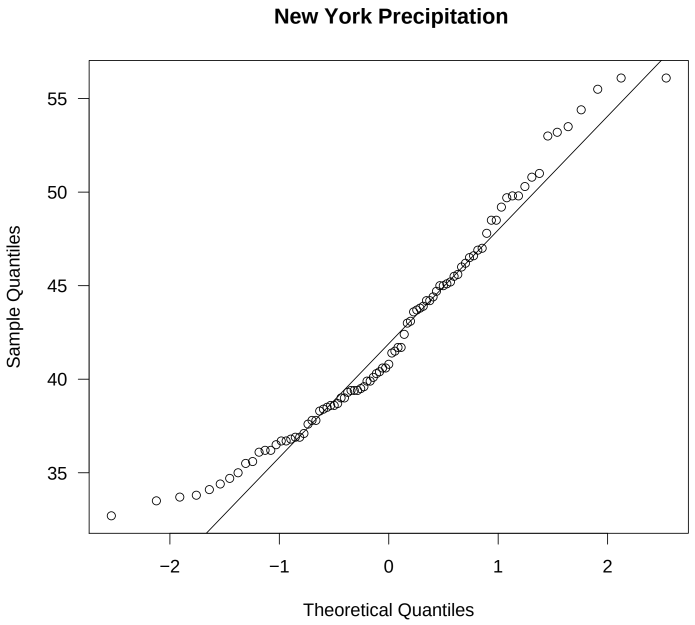
Quantile Based Plots 2

Theoretical Quantile-Quantile Plots

- A theoretical QQ plot examines whether or not a sample $X_1, \ldots, X_n$ has come from a distribution with a given distribution function $F(x)$.

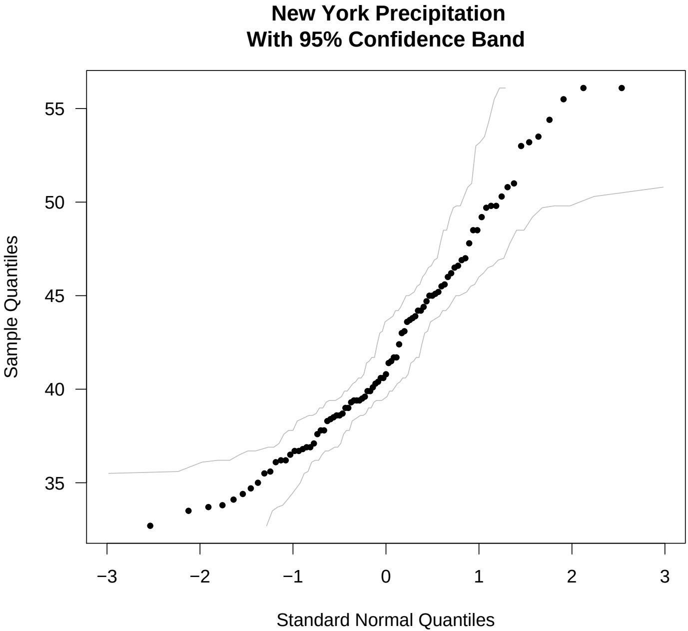
- The plot displays the sample quantiles $X_{(1)}, \ldots, X_{(n)}$ against the distribution quantiles $F^{-1}(p_1), \ldots, F^{-1}(p_n)$, where
$$p_i = \frac{i - 1/2}{n}.$$

- The next slide shows a normal q-q plot for the New York rainfall data.

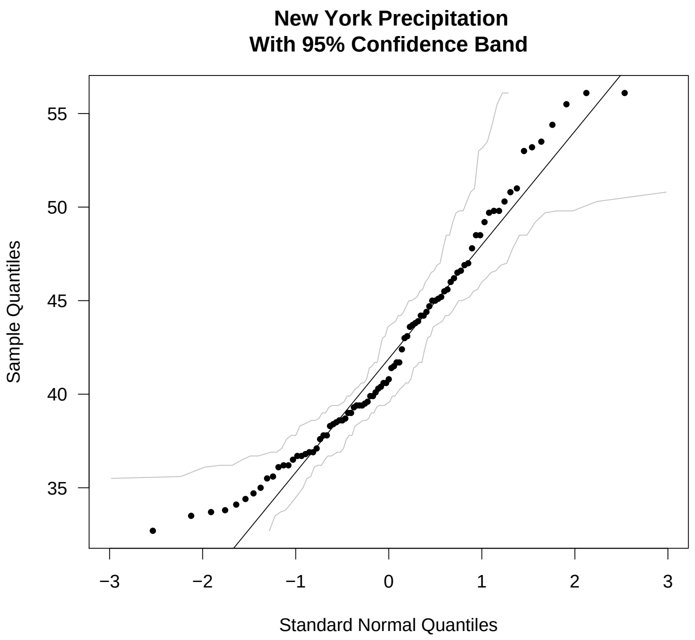

New York Precipitation
Sample Quantiles
Theoretical Quantiles
55
50
45
40
35
−2
−1
0
1
2

Notes

- The plot seems to show deviation from normality.
- The pattern of points seems to indicate that the distribution is skewed or perhaps light-tailed.
- However, we need be careful because sampling variability might cause the plot from linearity.
- We need to have some way of indicating the likely range of sampling variability in the plot.



New York Precipitation
With 95% Confidence Band
Sample Quantiles
55
50
45
40
35
−3
−2
−1
0
1
2
3
Standard Normal Quantiles


New York Precipitation
With 95% Confidence Band
Sample Quantiles
55
50
45
40
35
−3
−2
−1
0
1
2
3
Standard Normal Quantiles

Notes

- It is possible to draw a straight-line across the plot lying entirely within the confidence band.
- This means that the apparent deviation from normality is consistent with simple sampling variability.
- There is no reason to believe that the data is anything other than normally distrbuted.

Confidence Bands

- It is clearly useful to add confidence bands to qq plots.
- The methodology for doing this has been available for 30 years, but it isn't used very often.
- We will take a while to examine the theory which makes it possible to derive the formulae for the bands.

The Probability Integral Transform

If X is a random variable with distribution function $F(x)$ then $Y = F(X)$ is uniformly distributed on the interval $[0, 1]$.

Because $F(x)$ is a monotone increasing function

$$\begin{aligned} P[F(X) \leq F(x)] &= P[X \leq x] \\ &= F(x). \end{aligned}$$

Writing $Y = F(X)$ and $y = F(x)$, this becomes

$$P[Y \leq y] = y.$$

This means that Y is distributed as $U[0, 1]$.

The Empirical Distribution Function

For a random sample $X_1, \ldots, X_n$ from a distribution F, the empirical distribution function is defined to be

$$F_n(x) = \frac{\text{Number of } X_i \leq x}{n}.$$

This is a step function which has a jump of size $1/n$ at each of the X_i values.

Note that Number of $X_i \leq x$ is binomially distributed with n trials and probability of success $F(x)$.

This means that

$$E[F_n(x)] = F(x).$$

The Uniform Distribution is Enough

Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution with distribution function $F(x)$. Let p be a value in $[0, 1]$ with $x_p = F^{-1}(p)$ the pth quantile of the distribution.

The distribution of $F_n(x_p)$ is

$$\frac{\text{Binomial}(n, p)}{n},$$

independent of the distribution function $F(x)$.

This makes it possible to work out the theory for the Uniform$[0, 1]$ distribution and apply it to any distribution F.

The Kolmogorov-Smirnov Test

The Kolmogorov test uses a sample $X_1, \ldots, X_n$ with distribution function $F(x)$ to test the null hypothesis $H : F = F_0$.

The test is based on the Kolmogorov-Smirnov statistic:

$$S = \sup_x |F_n(x) - F|,$$

with the test being rejected if S is too large.

The distribution of S has been determined and critical points are available in tables.

From Test to Confidence Band

Suppose that k is the critical point for the KS statistic. Then we know that $P[S \leq k] = 1 - \alpha$.

This means that with probability $1 - \alpha$ each of the following is true.

$$|F_n(x) - F(x)| \leq k$$

$$-k \leq F(x) - F_n(x) \leq k$$

$$F_n(x) - k \leq F \leq F_n(x) + k$$

A $1 - \alpha$ simultaneous confidence band for F is thus

$$[F_n(x) - k, F_n(x) + k]$$

We could test for normality by seeing if a normal distribution function lies in this band.

Normal QQ Plot Confidence Bands

A normal distribution function lies within the bands if

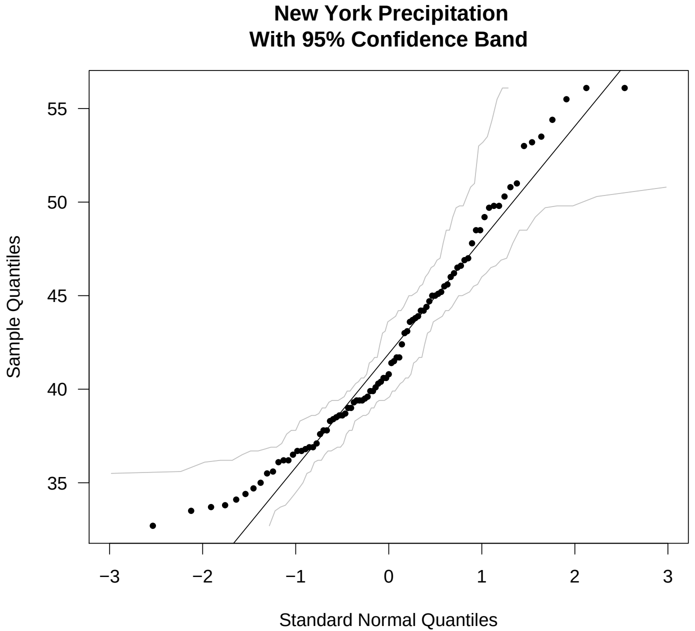
$$F_n(x) - k \leq \Phi(\frac{x-\mu}{\sigma}) \leq F_n(x) + k.$$

This is equivalent to having a straight-line lie between the transformed bounds:

$$\Phi^{-1}(F_n(x) - k) \leq \frac{x-\mu}{\sigma} \leq \Phi^{-1}(F_n(x) + k).$$

This provides the confidence bands for the normal QQ plot.

The normal QQ plot itself is based on the function $\Phi^{-1}(F_n(x))$. Both this function and the confidence bands are evaluated at the sample order statistics, $X_{(1)}, \ldots, X_{(n)}$.


New York Precipitation
With 95% Confidence Band
Sample Quantiles
Standard Normal Quantiles
55
50
45
40
35
−3
−2
−1
0
1
2
3

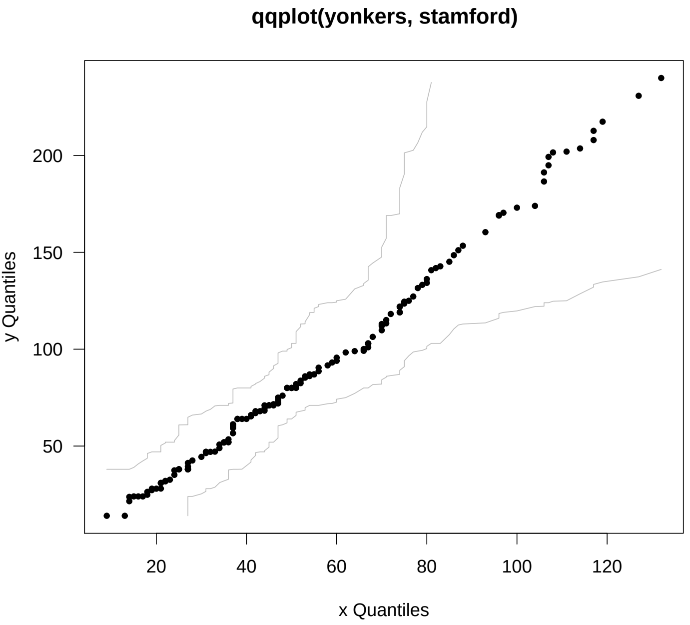
Empirical QQ Plot Confidence Bands

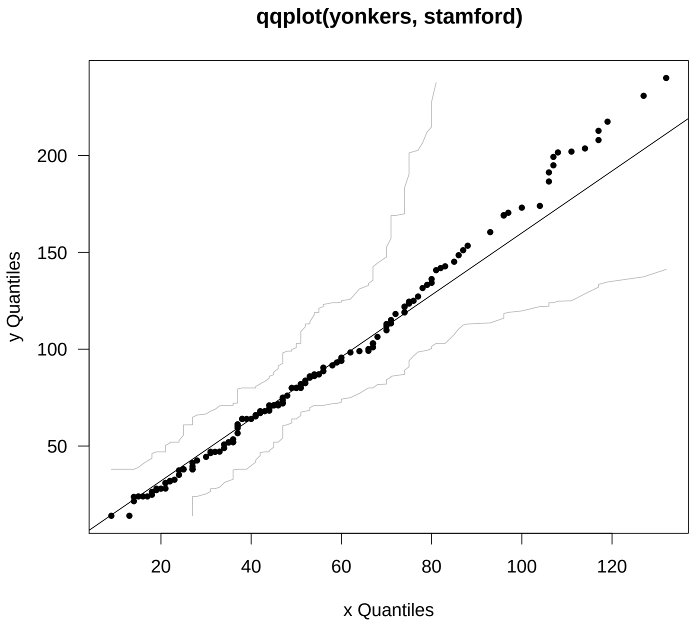
An empirical qq plot is a plot of the quantiles of one sample against the quantiles of another. This can be regarded as a plot of $F_{Yn}^{-1}(p)$ versus $F_{Xn}^{-1}(p)$ for p in the interval $[0, 1]$.

Confidence bands can be based on the two-sample Kolmogorov-Smirnov statistic

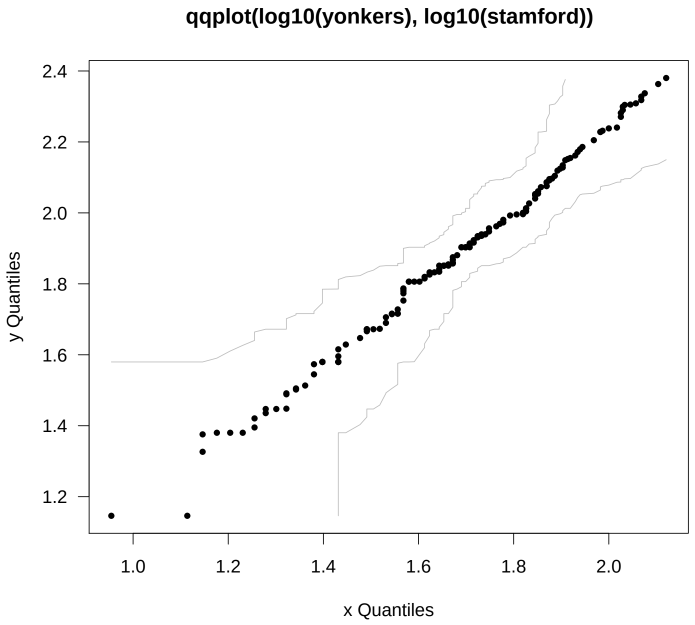
$$S = \sup_x |F_{Yn}(x) - F_{Yn}^{-1}(p)|.$$

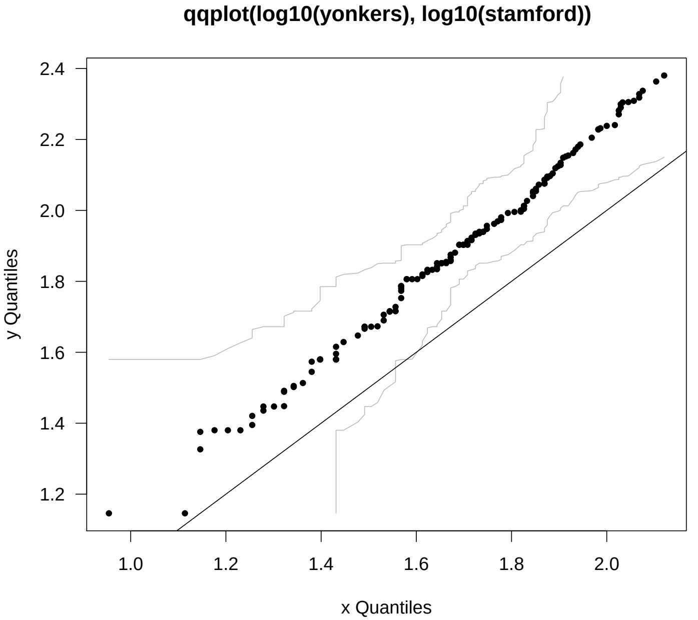
Again, this distribution has been studied and tabulated.


qqplot(yonkers, stamford)
y Quantiles
x Quantiles
200
150
100
50
20
40
60
80
100
120



qqplot(yonkers, stamford)
y Quantiles
x Quantiles
50
100
150
200
20
40
60
80
100
120


qqplot(log10(yonkers), log10(stamford))
2.4
2.2
2.0
1.8
1.6
1.4
1.2
y Quantiles
1.0
1.2
1.4
1.6
1.8
2.0
x Quantiles


qqplot(log10(yonkers), log10(stamford))
y Quantiles
x Quantiles
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
1.2
1.4
1.6
1.8
2.0

The Empirical Shift Function

The standard two sample t-test makes the hypothesis the the x sample and the y sample have the same distribution up to a constant shift. In other words:

$$F_X(x) = F_Y(x + \Delta).$$

The value $\Delta|$ measures how much the x distribution must be shifted to obtain the y distribution.

A more general model is the general *shift-function* model:

$$F_X(x) = F_Y(x + \Delta(x)).$$

Interpreting the Shift Function

Suppose that the x sample corresponds to a control group and the y sample corresponds to a treatment group. The shift function provides the effect of the treatment for different members of the population. Eg.

- Small members of the population are made bigger and large members are left unchanged (a selective effect).
- Small members of the population are made smaller and large members are made larger (an increase in spread).

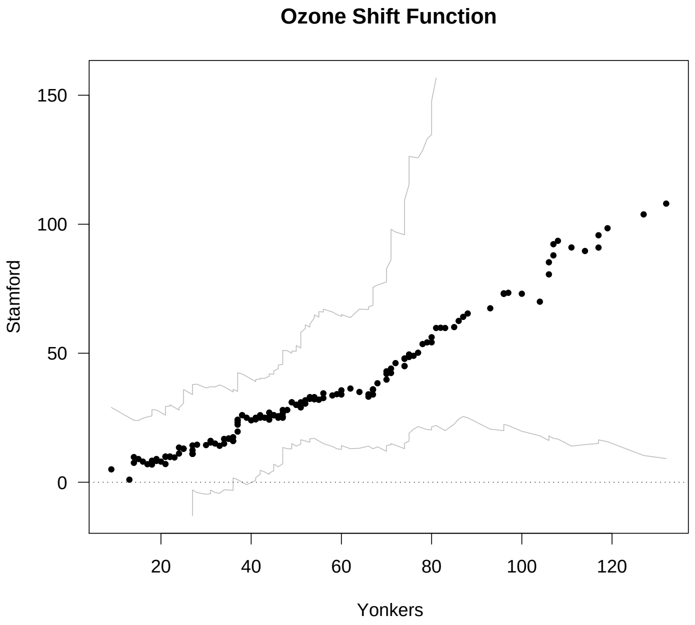
Estimating the Shift Function

The shift function satisfies:

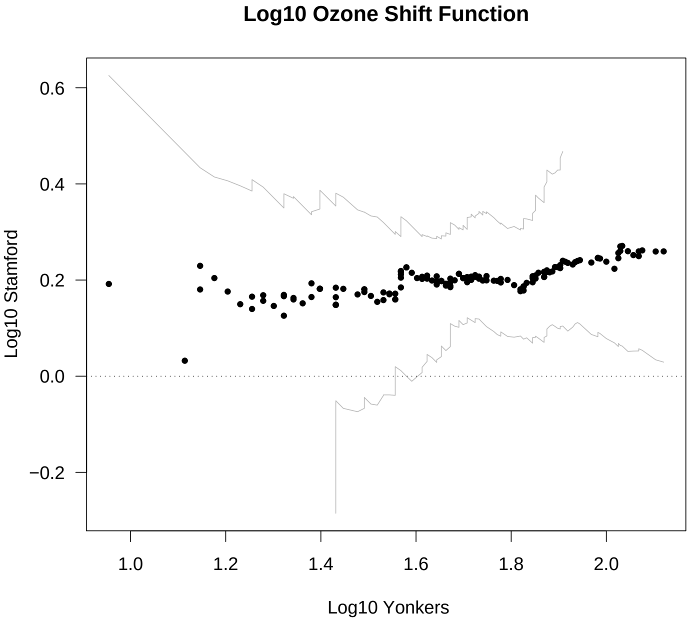
$$\Delta(x) = F_Y^{-1}(F_X(x)) - x$$

It can be estimated by replacing the theoretical distribution functions by the empirical ones based on the observed samples.

The shift function is closely related to the empirical quantile-quantile plot and confidence bands can be obtained in a similar way.



Ozone Shift Function
Stamford
Yonkers
150
100
50
0
20
40
60
80
100
120


Log10 Ozone Shift Function
Log10 Stamford
Log10 Yonkers
0.6
0.4
0.2
0.0
−0.2
1.0
1.2
1.4
1.6
1.8
2.0

An Example – Rats and Ozone

A group of young rats was randomly split into two groups. One group was used as a control and the other raised in an ozone enriched environment

The following weight gains were observed:

Control	41.0	38.4	24.4	25.9	21.9
	18.3	13.1	27.3	28.5	−16.9
	26.0	17.4	21.8	15.4	27.4
	19.2	22.4	17.7	26.0	29.4
	21.4	26.6	22.7		
Ozone	10.1	6.1	20.4	7.3	14.3
	15.5	−9.9	6.8	28.2	17.9
	−9.0	−12.9	14.0	6.6	12.1
	15.7	39.9	−15.9	54.6	−14.7
	44.1	−9.0			

A “Standard” Analysis

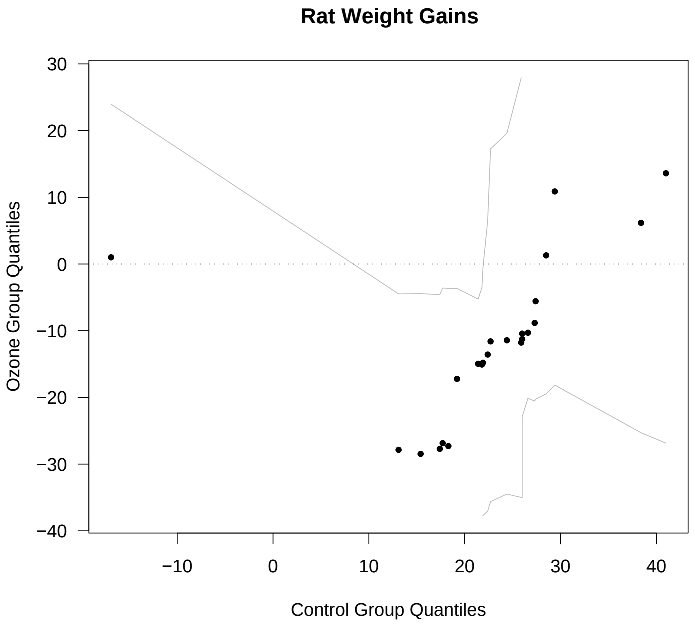
- A standard analysis would use a two-sample t-test to see whether ozone exposure has a significant effect on weight gain.

- The mean weight gains were:

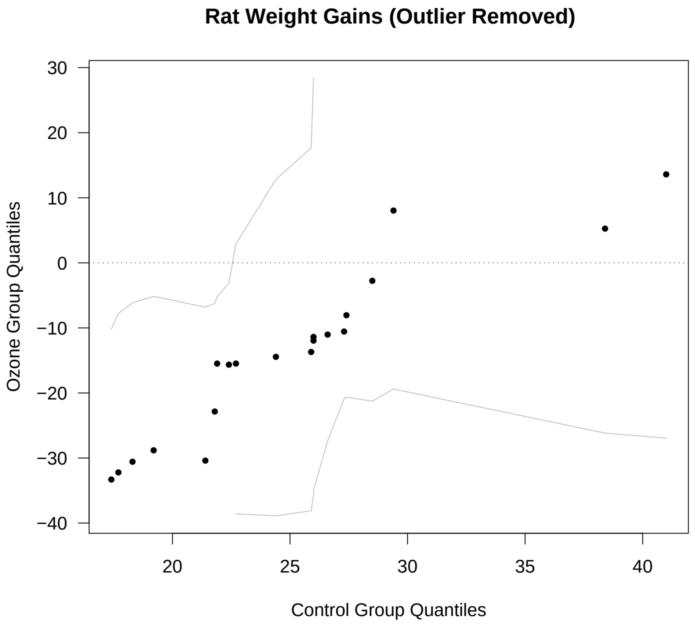
Control	22.4
Ozone	11.0

- The p-value for a two-sided test is 0.02.

- This is weak evidence that ozone exposure decreases the growth rates of juvenile rats.



Rat Weight Gains
Ozone Group Quantiles
Control Group Quantiles
30
20
10
0
−10
−20
−30
−40
−10
0
10
20
30
40


Rat Weight Gains (Outlier Removed)
Ozone Group Quantiles
Control Group Quantiles
30
20
10
0
−10
−20
−30
−40
20
25
30
35
40