HANDOUT #9: GOODNESS OF FIT AND BOX-COX TRANSFORMATIONS

I. GOF for Discrete Distributions

1. Completely Specified Distributions
   (a) Chi-square Measure of Fit
   (b) Binomial Distribution Example
   (c) Poisson Distribution Example
2. Distributions with Unspecified Parameters
   (a) Chi-square Measure of Fit - Reduced DF
   (b) Binomial Distribution Example
   (c) Poisson Distribution Example

II. GOF for Continuous Distributions

1. Completely Specified Distributions
   (a) Kolmogorov-Smirnov (KS) Measure
   (b) Cramer von Mises (CvM) Measure
   (c) Anderson-Darling (AD) Measure
   (d) Normal Distribution Example
   (e) Censored Data
2. Distributions With Unspecified Parameters
   (a) Shapiro-Wilk Measure for Normal Distributions
   (b) Correlation Measure for Normal Distributions
   (c) Modifications to KS, CvM, and AD Measures
   (d) Normal, Exponential, Weibull Distribution Examples
   (e) Censored Data

III. Box-Cox Transformation to Normality

GOODNESS OF FIT MEASURES

In many situations, the observations from a population or the outcomes from a process are described as being realizations of independent random variables from a specified distribution, such as Poisson or normal. It is not expected that the data are exactly generated from the specified distribution but that for practical purposes the specified probability distribution does well in describing the randomness in the observed outcomes. We have discussed how to use reference distribution plots to visually display the degree to which a specified distribution represents the observed data. However, it is often desirable to have a quantitative assessment of the degree to which the specified distribution fits the data. Our discussion will be separated into goodness-of-fit (gof) measures for discrete and for continuous distributions.

Discrete Goodness of Fit Measures

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of iid random variables (random sample) with a discrete distribution having pmf $f(y; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ is a vector of m parameters.

Let $f_o(y; \boldsymbol{\theta})$ be a pmf which the researcher conjectures fits the observed data.

We will develop a measure of the degree to which $f_o(y; \boldsymbol{\theta})$ models the observed data, i.e., the fit of the model to the data.

Example A rare but fatal disease of genetic origin occurring chiefly in infants and children is under investigation. If a couple are both carriers of the disease, a child of theirs has a probability .25 of being born with the disease. For 100 such couples having five children, a researcher recorded the number, Y, of children having the disease.

Diseased Children in Family	0	1	2	3	4	5	Total
Frequency	21	42	24	8	4	1	100

A proposed model for the distribution of Y would be a binomial model with $n = 5$ and $\theta = .25$.

That is, Y has a $B(5, .25)$ distribution with pmf having

$$\theta = .25, \quad f_o(y, .25) = \binom{5}{y}(.25)^y(.75)^{5-y} \quad \text{for} \quad y = 0, 1, 2, 3, 4, 5.$$

Does it appear that the binomial model provides a reasonable fit to this data set? If not, why?

GOF Measure for Discrete Distributions - Completely Specified Model

A general method for evaluating the fit of a discrete model will now developed.

Let $Y_1,\ Y_2,\ \ldots,\ Y_n$ be n iid observations from a discrete distribution represented by the r.v. Y.

Let $y_1 < y_2 < \cdots$ be the possible values of the discrete r.v. Y having pmf $f(y;\boldsymbol{\theta})$.

Let $f_o(y;\boldsymbol{\theta})$ be the proposed model for Y with $\boldsymbol{\theta}$, a vector of **known** parameters and

$$p_i = f_o(y_i;\boldsymbol{\theta}) = P[Y = y_i] \ \text{ for } \ i = 1, 2, \ldots, k-1 \text{ and } p_k = P[Y \geq y_k] = 1 - \sum_{i=1}^{k-1} p_i,$$

where k is selected based on the data.

A measure of how well $p_1, p_2, \ldots, p_k$ match the observed data is obtained by comparing the number of observations we would expect to observe having $Y = y_i$ if

$f_o(y;\boldsymbol{\theta})$ was the correct model for $f(y;\boldsymbol{\theta})$,

to the number of $Y_j = y_i$ in the data.

Let $O_i = \sum_{j=1}^{n} I(Y_j = y_i)$, for $i = 1,\ 2,\ \ldots,\ k$, that is,

O_i is the number of observations from the data, $Y_1,\ Y_2,\ \ldots,\ Y_n$, equal to y_i .

Under the model $f_o(y;\boldsymbol{\theta})$, the expected number of observations equal to y_i is given by $E_i = E[O_i] = np_i$.

This assumes that $f_o(y;\boldsymbol{\theta})$ is the correct model when calculating p_i's .

If the model is correct, then there should be a good match between the k pairs of values E_i and O_i.

A measure of the "fit of the model" is to measure how close are the $E_i s$ to the $O_i s$. An index of this fit is given by the **Chi-square Goodness-of-Fit Statistic:**

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The division by E_i is done to modulate the sample size. Otherwise, data sets which had large values for n would tend to have large values of Q even when there was a reasonable good match between O_i and E_i.

An assessment of the relative size of Q is obtained by noting that if $f_o(y;\boldsymbol{\theta})$ is the correct model then Q has approximately (large n) Chi-square distribution with $df = k - 1$. Therefore, we compute the probability of observing a value from the Chi-square distribution larger than the computed Q:

p-value $= P[\chi^2_{k-1} \geq Q] = 1 - F(Q) = 1 - pchisq(Q, k-1)$, where

F is the cdf of a Chi-square distribution with $df = k - 1$ and **pchisq(Q, k-1)** is an R-Function:

If p-value is large, ie, Q is relatively small then we conclude that $E_i s$ match $O_i s$ and hence that $f_o(y;\boldsymbol{\theta})$ is a reasonable model for $f(y;\boldsymbol{\theta})$.

Using the Chi-square distribution as an approximation to the sampling distribution of Q, requires a large value for the sample size n. The approximation by the chi-squared distribution is not very accurate if expected frequencies, $E_i = np_i$ are too low, that is n is too small. The appropriate size of n is assessed as follows:

- All the E_i must be larger than 1.0
- At most 20% of the E_i may be less than 5.0

The above two conditions are most often violated when the sample size n is too small. In fact, these conditions are used in the design stage of a study to determine the necessary sample size required to have a valid study.

When there is only 1 degree of freedom, the approximation is not reliable if expected frequencies are below 10. In this case, a better approximation, Yates's correction for continuity, should be used. The correction is to reduce the absolute value of each difference between observed and expected frequencies by 0.5 before squaring.

$$Q = \sum_{i=1}^{k} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

A method of overcoming the problem after data has been collected is to combine cells having low counts. This is one of the methods by which k is selected.

Example(cont.) From the $B(5, .25)$ distribution we compute the p_i and $E_i = np_i = 100p_i$ as given in the following table:

Diseased Children in Family	0	1	2	3	4	5	Total
p_i	.2373	.3955	.2637	.08789	.01465	.000977	1.00
E_i	23.73	39.55	26.37	8.789	1.465	.0977	100
O_i	21	42	24	8	4	1	100
$\frac{(O_i-E_i)^2}{E_i}$	.31	.15	.21	.07	4.39	8.33	13.46

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 13.46 \Rightarrow P[\chi^2_5 \geq 13.46] = 1 - F(13.46) = 1 - pchisq(13.46, 5) = .019,$$

where F is the cdf of a chi-square distribution with $df = 5$.

Use Chi-square table in textbook or R-function: **1-pchisq(13.46,5)**.

Is this a valid analysis? Have all the conditions been met to validly use the chi-square table in computing the p-value?

Guidelines for Assessing Fit

A general set of guidelines for using p-values to evaluate fit of model: (These rules are somewhat arbitrary.)

- $p - value > .25 \Rightarrow$ Excellent fit
- $.15 \leq p - value < .25 \Rightarrow$ Good fit
- $.05 \leq p - value < .15 \Rightarrow$ Moderately Good fit
- $.01 \leq p - value < .05 \Rightarrow$ Poor fit
- $p - value < .01 \Rightarrow$ Unacceptable fit

Based on the above criteria, we have a "Poor fit" of the Binomial model to the data. However, it appears the data fit rather well except for the last two cells. In fact, based on our rules for using the chi-square approximation, we should combine the last two cells and recompute the value of Q:

Diseased Children in Family	0	1	2	3	4 or 5	Total
p_i	.2373	.3955	.2637	.08789	.015625	1.00
E_i	23.73	39.55	26.37	8.789	1.5625	100
O_i	21	42	24	8	5	100
$\frac{(O_i - E_i)^2}{E_i}$	.31	.15	.21	.07	7.56	8.30

$$Q_{\text{new}} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 8.30 \Rightarrow P[\chi^2_4 \geq 8.30] = 1 - F(8.30) = 1 - pchisq(8.30, 4) = 0.08,$$

where F is the cdf of a chi-square distribution with $df = 4$.

We would thus conclude that the binomial model provides an "Moderately Good Fit" to the data.

One reason that the binomial model may not provide a better fit to the data is that the 5 Bernoulli trials in this situation may not be independent.

Note the relationship between the computations involved in obtaining Q:

$$Q_{\text{new}} = Q_{\text{old}} - E_{4,old} - E_{5,old} + E_{4,new} = 13.46 - 4.39 - 8.33 + \frac{(5 - 1.5625)^2}{1.5625} = 8.30$$

GOF Measure for Discrete Distributions
Incompletely Specified Model (Unknown Parameters)

In many situations, the distribution may not be completely specified in that not all the parameters in the proposed model have known values. That is, some or all the values in $\boldsymbol{\theta}$ may not be specified.

Fitting Binomial Model

Example: A certain brand of flashlight is sold with the four batteries included. A random sample of 150 flashlights is obtained and the number of defective batteries in each flashlight is determined, resulting in the following data:

Number of Defectives	0	1	2	3	4	Total
Frequency	26	51	47	16	10	150

Let D be the number of defective batteries in a randomly selected flashlight.

A reasonable model for the distribution of D would be a binomial model with $n = 4$ and θ, the probability that a randomly selected battery is defective, which is unspecified by the battery manufacturer. Thus, a possible model for D would be a $B(4, \theta)$ distribution with pmf:

$f_o(d, \theta) = \binom{4}{d}(\theta)^d(1-\theta)^{4-d}$ for $d = 0, 1, 2, 3, 4$.

Before we can answer the question, "Does it appear that the binomial model provides a reasonable fit to this data set?", we must first estimate the value of θ.

Let $\hat{\boldsymbol{\theta}}$ be an efficient estimator of the unknown parameters, such as MLE. Replace $\boldsymbol{\theta}$ in the formula for $f_o(d, \boldsymbol{\theta})$ with $\hat{\boldsymbol{\theta}}$ and compute Q as in the situation where $\boldsymbol{\theta}$ is known.

The distribution of Q is altered in that the degrees of freedom for the approximating chi-square distribution are now bounded by $k - 1 - w \leq df \leq k - 1$, where w is the number of parameters that must be estimated from the data. Therefore, we have that the p-value for Q is bounded by

$$P[\chi^2_{k-1-w} \geq Q] \leq \text{p-value} \leq P[\chi^2_{k-1} \geq Q]$$

Using $df = k - 1 - w$ will lead to a test which is more likely to declare that the proposed model for the data is incorrect and hence requires a better fit of the model to the data than does $df = k - 1$. For this reason,the use of $df = k - 1 - w$ in the calculation of p-values is preferred.

Example (continued)

We will illustrate these concepts by evaluating the fit of a binomial model to the flashlight data.

Let θ be the proportion of defective batteries.

The MLE of θ is

$$\begin{aligned}\hat{\theta} &= \frac{\text{Number of Defective Batteries}}{\text{Number of Batteries}} \\ &= \frac{(0)(26)+(1)(51)+(2)(47)+(3)(16)+(4)(10)}{(4)(150)} = 0.3883\end{aligned}$$

Let p_{i+1} be the probability that a randomly selected flashlight has i detective batteries for $i = 0, 1, 2, 3, 4$, that is, $p_{i+1} = P[D = i]$, where D has a Binomial(4,θ) distribution.

$$\widehat{p}_{i+1} = P[D = i] = f_o(i, \hat{\theta}) = \binom{4}{i}(.3883)^i(1 - .3883)^{4-i} \quad \text{for} \quad i = 0, 1, 2, 3, 4$$

$$\hat{E}_i = 150\hat{p}_i$$

,

$$\hat{Q} = \sum_{i=1}^{5} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

.

We will summarize these calculations in the following table:

i	1	2	3	4	5	Total
$\hat{p}_i$	.1400	.3555	.3385	.14339	.0227	1.00
$\hat{E}_i$	21.00	53.33	50.78	21.50	3.39	150
O_i	26	51	47	16	10	150
$\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$	1.19	0.10	0.28	1.41	12.89	15.87

Note: that $1 < E_5 = 3.39 < 5$ but $E_i > 5$ for $i = 1, 2, 3, 4$.

$$\hat{Q} = \sum_{i=1}^{5} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = 15.87, \quad df = 5 - 1 - 1 = 3 \Rightarrow$$

$$\text{p-value} = P[Q \geq 15.87] = 1 - F(15.87) = 1 - pchisq(15.87, 3) = 0.0012,$$

where F is the cdf of a chi-square distribution with $df = 3$.

With the p-value such a small number, we would thus conclude that the binomial model provides an "Unacceptable Fit" to the data.

One reason that the binomial model may not fit this situation is that the 4 batteries in a given flashlight are more likely to have a similar defective rate than 4 batteries in a different flashlight. That is, the Bernoulli trials in this situation may not be identically distributed.

Example of Fitting Poisson Model

Suppose $X_1, X_2, \ldots, X_n$ are iid random variables from a discrete distribution.

A Poisson model is proposed for the distribution.

That is, X_j has pmf

$$P[X = x] = f_o(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad x = 0, 1, 2, \ldots,$$

with λ unspecified.

The MLE of λ is $\hat{\lambda} = \bar{X}$.

Thus, the estimated pmf is given by

$$\hat{p}_{i+1} = f_o(i; \hat{\lambda}) = \frac{e^{-\bar{X}}\bar{X}^i}{i!}, \quad \text{for} \quad i = 0, 1, 2, \ldots.$$

Based on the data, we select k and then count the number of X_j equal to $0, 1, 2, \ldots, k-2$ and the number of $X_j \geq k-1$ that is,

$O_{i+1} =$ Number of $X_j = i$ for $i = 0, 1, \ldots, k-2$ and

$O_k =$ Number of $X_j \geq k-1$.

Next, compute $\hat{E}_i = n\hat{p}_i$ for $i = 1, 2, \ldots, k$ and

$\hat{Q} = \sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$.

We will illustrate with the following example:

Example

In a genetic experiment, investigators looked at 300 chromosomes of a particular type and counted the number of sister-chromatid exchanges on each chromosome. The data is from " On the Nature of Sister-Chromatied Exchanges in 5-Bromodeoxyuridine-Substituted Chromosomes", *Genetics*, 1979, pp. 1251-1264. The data is given next.

Number of Exchanges	0	1	2	3	4	5	6	7	8	9	Total
Observed Counts	7	24	42	59	62	44	41	14	5	2	300

A Poisson model was hypothesized for the distribution of the number of exchanges. An estimate of λ is

$$\begin{aligned}\widehat{\lambda} &= \bar{X} \\ &= \frac{1}{300}[(0)(7)+(1)(24)+(2)(42)+(3)(59)+(4)(62)+(5)(44)+(6)(41)+(7)(14)+(8)(5)+(9)(2)] \\ &= \frac{1155}{300} = 3.85\end{aligned}$$

The estimated pmf is given by

$\hat{p}_{i+1} = f_o(i;\hat{\lambda}) = \frac{e^{-3.85}(3.85)^i}{i!}$ = dpois(i,3.85), for $i = 0, 1, 2, \ldots, 7$ and

$\hat{p}_9 = 1 - \sum_{i=1}^{8} \hat{p}_i = 1 - P[X \leq 7] = 1 - ppois(7, 3.85)$.

$\hat{E}_i = 300\hat{p}_i$ for $i = 1, 2, \ldots, 9$

(A value of k=9 was selected in order to satisfy the requirement that all $\hat{E}_i s$ be greater than 1.)

The calculations for finding $\hat{Q}$ are given here:

Use the R-function: **dpois(i,3.85)** to compute the values of $\hat{p}_i$:

Number of Exchanges	0	1	2	3	4	5	6	7	≥ 8	Total
$\hat{p}_i$	.021	.082	.158	.202	.195	.150	.096	.053	.043	1.00
$\hat{E}_i$	6.4	24.6	47.3	60.7	58.4	45.0	28.9	15.9	12.8	300
O_i	7	24	42	59	62	44	41	14	7	300
$\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$	.06	.01	.60	.05	.22	.02	5.09	.22	2.64	8.91

$$\hat{Q} = \sum_{i=1}^{9} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = 8.91, \quad df = 9 - 1 - 1 = 7 \Rightarrow$$

$$P[\chi^2_7 \geq 8.91] = 1 - F(8.91) = 1 - pchisq(8.91, 7) = 0.259,$$

where F is the cdf of a Chi-square distribution with $df = 7$.

With a p-value of 0.259, we would thus conclude that the Poisson model provides an "Excellent Fit" to the data.

Goodness of Fit Measures for Continuous CDFs

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of iid random variables (random sample) with distribution having a continuous cdf $F(y;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ is a vector of m parameters.

Let $F_o(y;\boldsymbol{\theta})$ be a pdf which the researcher conjectures fits the observed data.

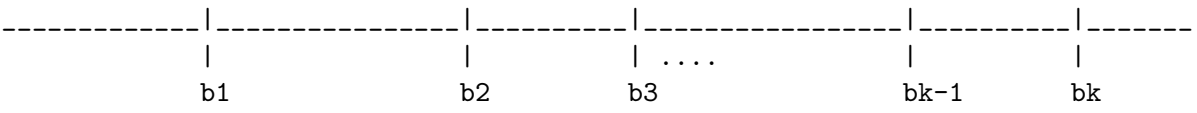
We will develop a measure of the degree to which $f_o(y;\boldsymbol{\theta})$ models the observed data.

The book, *Goodness-of-Fit Techniques*, by Ralph D'Agostino and Michael Stephens is the main reference for the following discussion.

In many books the Chi-square GOF statistic, Q, is used for continuous distributions as well as for discrete distributions. The interval of possible values for the distribution of Y is divided into k subintervals, bins, and thus the distribution of Y is discretized. The remainder of the chi-square gof calculations are then completed in the same manner as for a discrete distribution. The weakness of this procedure is that the selection of k and the assignment of the k bins to the interval of values is rather arbitrary. The conversion of a continuous model into a discrete model often results in a procedure which is not sensitive in detecting deviations of the observed data from the proposed model, especially in the tails of the distribution. This is very crucial in many statistical procedures, because it is the tail fit of a model that is most crucial. For example, power calculations in tests of hypotheses, and percentile determinations for confidence intervals and hypotheses testing rely on the ability to make accurate probability calculations in the tails of the selected model pdf.

These limitations in using the Chi-square gof statistic are very similar to the weaknesses of using a relative frequency histogram as an estimator of a pdf: the selection of the number of bins and their location may greatly affect the accuracy of the estimator.

For these reasons, it is not recommended to use the chi-square gof statistic when the data is from a continuous cdf.



$P_1 = P[Y \leq b_1]$

$P_2 = P[b_1 < Y \leq b_2]$

$P_3 = P[b_2 < Y \leq b_3]$

$\vdots$

$P_k = P[b_{k-1} < Y \leq b_k]$

$P_{k+1} = P[b_k < Y]$,

With O_i equal to the number of Y_i in each interval, and $E_i = nP_i$, compute the Chi-square statistic.

GOF Measure-Completely Specified Model

In this situation, the cdf F will be completely specified, that is, $F_o(y) = F_o(y; \boldsymbol{\theta})$ will have all values of its parameters given.

For example, $N(5, (4.3)^2)$, Exponential with $\beta = 4.2$, Weibull with $\gamma = 2.3$ and $\alpha = 3.9$.

Just stating Normal model or Weibull model would not provide a complete specification of the model because there are a number of unknown parameters.

The measures of gof will be based on the empirical cdf:

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) \quad = \text{proportion of } \ Y_i s \leq y$$

Kolmogorov-Smirnov (K-S) Measure

The Kolmogorov-Smirnov measure computes the maximum discrepancy between the proposed model cdf $F_o(y)$ and the sample cdf $F_n(y)$ over all values of y:

$$D_n = \sup_y [|F_n(y) - F_o(y)|] = max[D_n^-, D_n^+] \quad \text{where}$$

$$D_n^- = \sup_y [F_o(y) - F_n(y)] = \max_{1 \leq i \leq n} \left[F_o(Y_{(i)}) - \frac{i-1}{n} \right] = \quad \text{max discrepancy when } \ F_o(y) > F_n(y)$$

$$D_n^+ = \sup_y [F_n(y) - F_o(y)] = \max_{1 \leq i \leq n} \left[\frac{i}{n} - F_o(Y_{(i)}) \right] = \quad \text{max discrepancy when } \ F_o(y) < F_n(y)$$

where $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are the order statistics associated with $Y_1, Y_2, \ldots, Y_n$.

Thus, D_n measures the maximum difference between the proposed model for the population (process) and the cdf estimated from the observed data.

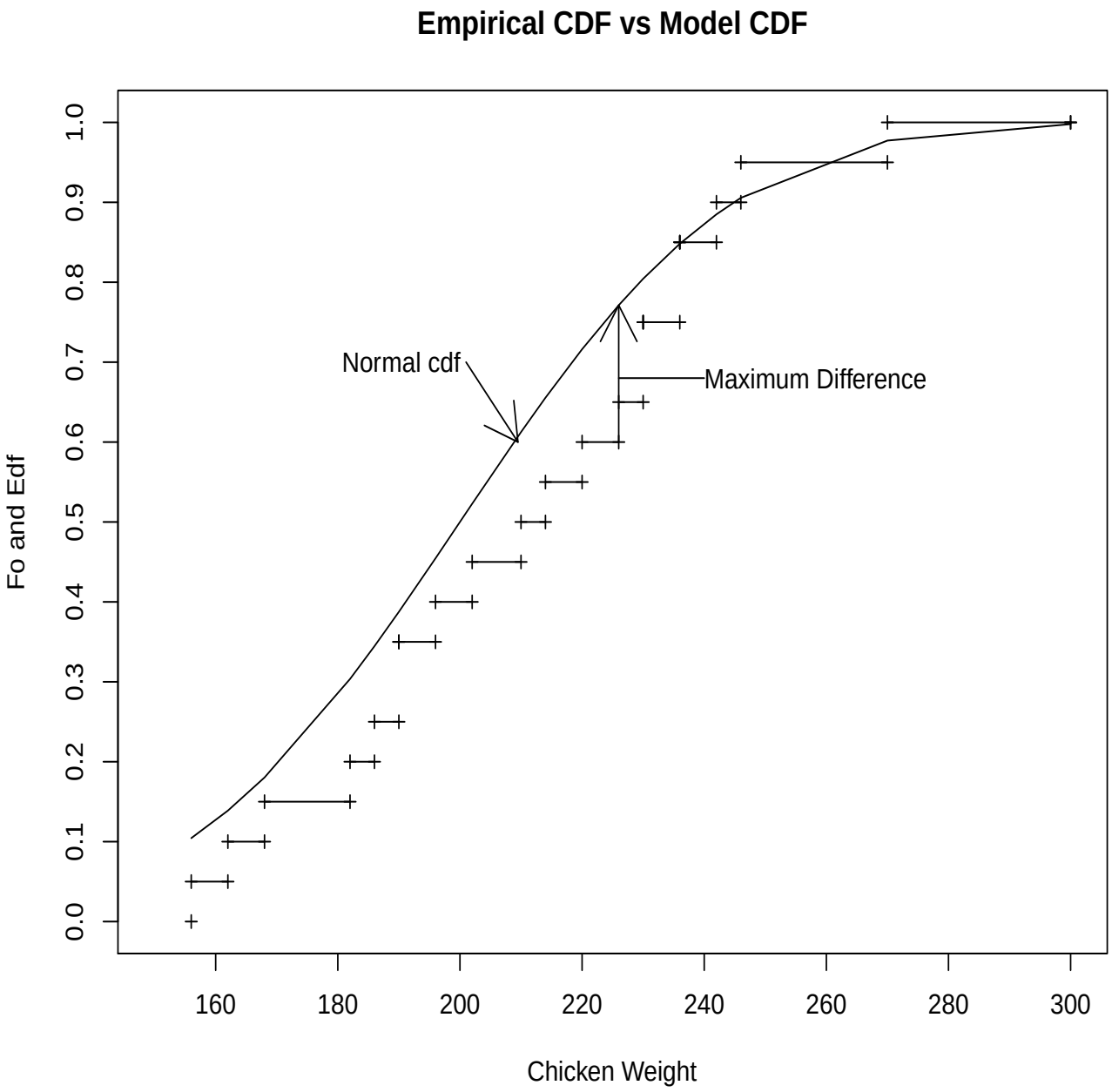
The following example will illustrate the above:

Example of Evaluating Fit of Data to Normal Distribution

Let W be the weights of 21-day-old Leghorn Chickens and let F be the cdf of W. We want to evaluate whether the cdf of W is a normal distribution with $\mu = 200$ and $\sigma = 35$, i.e., W is distributed $N(200, (35)^2)$. A random sample of $n = 20$ twenty-one-day-old Leghorn Chickens yields the following weights:

156	162	168	182	186	190	190	196	202	210
214	220	226	230	230	236	236	242	246	270

The following is a plot of the empirical cdf and the cdf for a normal distribution with $\mu = 200, \sigma = 35$:



To evaluate the relative size of D_n in order to determine the degree of fit of the model to the data, we need to determine the distribution of D_n under the assumption that F_o is the correct model for the population (process).

This would seem to require the determination of many such distributions depending on which model is proposed for the data.

That is, the distribution of D_n must depend on $F_o(y)$.

This would greatly limit the applicability of using D_n as a measure of gof.

However, the *probability integral transform theorem* states:

If the cdf of Y is a continuous cdf $F(\cdot)$, then the distribution of $X = F(Y)$ is Uniform on (0,1).

That is, the transformation, $X = F(Y)$, always results in X having a uniform distribution (0,1), provided F is the cdf of Y.

Therefore, the calculation of p-values for D_n as a gof measure can be computed using the distribution of order statistics from a uniform on (0,1) distribution:

$$D_n^- = \max_{1\le i\le n}\left[F_o(Y_{(i)}) - \frac{i-1}{n}\right] = \max_{1\le\ i\le n}\left[U_{(i)} - \frac{i-1}{n}\right]$$

$$D_n^+ = \max_{1\le i\le n}\left[\frac{i}{n} - F_o(Y_{(i)})\right] = \max_{1\le\ i\le n}\left[\frac{i}{n} - U_{(i)}\right],$$

where $U_{(1)} \le \cdots \le U_{(n)}$ are the order statistics from $U_1, \cdots, U_n$, iid uniform on (0,1) r.v.s.

The distribution of Kolmogorov-Smirnov gof measure, when $F = F_o$, does not depend on F_o and hence is called a **distribution-free** measure of goodness-of-fit.

One problem with the Kolmogorov-Smirnov gof measure is that the statistic D_n is not particularly efficient in detecting discrepancies between the true cdf F and the proposed model F_o in the tails of the distributions. This occurs because D_n only measures the maximum discrepancies between the two cdfs.

An alternative to the Kolmogorov-Smirnov gof measure for continuous cdfs is the Cramer-von Mises measure. It attempts to examine the overall discrepancies between the two cdfs.

Cramer-von Mises (CvM) Measure

Let $F_n(y)$ be the sample estimator of F, the true cdf and F_o be the proposed model for F. The Cramer-von Mises family of gof measures are given by

$$Q_n = n \int_{-\infty}^{\infty} \left[F_n(y) - F_o(y)\right]^2 \Psi(y) dF_o(y),$$

where $\Psi(\cdot)$ is a suitably selected weight function. The proper selection of $\Psi(\cdot)$ enables the Cramer-von Mises statistic to detect departures between F from F_o in the tails of the two distributions. The original Cramer-von Mises statistic had $\Psi(y) \equiv 1$ and was denoted as W_n^2:

$$\begin{aligned}
W_n^2 &= n \int_{-\infty}^{\infty} \left[F_n(y) - F_o(y)\right]^2 dF_o(y) \quad \text{let } u = F_o(y) \;\Rightarrow\; y = F_o^{-1}(u) \\
&= n \int_0^1 \left[F_n(F_o^{-1}(u)) - F_o(F_o^{-1}(u))\right]^2 du = n \int_0^1 \left[F_n(F_o^{-1}(u)) - u)\right]^2 du \\
&= n \sum_{i=1}^{n+1} \int_{U_{(i-1)}}^{U_{(i)}} \left[F_n(F_o^{-1}(u)) - u\right]^2 du \quad \text{with } U_{(i)} = F_o(Y_{(i)}), U_{(0)} = 0,\; U_{(n+1)} = 1 \\
&= n \sum_{i=1}^{n+1} \int_{U_{(i-1)}}^{U_{(i)}} \left[\frac{i-1}{n} - u\right]^2 du \quad = \frac{n}{3} \sum_{i=1}^{n+1} \left[\left(\frac{i-1}{n} - U_{(i-1)}\right)^3 - \left(\frac{i-1}{n} - U_{(i)}\right)^3\right] \\
&= \frac{n}{3} \sum_{i=1}^{n} \left[\left(\frac{i}{n} - U_{(i)}\right)^3 - \left(\frac{i-1}{n} - U_{(i)}\right)^3\right] \\
&= \sum_{i=1}^{n} \left[U_{(i)} - \frac{2i-1}{2n}\right]^2 + \frac{1}{12n} \\
&= \sum_{i=1}^{n} \left[U_{(i)} - \frac{i-1/2}{n}\right]^2 + \frac{1}{12n}
\end{aligned}$$

where $U_{(i)} = F_o(Y_{(i)})$ and thus $U_{(1)},\ U_{(2)},\ \ldots, U_{(n)}$ are the order statistics from a random sample of size n from a uniform on (0,1) distribution.

Therefore, W_n^2's distribution does not depend on the population distribution provided F_o is the correct distribution for the population.

W_n^2 is similar to the K-S statistic in that it is not very sensitive to departures in the tails of the distribution.

In order to rectify the CvM Measures inability to detect differences in the tails of the distributions, a third goodness of fit statistics is the Anderson-Darling (AD) Measure.

Anderson-Darling (AD) Measure

To rectify the CvM Measures inability to detect differences in the tails of the distributions the Anderson-Darling statistic, AD, uses the weight function

$\Psi(y) = [F_o(y)(1 - F_o(y))]^{-1}$.

The reason for selecting this function is that $E[F_n(y) - F(y)] = 0$ for all values of y but

$$Var[F_n(y) - F(y)] = \frac{1}{n}F(y)[1 - F(y)] \rightarrow \begin{cases} 0 & : \quad as \quad y \rightarrow -\infty \\ & : \\ \frac{1}{4n} & : \quad F(y) = \frac{1}{2} \\ 0 & : \quad as \quad y \rightarrow \infty \end{cases}$$

Thus, the variance of the difference, $[F_n(y) - F(y)]$ varies greatly for small n depending on how far y is from the center of the distribution.

To overcome this unequal variability in the statistic, the Anderson-Darling statistic uses the weights:

$\Psi(y) = [F_o(y)(1 - F_o(y))]^{-1}$,

which places greater weight on the differences $[F_n(y) - F(y)]$ for values of y where the variance is small and smaller weight on the differences for values of y where the variance is large.

The Anderson-Darling thus has greater sensitivity to detect departures between F and F_o in the tails of the distributions than do either the K-S statistic or Cramer-von Mises statistic.

$$\begin{aligned} A_n^2 &= n \int_{-\infty}^{\infty} [F_n(y) - F_o(y)]^2 [F_o(y)(1 - F_o(y))]^{-1} dF_o(y) \\ &= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left\{log[F_o(Y_{(i)})] + log[1 - F_o(Y_{(n+1-i)})]\right\} \\ &= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)log[F_o(Y_{(i)})] - \frac{1}{n}\sum_{i=1}^{n}(2n+1-2i)log[1 - F_o(Y_{(i)})]. \\ &= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)log[U_{(i)}] - \frac{1}{n}\sum_{i=1}^{n}(2n+1-2i)log[1 - U_{(i)}]. \end{aligned}$$

Distribution-Free GOF Statistics

Just as was done in the K-S statistics, the probability integral transform theorem allows the derivation of the distribution of the Cramer-von Mises and Anderson-Darling statistic. This results in distributions which depend only on the sample size n and not on the specific form of F. More specifically, we have the following:

- Let F_o be the proposed cdf for the population cdf
- Let $U_i = F_o(Y_i), \quad i = 1, \ldots, n$, then $U_1, U_2, \ldots, U_n$ are iid with cdf H
- If $F(y) = F_o(y)$, for all y, then H is a uniform on $(0, 1)$ cdf, that is,
- $H(y) = y$ for $0 \leq y \leq 1$.
- If $F(y) \neq F_o(y)$, for all y, then H is not a uniform on $(0, 1)$ cdf
- The K-S, CvM, and AD statistics all attempt to measure the degree to which H is not a uniform on (0,1) cdf.
- The p-value is calculated under assumption that $F = F_o$ and hence the p-values associated with K-S, CvM, and AD statistics are all computed as if $U_i s$ have a uniform on (0,1) distribution.
- It is important to note that the distribution of all three GOF statistics, K-S, CvM, and AD statistics, depend on the form of F if F_o is not the correct model for F.

Tables for Computing p-values

In the following table, approximations to the upper percentiles of the Kolmogorov-Smirnov (D_n), Cramer-von Mises (W_n^2), and Anderson-Darling (A_n^2) statistics are given.

The approximations allow a single table for all values of the sample size n. Without the use of the approximation, it would be necessary to have a separate table for each value of n.

Table 1: Percentiles for GOF Measures (Completely Specified Distributions)

		Upper Percentiles							
Statistic	Modified Statistic	.25	.15	.10	.05	.025	.01	.005	.001
D_n	$D_n(\sqrt{n} + .12 + .11/\sqrt{n})$	1.019	1.138	1.224	1.358	1.480	1.628	1.731	1.950
W_n^2	$(W_n^2 - \frac{.4}{n} + \frac{.6}{n^2})(1 + \frac{1}{n})$	0.209	0.284	0.347	0.461	0.581	0.743	0.869	1.167
A_n^2	For all $n \geq 5$	1.248	1.610	1.933	2.492	3.070	3.857	4.500	6.000

The following table given the cdf of the Anderson-Darling measure when the proposed model for F is completely specified, that is, $G(z) = P[A_n^2 \leq z]$, for $n \geq 5$.

Table 2: CDF for Anderson-Darling (Completely Specified Distributions)

z	G(z)	z	G(z)	z	G(z)	z	G(z)	z	G(z)	z	G(z)
0.05	0.0000	0.75	0.4815	1.45	0.8111	2.15	0.9239	2.85	0.9674	3.80	0.9891
0.10	0.0000	0.80	0.5190	1.50	0.8235	2.20	0.9285	2.90	0.9692	3.90	0.9902
0.15	0.0000	0.85	0.5537	1.55	0.8350	2.25	0.9328	2.95	0.9710	4.00	0.9913
0.20	0.0096	0.90	0.5858	1.60	0.8457	2.30	0.9368	3.00	0.9726	4.25	0.9934
0.25	0.0296	0.95	0.6154	1.65	0.8556	2.35	0.9405	3.25	0.9795	4.50	0.9950
0.30	0.0618	1.00	0.6427	1.70	0.8648	2.40	0.9441	3.30	0.9807	4.60	0.9955
0.35	0.1036	1.05	0.6680	1.75	0.8734	2.45	0.9474	3.35	0.9818	4.70	0.9960
0.40	0.1513	1.10	0.6912	1.80	0.8814	2.50	0.9504	3.40	0.9828	4.80	0.9964
0.45	0.2019	1.15	0.7127	1.85	0.8888	2.55	0.9534	3.45	0.9837	4.90	0.9968
0.50	0.2532	1.20	0.7324	1.90	0.8957	2.60	0.9561	3.50	0.9846	5.00	0.9971
0.55	0.3036	1.25	0.7503	1.95	0.9021	2.65	0.9586	3.55	0.9855	5.50	0.9983
0.60	0.3520	1.30	0.7677	2.00	0.9082	2.70	0.9610	3.60	0.9863	6.00	0.9990
0.65	0.3930	1.35	0.7833	2.05	0.9138	2.75	0.9633	3.65	0.9870	7.00	0.9997
0.70	0.4412	1.40	0.7973	2.10	0.9190	2.80	0.9654	3.70	0.9878	8.00	0.9999

Example of Evaluating Fit of Data to Normal Distribution

Let W be the weights of 21-day-old Leghorn Chickens and let F be the cdf of W. We want to evaluate whether the cdf of W is a normal distribution with $\mu = 200$ and $\sigma = 35$, i.e., W is distributed $N(200, (35)^2)$. A random sample of $n = 20$ twenty-one-day-old Leghorn Chickens yields the following weights:

156	162	168	182	186	190	190	196	202	210
214	220	226	230	230	236	236	242	246	270

The following R code yields the necessary values for the D_n W_n^2 A_n^2:

```
Calculations for GOF for Weight of Chickens Example: gofnormex.R
The above program is for the situation where the mean
and standard deviation of the normal distribution are specified.

x = c(156,162,168,182,186,190,190,196,202,210,214,220,226,230,230,236,236,242,246,270)
n = 20
m = 200
a = 35
x = sort(x)
z = pnorm(x,m,a)   #computes F0(X(i))
i = seq(1,n,1)

# K-S Computations:

d1 = i/n - z

dp = max(d1)

d2 = z - (i - 1)/n

dm = max(d2)

ks = max(dp,dm)

KS = ks*(sqrt(n)+.12+0.11/sqrt(n))

#reject normality at 0.05 level if KS > 1.358
```

From the above output we obtain:

Kolmogorov-Smirnov: $D_n = 0.1712159$ with modified value

$D_n^* = D_n(\sqrt{n} + 0.12 + 0.11/\sqrt{n}) = 0.7904581.$

The p-value is given by $P[D_n^* \geq 0.790]$

which from Table 1, we conclude $p - value \geq 0.25$.

```
# Cramer-von Mises Computations:

wi = (z-(2*i-1)/(2*n))^2

s = sum(wi)

cvm = s + 1/(12*n)

CvM = (cvm-.4/n+.6/n**2)*(1+1/n)

#reject normality at 0.05 level if CvM > 0.461
```

Cramer-von Mises: $W_n^2 = 0.1874563$ with modified value
$W_n^{2*} = (W_n^2 - .4/n + .6/n**2) * (1 + 1/n) = 0.1774$.
The p-value is given by $P[W_n^{2*} \geq 0.1774]$ which from Table 1,
we conclude $p - value \geq 0.25$.

```
# Anderson-Darling Computations:

a1i = (2*i-1)*log(z)

a2i = (2*n+1-2*i)*log(1-z)

s1 = sum(a1i)

s2 = sum(a2i)

AD = -n-(1/n)*(s1+s2)

#reject normality at 0.05 level if AD > 2.492
```

Anderson-Darling: $A_n^2 = 1.016849$.
The p-value is given by $P[A_n^2 \geq 1.016849] = 1 - G(1.02)$,
using Table 2, we obtain $p - value \approx 1 - .65 \approx 0.35$.

The R function **ks.test(x,"pnorm",mu,sigma)**

calculates the K-S test for normal with specified values for "mu" and "sigma". For our example,

ks.test(x,"pnorm",200,35) yields:

One-sample Kolmogorov-Smirnov test

data: x

D = 0.1712, p-value = 0.6008

Other distributions can be used in the **ks.test** function, for example,

- ks.test(x,"pexp",s); where s is the reciprocal of β: $s = \frac{1}{\beta}$
- ks.test(x,"pgamma",c,s), where c is the value of the shape parameter, α), s is the value of the scale parameter, β)
- ks.test(x,"pweibull",c,s), where c is the value of the shape parameter (γ), s is the value of the scale parameter, α)
- many other distributions can be specified but the values of the location, shape and scale parameters must be specified.

GOF Measure - Model Completely Specified - Censored Data

Let F_o be a completely specified cdf or

the standard member of a location scale family of distributions.

Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics from an iid set of r.v.s with cdf F.

Suppose we want to evaluate if F_o is an appropriate model for F, that is,

Evaluate the statement: $F = F_o$.

Probability Plot - Censored Data

Suppose $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$ are the m uncensored observations and

suppose $Y_{(m+1)} \leq Y_{(m+2)} \leq \cdots \leq Y_{(n)}$ are the $n - m$ censored observations.

In the probability plot, plot just the m uncensored values:

$$\left(Q_o(u_1), Y_{(1)}\right), \left(Q_o(u_2), Y_{(2)}\right), \cdots, \left(Q_o(u_m), Y_{(m)}\right)$$

where $u_i = \frac{i-.5}{n}$. Note, that the denominator of u_i is n and not m.

Anderson-Darling GOF Measure - Censored Data

Type I or Type II Censoring

Suppose $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$ are the m uncensored observations and
$Y_{(m+1)} \leq Y_{(m+22)} \leq \cdots \leq Y_{(n)}$ are censored.

In the case of Type I censoring, the only knowledge about the censored values is that
$Y_{(i)} \geq t$ for $i = m+1, \ldots, n$.

In the case of Type II censoring, the only knowledge about the censored values is that $Y_{(i)} \geq Y_{(m)}$ for $i = m+1, \ldots, n$.

Let $T_{(i)} = F_o(Y_{(i)}$, where F_o is the proposed cdf for the population from which the data was obtained.

Let R be the proportion of the distribution which is sampled, that is,
$R = F_o(t)$ for Type I censoring and $R = F_o(Y_{(m)})$ for Type II censoring.

Calculate a modified A-D statistic which takes into account right censoring:

$$A_n^2 = -\frac{1}{n}\sum_{i=1}^{m}(2i-1)\left[log(T_{(i)}) - log(1-T_{(i)})\right] - 2\sum_{i=1}^{m} log(1-T_{(i)})$$
$$-\frac{1}{n}\left[(n-m)^2 log(1-R)\right] + \frac{m^2}{n}\ log(R) - nR$$

Tables for calculating the p-values for the A-D statistics are contained in the following table which were taken from an article by Pettit and Stephens published in *Biometrika*, Vol. 63, pp. 291-298. The p-value is obtained by first computing the value of R, then the value of A_n^2, then the p-value is obtained by p=value = 1-Percentage Point in Table.

Percentage Points	Proportion of population sampled, R				
	0.75	0.80	0.85	0.90	0.95
.01	0.043	0.061	0.083	0.110	0.145
.025	0.055	0.078	0.105	0.137	0.179
.05	0.071	0.098	0.130	0.168	0.216
.10	0.095	0.129	0.169	0.216	0.271
.50	0.321	0.402	0.488	0.578	0.674
.90	1.134	1.322	1.498	1.661	1.810
.95	1.546	1.784	2.000	2.194	2.362
.975	1.976	2.266	2.525	2.752	2.940
.99	2.562	2.925	3.243	3.513	3.730

Note this table is only for the case of goodness of fit when the proposed cdf is completely specified, no unknown parameters in F_o. For the situations where F_o contains unknown parameters which are replaced with the MLE's, the p-values will only be approximate.

Case 3: Random Censoring: Suppose there are m uncensored and $n-m$ randomly censored observations. Compute A^2_m using just the uncensored observations and use m as the sample size in tables for determining p-values.

GOF for Censored Data Based on PL Estimator $\hat{S}(t)$

The following discussion is based on material from *Survival Analysis* by Rupert Miller.

From the censored data obtain the Kaplan-Meier Product Limit estimator of the survival function: $\hat{S}(t)$.

1. Plot $log(\hat{S}(t))$ vs t, if the plotted points are close to the line $y = t/C$ then an exponential($\beta = C$) model would be appropriate
2. Plot $log(-log(\hat{S}(t)))$ vs $log(t)$, if the plotted points are close to the line $y = C_1 + C_2\ log(t)$ then a Weibull($\gamma = C_2,\ \alpha = e^{-C_1/C_2}$) model would be appropriate
3. Plot $\hat{S}(t)$ vs t, if the plotted points are close to the line $y = 1 - \Phi\left(\frac{log(t)-C_1}{C_2}\right)$, where $\Phi()$ is the N(0,1) cdf, then a LogNormal($\mu = C_1, \sigma = C_2$) model would be appropriate

For evaluating the fit of other distributions the standard reference distribution plot could be implemented, that is, plot $Q_o(u)$ vs $\hat{Q}(u)$ where

$Q_o(u)$ is the quantile function from the specified distribution, for example, Gamma($\alpha = 2, \beta = 5$) and

$\hat{Q}(u)$ are estimated quantiles obtained from the PL estimator, that is, $\hat{Q}(u) = inf\{y:\ S(y) \leq 1-u\}$

Plot using $u_i = \frac{i-.5}{n}$. If the plotted points are close to a straight line then the selected model is appropriate.

Complete Data Sets - No Censored Values

GOF Measure - Model Not Completely Specified

Suppose there is no censoring in the data and the cdf F is not be completely specified, that is, $F_o(y) = F_o(y;\boldsymbol{\theta})$, where some or all the values of $\boldsymbol{\theta}$ are not given a specific value.

For example, normal model but with μ and σ not given,

Exponential with the value of β not specified,

Gamma with α and/or β not specified.

We will begin with a measure for the normal distribution which is not a function of the edf and then develop measures based on the edf.

GOF Measure for the Normal Distribution: Shapiro-Wilk Measure

Shapiro and Wilks W statistic is one of the most powerful procedures for assessing the fit of the normal distribution. The W statistic is a measure of the straightness of the normal reference plot, and small values of W indicate a departure from normality. The values of μ and σ do not need to be specified for the computation of the W statistic:

$$W = \frac{\left(\sum_{i=1}^{k} a_{n-i+1}[X_{(n-i+1)} - X_{(i)}]\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2},$$

where $k = \frac{n}{2}$ if n is even and

$k = \frac{(n-1)}{2}$ if n is odd,

$X_{(i)}s$ are the order statistics of the X_is and

the coefficients a_is are given in Table A28 on the following page.

We can use the percentiles in Table A29 on the following page to assess the p-value associated with a computed value of W. The computation of W can also be obtained from SAS and

R: **shapiro.test(y)**.

See the following example for the necessary SAS code.

EXAMPLE - Using Shapiro-Wilk GOF Measure

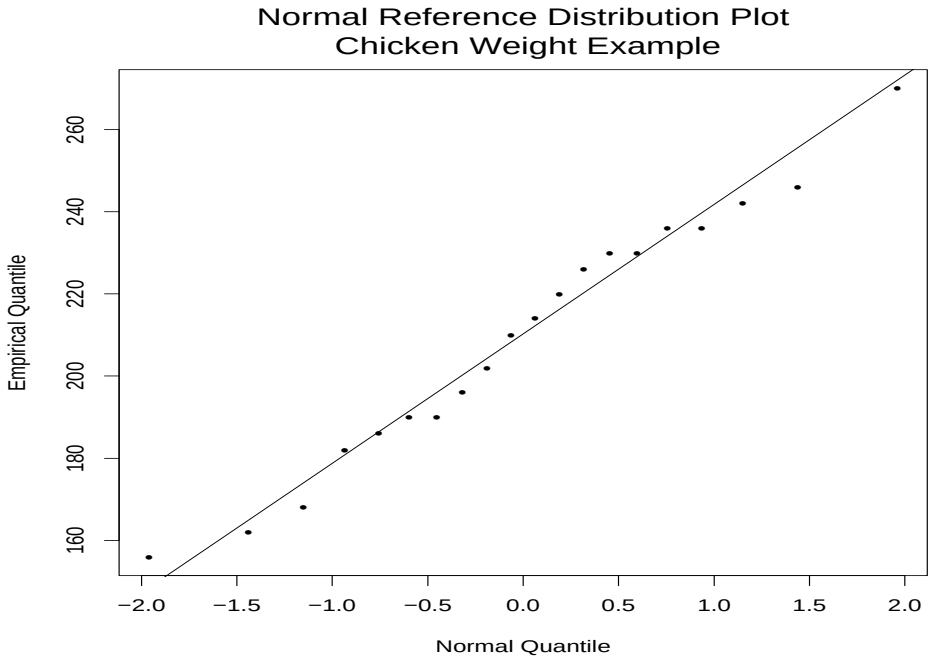
Suppose in the chicken weight example the researcher did not specify the values for μ and σ. Evaluate the degree of fit of the normal distribution to these 20 data values using the S-W statistic. The data in ordered form is given by:

156	162	168	182	186	190	190	196	202	210
214	220	226	230	230	236	236	242	246	270

$$W = \frac{\left(\sum_{i=1}^{k} a_{n-i+1}[X_{(n-i+1)} - X_{(i)}]\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} =$$

$$= \frac{(.4734 * 270 + .3211 * 246 + \cdots + .0140 * 214 - .0140 * 210 - .0422 * 202 - \cdots - .3211 * 162 - .4734 * 156)^2}{17844.8} = .975629$$

From Table A29, $p - value = Pr[W < .975629] > Pr[W < .959] = .50$. Therefore, we can conclude that the normal distribution provides an excellent fit to the chicken weight data.
A normal reference distribution plot of the data confirms this calculation



The SAS code to compute the Shapiro-Wilk statistic is given here:

```
*Calculations for GOF for Weight of Chickens Example:
~longneck/meth1/gofnorm.sas;
options ps=72 ls=65;
data;
    input y @@;
    cards;
156 162 168 182 186 190 190 196 202 210 214 220 226
    230 230 236 236 242 246 270
run;
proc univariate plot normal;
    run;
```

```
                  Tests for Normality

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

Shapiro-Wilk           W     0.975658    Pr < W      0.8667
Kolmogorov-Smirnov     D     0.103722    Pr > D     >0.1500
Cramer-von Mises       W-Sq  0.033776    Pr > W-Sq  >0.2500
Anderson-Darling       A-Sq  0.214174    Pr > A-Sq  >0.2500
```

Note that the value for W is slightly different from our calculation.

The R function

shapiro.test(x),

where x is the data vector, computes the Shapiro-Wilk statistic.

For this example, W = 0.9757 with

p-value = 0.8667,

the identical values as obtained from SAS.

The R package **"nortest"** contains the functions:

ad.test(x); cvm.test(x); lillie.test(x)

which are the Anderson-Darling Test, Cramer von Mises, Kolmogorov-Smirnov tests, respectively. In each of these tests, the values of μ and σ are replaced with $\bar{x}$ and s computed from the data.

For the above examples, we obtain the following values from R:

Test	T.S.	p-value
Shapiro-Wilk	W=.9759	.8667
Kolmogorov-Smirnov	SD=.1037	.8288
Cramer-von-Mises	W=.0338	.7776
Anderson-Darling	AD=.2142	.8260

Table A28 Coefficients Used in the Shapiro–Wilk Test for Normality*

i	a_{n-i+1}											
	$n=3$	4	5	6	7	8	9	10	11	12	13	14
1	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
2		0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
3				0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
4						0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
5								0.0399	0.0695	0.0922	0.1099	0.1240
6										0.0303	0.0539	0.0727
7												0.0240
	$n=15$	16	17	18	19	20	21	22	23	24	25	26
1	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407
2	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043
3	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533
4	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151
5	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836
6	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563
7	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316
8		0.0196	0.0359	0.0496	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089
9				0.0163	0.0303	0.0422	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876
10						0.0140	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672
11								0.0122	0.0228	0.0321	0.0403	0.0476
12										0.0107	0.0200	0.0284
13												0.0094
	$n=27$	28	29	30	31	32	33	34	35	36	37	38
1	0.4366	0.4328	0.4291	0.4254	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015
2	0.3018	0.2992	0.2968	0.2944	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774
3	0.2522	0.2510	0.2499	0.2487	0.2475	0.2463	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391
4	0.2152	0.2151	0.2150	0.2148	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110
5	0.1848	0.1857	0.1864	0.1870	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881
6	0.1584	0.1601	0.1616	0.1630	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686
7	0.1346	0.1372	0.1395	0.1415	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513
8	0.1128	0.1162	0.1192	0.1219	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356
9	0.0923	0.0965	0.1002	0.1036	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211
10	0.0728	0.0778	0.0822	0.0862	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075
11	0.0540	0.0598	0.0650	0.0697	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947
12	0.0358	0.0424	0.0483	0.0537	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824
13	0.0178	0.0253	0.0320	0.0381	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706
14		0.0084	0.0159	0.0227	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592
15				0.0076	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481
16						0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372
17								0.0062	0.0119	0.0172	0.0220	0.0264
18										0.0057	0.0110	0.0158
19												0.0053
	$n=39$	40	41	42	43	44	45	46	47	48	49	50
1	0.3989	0.3964	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2755	0.2737	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2380	0.2368	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2104	0.2098	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1880	0.1878	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1689	0.1691	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1520	0.1526	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1366	0.1376	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1225	0.1237	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1092	0.1108	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.0967	0.0986	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0848	0.0870	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0733	0.0759	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0622	0.0651	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0515	0.0546	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0409	0.0444	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0305	0.0343	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0203	0.0244	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0101	0.0146	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20		0.0049	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21				0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22						0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23								0.0039	0.0076	0.0111	0.0143	0.0174
24										0.0037	0.0071	0.0104
25												0.0035

*$a_i = -a_{n-i+1}$ for $i = 1, 2, \ldots, k$ where $k = n/2$ if n is even and $k = (n-1)/2$ if n is odd.

Source: Shapiro, S. S. and Wilk, M. B. (1965). "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611. Copyright Biometrika Trustees. Reprinted with permission.

Table A29 Critical Values for the Shapiro–Wilk Test for Normality

n	Critical Value $\alpha = 1\%$	2%	5%	10%	50%
3	0.753	0.756	0.767	0.789	0.959
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
6	0.713	0.743	0.788	0.826	0.927
7	0.730	0.760	0.803	0.838	0.928
8	0.749	0.778	0.818	0.851	0.932
9	0.764	0.791	0.829	0.859	0.935
10	0.781	0.806	0.842	0.869	0.938
11	0.792	0.817	0.850	0.876	0.940
12	0.805	0.828	0.859	0.883	0.943
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
16	0.844	0.863	0.887	0.906	0.952
17	0.851	0.869	0.892	0.910	0.954
18	0.858	0.874	0.897	0.914	0.956
19	0.863	0.879	0.901	0.917	0.957
20	0.868	0.884	0.905	0.920	0.959
21	0.873	0.888	0.908	0.923	0.960
22	0.878	0.892	0.911	0.926	0.961
23	0.881	0.895	0.914	0.928	0.962
24	0.884	0.898	0.916	0.930	0.963
25	0.888	0.901	0.918	0.931	0.964
26	0.891	0.904	0.920	0.933	0.965
27	0.894	0.906	0.923	0.935	0.965
28	0.896	0.908	0.924	0.936	0.966
29	0.898	0.910	0.926	0.937	0.966
30	0.900	0.912	0.927	0.939	0.967
31	0.902	0.914	0.929	0.940	0.967
32	0.904	0.915	0.930	0.941	0.968
33	0.906	0.917	0.931	0.942	0.968
34	0.908	0.919	0.933	0.943	0.969
35	0.910	0.920	0.934	0.944	0.969
36	0.912	0.922	0.935	0.945	0.970
37	0.914	0.924	0.936	0.946	0.970
38	0.916	0.925	0.938	0.947	0.971
39	0.917	0.927	0.939	0.948	0.971
40	0.919	0.928	0.940	0.949	0.972
41	0.920	0.929	0.941	0.950	0.972
42	0.922	0.930	0.942	0.951	0.972
43	0.923	0.932	0.943	0.951	0.973
44	0.924	0.933	0.944	0.952	0.973
45	0.926	0.934	0.945	0.953	0.973
46	0.927	0.935	0.945	0.953	0.974
47	0.928	0.928	0.946	0.954	0.974
48	0.929	0.937	0.947	0.954	0.974
49	0.929	0.937	0.947	0.955	0.974
50	0.930	0.938	0.947	0.955	0.974

Source: Adapted from Shapiro, S. S. and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, **52**, 591–611. Copyright Biometrika Trustees. Reprinted with permission.

GOF Test for Normal Dist. Using Reference Distribution Plot

The test is based on determining the Pearson correlation coefficient for the points plotted in the normal reference distribution plot from a random sample $Y_1, Y_2, \ldots, Y_n$ from a population which is proposed to have a normal distribution with unspecified parameters (μ, σ):

$$\left(Q_Z(u_1), Y_{(1)}\right),\ \left(Q_Z(u_2), Y_{(2)}\right),\ \ldots,\ \left(Q_Z(u_n), Y_{(n)}\right)$$

It has been shown by Looney and Gulledge (1985), (" Use of the Correlation Coefficient With Normal Probability Plots", *The American Statistician*, 1985, Vol. 39, No. 1), that using the Blom plotting points:

$$u_i = \frac{i - .375}{n + .25} \quad \text{for} \quad i = 1, 2, \ldots, n$$

yields a slightly more powerful test compared to the test using $u_i = \frac{i-.5}{n}$

The measure of closeness of the n plotted points is computed as

$$R = \frac{\sum_{i=1}^{n} \left(Q_Z(u_i) - \bar{Q}\right)\left(Y_{(i)} - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(Q_Z(u_i) - \bar{Q}\right)^2}\sqrt{\sum_{i=1}^{n} \left(Y_(i) - \bar{Y}\right)^2}}$$

R is a measure of the closeness of the n plotted points to a straight line. The larger the value of R the closer the points are to a straight line and hence the better the fit of a normal distribution to the data.

To determine the p-values associated with R we need a special set of tables because the standard tables for R are invalid in this situation due to the $Y_{(i)}$'s being correlated and the $Q_Z(u_i)$'s being non-random. The tables are contained in the article by Looney and Gulledge (1985) and are reproduced on the next page.

We can apply the above method to the Chicken Weight example using the following R code:

```
# Correlation Test

x = c(156,162,168,182,186,190,190,196,202,210,214,220,226,
      230,230,236,236,242,246,270)
y = sort(x)
n = length(x)
i = seq(1,n,1)
u = (i-.375)/(n+.25)
q = qnorm(u)
r = cor.test(q,y)
```

From the above R code we obtain

$r = 0.991$ which from the tables with n=20 we obtain a p-value of 0.900 which is nearly the same as the p-value from the SW test, 0.8667.

Table 2. Empirical Percentage Points for Correlation Coefficient Test Based on Blom's Plotting Position

	Level													
n	*.000*	*.005*	*.010*	*.025*	*.050*	*.100*	*.250*	*.500*	*.750*	*.900*	*.950*	*.975*	*.990*	*.995*
3	.866	.867	.869	.872	.879	.891	.924	.966	.992	.999	.9997	.9999	1.000	1.000
4	.785	.813	.824	.846	.868	.894	.931	.958	.979	.992	.996	.998	.999	1.000
5	.729	.807	.826	.856	.880	.903	.934	.960	.977	.988	.992	.995	.997	.998
6	.686	.820	.838	.866	.888	.910	.939	.962	.977	.986	.990	.993	.996	.997
7	.651	.828	.850	.877	.898	.918	.944	.964	.978	.986	.990	.992	.995	.996
8	.623	.840	.861	.887	.906	.924	.948	.966	.978	.986	.990	.992	.994	.995
9	.599	.854	.871	.894	.912	.930	.952	.968	.980	.986	.990	.992	.994	.995
10	.578	.862	.879	.901	.918	.934	.954	.970	.980	.987	.990	.992	.994	.995
11	.560	.870	.886	.907	.923	.938	.957	.972	.981	.987	.990	.992	.994	.995
12	.544	.876	.892	.912	.928	.942	.960	.973	.982	.988	.990	.992	.994	.995
13	.529	.885	.899	.918	.932	.945	.962	.974	.983	.988	.991	.992	.994	.995
14	.516	.890	.905	.923	.935	.948	.964	.976	.984	.989	.991	.992	.994	.995
15	.504	.896	.910	.927	.939	.951	.965	.977	.984	.989	.991	.993	.994	.995
16	.493	.899	.913	.929	.941	.953	.967	.978	.985	.989	.991	.993	.994	.995
17	.483	.905	.917	.932	.944	.954	.968	.979	.986	.990	.992	.993	.994	.995
18	.473	.908	.920	.935	.946	.957	.970	.979	.986	.990	.992	.993	.9945	.9952
19	.465	.914	.924	.938	.949	.958	.971	.980	.987	.990	.992	.993	.9946	.9953
20	.457	.916	.926	.940	.951	.960	.972	.981	.987	.991	.992	.994	.9947	.9954
21	.449	.918	.930	.943	.952	.961	.973	.982	.987	.991	.993	.994	.995	.996
22	.442	.923	.933	.945	.954	.963	.974	.982	.988	.991	.993	.994	.995	.996
23	.435	.925	.935	.947	.956	.964	.975	.983	.988	.991	.993	.994	.995	.996
24	.429	.927	.937	.949	.957	.965	.976	.983	.988	.992	.993	.994	.995	.996
25	.422	.929	.939	.951	.959	.966	.976	.984	.989	.992	.993	.994	.995	.996
26	.417	.932	.941	.952	.960	.967	.977	.984	.989	.992	.993	.994	.995	.996
27	.411	.934	.943	.953	.961	.968	.978	.985	.989	.992	.994	.995	.9955	.9960
28	.406	.936	.944	.955	.962	.969	.978	.985	.990	.992	.994	.995	.9955	.9960
29	.401	.939	.946	.956	.963	.970	.979	.985	.990	.993	.994	.995	.9956	.9961
30	.396	.939	.947	.957	.964	.971	.979	.986	.990	.993	.994	.995	.9957	.9962
31	.392	.942	.950	.958	.965	.972	.980	.986	.990	.993	.994	.995	.9957	.9962
32	.387	.943	.950	.959	.966	.972	.980	.987	.991	.993	.994	.995	.9958	.9963
33	.383	.944	.951	.961	.967	.973	.981	.987	.991	.993	.994	.995	.9959	.9963
34	.379	.946	.953	.962	.968	.974	.981	.987	.991	.993	.994	.995	.996	.997
35	.375	.947	.954	.962	.969	.974	.982	.987	.991	.994	.9945	.9953	.996	.997
36	.371	.948	.955	.963	.969	.975	.982	.988	.991	.994	.9946	.9954	.996	.997
37	.368	.950	.956	.964	.970	.976	.983	.988	.991	.994	.995	.9955	.9962	.997
38	.364	.951	.957	.965	.971	.976	.983	.988	.992	.994	.995	.9956	.9963	.997
39	.361	.951	.958	.966	.971	.977	.983	.988	.992	.994	.995	.9957	.9963	.997
40	.358	.953	.959	.966	.972	.977	.984	.989	.992	.994	.995	.9957	.9964	.997
41	.354	.953	.960	.967	.973	.977	.984	.989	.992	.994	.995	.996	.9965	.9968
42	.351	.954	.961	.968	.973	.978	.984	.989	.992	.994	.995	.996	.9965	.9969
43	.348	.956	.961	.968	.974	.978	.984	.989	.992	.994	.995	.996	.9966	.9969
44	.346	.957	.962	.969	.974	.979	.985	.989	.993	.9945	.9953	.996	.9966	.9970
45	.343	.957	.963	.969	.974	.979	.985	.990	.993	.9945	.9954	.996	.9966	.9970
46	.340	.958	.963	.970	.975	.980	.985	.990	.993	.995	.9955	.9961	.9968	.9971
47	.337	.959	.965	.971	.976	.980	.986	.990	.993	.995	.9956	.9962	.9968	.9972
48	.335	.959	.965	.971	.976	.980	.986	.990	.993	.995	.9956	.9962	.9968	.9972
49	.332	.961	.966	.972	.976	.981	.986	.990	.993	.995	.9957	.9963	.9968	.9972
50	.330	.961	.966	.972	.977	.981	.986	.990	.993	.995	.9957	.9963	.9969	.9972
55	.319	.965	.969	.974	.979	.982	.987	.991	.994	.995	.996	.9966	.9971	.9974
60	.309	.967	.971	.976	.980	.984	.988	.992	.994	.9956	.9963	.9968	.9973	.9975
65	.300	.969	.973	.978	.981	.985	.989	.992	.994	.996	.9965	.9969	.9974	.9977
70	.292	.971	.975	.979	.983	.986	.990	.993	.995	.996	.9966	.9971	.9975	.9978
75	.284	.973	.976	.981	.984	.987	.990	.993	.995	.996	.9968	.9972	.9976	.9979
80	.277	.975	.978	.982	.985	.987	.991	.993	.995	.996	.9970	.9974	.9978	.9980
85	.271	.976	.979	.983	.985	.988	.991	.994	.996	.9966	.9971	.9975	.9979	.9981
90	.266	.977	.980	.984	.986	.988	.992	.994	.996	.9967	.9972	.9976	.9979	.9981
95	.260	.979	.981	.984	.987	.989	.992	.994	.996	.9969	.9973	.9977	.9980	.9982
100	.255	.979	.982	.985	.987	.989	.992	.995	.996	.9970	.9974	.9978	.9981	.9983

K-S, CvM, A-D Measures for the Normal Distribution

When the parameters of the distribution are not specified, the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics cannot be calculated directly because in making the transformation to $X = F_o(Y)$, F_o is not completely specified, it has unknown parameters.

An approach which overcomes this problem is to estimate the unknown parameters in F_o using maximum likelihood estimators and then calculate the K-S, or CvM, or A-D statistics.

The percentage points given in Table 1 and Table 2 would be good approximations only if the sample size n is large.

For small n, D'Agostino and Stephens provide in their book, *Goodness-of -Fit Techniques*, modifications and percentage points for the statistics. However, a separate table must be developed for each family of distributions, that is, there are separate tables for Normal, Exponential, Weibull, extreme-value distribution, etc.

The following table provides modifications and percentage points for the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics for measuring fit of a normal distribution when the parameters are not specified.

Table 3: Modifications and Percentiles for GOF Measures for Normal Distributions with μ and σ Unknown

		Upper Percentiles							
Statistic	Modified Statistic	.50	.25	.15	.10	.05	.025	.01	.005
D_n	$D_n(\sqrt{n} - .01 + .85/\sqrt{n})$	-	-	0.775	0.819	0.895	0.995	1.035	-
W_n^2	$W_n^2(1 + \frac{.5}{n})$	0.051	0.074	0.091	0.104	0.126	0.148	0.179	0.201
A_n^2	$A_n^2(1 + \frac{.75}{n} + \frac{2.25}{n^2})$	0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159

For the chicken weight example, we have the following values for the gof measures:

$$\text{For KS:} \quad D_n = .10372 \Rightarrow D_n(\sqrt{n} - .01 + .85/\sqrt{n}) = .4825 \Rightarrow p - value > 0.15$$

$$\text{For CvM:} \quad W_n^2 = .03378 \Rightarrow W_n(1 + \frac{.5}{n}) = .03462 \Rightarrow p - value > 0.50$$

$$\text{For AD:} \quad A_n^2 = .2142 \Rightarrow A_n(1 + \frac{.75}{n} + \frac{2.25}{n^2}) = .2234 \Rightarrow p - value > .50$$

These values are consistent with the values obtained from the SAS output.

K-S, CvM, A-D Measures for the Exponential Distribution

Let $Y_1, Y_2, \cdots, Y_n$ be iid r.v.s with continuous cdf $F(\cdot)$.

The following modifications provide gof measures for the exponential distribution when β is not specified.

The exponential cdf is given by $F_o(y) = 1 - e^{-\frac{y}{\beta}}$.

The MLE of β is given by $\hat{\beta} = \bar{Y}$.

In our formulas for Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling, replace $F_o(Y_{(i)})$ with $\widehat{F_o}(Y_{(i)}) = 1 - e^{-Y_{(i)}/\bar{Y}}$

then compute the modified forms as given below

$$\text{For KS:} \quad \left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$$

$$\text{For CvM:} \quad W_n^2\left(1.0 + \frac{0.16}{n}\right)$$

$$\text{For AD:} \quad A_n^2\left(1.0 + \frac{0.6}{n}\right)$$

The percentiles are given in the following table:

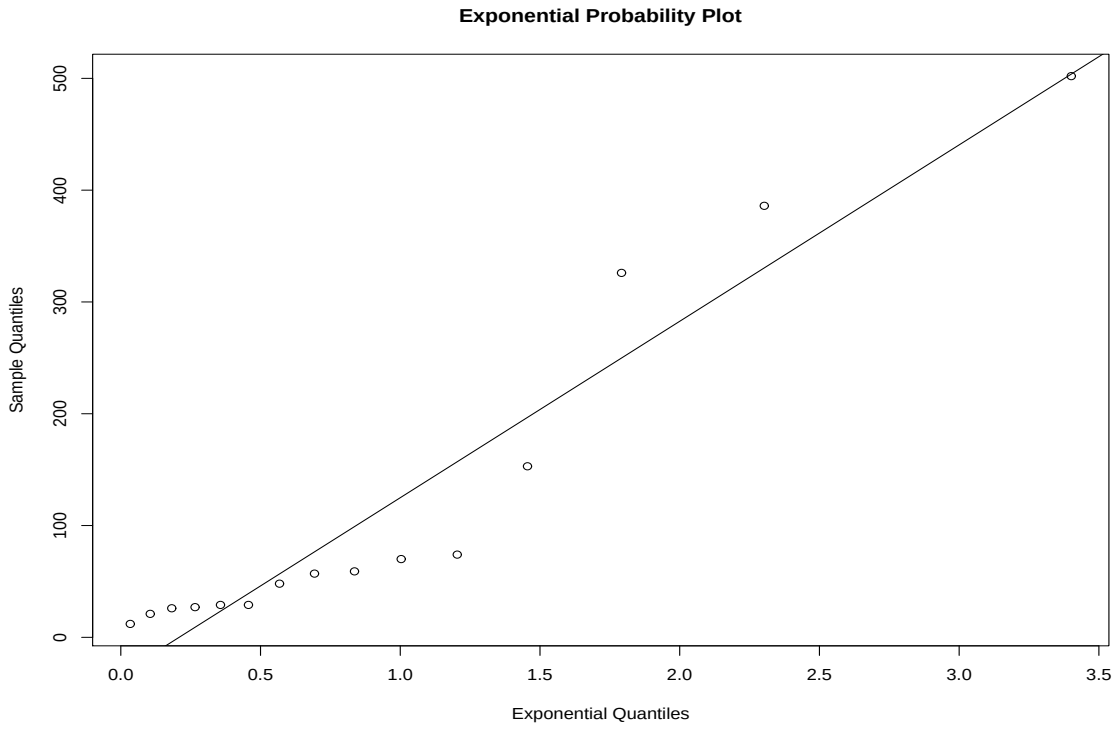
Table 4: Modifications and Percentiles for GOF Measures for Exponential Distribution with β Unknown

		Upper Percentiles								
Statistic	Modified Statistic	.25	.20	.15	.10	.05	.025	.01	.005	.0025
D_n	$(D_n - \frac{0.2}{n})(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}})$	-	-	0.926	0.995	1.094	1.184	1.298	-	-
W_n^2	$W_n^2(1.0 + \frac{0.16}{n})$	0.116	0.130	0.148	0.175	0.222	0.271	0.338	0.390	0.442
A_n^2	$A_n^2(1.0 + \frac{0.6}{n})$	0.736	0.816	0.916	1.062	1.321	1.591	1.959	2.244	2.534

Example Let F be the cdf for the time to failure of air conditioners of a particular brand. A random sample of 15 units are put on an accelerated failure test and the times to failure(in hours) are given here:

12	21	26	27	29	29	48	57
59	70	74	153	326	386	502	

An Exponential Reference Distribution plot is given here.



From the data compute: $\bar{Y} = 121.2$ and let $\widehat{F_o}(Y_{(i)}) = 1 - e^{-Y_{(i)}/121.2}$. The following R code performs the necessary calculations.

```
w = c(12,21,26,27,29,29,48,57,59,70,74,153,326,386,502)
n = 15
lam = mean(w)
w = sort(w)

z = 1-exp(-w/lam)   #computes F0(X(i))

i = seq(1,n,1)

# K-S Computations:

d1 = i/n - z

dp = max(d1)

d2 = z - (i - 1)/n

dm = max(d2)

KS = max(dp,dm)

KSM = (KS-.2/n)*(sqrt(n)+.26+.5/sqrt(n))
```

```
# Cramer-von Mises Computations:

wi = (z-(2*i-1)/(2*n))^2

s = sum(wi)

cvm = s + 1/(12*n)

cvmM = cvm*(1+.16/n)



# Anderson-Darling Computations:

a1i = (2*i-1)*log(z)

a2i = (2*n+1-2*i)*log(1-z)

s1 = sum(a1i)

s2 = sum(a2i)

AD = -n-(1/n)*(s1+s2)

ADM = AD*(1+.6/n)
```

From the output we obtain the following values for the modified statistics and use these values to obtain approximate p-values from Table 4:

$$KSM: \quad \left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right) = 1.122 \Rightarrow 0.025 < p - value < 0.05$$

$$cvmM: \quad W_n^2\left(1.0 + \frac{0.16}{n}\right) = 0.221 \Rightarrow p - value = 0.05$$

$$ADM: \quad A_n^2\left(1.0 + \frac{0.6}{n}\right) = 1.210 \Rightarrow 0.05 < p - value < 0.10$$

From the exponential reference distribution plot and the gof statistics, we would conclude that the exponential model does not fit the data very well.

A-D Measure for the Extreme Value Distribution

Let $Y_1, Y_2, \cdots, Y_n$ be iid r.v.s with continuous cdf $F(\cdot)$.

The following modifications provide gof measures for the extreme value distribution when parameters are not specified.

The extreme value cdf is given by

$$F_o(y) = e^{-e^{-(y-\phi)/\theta}}.$$

The MLEs of the parameters are given by

$$\hat{\phi} = -\hat{\theta} ln \left[\frac{1}{n} \sum_{i=1}^{n} e^{-Y_i/\hat{\theta}} \right]$$

Solve iteratively for $\hat{\theta}$:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i - \left[\sum_{i=1}^{n} X_i e^{-X_j/\hat{\theta}} \right] \left[\sum_{i=1}^{n} e^{-X_j/\hat{\theta}} \right]$$

In our formula for Anderson-Darling, replace $F_o(Y_{(i)})$ with $\widehat{F_o}(Y_{(i)})$ where ϕ and θ are replaced with their MLEs.

Then make the following modification to adjust for the value of n:

$$A_n^2 \left(1.0 + \frac{0.2}{\sqrt{n}} \right)$$

The percentiles for the Anderson-Darling Statistic are given in the following table:

Table 5: Modifications and Percentiles for A-D Measure for Extreme Value Distribution with Unspecified Parameters

		Upper Percentiles				
Statistic	Modified Statistic	.25	.10	.05	.025	.01
A_n^2	$A_n^2(1.0 + \frac{0.2}{\sqrt{n}})$	0.474	0.637	0.757	0.877	1.038

A-D Measure for the Weibull Distribution

Let $X_1, X_2, \cdots, X_n$ be iid r.v.s with continuous cdf $F(\cdot)$.

The following modifications provide gof measures for the Weibull distribution when parameters are not specified.

The Weibull cdf is given by

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

If a rv X has a Weibull cdf, then the transformation, $Y = -log(X)$ results in the rv Y having cdf

$$F_o(y) = e^{-e^{-(y-\phi)/\theta}},$$

where $\theta = \frac{1}{\gamma}$ and $\phi = -log(\alpha)$.

The Anderson-Darling gof statistic for the extreme value distribution is then used to measure the fit of the Weibull distribution using $Y_i = -log(X_i)$ as the observed data values.

R Code for GOF for Weibull

The following program files yield the mle estimates for a Weibull distribution and then computes the Anderson-Darling Statistics for testing goodness of the fit of a Weibull Distribution with unspecified parameters.

The statistics include the modification needed to use the Tables included in this handout.

The example used to illustrate the computation is based on a random sample of n=23 observations on the number of revolutions to failure of ball bearings:

```
17.88  28.92   33.00    41.52   42.12   45.60   48.40  51.84
51.96  54.12   55.56    67.80   68.64   68.64   68.88  84.12
93.12  98.64  105.12   105.84  127.92  128.04    173.40
```

```
*R Code to find MLE:

library(MASS)

x <- c(
17.88 , 28.92 ,  33.00 ,   41.52 ,  42.12 ,  45.60 ,  48.40,  51.84 ,
51.96 , 54.12 ,  55.56   , 67.80 ,  68.64  , 68.64 ,  68.88 , 84.12 ,
93.12 , 98.64 , 105.12  , 105.84 , 127.92 , 128.04 , 173.40)

fitdistr(x,"weibull")

output from R code:

    shape        scale
   2.1011178   81.8324383
 ( 0.3285826) ( 8.5971353)
```

From the R code we obtain our parameter estimates:

Estimate of γ is *Weibull Shape*, that is, $\hat{\gamma} = 2.1012$

Estimate of α is *Weibull Scale*, that is, $\hat{\alpha} = 81.8324$

In the Extreme-value distribution form we have

$$\hat{\phi} = -log(\hat{\alpha}) = -log(WeibullScale) = -log(81.8324) = -4.4047$$

$$\hat{\theta} = 1/\hat{\gamma} = 1/2.1012 = 0.4759$$

Anderson-Darling: Let

$$Y_i = -log(X_i); \quad U_i = \widehat{F}_o(Y_i) = e^{-e^{-(Y_i - \hat{\phi})/\hat{\theta}}}$$

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)log(U_{(i)}) + (2n+1-2i)log(1-U_{(i)}))\right] = 0.3276$$

Modify the AD statistic because parameters were estimated:

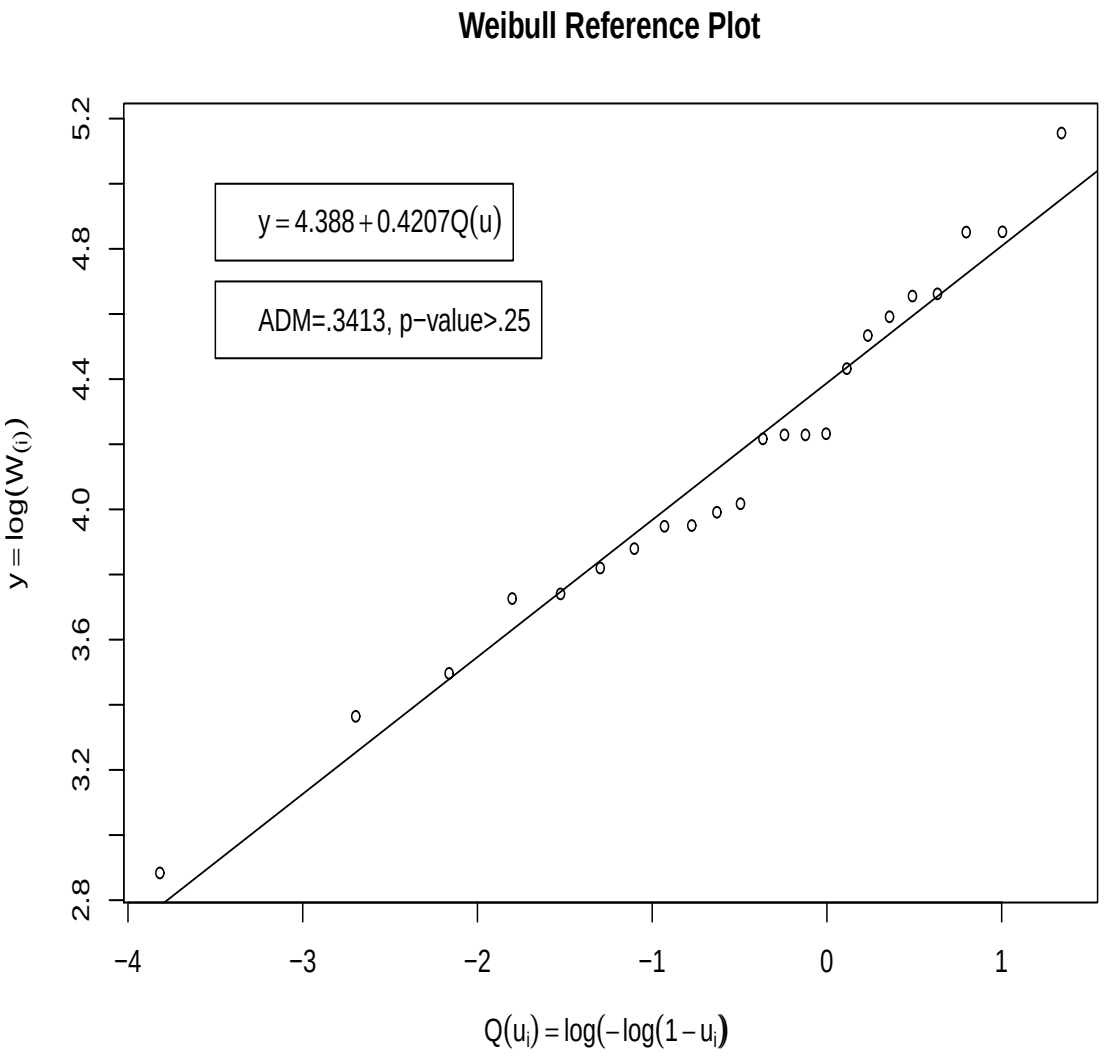
$$AD = AD\left[1 + \frac{.2}{\sqrt{n}}\right] = (.3277)\left[1 + \frac{.2}{\sqrt{23}}\right] = 0.3413$$

Compute p-value using the percentiles in Table 5 yielding $p - value > .25$. Thus, we have a very good fit of a Weibull distribution to the data.

Weibull Probability Plot and GOF Using R

```
# gofweibmle.R
# The following program computes the Anderson-Darling Statistics
# for testing goodness of the fit of a
# Weibull Distribution
# with unspecified parameters (need to supply MLE's).
# The statistics include the modification needed to use the Tables included
# in the GOF handout.
# This example is based on a random sample of n=23 observations:
x = c(17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.40, 51.84,
51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12,
93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40)
n = length(x)
i = seq(1,n,1)
y = -log(x)
y = sort(y)
# Anderson-Darling: For Weibull Model
library(MASS)
mle <- fitdistr(x,"weibull")
shape = mle$estimate[1]
scale = mle$estimate[2]
a = -log(scale)
b = 1/shape
z   = exp(-exp(-(y-a)/b))
A1i = (2*i-1)*log(z)
A2i = (2*n+1-2*i)*log(1-z)
s1  = sum(A1i)
s2  = sum(A2i)

AD  = -n-(1/n)*(s1+s2)
ADM  = AD*(1+.2/sqrt(n))
AD
ADM
n
n = length(y)
weib= -y
weib= sort(weib)
i= 1:n
ui= (i-.5)/n
QW= log(-log(1-ui))
postscript("u:/meth1/psfiles/weibgofmle.ps",horizontal=FALSE)
plot(QW,weib,abline(lm(weib~QW)),
       main="Weibull Reference Plot",cex=.75,lab=c(7,11,7),
       xlab="Q=ln(-ln(1-ui))",
       ylab="y=ln(W(i))")
legend(-3.5,5.0,"y=4.388+.4207Q")
legend(-3.5,4.7,"AD=.3721, p-value>.25")
graphics.off()

-------------------------------------------------------------------------
 OUTPUT: from R:

 AD  = 0.3276
 ADM = 0.3413
```

A Weibull Probability plot of the data is given here:

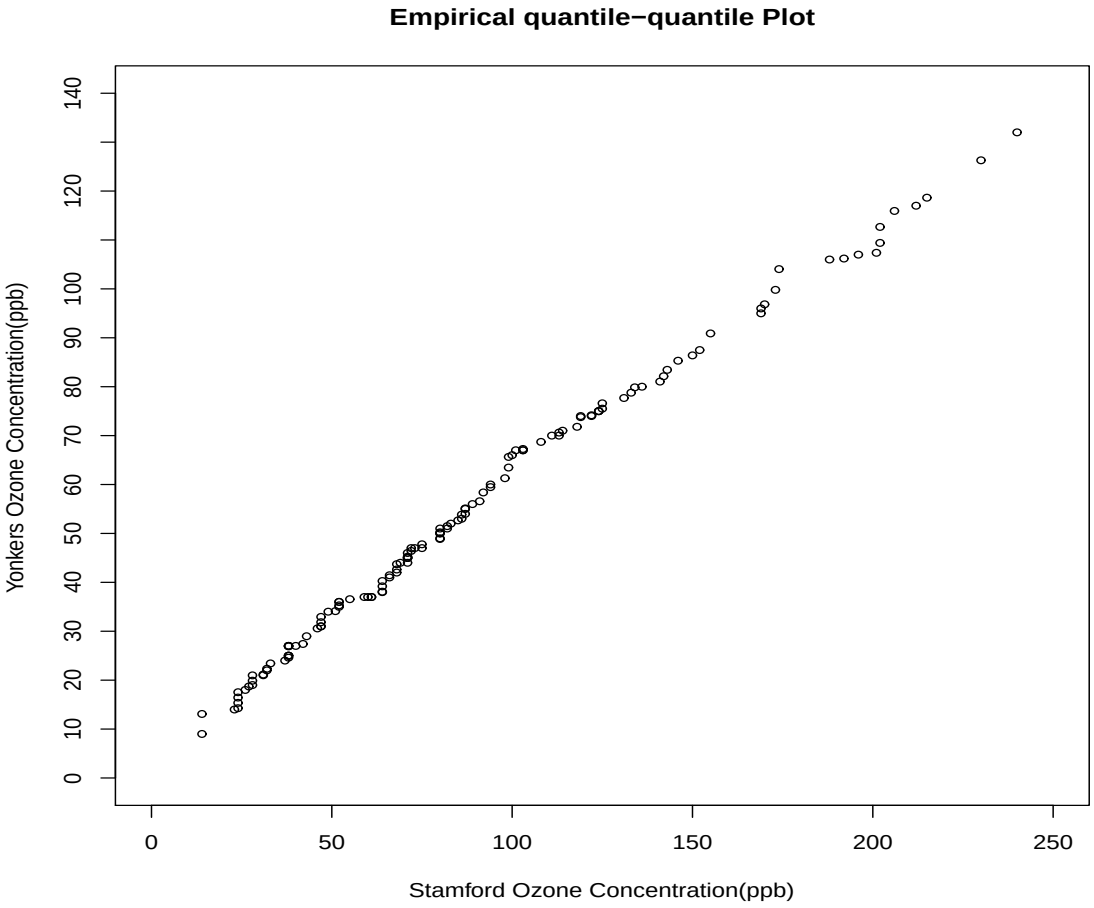


From the R output we have ADM = 0.3413 which implies from Table 5 that $p - value > 0.25$.

From the p-value and the Weibull Reference Distribution plot we would conclude that the Weibull distribution provides an excellent fit to the bearing data.

Goodness of Fit of the Weibull Model to the Ozone Data

In Handout 8, we produced the following q-q plot for the Ozone data from Stamford and Yonkers:

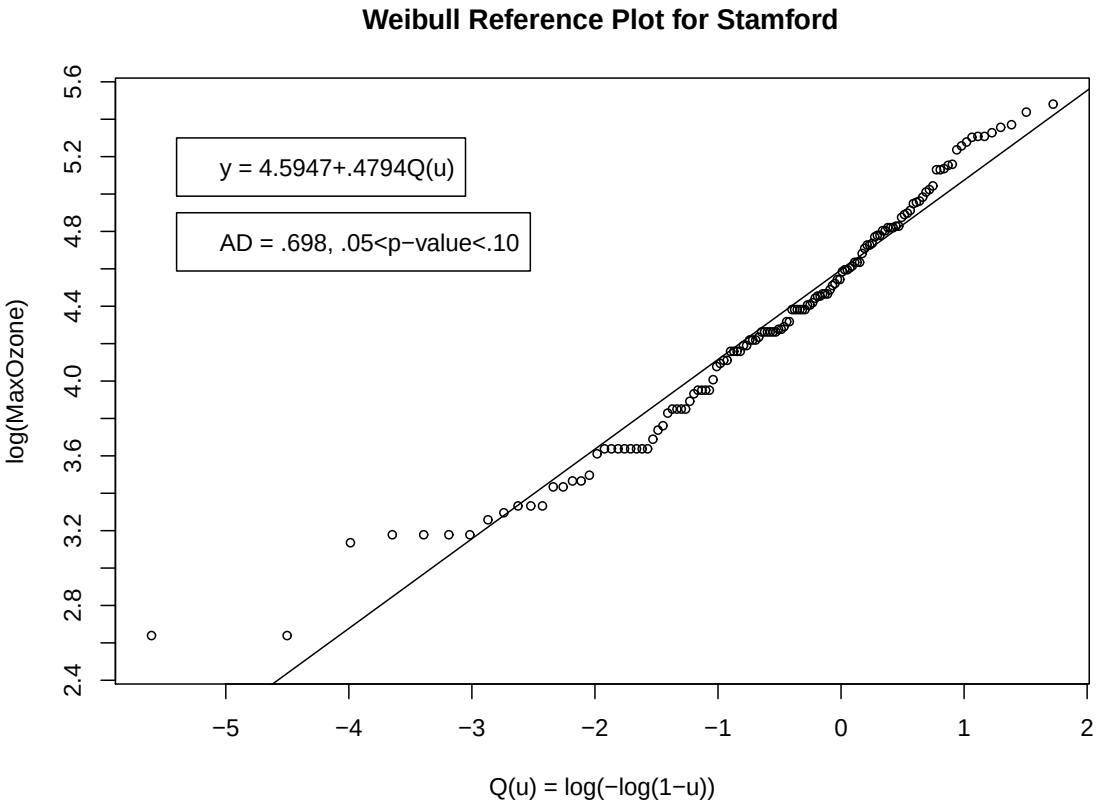


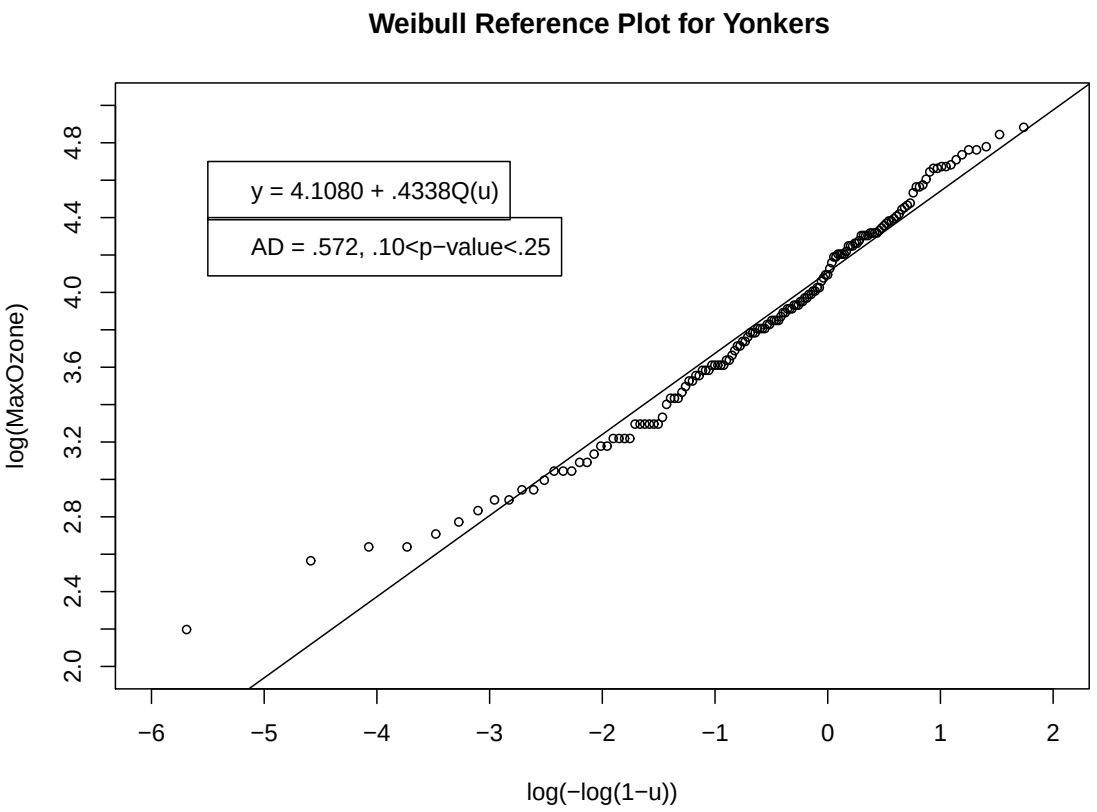
Based on this plot it would appear that the ozone data from the two cities belong to the same parametric family. Because the ozone measurements are the maximum ozone readings on given days, we postulated that a Weibull model may be an appropriate model for these two cities. We will next perform a GOF test for the Weibull model for the two data sets.

From the Anderson-Darling test we have AD=.698 with $.05 < p - value < .10$ for the Stamford ozone data and AD=.572 with $.10 < p - value < .25$ for the Yonkers ozone data. Thus, there is a moderately good fit of a Weibull model for the Stamford data and a good fit for the Yonkers data. This is depicted in the following Weibull reference distribution plots.

Stamford AD=0.690 .05 < p-value < .10

Yonkers AD=0.572 .10 < p-value < .25


Weibull Reference Plot for Stamford
y = 4.5947+.4794Q(u)
AD = .698, .05<p−value<.10
log(MaxOzone)
Q(u) = log(−log(1−u))


Weibull Reference Plot for Yonkers
y = 4.1080 + .4338Q(u)
AD = .572, .10<p−value<.25
log(MaxOzone)
log(−log(1−u))

Box-Cox Transformation

Many of the standard statistical procedures require the data to have a normal distribution.

Suppose the data $Y_1, Y_2, \ldots, Y_n$ is iid r.v.'s with positive values and a pdf f_Y which is skewed.

A power transformation defined by

$$y^{(\theta)} = \begin{cases} \frac{(y^\theta - 1)}{\theta} & \text{if } \theta \neq 0 \\ log(y) & \text{if } \theta = 0 \end{cases}$$

can sometimes produce 'nearly' a normal distribution for $y^{(\theta)}$. That is, the pdf of $y^{(\theta)}$ is a $N(\mu, \sigma^2)$ pdf.

Note: $\lim_{\theta \to 0} \frac{(y^\theta - 1)}{\theta} = log(y)$.

If in fact the power transformation is successful, and $y^{(\theta)}$ has a normal distribution, $N(\mu, \sigma^2)$ then the pdf of y, f_Y, is given by

$$f_Y(y) = y^{\theta-1} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y^{(\theta)} - \mu)^2} = y^{\theta-1} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}\left[\frac{(y^\theta - 1)}{\theta} - \mu\right]^2}$$

Determination of θ:

1. Try several values of θ, do normal-quantile plots and select value of θ which most nearly produces straight line in plot.

2. Maximum Likelihood Estimation: If $y_1^{(\theta)}, ..., y_n^{(\theta)}$ are iid $N(\mu, \sigma^2)$, then the log-likelihood function of $y_1, ..., y_n$ is given by

$$l(\mu, \sigma^2, \theta) = log\left(\prod_{i=1}^{n} f_Y(y_i)\right) = (\theta - 1)\sum_{i=1}^{n} log(y_i) - \frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^{(\theta)} - \mu)^2$$

   For fixed values of θ, maximizing $l(\mu, \sigma^2, \theta)$ over μ and σ yields

$$l_{max}(\theta) = (\theta - 1)\sum_{i=1}^{n} log(y_i) - \frac{n}{2}\left[log(2\pi\hat{\sigma}^2(\theta)) + 1\right]$$

   where,

$$\hat{\sigma}^2(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(\theta)} - \overline{y^{(\theta)}}\right)^2$$

   Select the value of θ which maximizes

$$l^*_{max}(\theta) = (\theta - 1)\sum_{i=1}^{n} log(y_i) - \frac{n}{2}\left[log(2\pi\hat{\sigma}^2(\theta)) + 1\right]$$

a. Generally, we take $\hat{\theta}$, the value which maximizes $L^*(\theta)$, to be one of

$$\ldots, \quad -2, \quad -\frac{3}{2}, \quad -1, \quad -\frac{1}{2}, \quad 0, \quad \frac{1}{2}, \quad 1, \quad \frac{3}{2}, \quad \ldots$$

or "Quarters" of "Thirds" provided the selected value of $\hat{\theta}$ falls in the C.I. for θ.

b. An approximate $100(1-\alpha)\%$ C.I. for θ consists of those values of θ satisfying

$$l_{max}(\hat{\theta}) - l_{max}(\theta) \leq \frac{1}{2}\chi^2(1-\alpha)$$

where $\chi^2(1-\alpha)$ is the upper $100(1-\alpha)$ percentile of the Chi-square distribution with d.f. $= 1$.

Note: Just draw a plot of $l_{max}(\theta)$ vs. θ. Draw a horizontal line at the level

$$l_{max}(\hat{\theta}) - \frac{1}{2}\chi^2(1-\alpha)$$

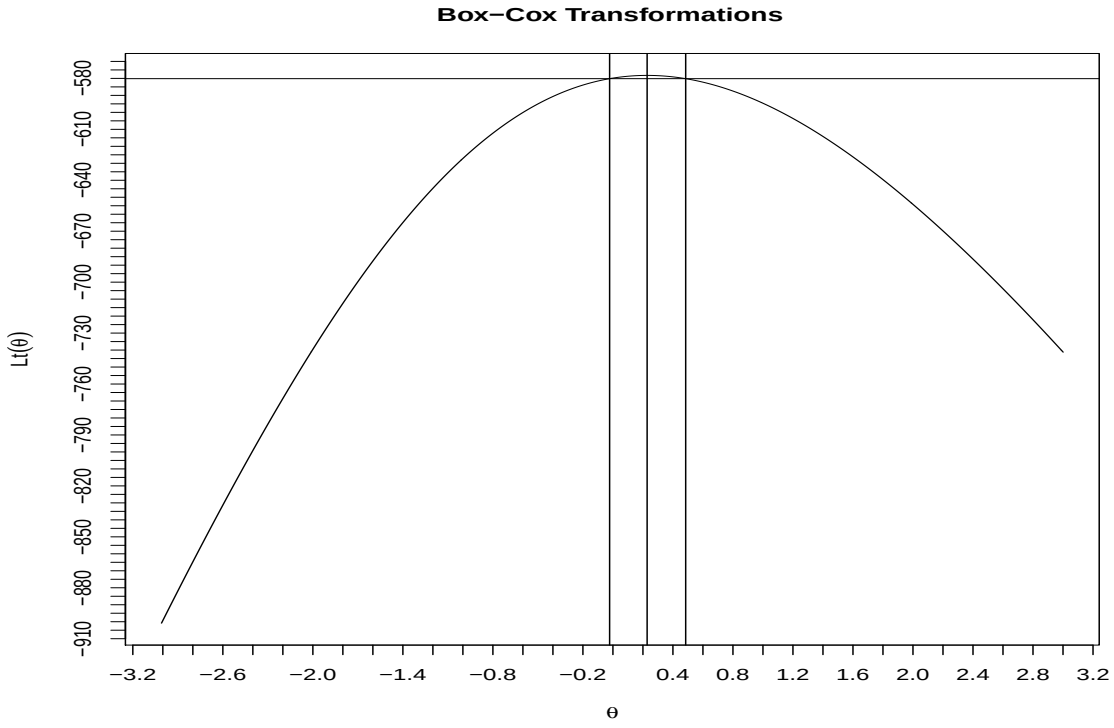
This line in "most cases" cut the curve at two values of θ and these values will be the endpoints of the approximate C.I.

We will apply the Box-Cox transformations to the Stamford ozone data using the following R Code: boxcox,samozone.R

```
y = scan("u:/meth1/sfiles/ozone1.DAT")
n = length(y)
yt0 = log(y)
s = sum(yt0)
varyt0 = var(yt0)
Lt0 = -1*s - .5*n*(log(2*pi*varyt0)+1)
th = 0
Lt = 0
t = -3.01
i = 0
while(t < 3)
{t = t+.001
i = i+1
th[i] = t
yt = (y^t -1)/t
varyt = var(yt)
Lt[i] = (t-1)*s - .5*n*(log(2*pi*varyt)+1)
if(abs(th[i])<1.0e-10)Lt[i]<-Lt0
if(abs(th[i])<1.0e-10)th[i]<-0
}
# The following outputs the values of the likelihood and theta and yields
# the value of theta where likelihood is a maximum
out = cbind(th,Lt)
Ltmax= max(Lt)
imax= which(Lt==max(Lt))
thmax= th[imax]

postscript("boxcox,plotsam.ps",height=8,horizontal=FALSE)

plot(th,Lt,lab=c(30,50,7),main="Box-Cox Transformations",
           xlab=expression(theta),
           ylab=expression(Lt(theta)))
```

```
#the following plots a 95\% c.i. for theta

cic = Ltmax-.5*qchisq(.95,1)

del= .01
iLtci = which(abs(Lt-cic)<=del)
iLtciL= min(iLtci)
iLtciU= max(iLtci)
thLci= th[iLtciL]
thUci= th[iLtciU]
abline(h=cic)
abline(v=thLci)
abline(v=thUci)
abline(v=thmax)
```

The plot of the likelihood function is given here with lines indicating a 95% confidence interval on values of θ which maximize the likelihood function.

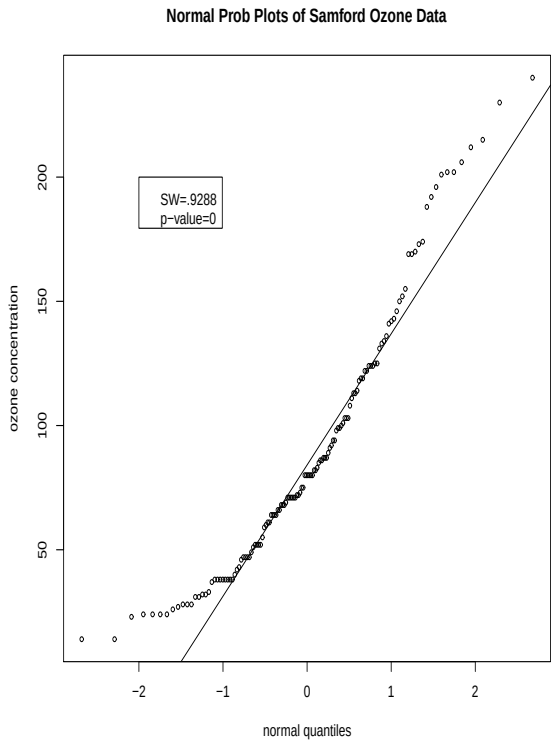


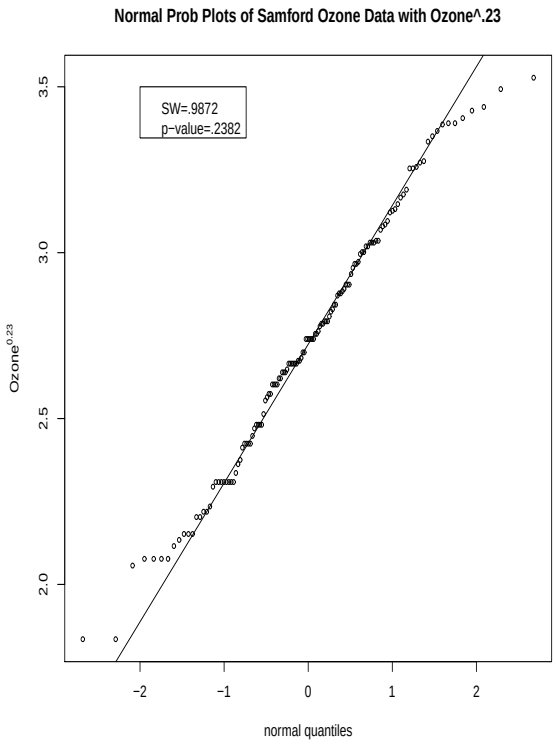
from the R output we obtain a 95% C.I. for θ (thLci,thUci)= $(-.022, .485)$.

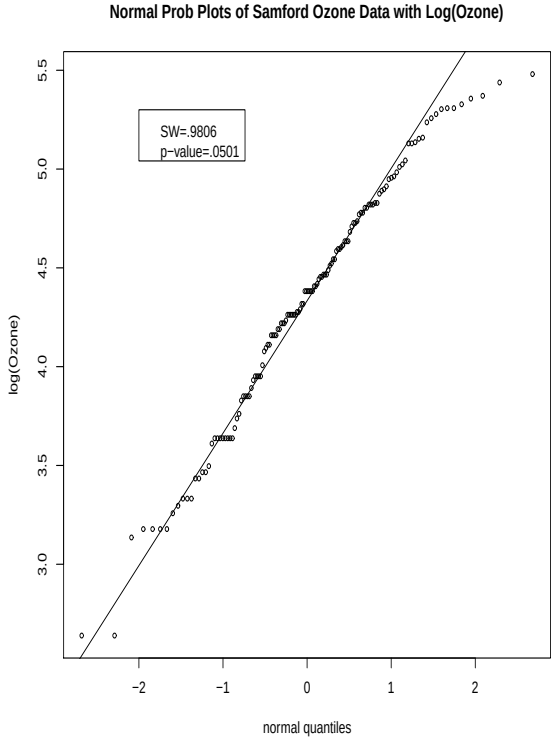
We will now present normal reference distribution plots for the ozone data and three transformations along with their values from the Shapiro-Wilk statistics:

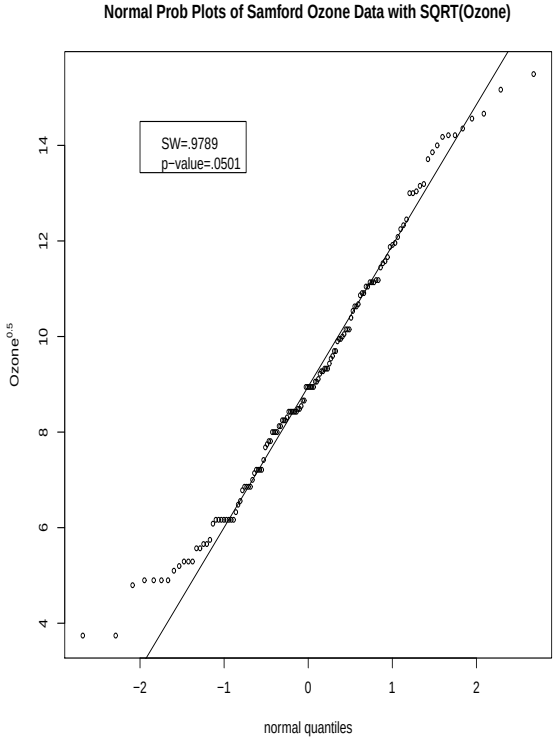
$$\theta = .23; \quad \theta = 0 \Rightarrow \quad \text{log transformation}; \quad \theta = .5 \Rightarrow \quad \text{square root transformation}$$

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

qqnorm(x,main="Normal Prob Plots of Samford Ozone Data",
                xlab="normal quantiles",ylab="ozone concentration",cex=.65)
qqline(x)
text(-2,200,"SW=.9288")
text(-2,190,"p-value=0")

y1= log(x)
y2= x^.23
y3= x^.5
s = shapiro.test(x)
s1 = shapiro.test(y1)
s2 = shapiro.test(y2)
s3 = shapiro.test(y3)



postscript("u:/meth1/psfiles/boxcox_ozone2.ps",height=8,horizontal=F)


qqnorm(y2,main="Normal Prob Plots of Samford Ozone Data with (Ozone)^.23",
                xlab="normal quantiles",ylab=expression(Ozone^.23),cex=.65)
          qqline(y2)
text(-2,3.5,"SW=.9872")
text(-2,3.4,"p-value=.2382")


postscript("u:/meth1/psfiles/boxcox_ozone1.ps",height=8,horizontal=F)


qqnorm(y1,main="Normal Prob Plots of Samford Ozone Data with Log(Ozone)",
                xlab="normal quantiles",ylab="Log(Ozone)",cex=.65)
          qqline(y1)
text(-2,5.0,"SW=.9806")
text(-2,4.85,"p-value=.0501")

postscript("u:/meth1/psfiles/boxcox_ozone3.ps",height=8,horizontal=F)



qqnorm(y3,main="Normal Prob Plots of Samford Ozone Data with SQRT(Ozone)",
                xlab="normal quantiles",ylab=expression(Ozone^.5),cex=.65)
          qqline(y3)
text(-2,14.5,"SW=.9789")
text(-2,13.5,"p-value=.0501")

#graphics.off()
```

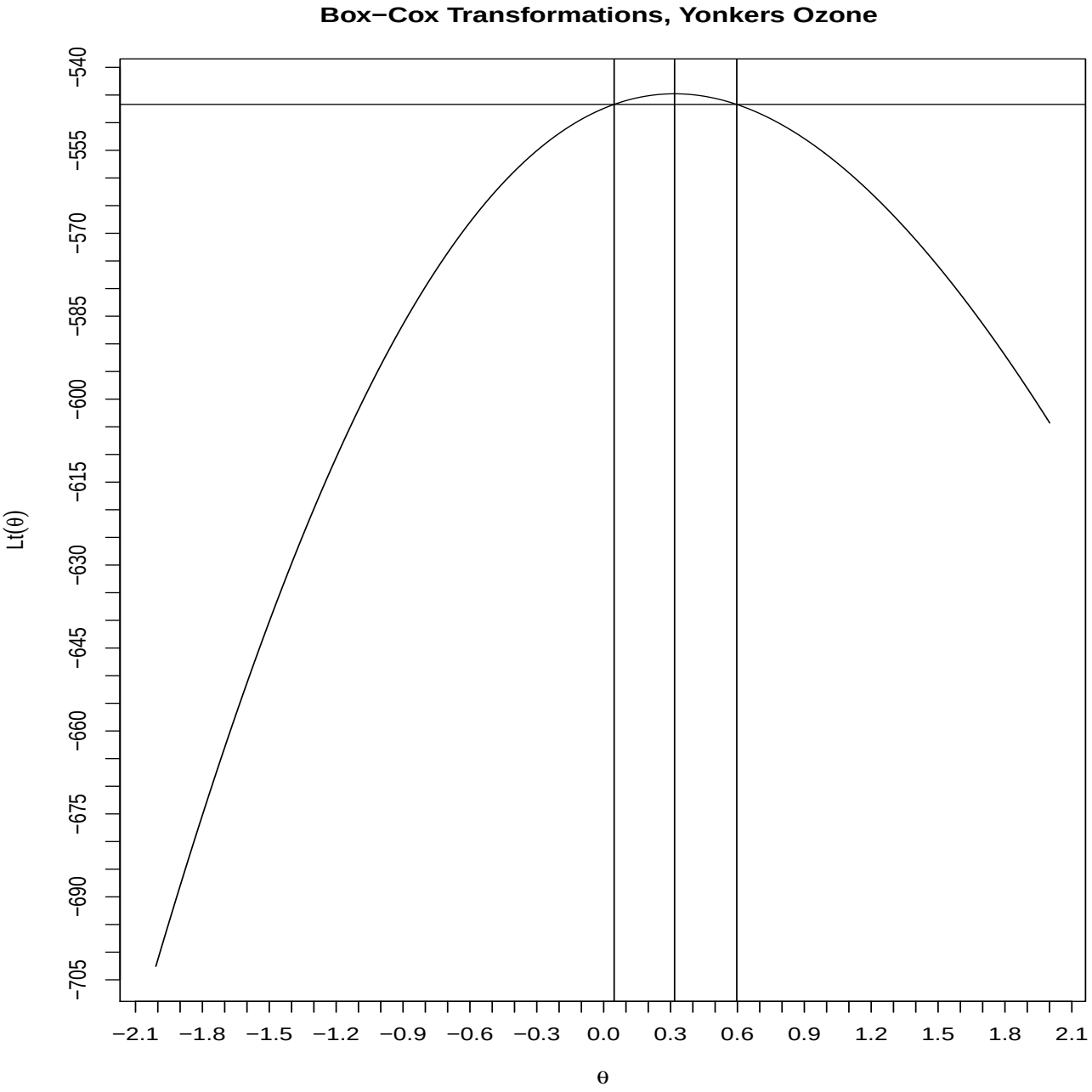

Normal Prob Plots of Samford Ozone Data
SW=.9288
p−value=0
ozone concentration
normal quantiles


Normal Prob Plots of Samford Ozone Data with Ozone^.23
SW=.9872
p−value=.2382
Ozone^0.23
normal quantiles

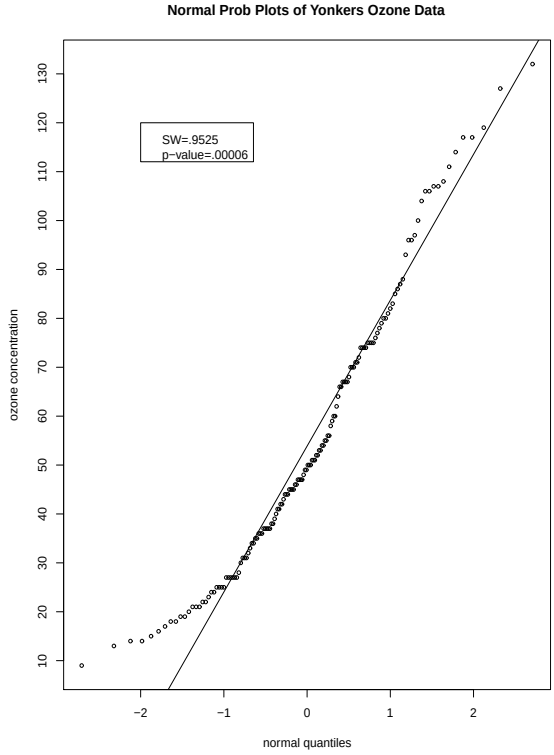

Normal Prob Plots of Samford Ozone Data with Log(Ozone)
SW=.9806
p−value=.0501
log(Ozone)
normal quantiles

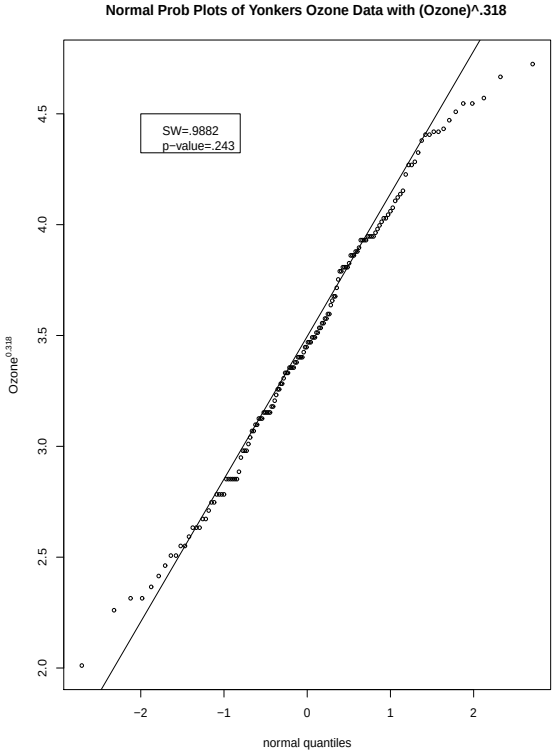

Normal Prob Plots of Samford Ozone Data with SQRT(Ozone)
SW=.9789
p−value=.0501
Ozone^0.5
normal quantiles

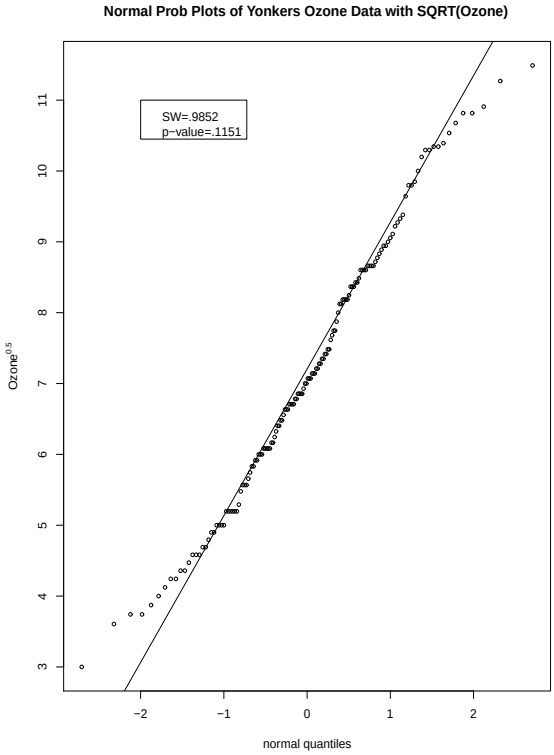
The plot of the likelihood function for the Yonkers ozone readings is given here with lines indicating a 95% confidence interval on values of θ which maximize the likelihood function. We have that $\hat{\theta} = .32$

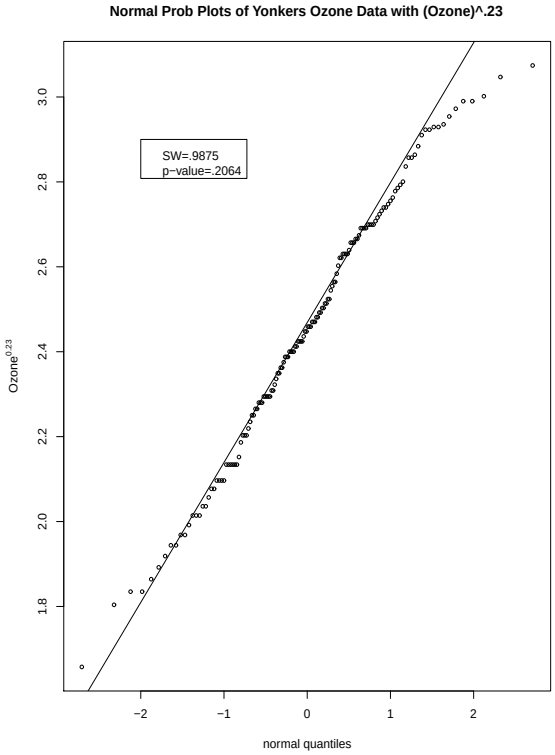


from the R output we obtain a 95% C.I. for θ (thLci,thUci)= $(.047, .60)$.


Normal Prob Plots of Yonkers Ozone Data
SW=.9525
p−value=.00006
ozone concentration
normal quantiles


Normal Prob Plots of Yonkers Ozone Data with (Ozone)^.318
SW=.9882
p−value=.243
Ozone0.318
normal quantiles


Normal Prob Plots of Yonkers Ozone Data with SQRT(Ozone)
SW=.9852
p−value=.1151
Ozone0.5
normal quantiles


Normal Prob Plots of Yonkers Ozone Data with (Ozone)^.23
SW=.9875
p−value=.2064
Ozone0.23
normal quantiles

From the following fits we have that the MLE for Stamford is $\hat{\theta} = .23$ and the MLE for Yonkers is $\hat{\theta} = .32$. The problem with any statistical analysis is if we use a different transformation for each of the cities then it will be impossible to compare the means, medians, or any other parameter associated with the distributions because the measurements of the transformed data will be in different scales. In this example, note that if we used $\hat{\theta} = .23$ for both cities we obtain reasonably good fits to a normal distribution. Thus, we could transform both cities' data using $\hat{\theta} = .23$ and then make valid comparisons with respect to the transformed data.

Summarizing GOF Measures

1. Discrete Distributions with all parameters specified and k cells:

   Use Q - Chi-square GOF test with $df = k - 1$

2. Discrete Distributions with some of the parameters unspecified and k cells:

   Use Q - Chi-square GOF test with $df = k - 1 - m$, where m is number of estimated parameters in the model

3. Continuous Distribution with all parameters specified

   Use the K-S or A-D GOF tests

4. Continuous Distribution with some of the parameters unspecified

   Case 1: For the Normal Distribution use Shapiro-Wilk test

   Case 2: For the Exponential, Weibull, etc., Distributions use modifications given by Stephens and D'Agnostino

   Case 3: For those cases not covered by Stephens and D'Agnostino, estimated parameters using MLEs and then use K-S or A-D GOF measures assuming the distribution is completely specified. However, if n is not large, these procedures may be inaccurate due to the inaccuracy in estimating the unknown parameters.

5. For Censored data:

   Use the modified A-D procedure.

   Alternatively, use all the data, both censored and uncensored, to estimate the unknown parameters using MLE with the likelihood function modified for censored data as was done in Handout 7. Compute the A-D statistic using just the uncensored data and compute the p-value using the tables for the uncensored case. In these tables, the sample size is the number of uncensored data values.

6. Use graphical procedures with the appropriate modifications to accommodate the censoring.