(Trt))
legend(20,60,c("Placebo","Treated"),pch=0:1)
av=seq(0,151,.1)
bv = predict(m5,list(Base=av))
lines(av,bv)

postscript("u:/meth1/psfiles/RegY4Age.ps")

#Regression Plot of Y4 vs Age
m6 = lm(Y4~Age+I(Age^2)+I(Age^3))
summary(m6)
plot(Age,Y4,main="Seizures in Epilepsy Study  (RSqAdj = 0)",xlab="Age of Patient",
ylab="Y4 = Seizure Counts - Fourth 8 weeks Period",pch=as.numeric(Trt))
legend(36,98,c("Placebo","Treated"),pch=0:1)
av=seq(0,42,.1)
bv = predict(m6,list(Age=av))
lines(av,bv)

postscript("u:/meth1/psfiles/RegY4BaseWO.ps")

#Regression Plot of Y4 vs Base with outlier removed
m7 = update(m5,subset=(1:59)[-c(49)])
summary(m7)
plot(Base[-c(49)],Y4[-c(49)],main="Seizures in Epilepsy Study Without Outlier
(RSqAdj = .551)",xlab="Base Seizure Counts",
ylab="Y4 = Seizure Counts - Fourth 8 wks",pch=as.numeric(Trt))
legend(13,28,c("Placebo","Treated"),pch=0:1)
av=seq(0,112,.1)
bv = predict(m7,list(Base=av))
lines(av,bv)
```

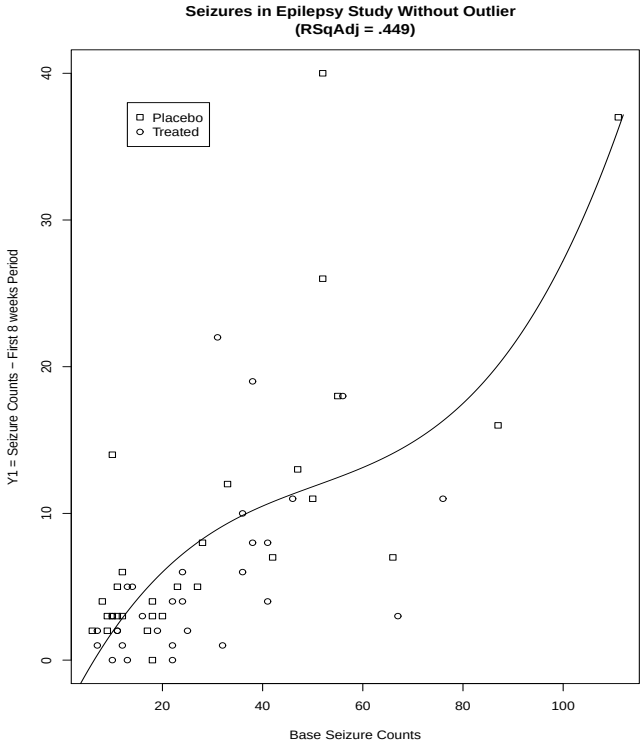
```
postscript("u:/meth1/psfiles/RegY4AgeWO.ps")

#Regression Plot of Y4 vs Age with outlier removed
m8 = update(m6,subset=(1:59)[-c(49)])
summary(m8)
plot(Age[-c(49)],Y4[-c(49)],main="Seizures in Epilepsy Study Without Outlier
(RSqAdj = 0)",xlab="Age of Patient",
ylab="Y4 = Seizure Counts - Fourth 8 wks",pch=as.numeric(Trt))
legend(20,38,c("Placebo","Treated"),pch=0:1)
av=seq(0,42,.1)
bv = predict(m8,list(Age=av))
lines(av,bv)
```

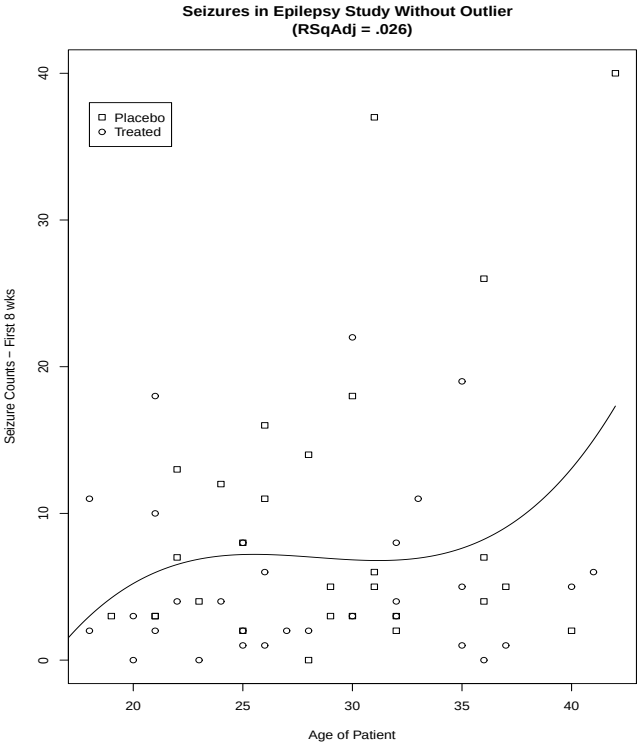

Y1
Y2
Y3
Y4
Trt
Base
Age

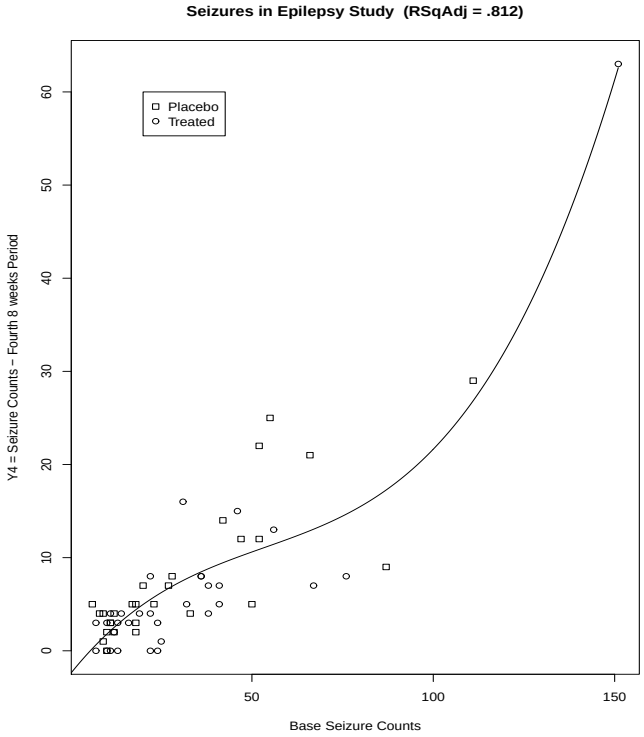

Y1
Y2
Y3
Y4
Base
0 20 40 60
0 20 40 60
0 20 40 60 80
0 20 40 60
50 100 150

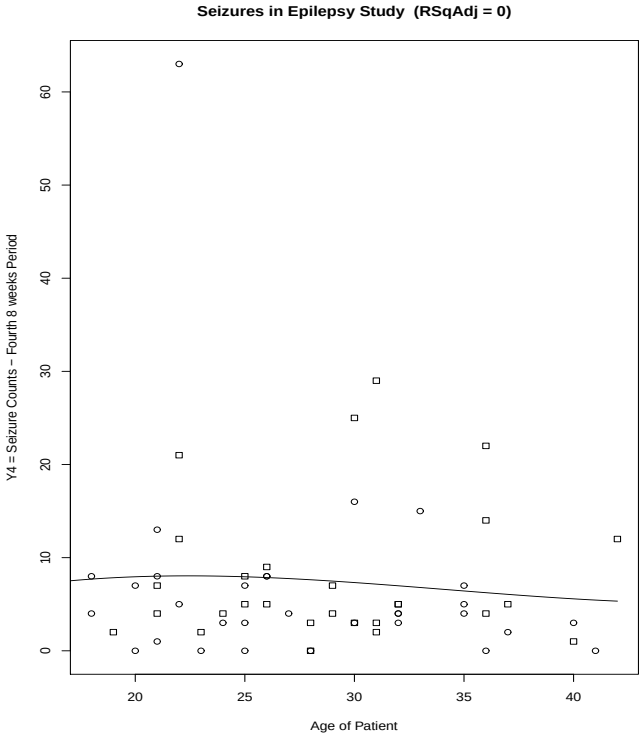

Seizures in Epilepsy Study (RSqAdj = .828)
Placebo
Treated
Y1 = Seizure Counts – First 8 weeks Period
Base Seizure Counts
0
20
40
60
80
100
50
100
150

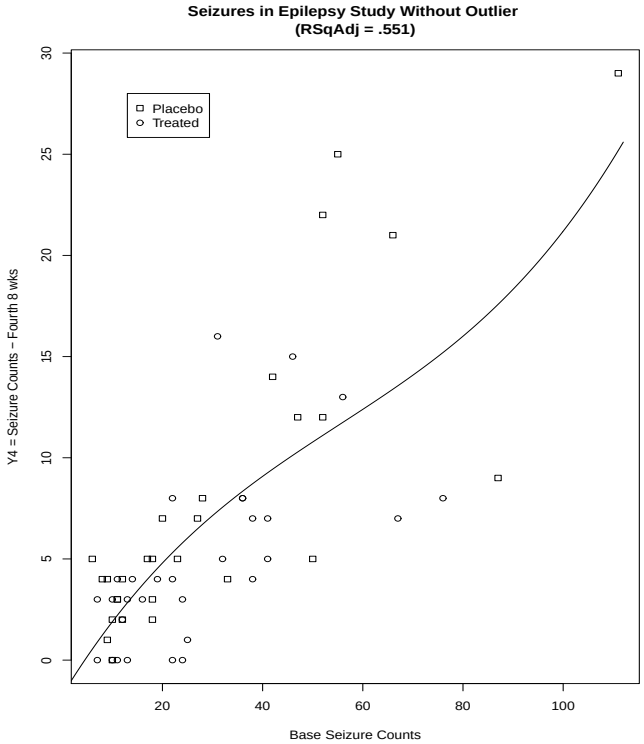

Seizures in Epilepsy Study (RSqAdj = 0)
Placebo
Treated
Y1 = Seizure Counts – First 8 weeks Period
Age of Patient
0
20
40
60
80
100
20
25
30
35
40

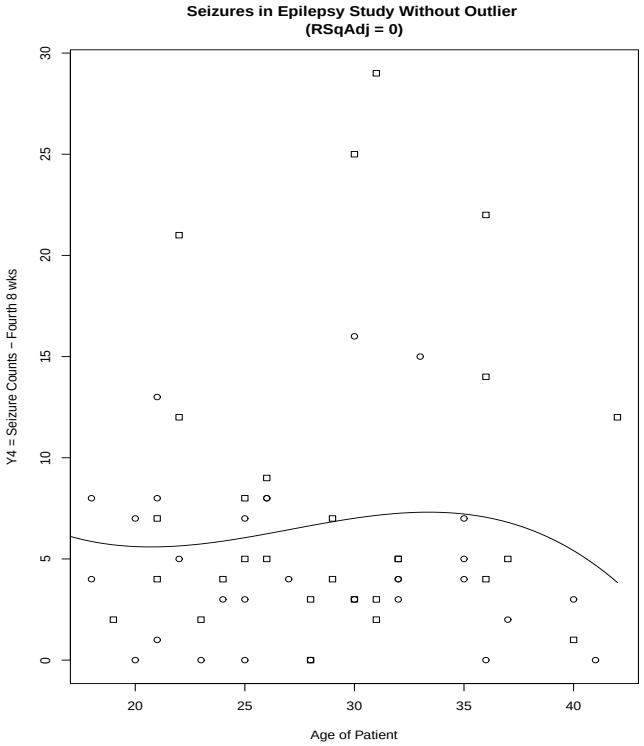

Seizures in Epilepsy Study Without Outlier
(RSqAdj = .449)
Placebo
Treated
Y1 = Seizure Counts – First 8 weeks Period
Base Seizure Counts
0
10
20
30
40
20
40
60
80
100


Seizures in Epilepsy Study Without Outlier
(RSqAdj = .026)
Placebo
Treated
Seizure Counts – First 8 wks
Age of Patient
0
10
20
30
40
20
25
30
35
40


Seizures in Epilepsy Study (RSqAdj = .812)
Placebo
Treated
Y4 = Seizure Counts – Fourth 8 weeks Period
Base Seizure Counts
0
10
20
30
40
50
60
50
100
150


Seizures in Epilepsy Study (RSqAdj = 0)
Y4 = Seizure Counts – Fourth 8 weeks Period
Age of Patient
0
10
20
30
40
50
60
20
25
30
35
40


Seizures in Epilepsy Study Without Outlier
(RSqAdj = .551)
Placebo
Treated
Y4 = Seizure Counts – Fourth 8 wks
Base Seizure Counts
0
5
10
15
20
25
30
20
40
60
80
100


Seizures in Epilepsy Study Without Outlier
(RSqAdj = 0)
Y4 = Seizure Counts – Fourth 8 wks
Age of Patient
0
5
10
15
20
25
30
20
25
30
35
40