START 10/4/21 Monday

HANDOUT #8: GRAPHICAL SUMMARIES OF DATA AND COMPARISON GRAPHS

TYPES OF GRAPHS

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

Supplemental Reading:

- Chapter 5, Sections 6.1, 14.6.2, and 15.1 in Tamhane/Dunlop book

Reference Distribution Plots:

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

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

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

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

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

Case 1: F_o Completely Specified

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

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

1. plot $\left(Q_o(u_i), Y_{(i)}\right)$, where $Y_{(i)} = \text{sort } (Y_1, Y_2, \ldots, Y_n)$

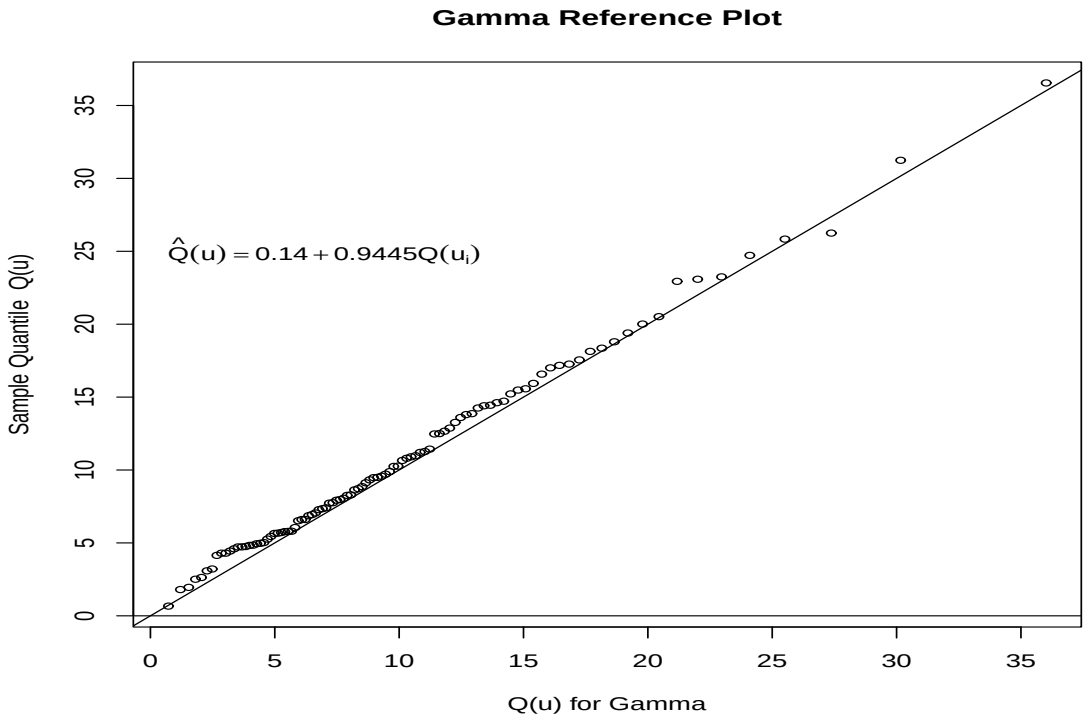
   Note: $Y_{(i)} = \widehat{Q}(u_i)$, where $\widehat{Q}(u_i)$ is the sample quantile evaluated at

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

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

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

STOP MON 10/4/21



Case 2: F_o **Not Completely Specified**

Let Q_o be the quantile function associated with F_o and $\widehat{Q}$ be the sample quantile. Suppose F_o is stated to be from a specific parametric family but the parameters are unspecified. For example, F_o is from a Weibull family but β and λ are not specified. We could just estimate the parameters using the observed data and use these values in our $Q-Q$ plot. However, this could result in varying degrees of inaccuracy in specifying Q_o and could result in a misstatement of the match between the Q_o and the sample quantile $\widehat{Q}$. This approach should be used only if one of the following approaches is not feasible and it should be emphasized in the documentation that you are using estimated parameters in the procedure, not the true parameters. Furthermore, n should be large, $n > 100$, and the estimators should be MLEs.

Case 3: F_o **Member of Location-Scale Family**

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

For example,

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

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

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

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

START WED 10/6/21

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

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

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

Then

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

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

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

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

Therefore,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rearranging terms we obtain:

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

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

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

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

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

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

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

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

Case 1: h is an increasing function.

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

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

Case 2: h is a decreasing function.

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

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

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

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

Example: Weibull Distribution:

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

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

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

Recall, that for Y having a Weibull distribution:

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

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

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

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

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

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

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

Assessing the Fit of Line in a Reference Distribution Plot

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

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

obtain the least squares line

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

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

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

is often placed on the plot to assess the fit.

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

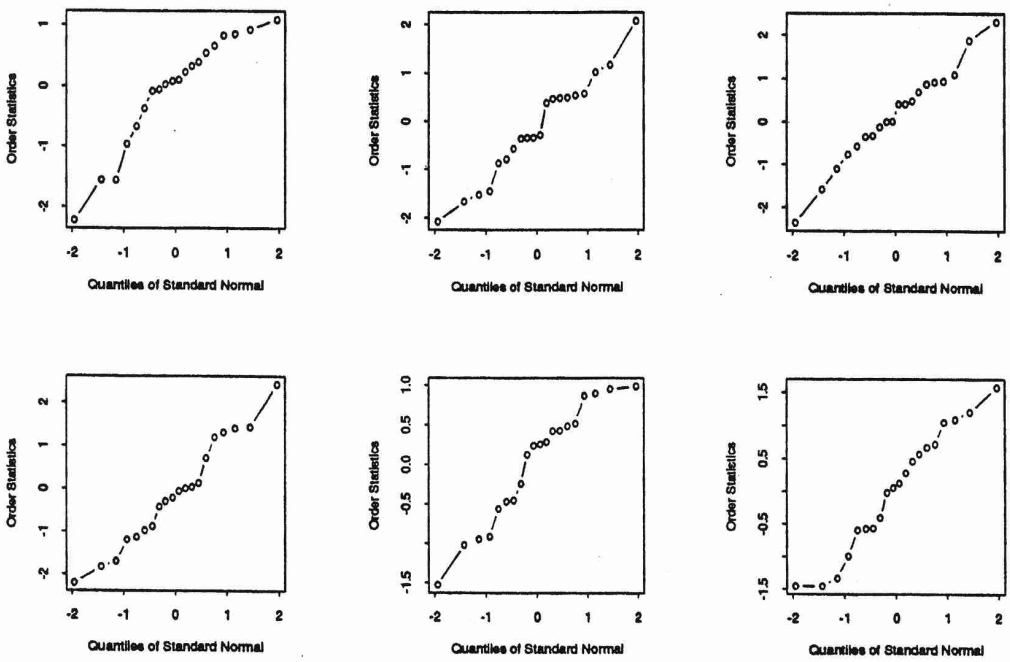
This line measures the fit of the model to the observed data in the middle portion of the values. Depending on the size of n, the points may or not fit very close to a straight line as will be demonstrated with the series of plots given on the next four pages. A short description of the graphs will now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

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

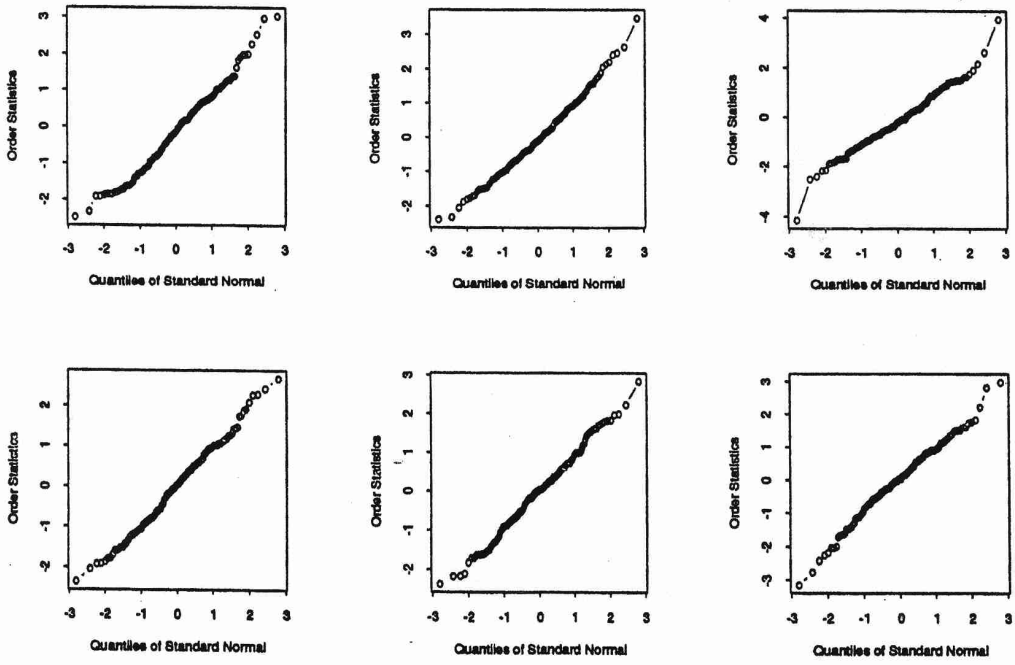
1. Normal Distribution - all 12 plots should be relatively close to line
2. Chi-squared Distribution(df=4) - Right skewed distribution with short tail on the left end of distribution - 12 plots should have plotted points curving away from line with most points above the line on the left and curving away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

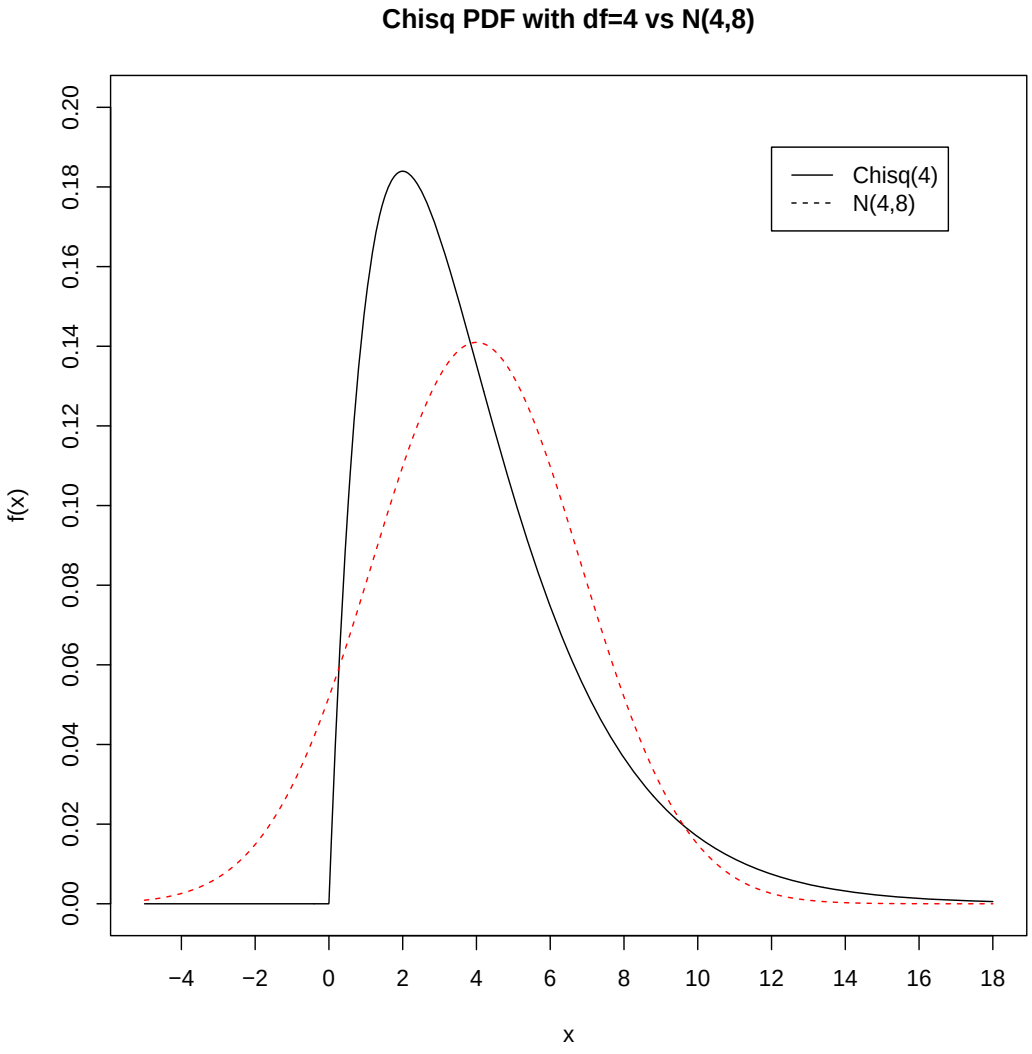
Normal Probability Plots, Normal Data, n=20/200

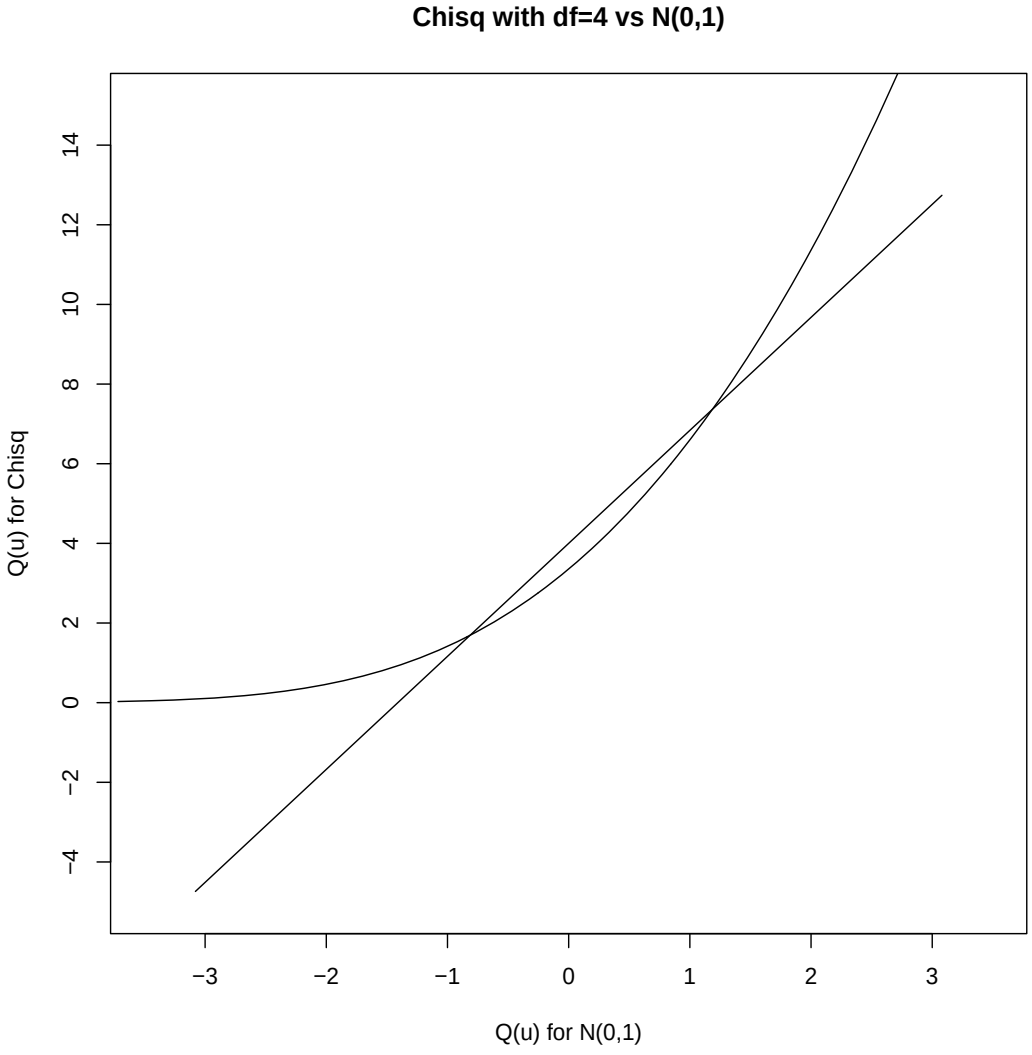
$N = 20$



$N = 200$

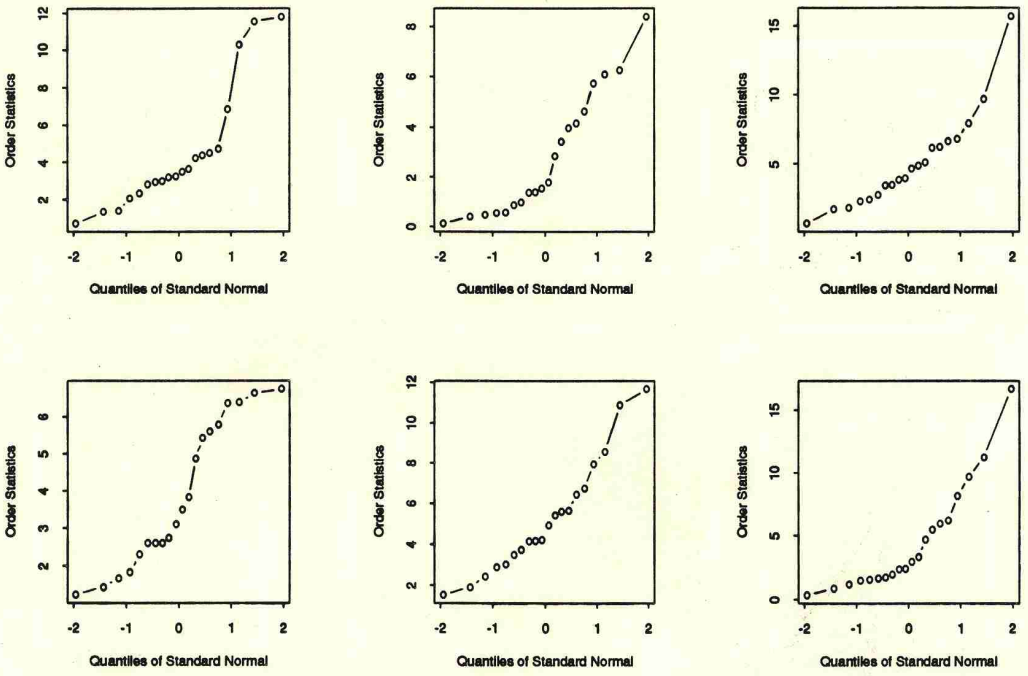



Chisq PDF with df=4 vs N(4,8)
f(x)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
−4
−2
0
2
4
6
8
10
12
14
16
18
x
Chisq(4)
N(4,8)

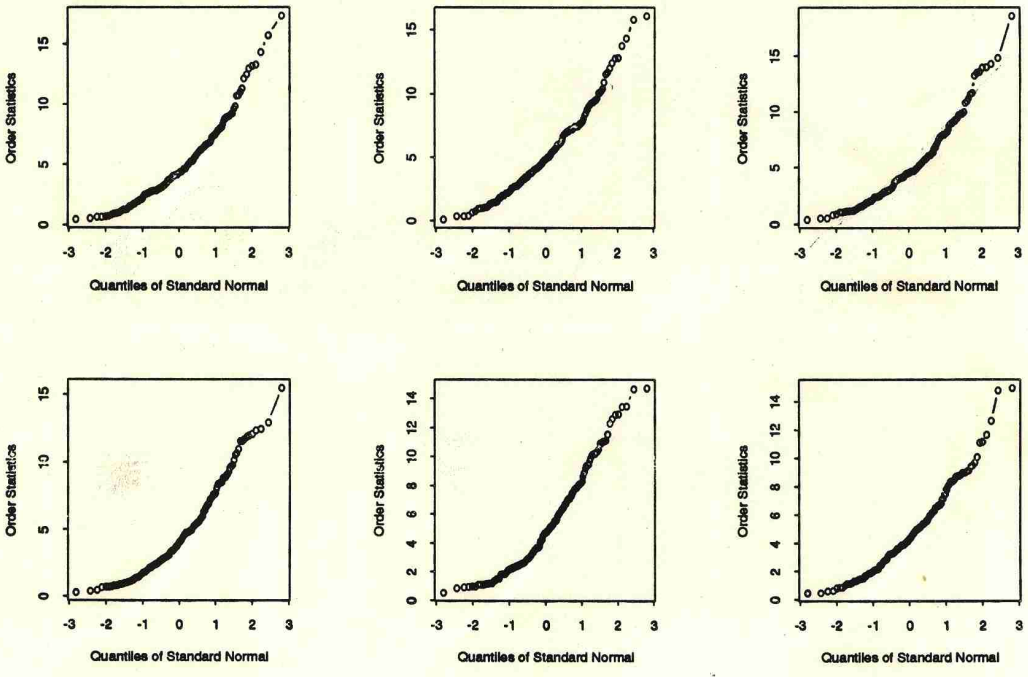

Chisq with df=4 vs N(0,1)
Q(u) for Chisq
Q(u) for N(0,1)
−4
−2
0
2
4
6
8
10
12
14
−3
−2
−1
0
1
2
3

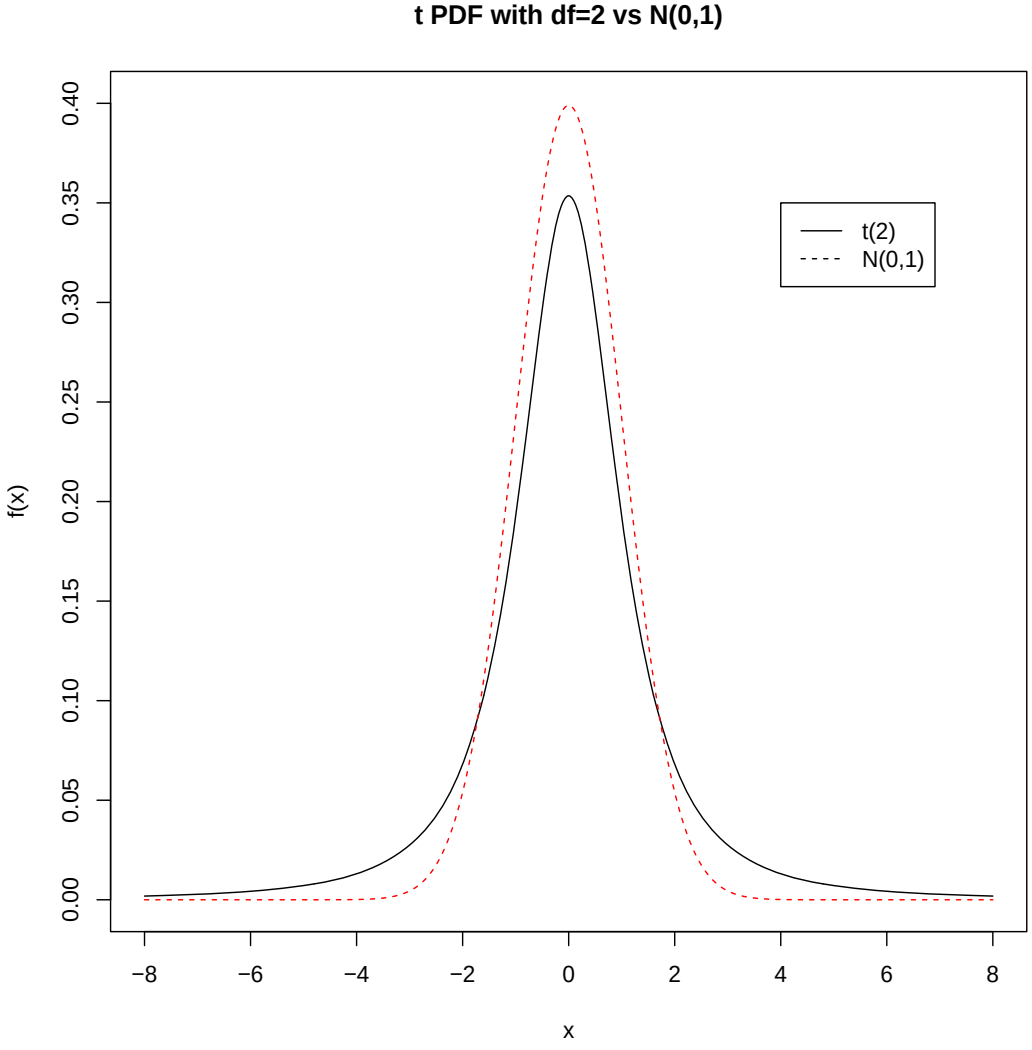
Normal Probability Plots, Chi-Squared (4 df) Data, n=20/200

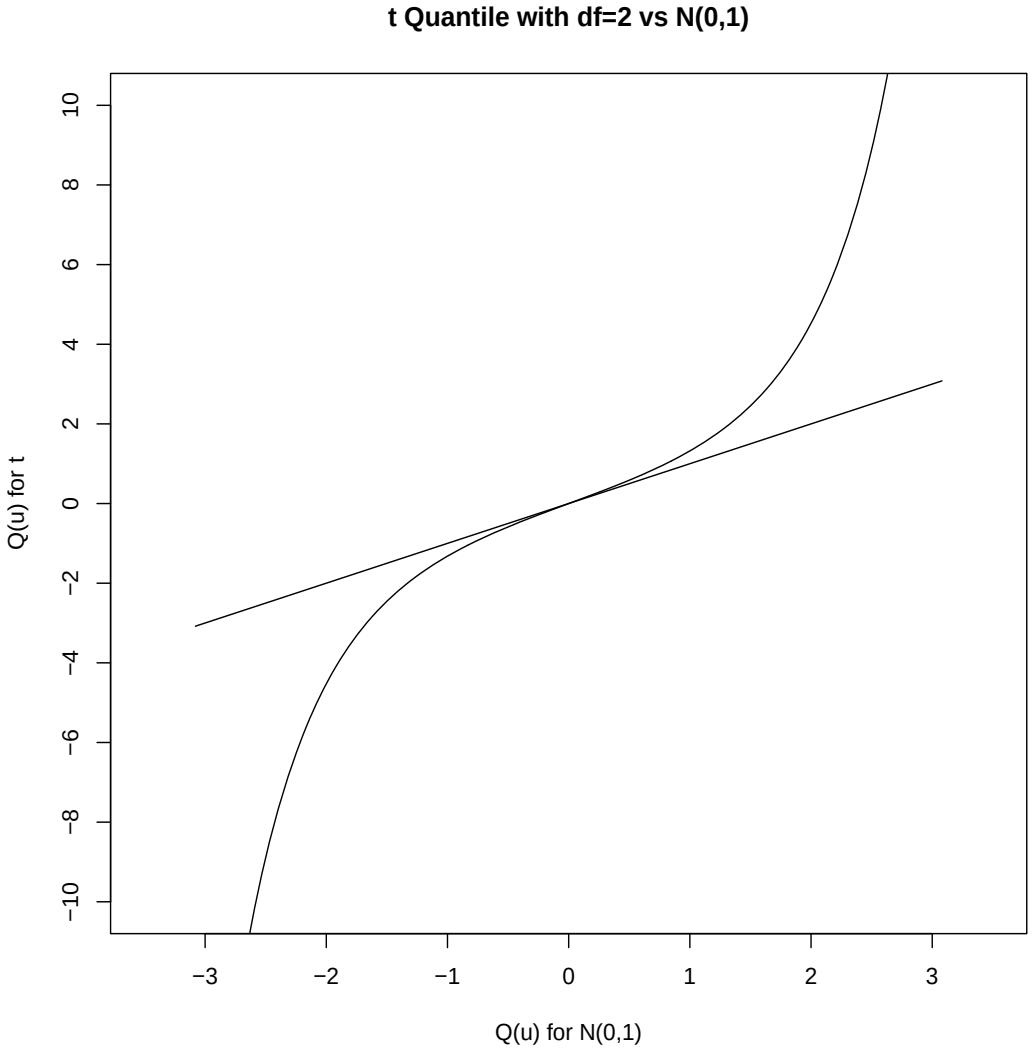
$N = 20$



$N = 200$

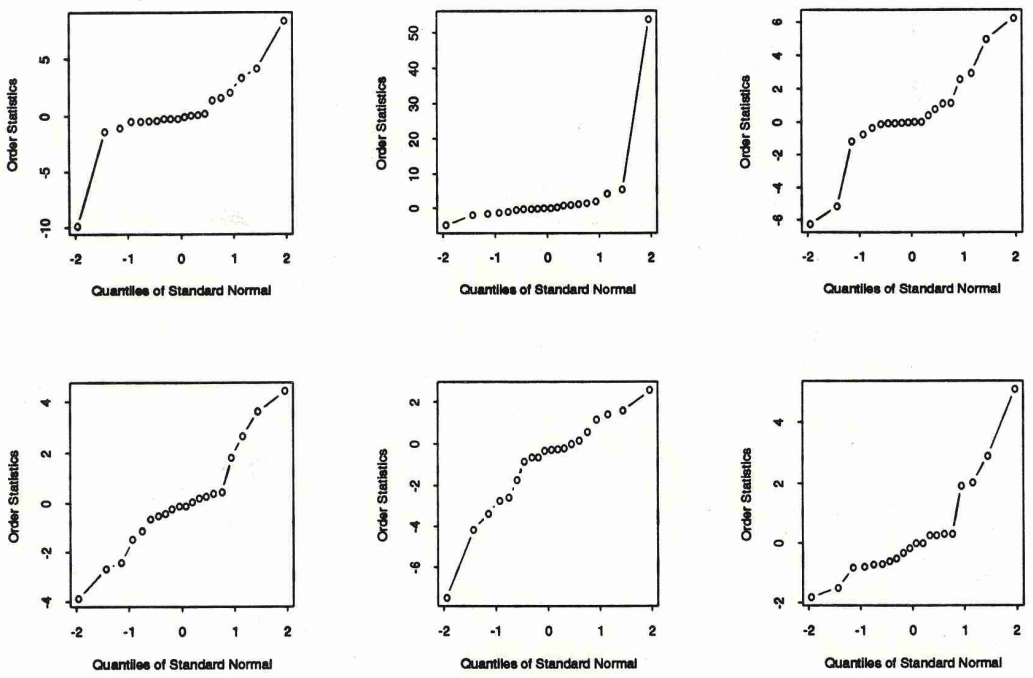



t PDF with df=2 vs N(0,1)
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−8
−6
−4
−2
0
2
4
6
8
x
t(2)
N(0,1)

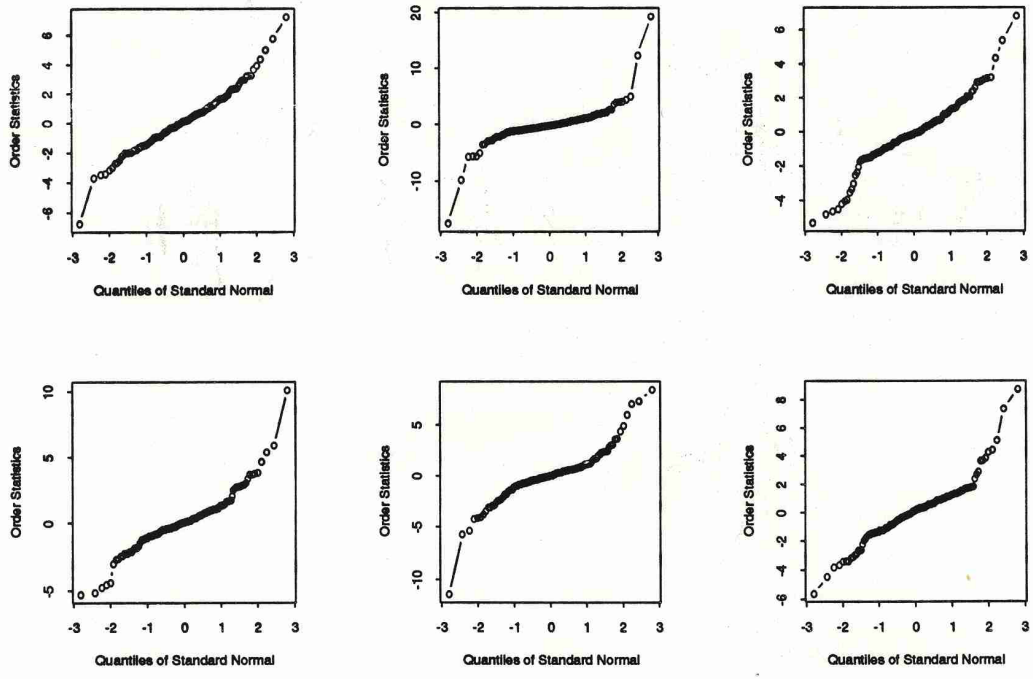

t Quantile with df=2 vs N(0,1)
Q(u) for t
10
8
6
4
2
0
−2
−4
−6
−8
−10
−3
−2
−1
0
1
2
3
Q(u) for N(0,1)

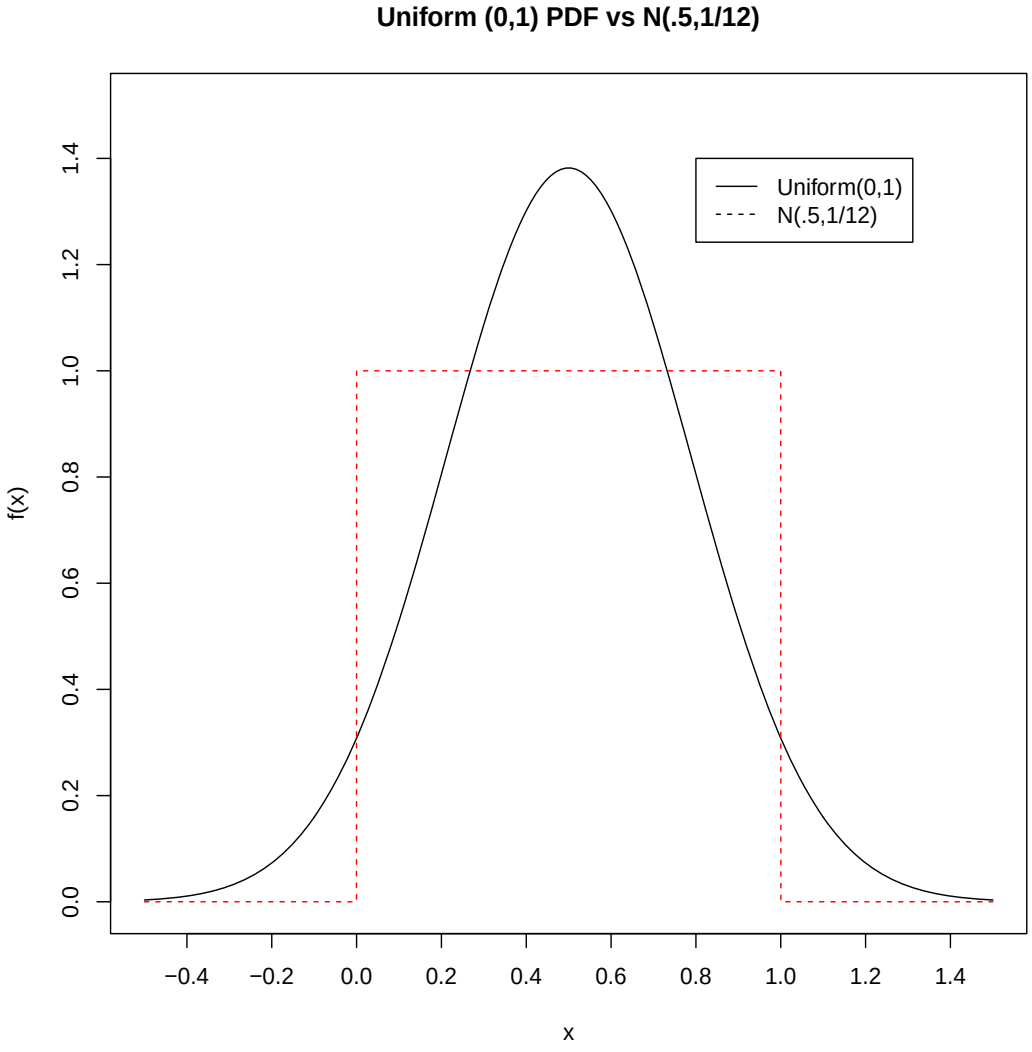
Normal Probability Plots, Student's t (2 df) Data, n=20/200

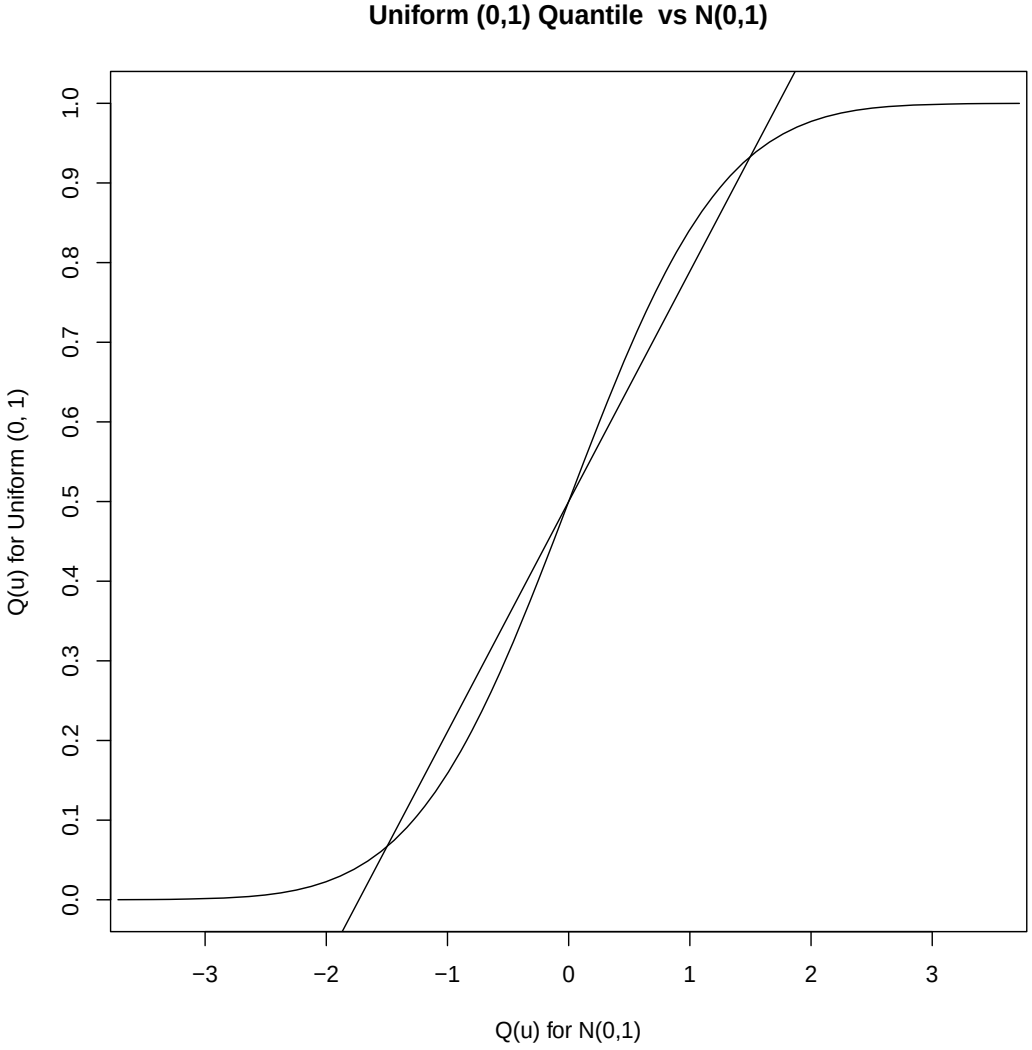
$N = 20$



$N = 200$

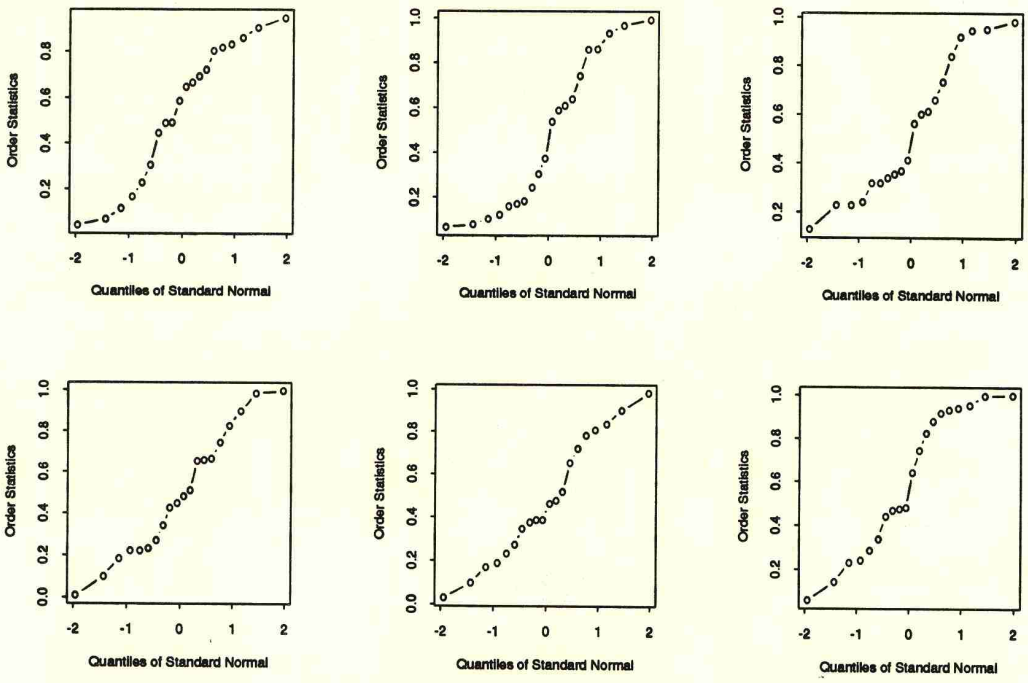



Uniform (0,1) PDF vs N(.5,1/12)
Uniform(0,1)
N(.5,1/12)
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

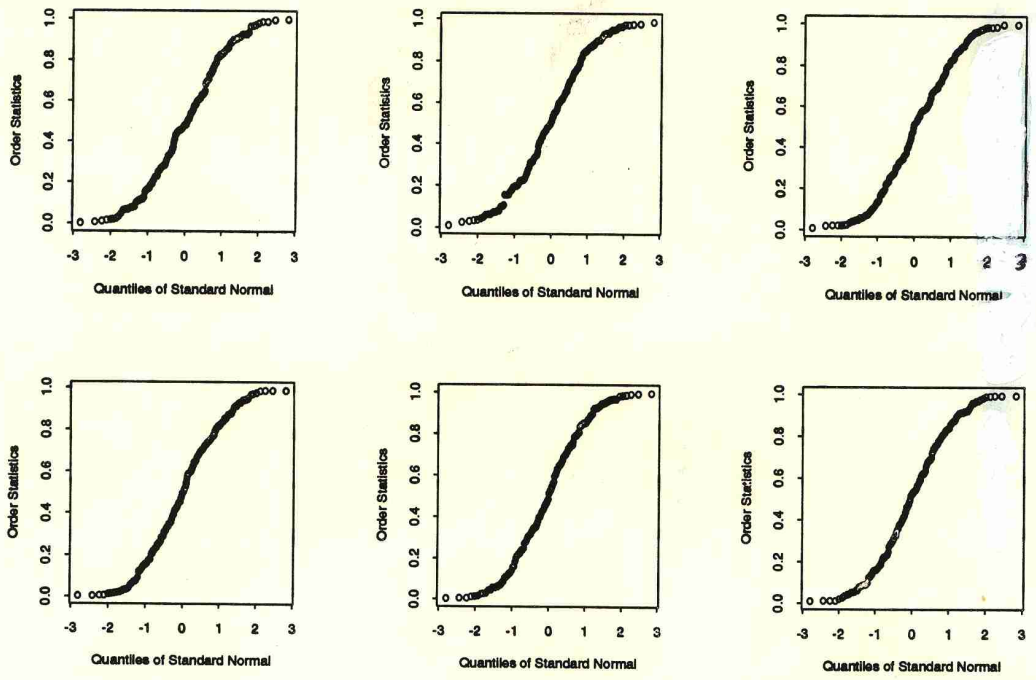

Uniform (0,1) Quantile vs N(0,1)
Q(u) for Uniform (0, 1)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
−3
−2
−1
0
1
2
3
Q(u) for N(0,1)

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

$N = 20$

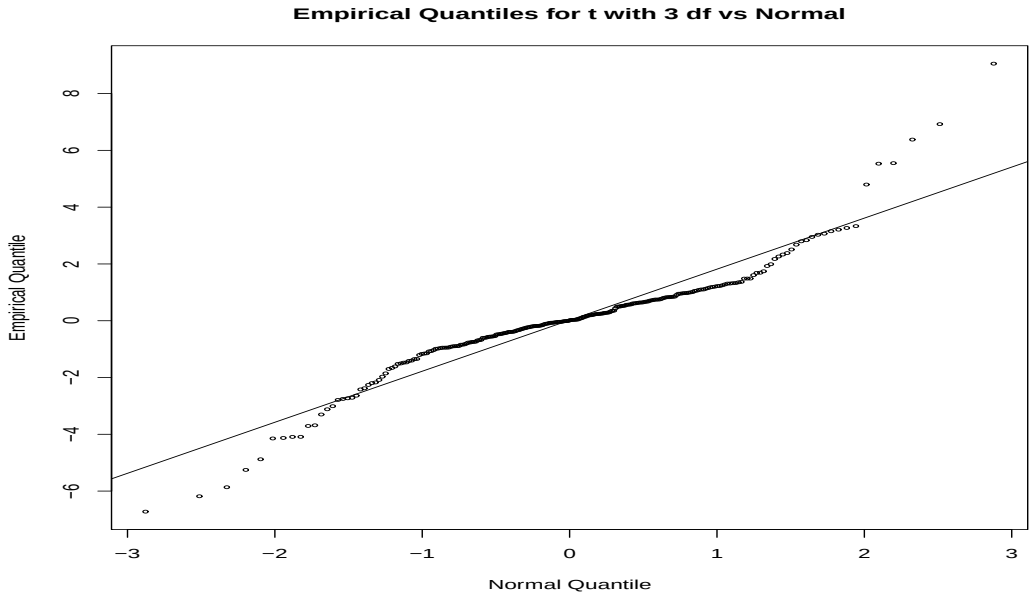


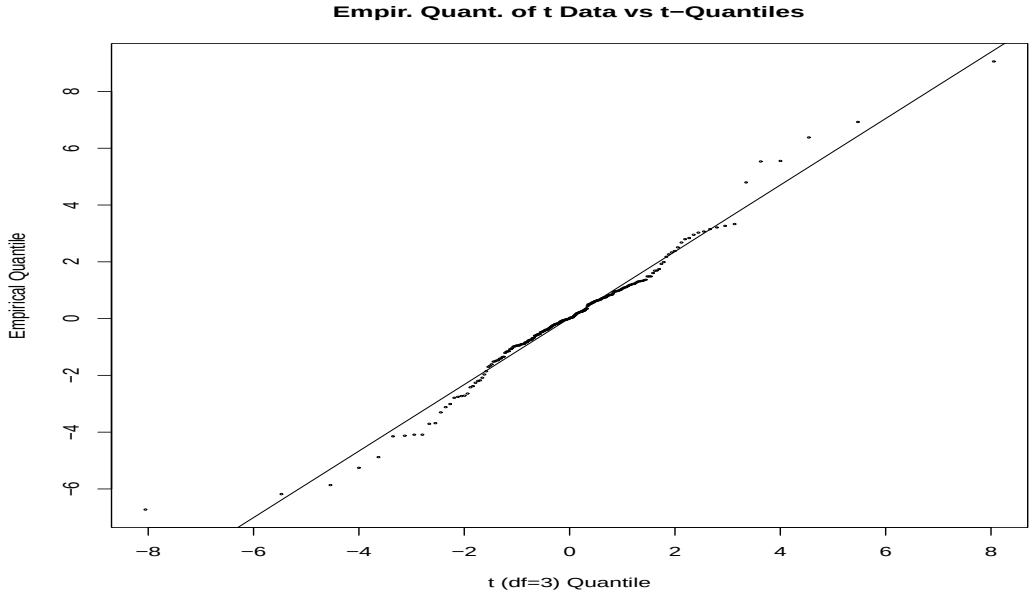
$N = 200$

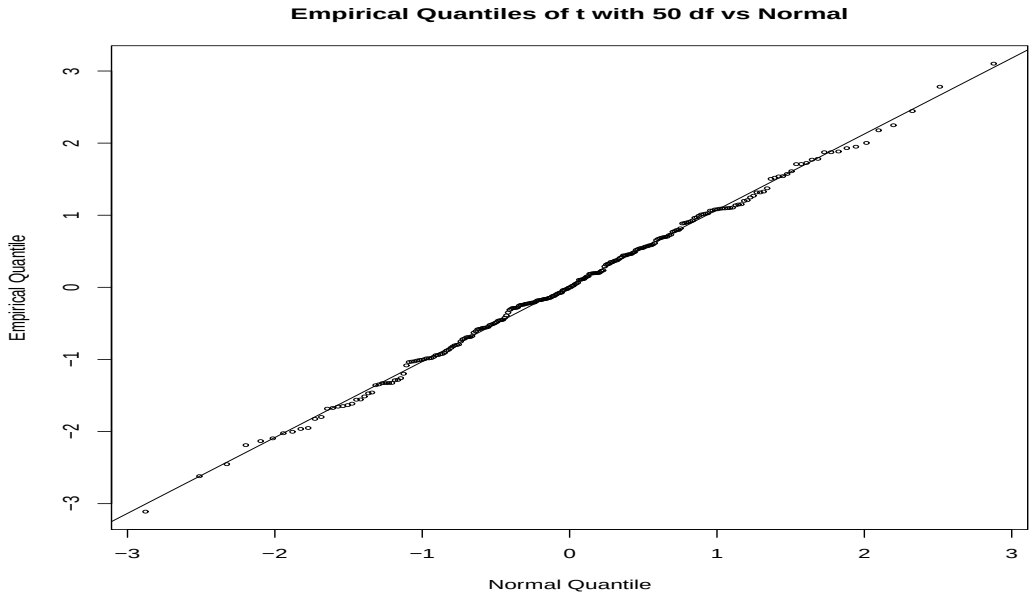


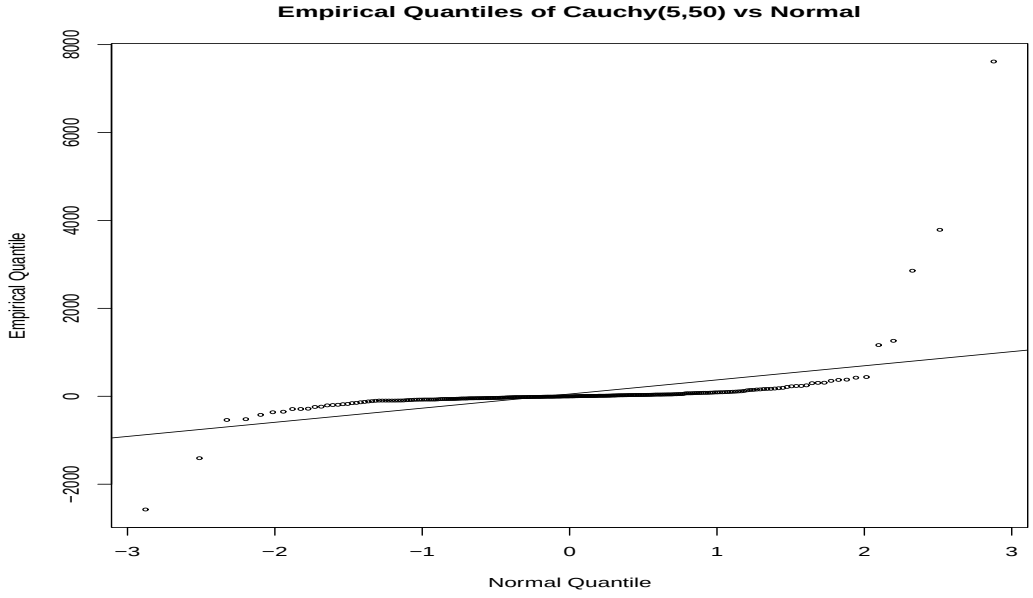
```
#The following R program generates data from various specified distributions
#and the plots the generated data various various reference distributions.
# This program is refdist.R in R Files on eCampus
#-------------------------------------------------------------------------

#generates 250 observations from t with df=3 and 50, Cauchy, Gamma with
#shape=2 and scale=1/3, weibull with scale=16 and shape=2, uniform on (-2,5):
#Note in the gamma function, gamma(a,b): a=shape, b=1/scale


n=250

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

t3  = sort(rt(n,3))
t50 = sort(rt(n,50))
cau = sort(rcauchy(n,5,50))
wei = sort(rweibull(n,2,16))
gam = sort(rgamma(n,2,3))
uni = sort(runif(n,-2,5))

#The following commands will generate various normal probability plots:


# Empirical Quantile of t with df=3  vs Normal Quantiles:

plot(z,t3,xlab="Normal Quantile",ylab="Empirical Quantile",
       lab=c(7,8,7),main="Empirical Quantiles for t with 3 df vs Normal",cex=.5)
abline(lm(t3~z))

# Empirical t with df= 3-Quantile vs t with df=3 Quantiles:


t=sort(qt(u,3))
plot(t,t3, xlab="t (df=3) Quantile",
     ylab="Empirical Quantile",lab=c(6,9,7), main=
"Empir. Quant. of t Data vs t-Quantiles",cex=.25)
abline(lm(t3~t))




# empirical t with df=50-Quantile vs normal quantiles:

plot(z,t50,xlab="Normal Quantile",ylab="Empirical Quantile",
       lab=c(7,8,7),main="Empirical Quantiles of t with 50 df vs Normal",cex=.5)
abline(lm(t50~z))
```

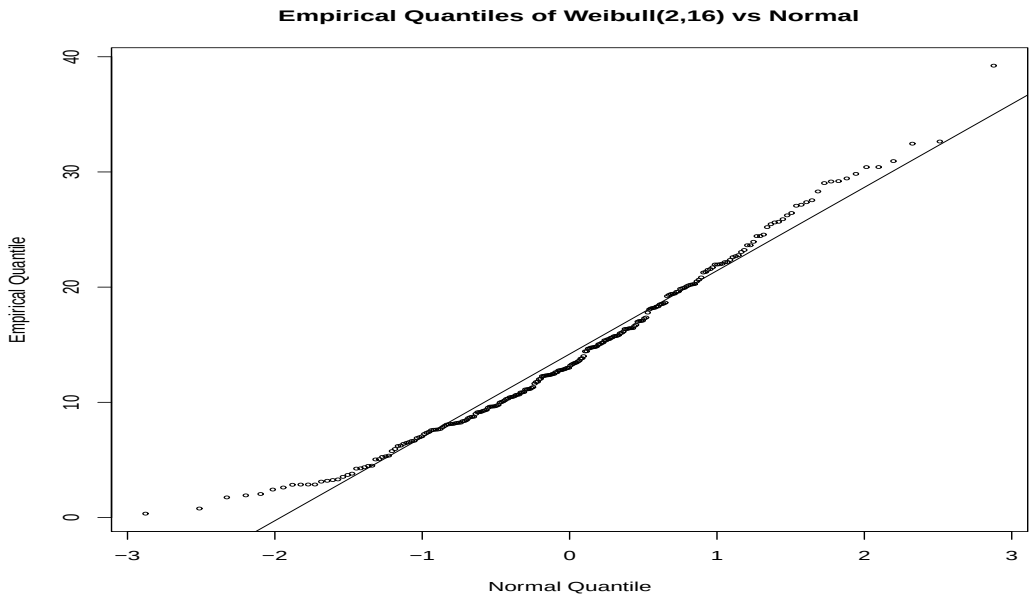
```
# empirical Cauchy-Quantile vs Normal Quantiles:

plot(z,cau,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
main="Empirical Quantiles of Cauchy(5,50) vs Normal",cex=.5)
abline(lm(cau~z))
graphics.off()
par(mfrow=c(2,1))



# empirical Weibull-Quantile vs Normal Quantiles:

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

# empirical Weibull-Quantile vs Weibull-Quantiles:

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


par(mfrow=c(2,1))



# empirical Gamma-Quantile vs Normal Quantiles:

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

# empirical Uniform-Quantile vs Normal Quantiles:

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

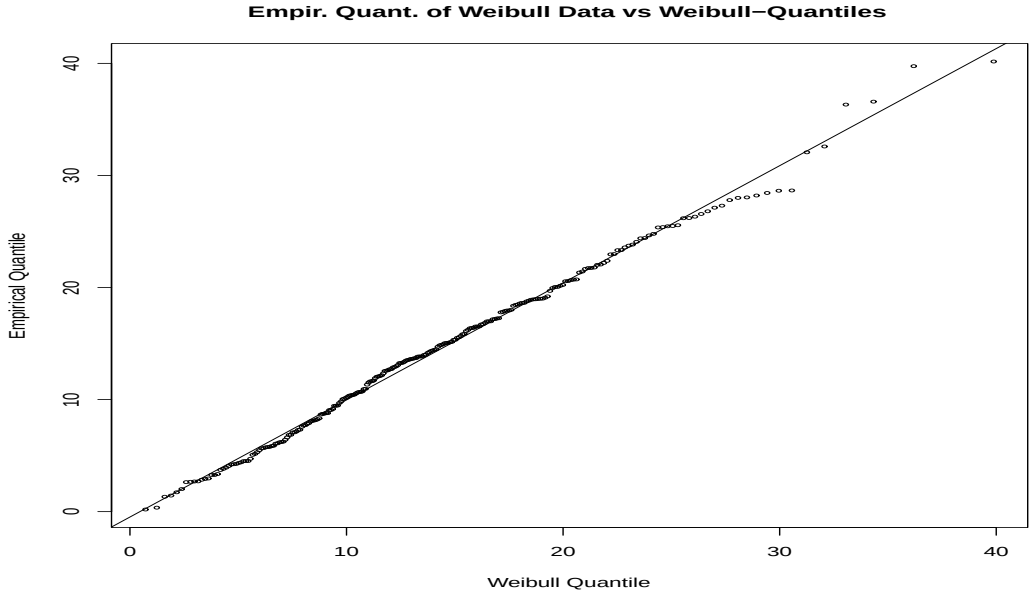

Empirical Quantiles for t with 3 df vs Normal
Empirical Quantile
Normal Quantile
−3
−2
−1
0
1
2
3
−6
−4
−2
0
2
4
6
8

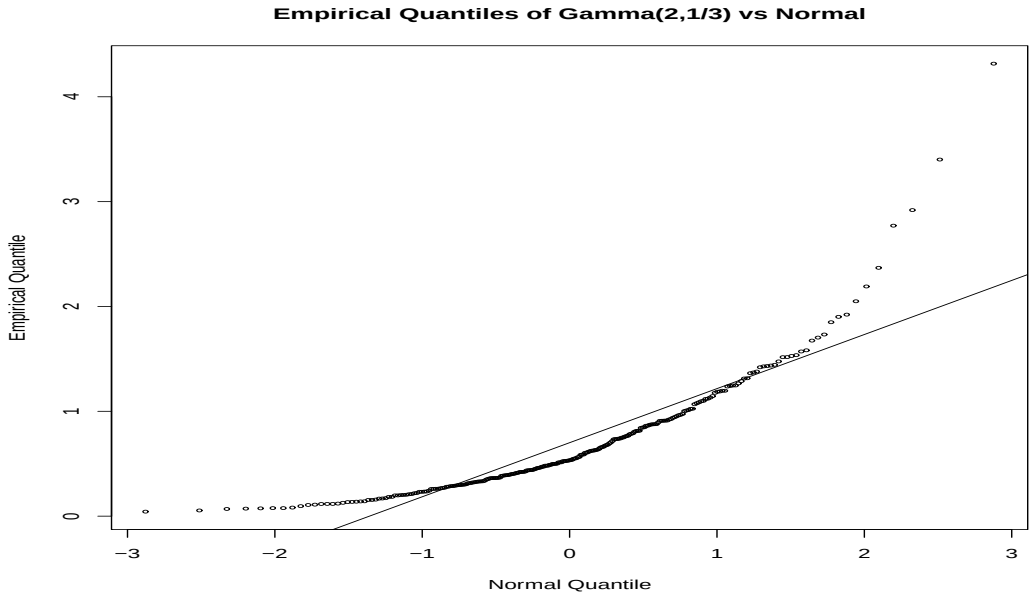

Empir. Quant. of t Data vs t−Quantiles
Empirical Quantile
t (df=3) Quantile
−8
−6
−4
−2
0
2
4
6
8

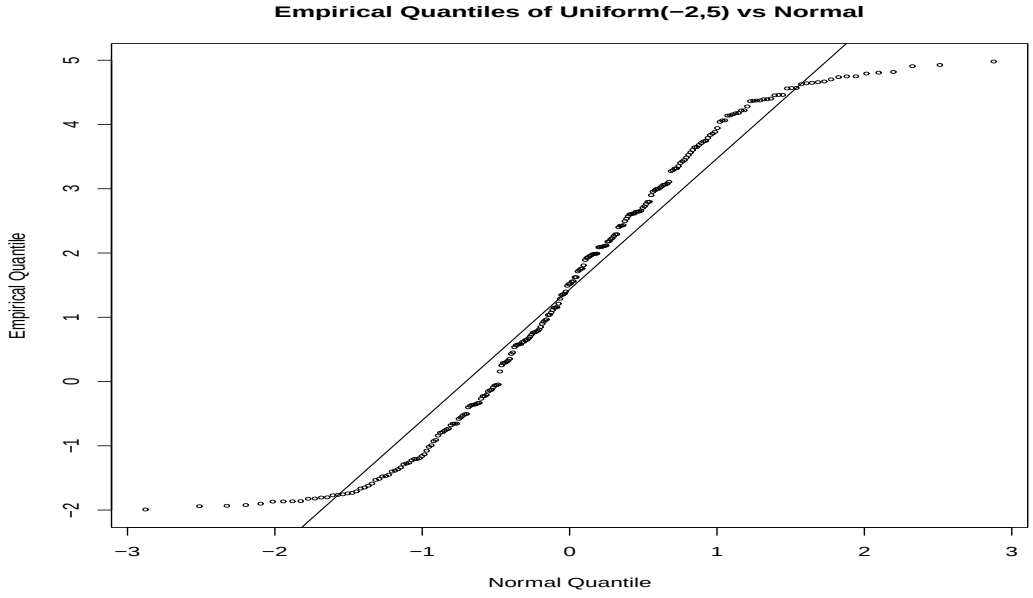

Empirical Quantiles of t with 50 df vs Normal
Empirical Quantile
Normal Quantile
−3
−2
−1
0
1
2
3


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


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


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


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


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

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

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

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

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

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

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

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

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

That is, suppose we have

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

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

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

Plot the n points

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

For example, suppose we have data from two populations:
$(X_1, X_2, X_3, X_4, X_5) = 3,\ 6,\ 7,\ 10,\ 12$ and $(W_1, W_2, \ldots, W_8) = 2,\ 4,\ 6,\ 10,\ 14,\ 22,\ 26,\ 29$
We want to construct a Q-Q plot to compare the two distributions: $Q_X(u)$ and $Q_W(u)$.
Take n = min(5, 8) = 5.
For i = 1, 2, 3, 4, 5; $u_i = (i - .5)/5 = 1/10,\ 3/10,\ 5/10,\ 7/10,\ 9/10$;

$\widehat{Q}_X(u_i) = X_{(n_1 u_i + .5)} = X_{(5u_i + .5)} = X_{(i)} = (3,\ 6,\ 7,\ 10,\ 12)$

$\widehat{Q}_W(u_i) = W_{(n_2 u_i + .5)} = W_{(8u_i + .5)} = (W_{(1.3)}, W_{(2.9)}, W_{(4.5)}, W_{(6.1)}, W_{(7.7)}) = (2.6,\ 5.8,\ 12,\ 22.4,\ 28.1)$

Thus, the 5 plotted points are

$(\widehat{Q}_X(u_i),\ \widehat{Q}_W(u_i)) = (3,\ 2.6),\ (6,\ 5.8),\ (7,\ 12),\ (10,\ 22.4),\ (12,\ 28.1)$

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

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

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

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

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

(Does this prove our point? No, we need the converse).

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

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

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

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

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

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

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

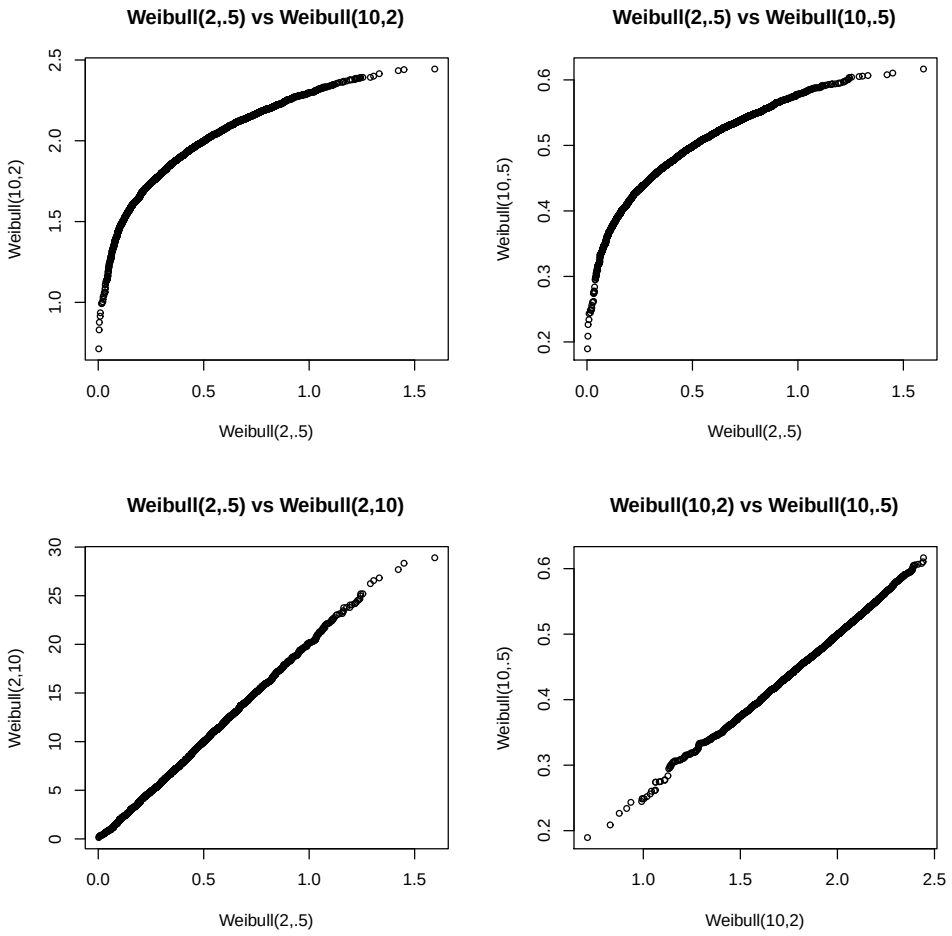
However, is the converse true? No, we are relating the distribution of the random variables X and Y, not their actual values.

Y = tensile strength of the allow used in a brand of golf clubs

X = pulse rate in beats per second of a patient on a new drug

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

See the Q-Q plot on the following page.


Weibull(2,.5) vs Weibull(10,2)
Weibull(10,2)
Weibull(2,.5)
Weibull(2,.5) vs Weibull(10,.5)
Weibull(10,.5)
Weibull(2,.5)
Weibull(2,.5) vs Weibull(2,10)
Weibull(2,10)
Weibull(2,.5)
Weibull(10,2) vs Weibull(10,.5)
Weibull(10,.5)
Weibull(10,2)

Plots Associated with Mixture Distributions

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

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

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

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

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

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

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

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

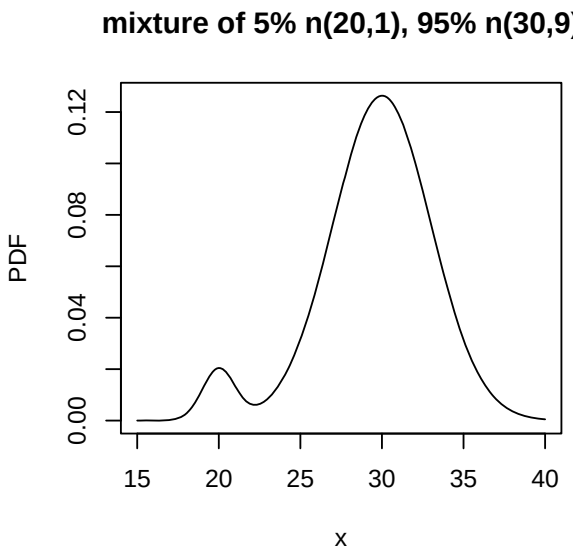
If we have a random sample from a population $Y_1, Y_2, \cdots, Y_n$, how can we tell whether or not the population has distinct subpopulations?

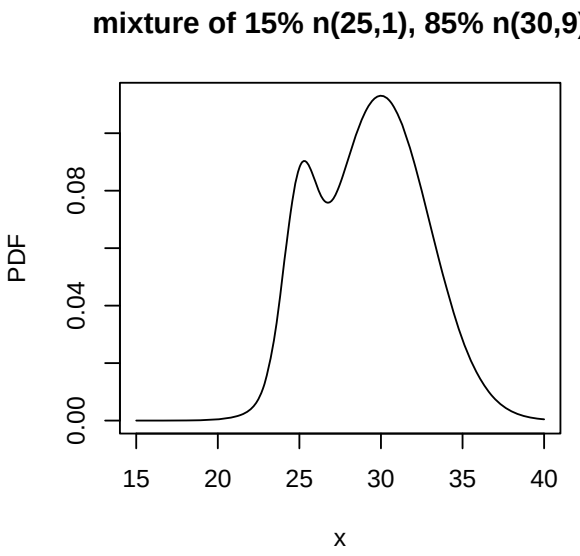
1. We can plot the kernel density estimator $\widehat{f_Y}(\cdot)$ and look for distinct modes in the plot. A problem with this method is that modes can appear and disappear as we vary the bandwidth. Thus it is very crucial to select an optimal bandwidth when using the kernel density estimator in such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

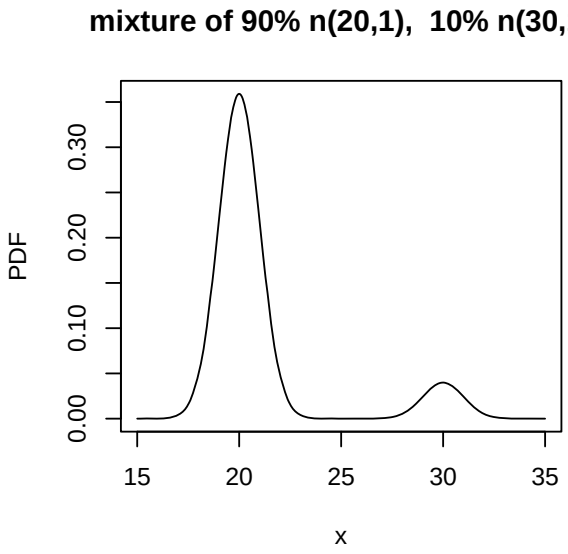
2. We can plot $\widehat{Q}(y)$ versus y and look for jumps in the plot. Once again, the question arises *when is a jump in the sample quantile plot a true jump in the population quantile.*

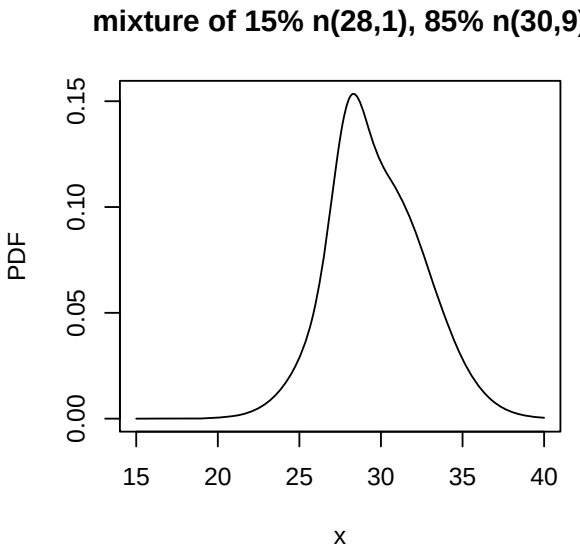
The plots on the next two pages will attempt to illustrate these problems in the simple case of mixing two normal distributions. We will consider four cases:

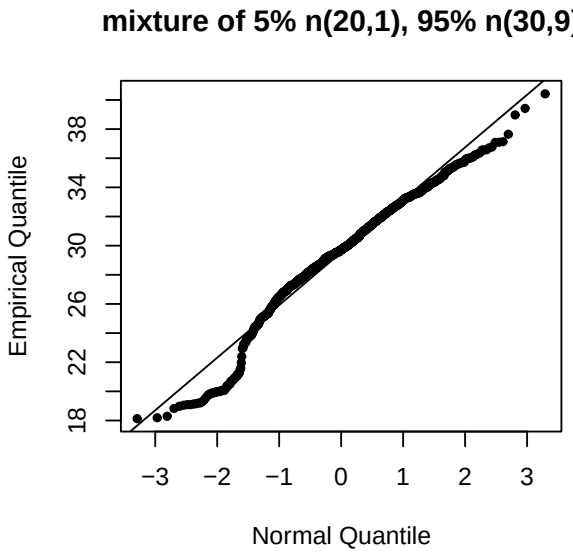
1. Case 1: 5% N(20,1) and 95% N(30,9)
2. Case 2: 15% N(25,1) and 85% N(30,9)
3. Case 3: 90% N(20,1) and 10% N(30,1)
4. Case 4: 15% N(28,1) and 85% N(30,9)

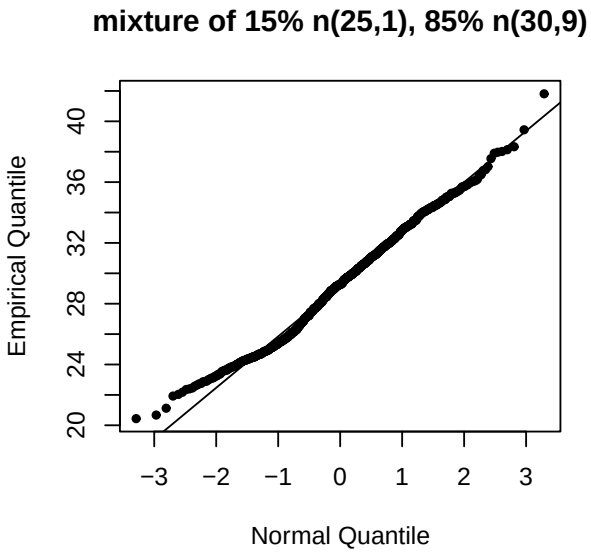

mixture of 5% n(20,1), 95% n(30,9)
PDF
0.00
0.04
0.08
0.12
15
20
25
30
35
40
x

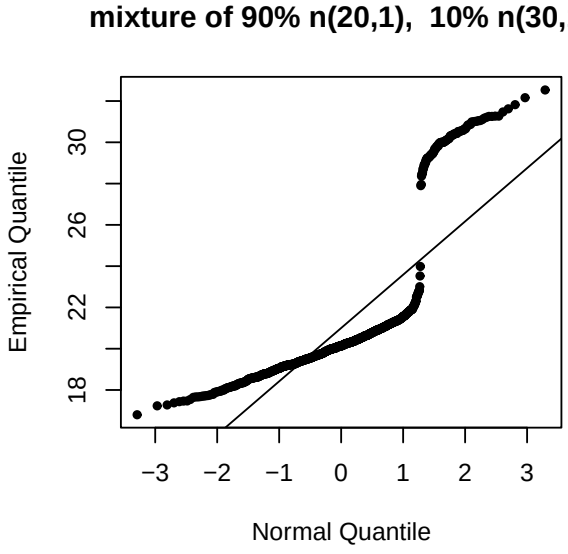

mixture of 15% n(25,1), 85% n(30,9)
PDF
0.00
0.04
0.08
15
20
25
30
35
40
x

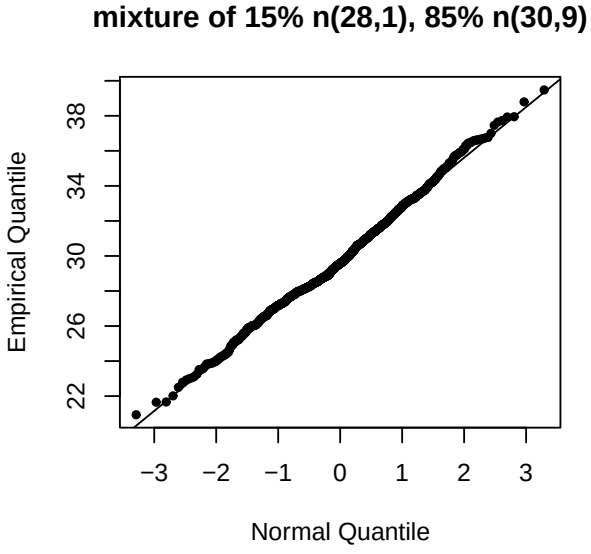

mixture of 90% n(20,1), 10% n(30,1)
PDF
0.00
0.10
0.20
0.30
15
20
25
30
35
x


mixture of 15% n(28,1), 85% n(30,9)
PDF
0.00
0.05
0.10
0.15
15
20
25
30
35
40
x


mixture of 5% n(20,1), 95% n(30,9)
Empirical Quantile
Normal Quantile
18
22
26
30
34
38
−3
−2
−1
0
1
2
3


mixture of 15% n(25,1), 85% n(30,9)
Empirical Quantile
Normal Quantile
20
24
28
32
36
40
−3
−2
−1
0
1
2
3


mixture of 90% n(20,1), 10% n(30,1)
Empirical Quantile
Normal Quantile
18
22
26
30
−3
−2
−1
0
1
2
3


mixture of 15% n(28,1), 85% n(30,9)
Empirical Quantile
Normal Quantile
22
26
30
34
38
−3
−2
−1
0
1
2
3

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

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

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

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

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

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

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

We will consider several cases:

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

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

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

The graphs on the next page will illustrate these ideas.

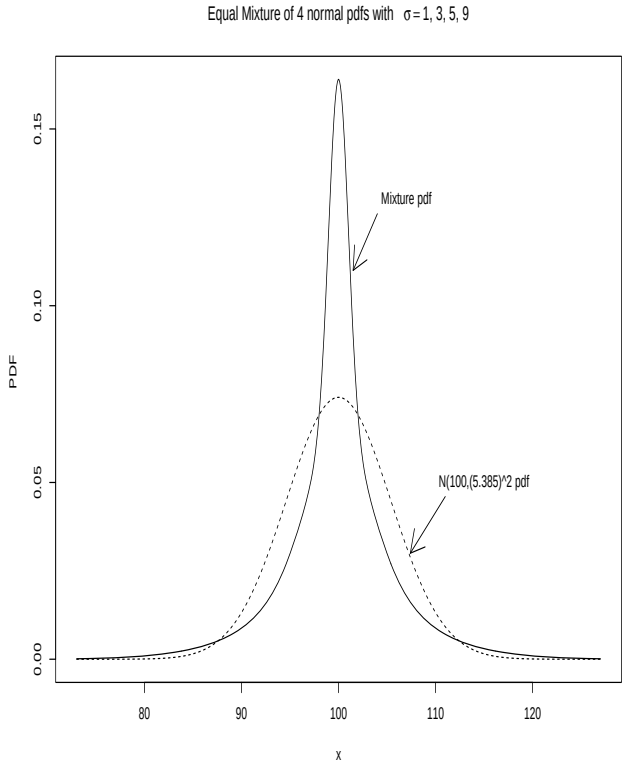
We will consider the situations depicted in the following table with $\mu_1 = \mu_2 = \mu_3 = \mu_4$. In this case, recall

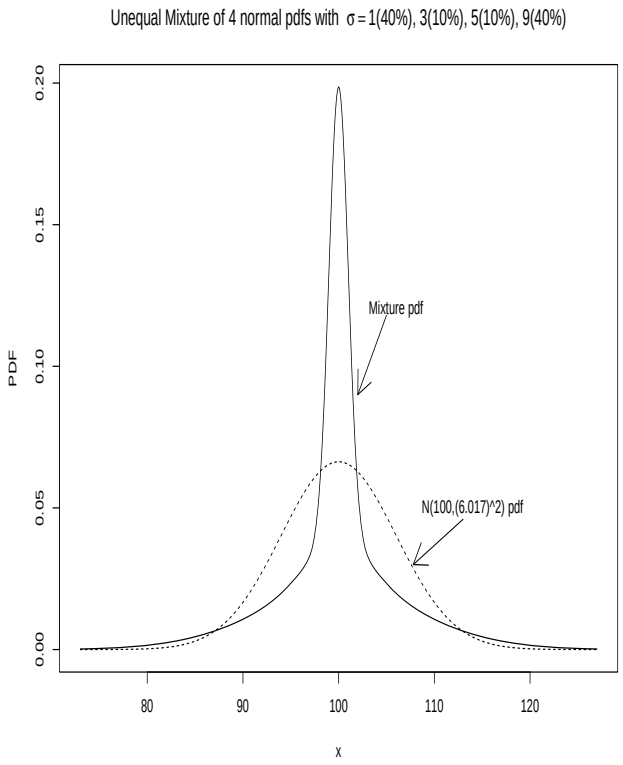
$$\sigma_Y^2 = \sum_{i=1}^{k} p_i \left(\sigma_i^2 + \mu_i^2\right) - \left(\sum_{i=1}^{k} p_i \mu_i\right)^2 = \sum_{i=1}^{k} p_i \sigma_i^2$$

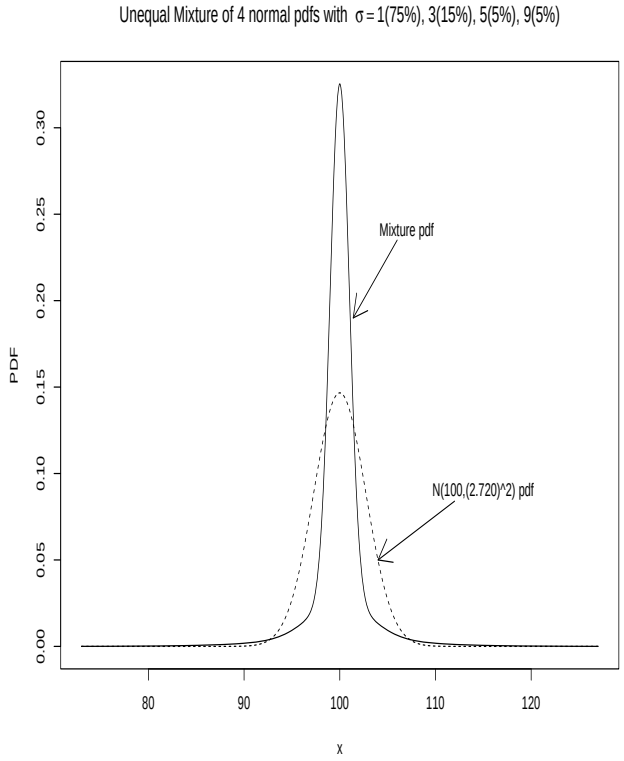
The following table contains four sets of values for the $p_i s$:

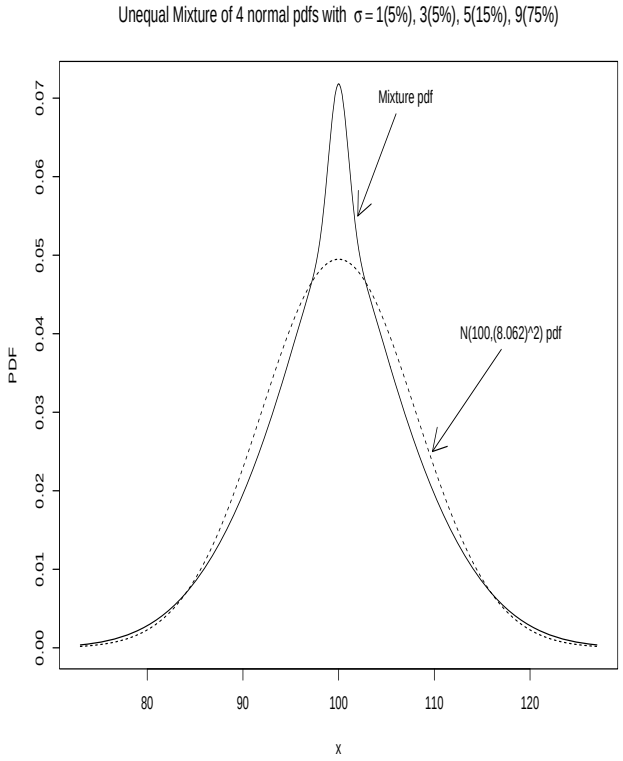
Four sets of values for $p_i' s$

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









END Wednesday 10/6/21

START Friday 10/8/21

Box Plots and Their Shapes

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

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

.

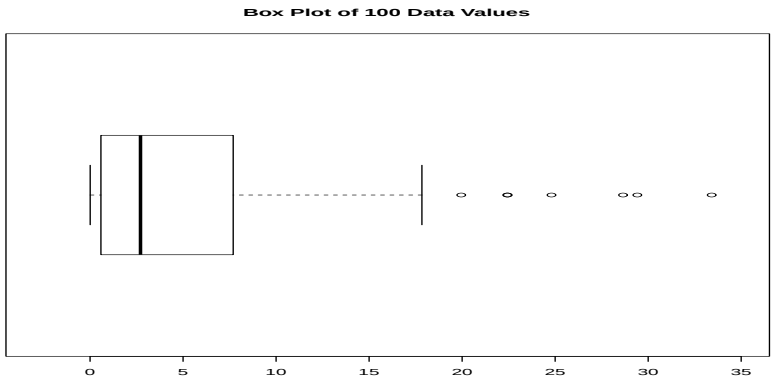
The box plot has the following features:

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

   Lower Fence: $\widehat{Q}_1 - 1.5(IQR)$ or $X_{(1)}$

   and

   Upper Fence: $\widehat{Q}_3 + 1.5(IQR)$ or $X_{(n)}$
6. $IQR = \widehat{Q}_3 - \widehat{Q}_1$ is referred to as the InterQuartile Range.
7. Two lines denoted as the **Whiskers** of the plot are drawn from $\widehat{Q}_1$ and $\widehat{Q}_3$ to the most extreme data values that are still *inside* the fences.
8. Data values having values outside the fences are regarded as *outliers*, that is, values which are unusual relative to the remaining data values. They are often denoted by asterisks.



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

$$\text{Lower Outer Fence:} \quad \widehat{Q}_1 - 3(IQR) \quad \text{or} \quad X_{(1)}$$

and

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

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

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

Some comments on the expected shape of box plots:

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

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

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

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

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

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

Outlier Detection

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

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

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

Case 1: The cdf of Y, F_Y is Known

The probability can be directly calculated:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Consider the following four cases:

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

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

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

If γ is the same the probability is equal.

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

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

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

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

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

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

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

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

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

What can we conclude from the above calculations?

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

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

Expected Number of Outliers in n Data Values

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

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

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

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

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

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

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

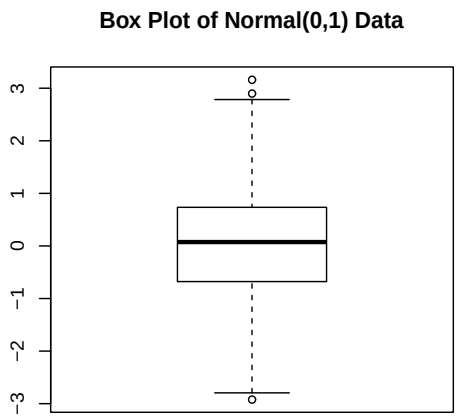
Note: For a symmetric distribution with location parameter = 0,

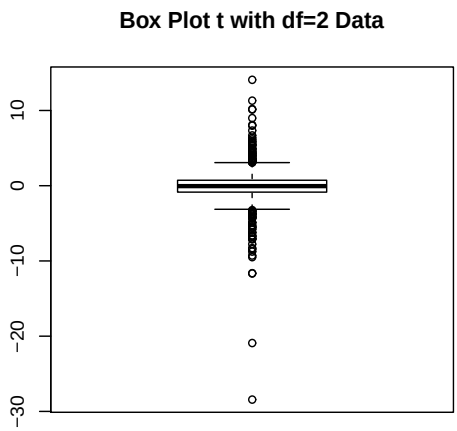
$$Q(.75) = -Q(.25) \Rightarrow IQR = -2Q(.25) \Rightarrow$$

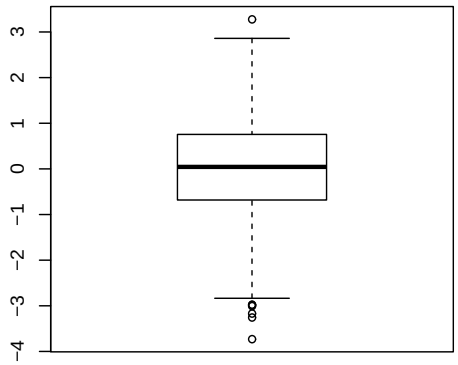
$$\begin{aligned} P[Y \text{ is an Outlier}] &= P\left[Y < Q_Y(.25) - 1.5(IQR_Y)\right] + P[Y > Q_Y(.75) + 1.5(IQR_Y)] \\ &= 2P\left[Y < 4Q(.25)\right] = 2F_Y(4Q(.25)) \end{aligned}$$

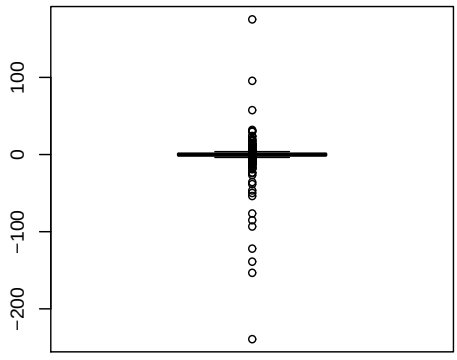
$$Q_{Normal}(.25) = -.6745 \quad Q_{logistic}(.25) = -1.0986 \quad Q_{DouExp}(.25) = -.9093 \quad Q_{Cauchy}(.25) = -1.0$$

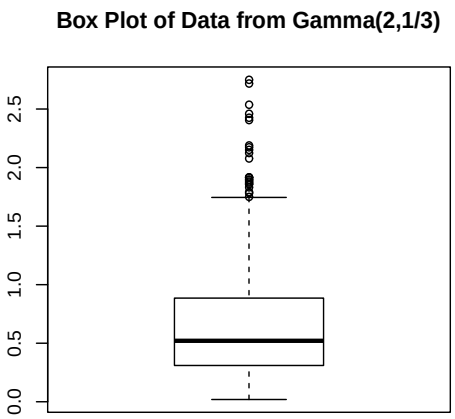
Box plots for six distributions were generated based on 1000 observations from each of the six distributions. The plots are displayed on the following page.

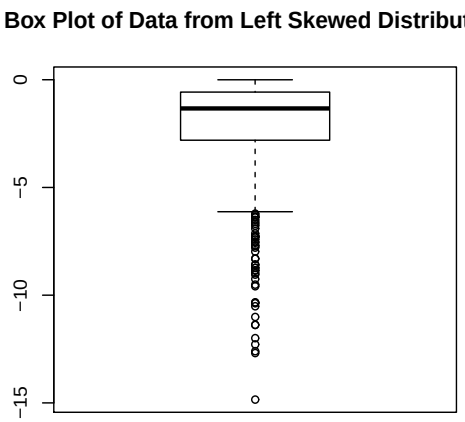

Box Plot of Normal(0,1) Data
3
2
1
0
−1
−2
−3

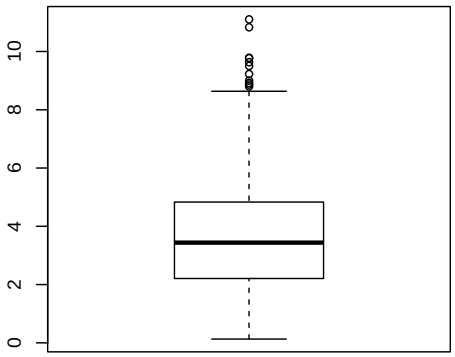

Box Plot t with df=2 Data
10
0
−10
−20
−30

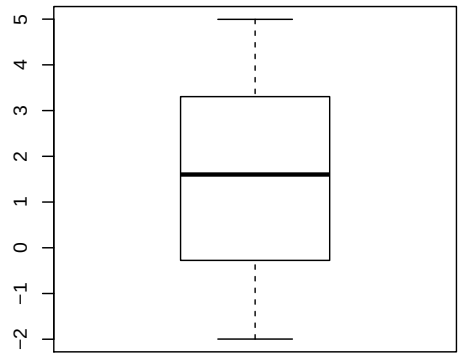

Box Plot t with df=30 Data
3
2
1
0
−1
−2
−3
−4


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


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


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


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

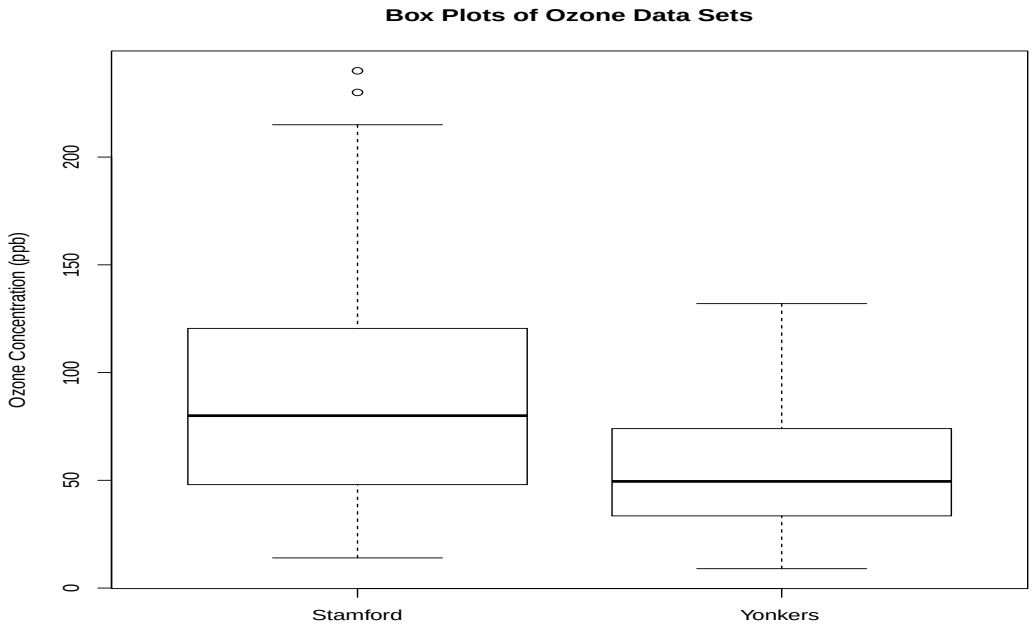

Box Plot of Data from Uniform(−2,5)
−2
−1
0
1
2
3
4
5

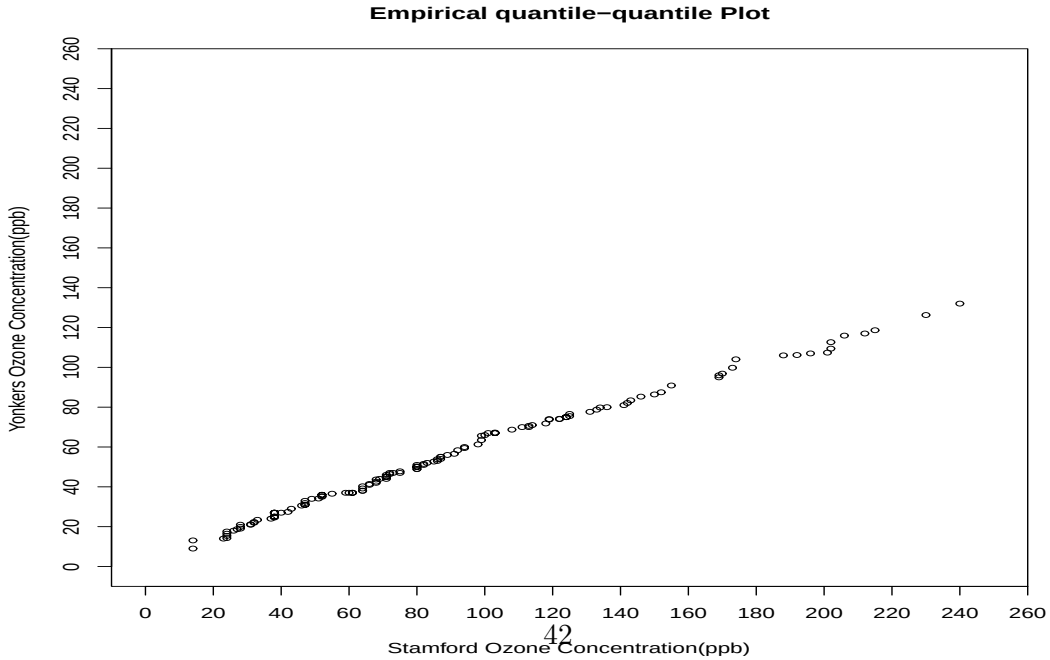
Box Plots, Quantile Plots, and Time Series Plots for Ozone Data

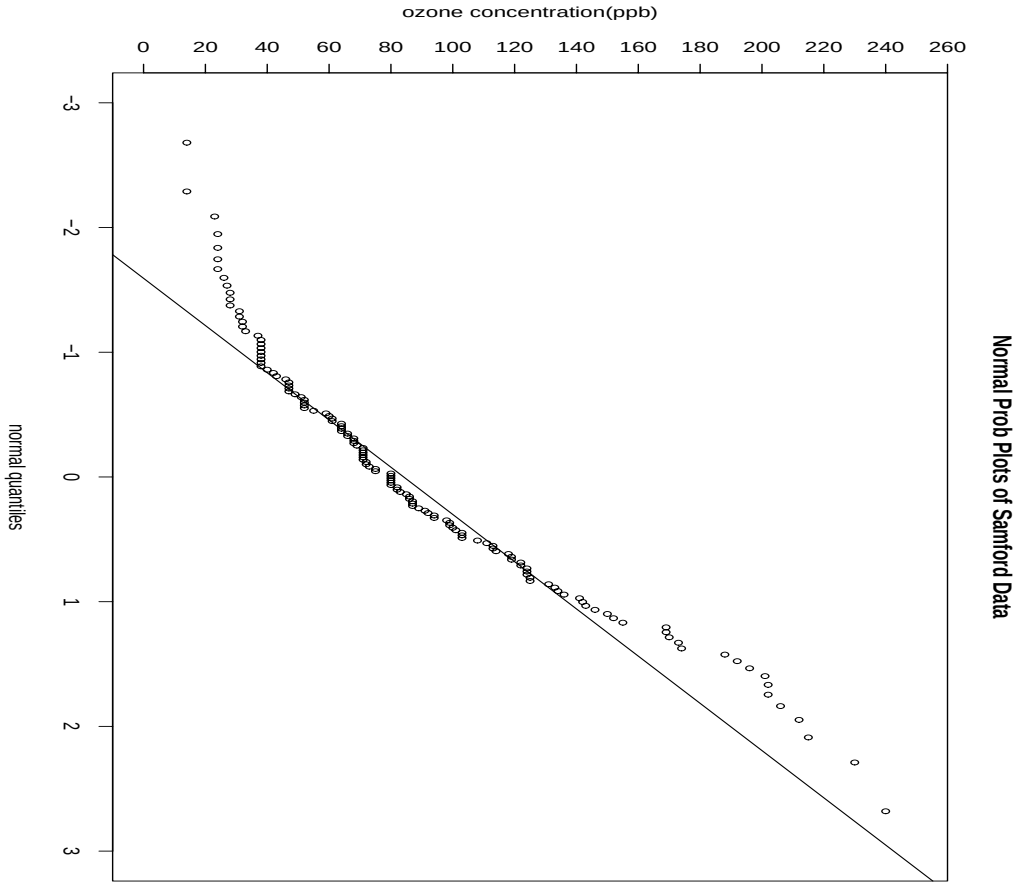
The following R code, generates various graphical comparisons of the Ozone data.

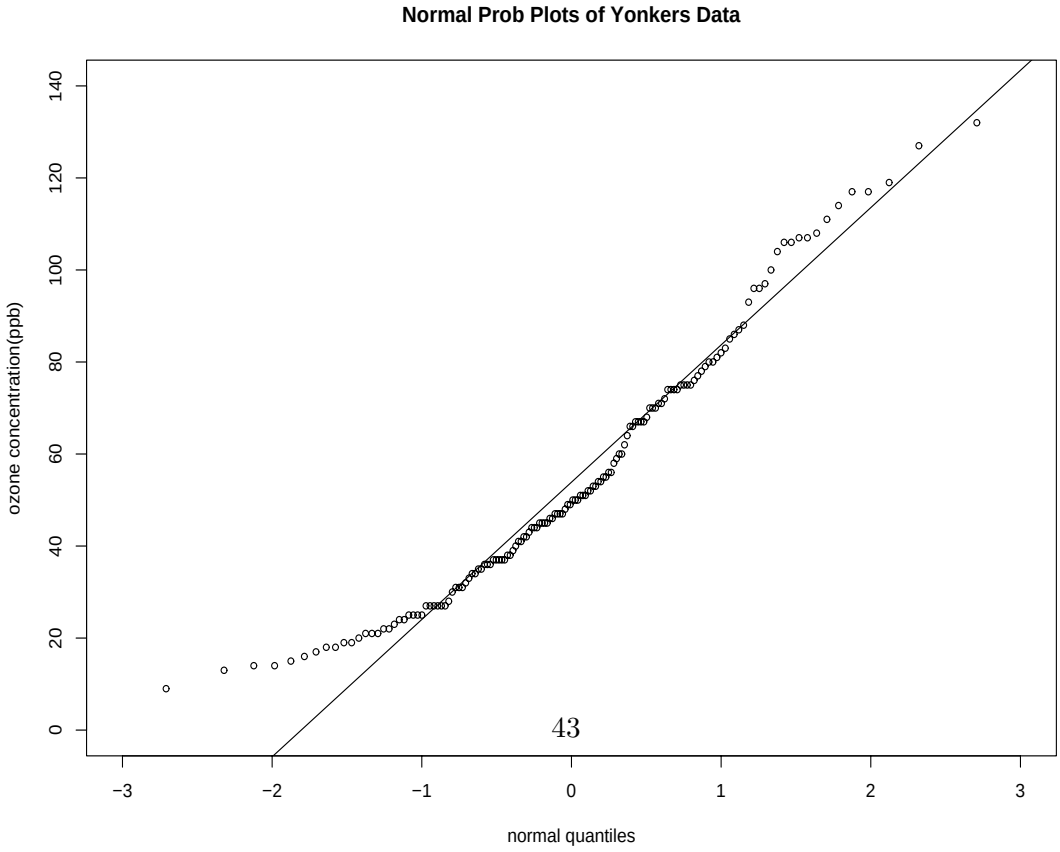
```
#The following R program generates various graphical comparisons of the
#Ozone data. The ozone data is in the files ozone1.DAT and ozone2.DAT
#The code is in R Files folder as ozonecompare.R
#------------------------------------------------------------------------------
#input data:

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

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

#creates side-by-side box plots:


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


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


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









#creates a normal Reference Distribution plot:
```

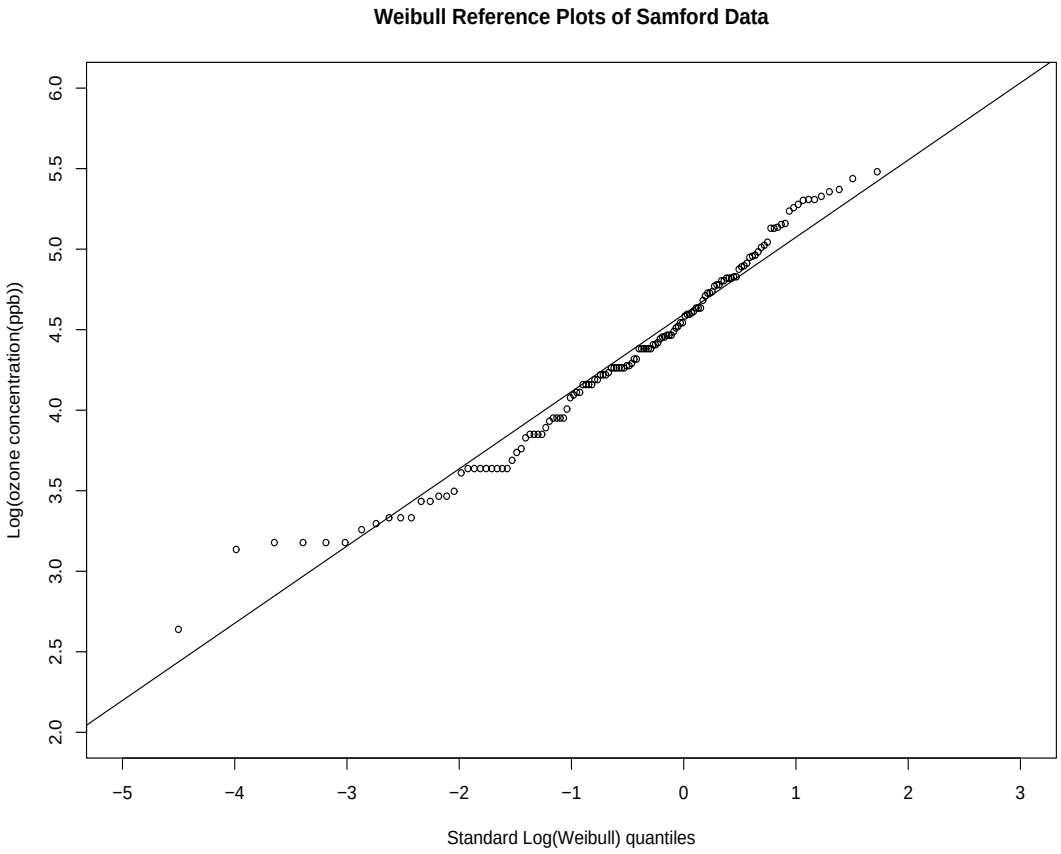
```
postscript("u:meth1/psfiles/ozonecomparenormalSamford.ps",height=8,horizontal=F)


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



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



#creates Weibull Reference Distribution Plots:

   plot(QZ1,w1,main="Weibull Reference Plots of Samford Data",
          xlab="Standard Log(Weibull) quantiles",
          ylab="Log(ozone concentration(ppb))",
          ylim=c(2,6),xlim=c(-5,3),lab=c(7,10,7),cex=.75)
    abline(lm(w1~QZ1))


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

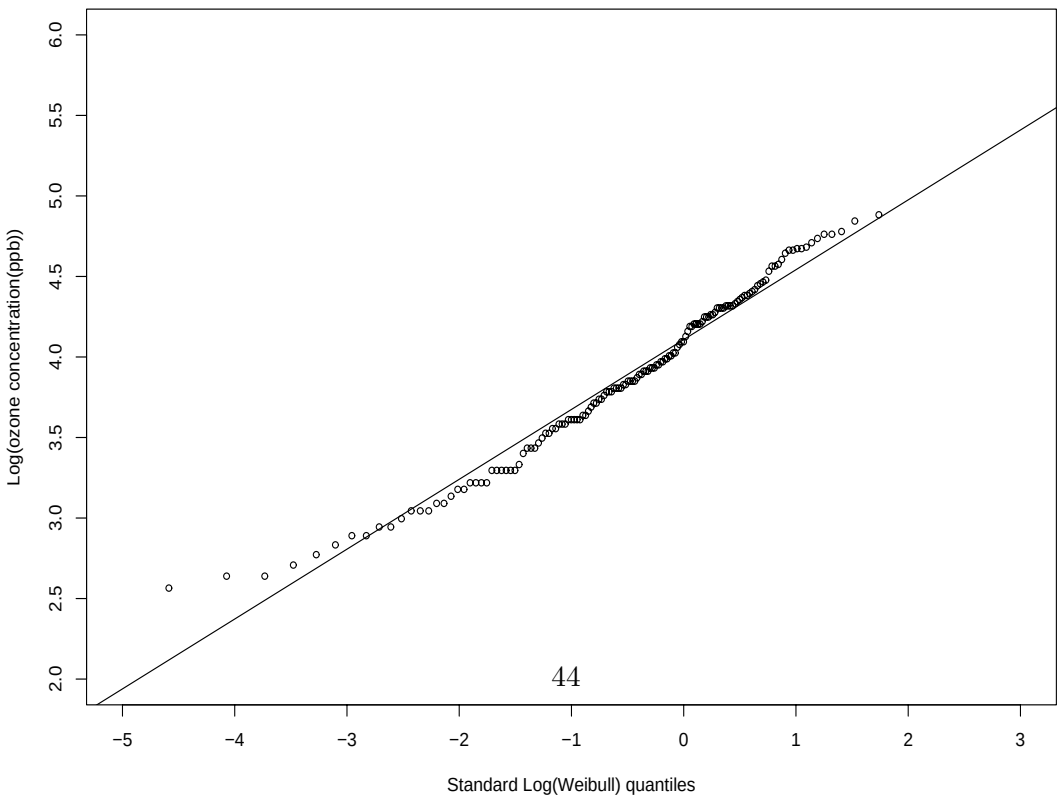

Box Plots of Ozone Data Sets
Ozone Concentration (ppb)
0
50
100
150
200
Stamford
Yonkers


Empirical quantile−quantile Plot
Yonkers Ozone Concentration(ppb)
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Stamford Ozone Concentration(ppb)


ozone concentration(ppb)
0 20 40 60 80 100 120 140 160 180 200 220 240 260
normal quantiles
-3 -2 -1 0 1 2 3
Normal Prob Plots of Samford Data


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


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


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

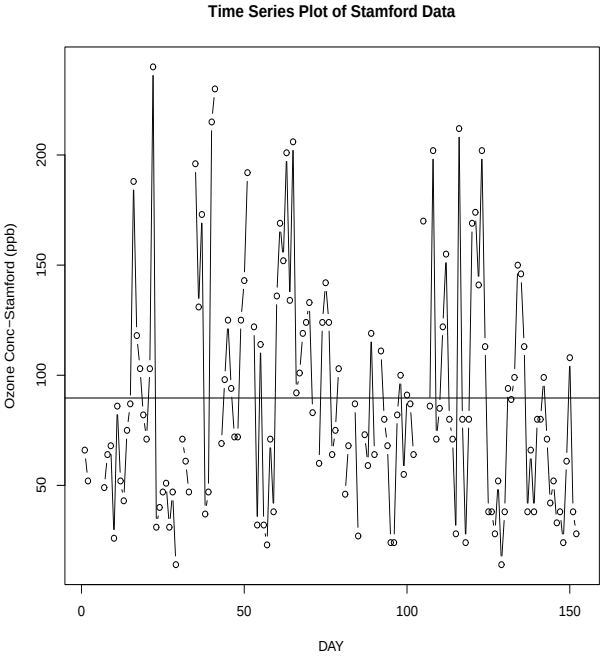
Time Series Plots

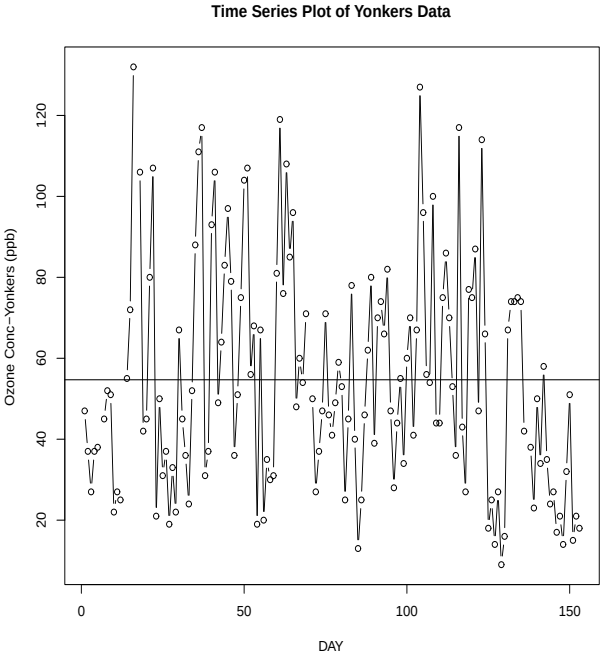
The following R code in the R Files folder as ozonecompare,time.R produces plots which take into account the time element in the ozone data.

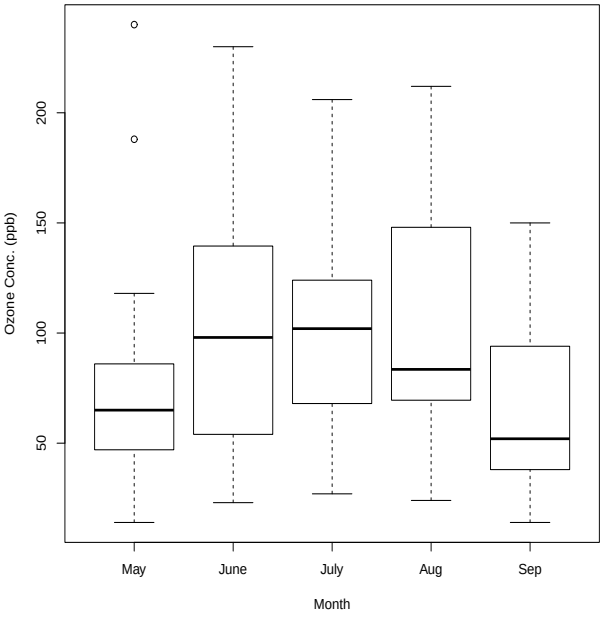
```
#input data:
         y1 = scan("u:/meth1/Rfiles/ozone1.DAT")
          y2 = scan("u:/meth1/Rfiles/ozone2.DAT")

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

#create time index

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

#classify data by month

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


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

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







       plot(y1na,type="b",ylim=c(0,250),
```

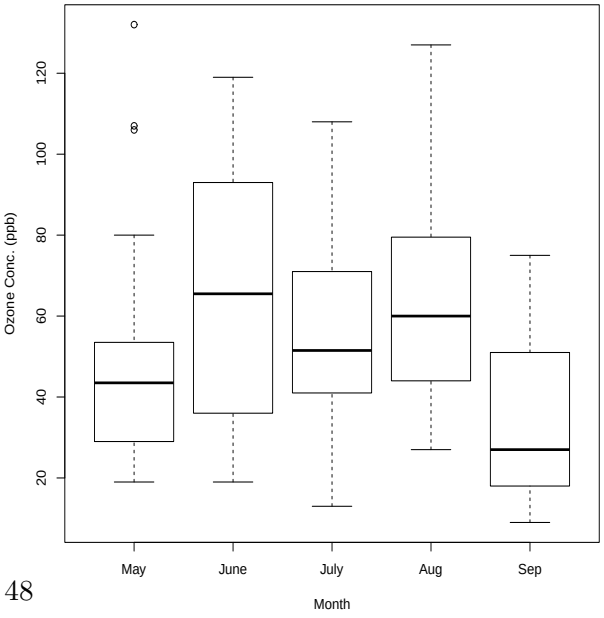
```
          ylab="Ozone Conc-Stamford (ppb)",xlab="DAY",
          main="Time Series Plot of Stamford Data",cex=.9)
          abline(h=mean(y1))

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


#side by side boxplots for the various months:


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

          boxplot(y2m,y2jn,y2jl,y2a,y2s,xlab="Month",ylab="Ozone Conc. (ppb)",
          main="Boxplots of Ozone Conc. for Yonkers by Month",
               names=c("May","June","July","Aug","Sep"))
g
          boxplot(y1m,y1jn,y1jl,y1a,y1s,xlab="Month",ylab="Ozone Conc. (ppb)",
          main="Ozone Conc. for Stamford",ylim=c(0,250),
               names=c("May","June","July","Aug","Sep"),cex=.75)

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

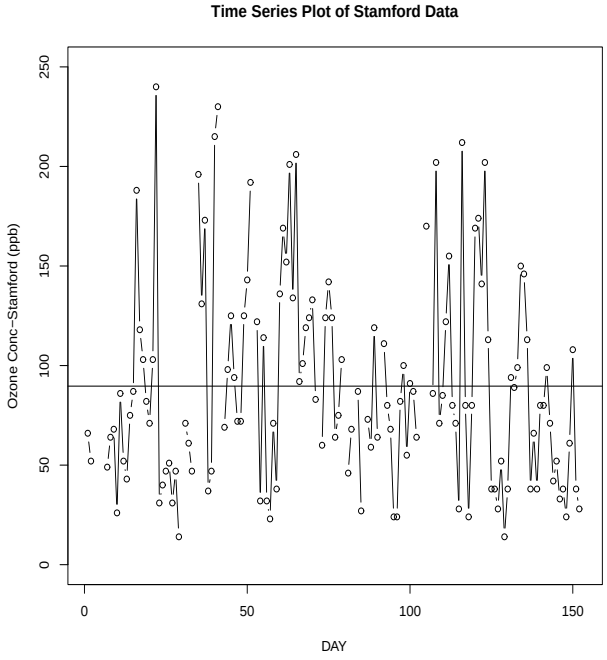
```
Call: acf(x = StamfordOzone)


AUTOCORRELATION MATRIX:

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

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

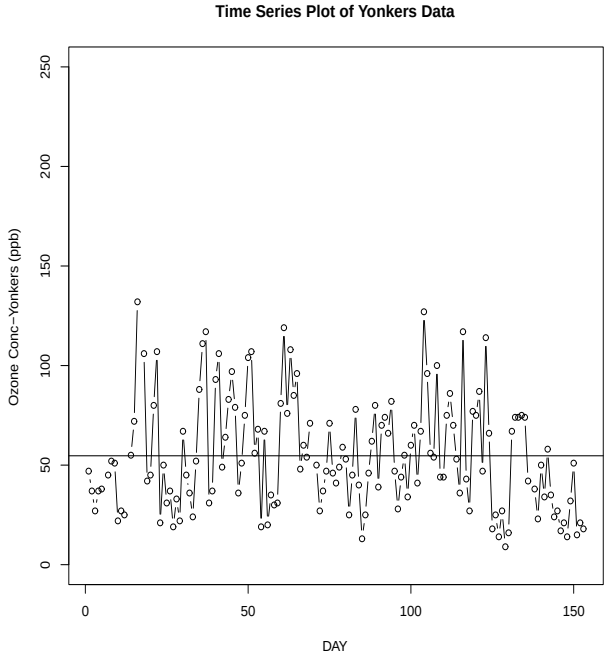

Time Series Plot of Stamford Data
Ozone Conc−Stamford (ppb)
50
100
150
200
0
50
100
150
DAY

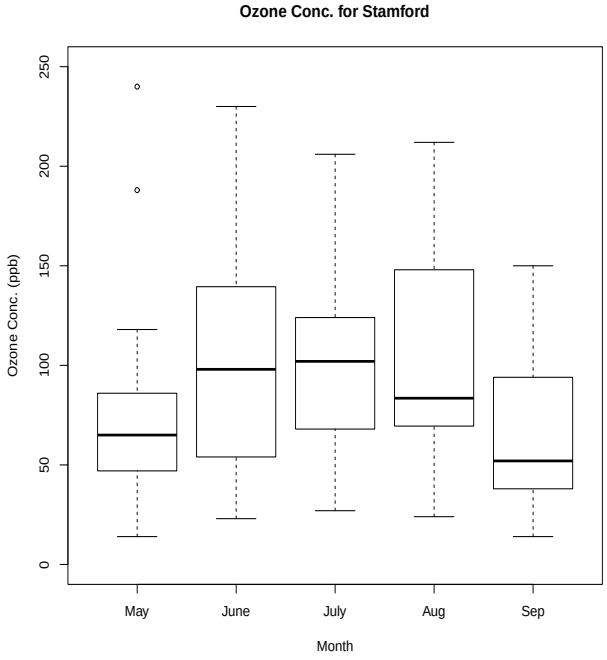

Time Series Plot of Yonkers Data
Ozone Conc−Yonkers (ppb)
20
40
60
80
100
120
0
50
100
150
DAY

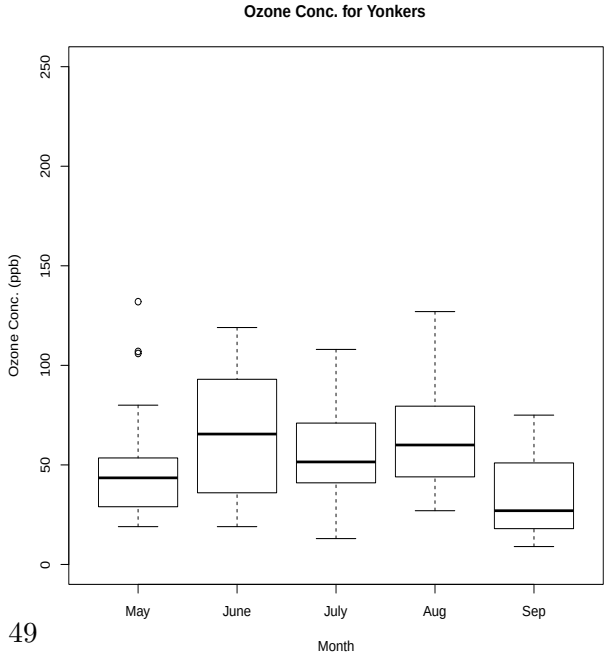

Boxplots of Ozone Conc. for Stamford by Month
Ozone Conc. (ppb)
50
100
150
200
May
June
July
Aug
Sep
Month


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


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


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


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


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

Scatterplots of Multiple Response Data

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

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

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

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

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

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

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

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

TABLE 3.7

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

TABLE 3.8

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

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

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

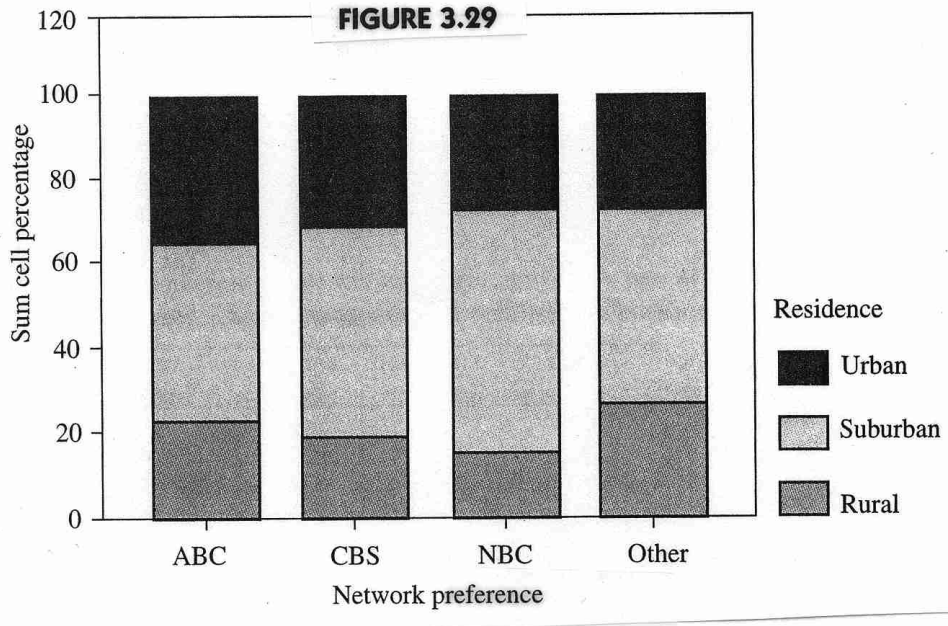


FIGURE 3.29

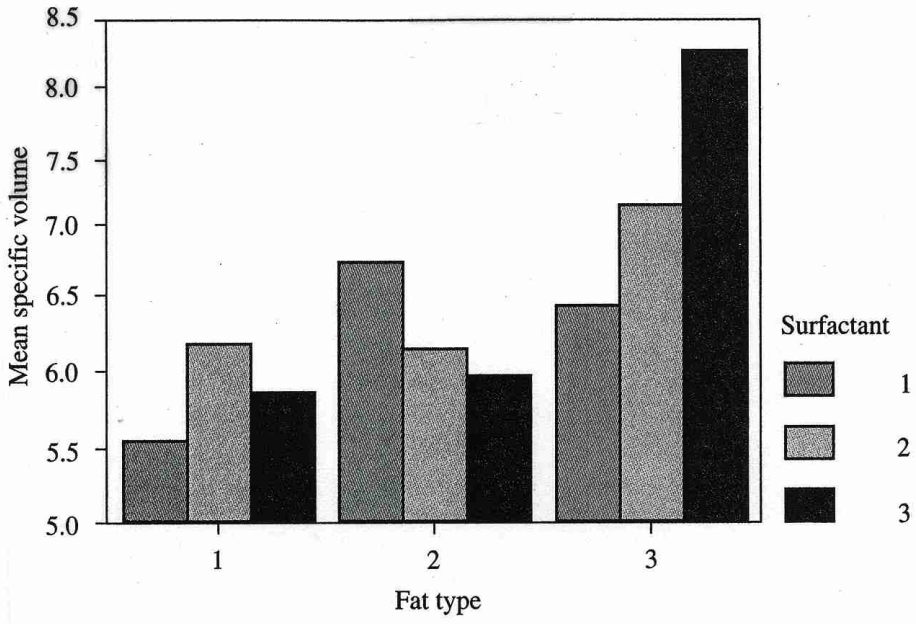
A second extension of the bar graph provides a convenient method for displaying the relationship between a single quantitative and a qualitative variable. A food scientist is studying the effects of combining different types of fats with different surfactants on the specific volume of baked bread loaves.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

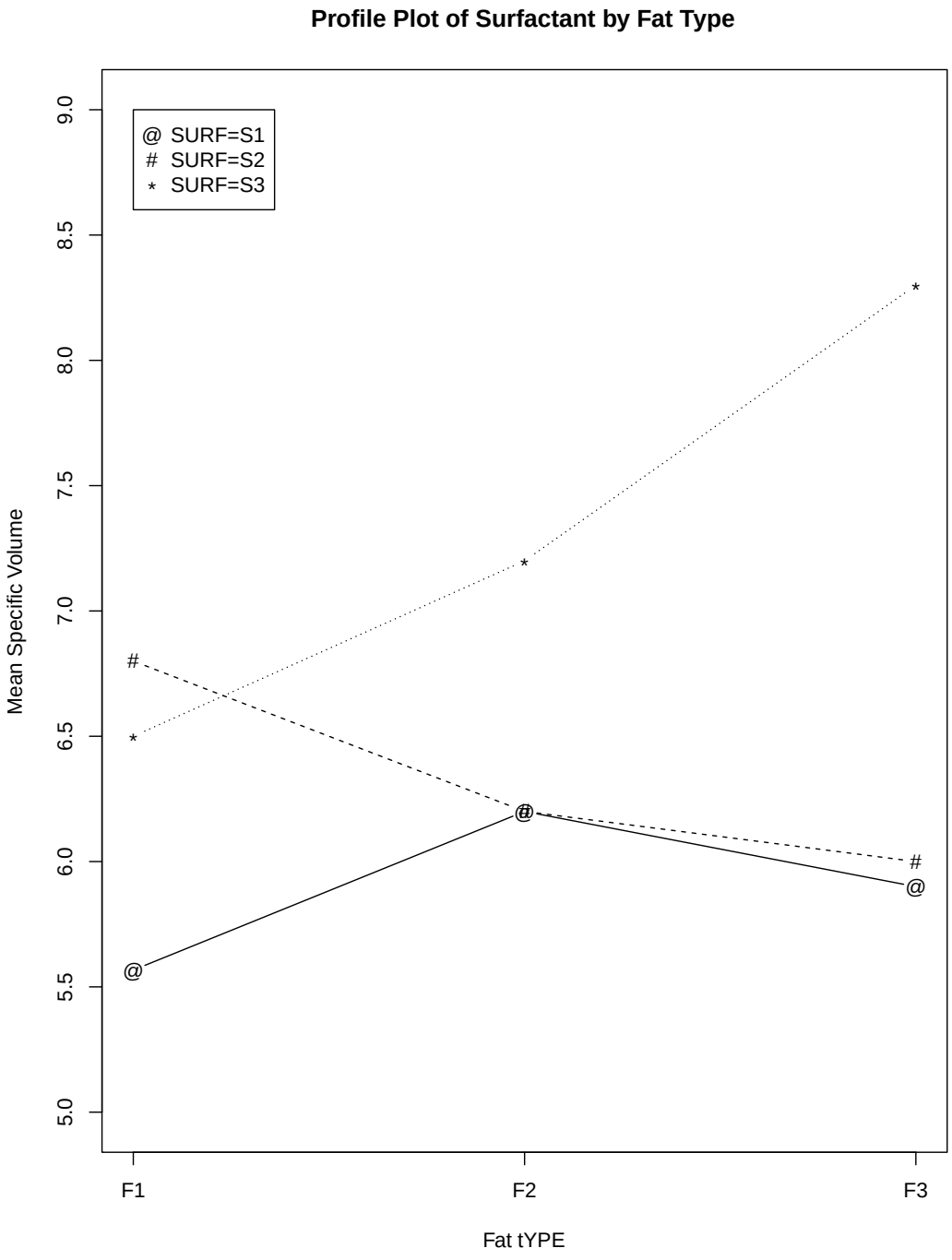
In this experiment, the scientist wants to make inferences from the results of the experiment to the commercial production process. Figure 3.30 is a **cluster bar graph** from the baking experiment. This type of graph allows the experimenter to examine the simultaneous effects of two factors, type of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

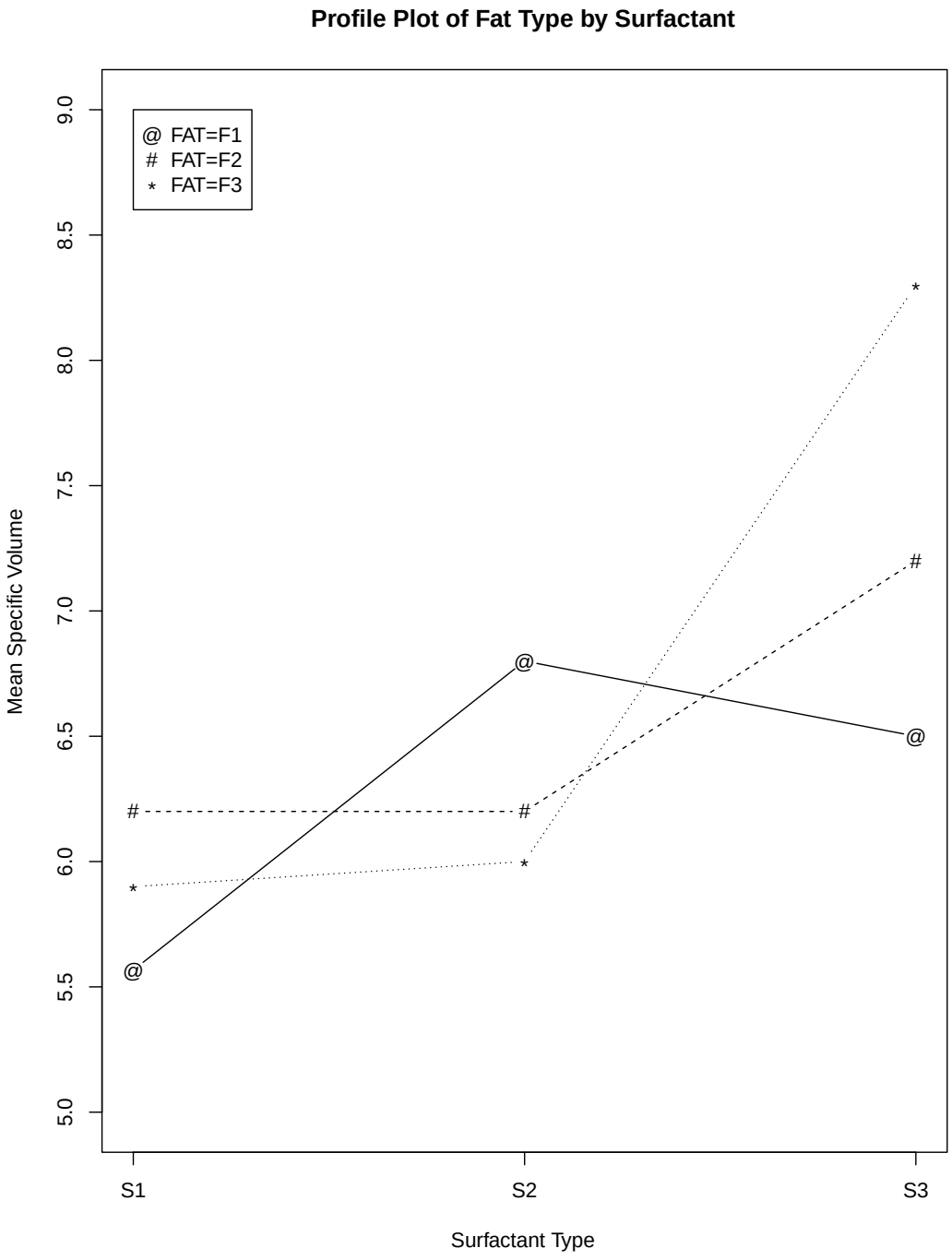
TABLE 3.9

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

FIGURE 3.30



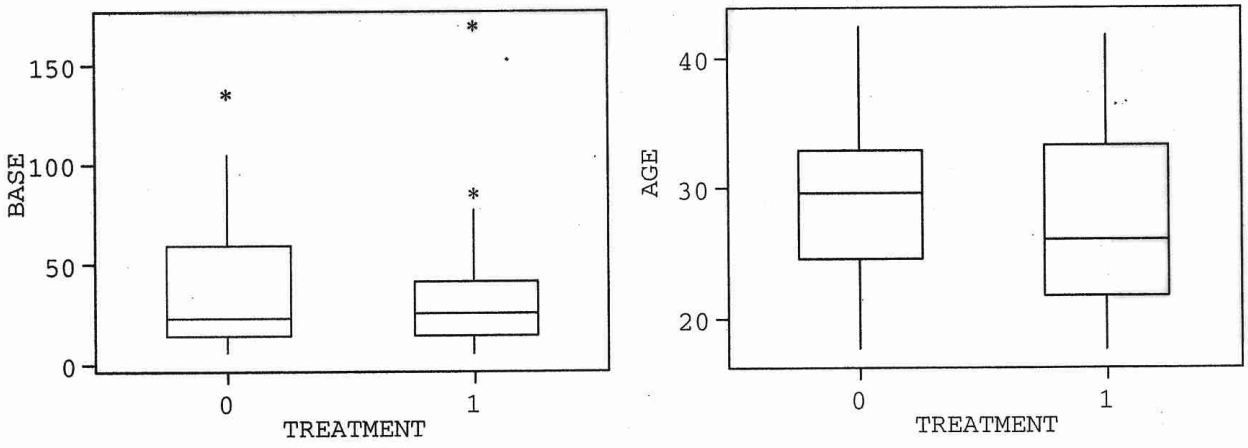

Profile Plot of Surfactant by Fat Type
@ SURF=S1
# SURF=S2
* SURF=S3
Mean Specific Volume
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
F1
F2
F3
Fat tYPE


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

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

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

FIGURE 3.32



(a) Boxplot of base by treatment

(b) Boxplot of age by treatment

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

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

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

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

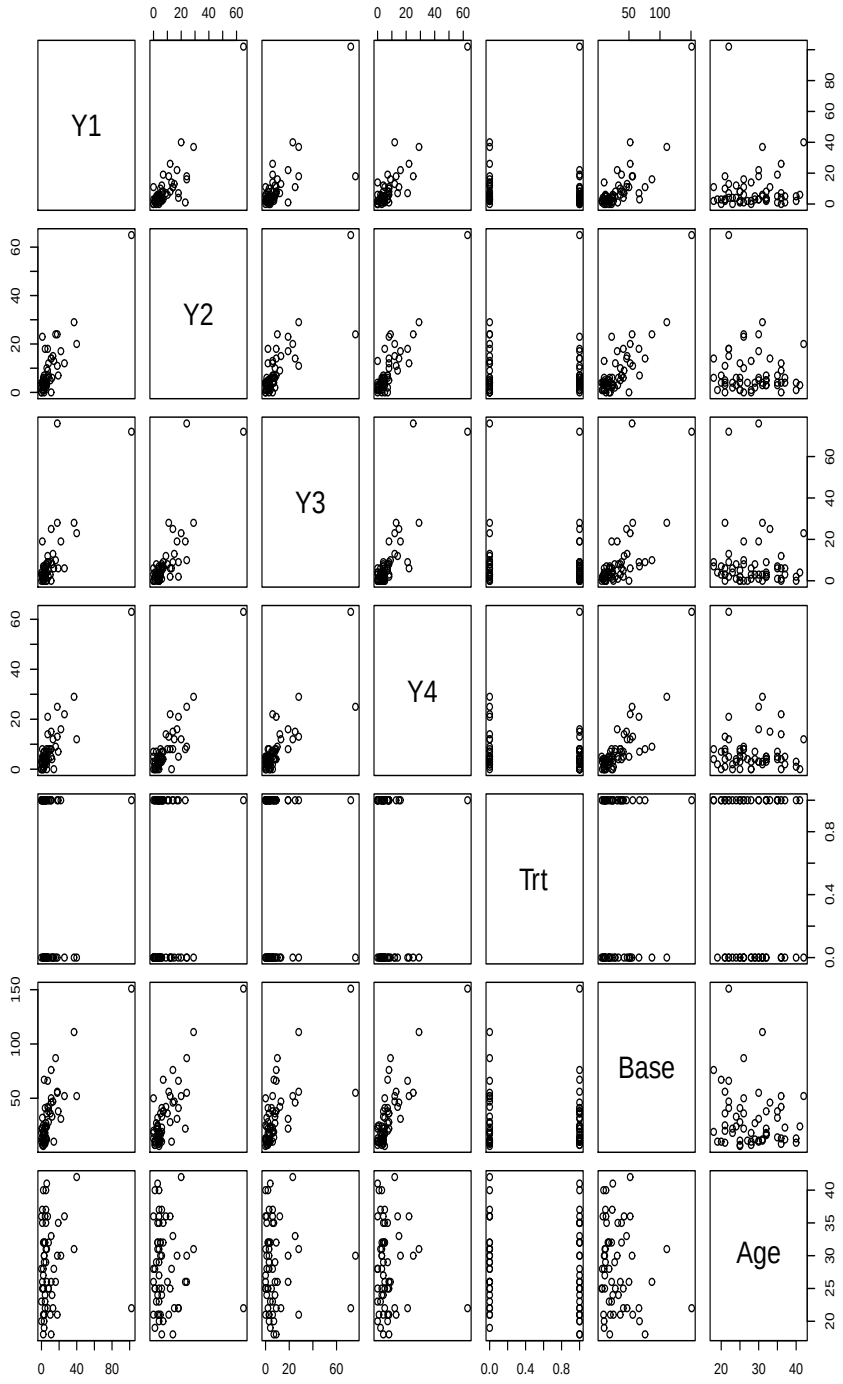
Cell Contents: Pearson correlation
P-Value

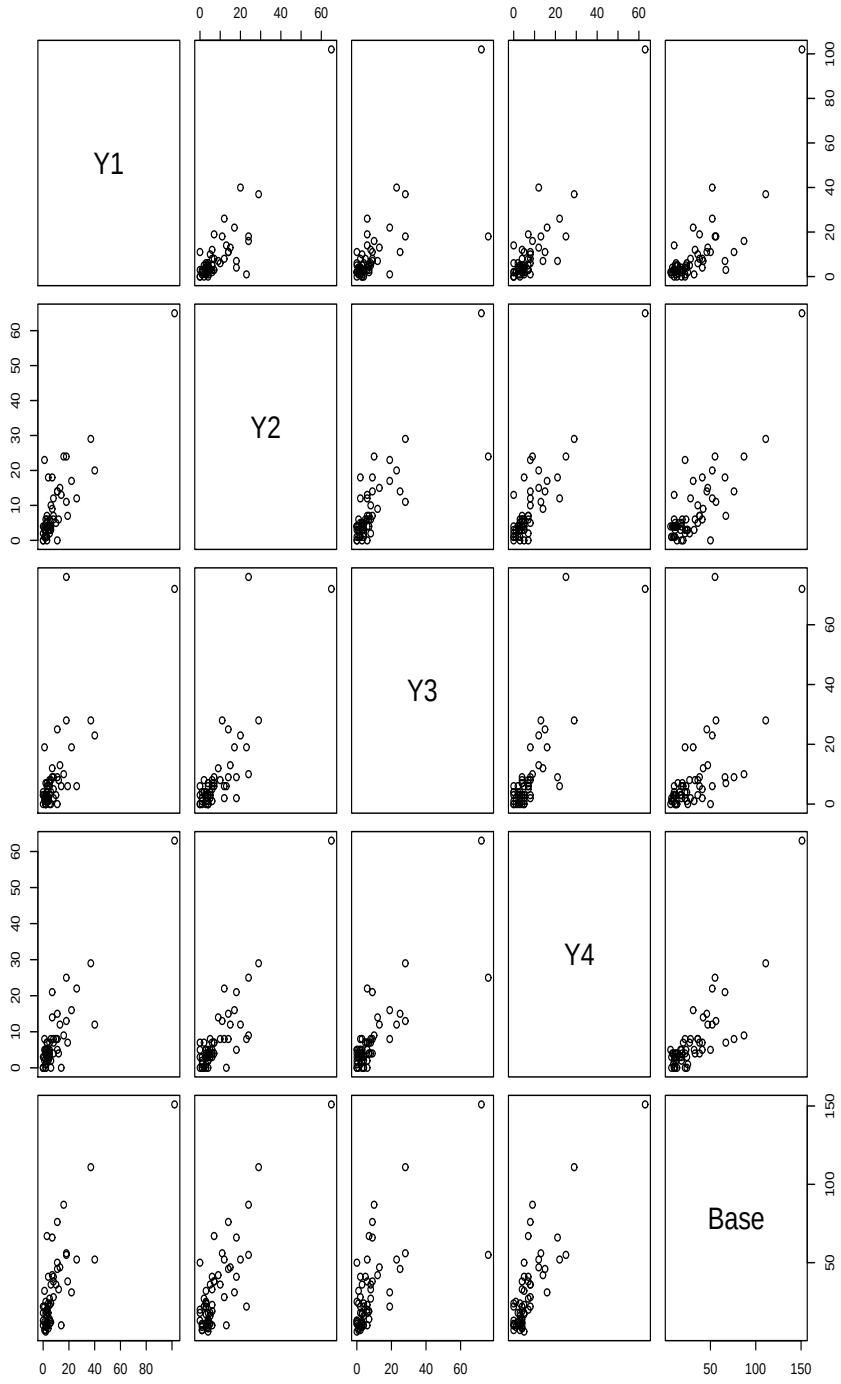
The following R code in R Files as epilepsy-plots.R will produce various plots of the Epilepsy data:

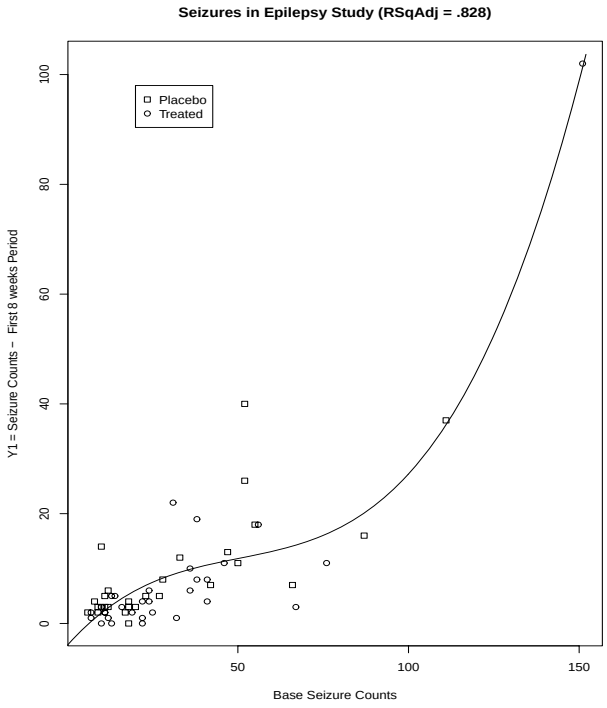
```
par(ask=TRUE)

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


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


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


Y1 = x[,2]
Y2 = x[,3]
Y3 = x[,4]
Y4 = x[,5]
Base = x[,6]
Age = x[,7]
Y <- cbind(Y1,Y2,Y3,Y4,Base,Age)
c = cor(Y)
c = round(c,3)


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


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


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


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

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

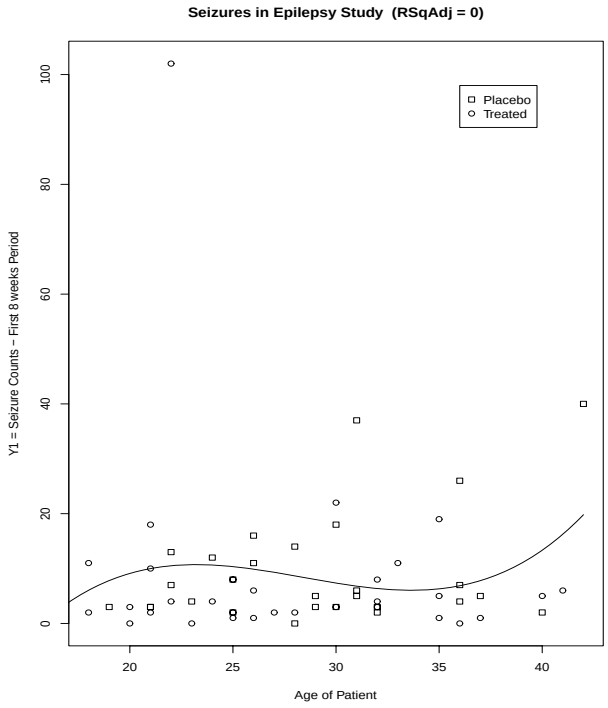
```
plot(Base[-c(49)],Y1[-c(49)],main="Seizures in Epilepsy Study Without Outlier
(RSqAdj = .449)",xlab="Base Seizure Counts",
ylab="Y1 = Seizure Counts - First 8 weeks Period",pch=as.numeric(Trt))
legend(13,38,c("Placebo","Treated"),pch=0:1)
av=seq(0,112,.1)
bv = predict(m1,list(Base=av))
lines(av,bv)


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


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


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


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

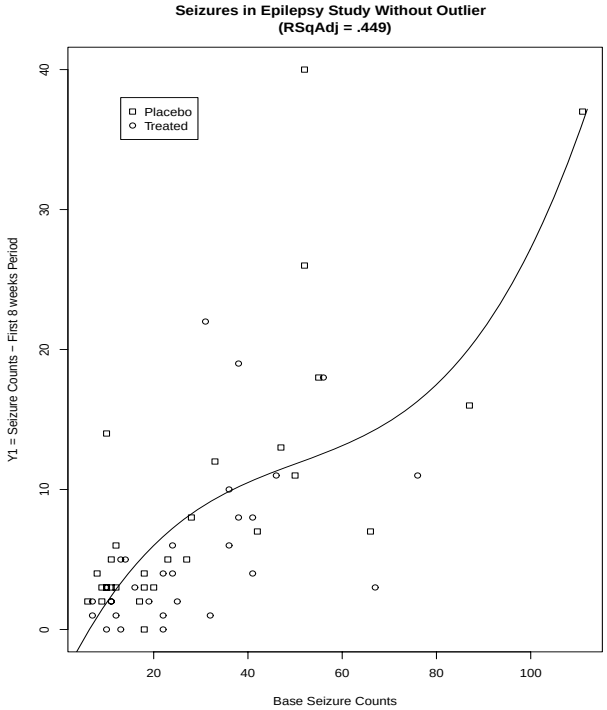
```
#Regression Plot of Y4 vs Age with outlier removed
m8 = update(m6,subset=(1:59)[-c(49)])
summary(m8)
plot(Age[-c(49)],Y4[-c(49)],main="Seizures in Epilepsy Study Without Outlier
(RSqAdj = 0)",xlab="Age of Patient",
ylab="Y4 = Seizure Counts - Fourth 8 wks",pch=as.numeric(Trt))
legend(20,38,c("Placebo","Treated"),pch=0:1)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

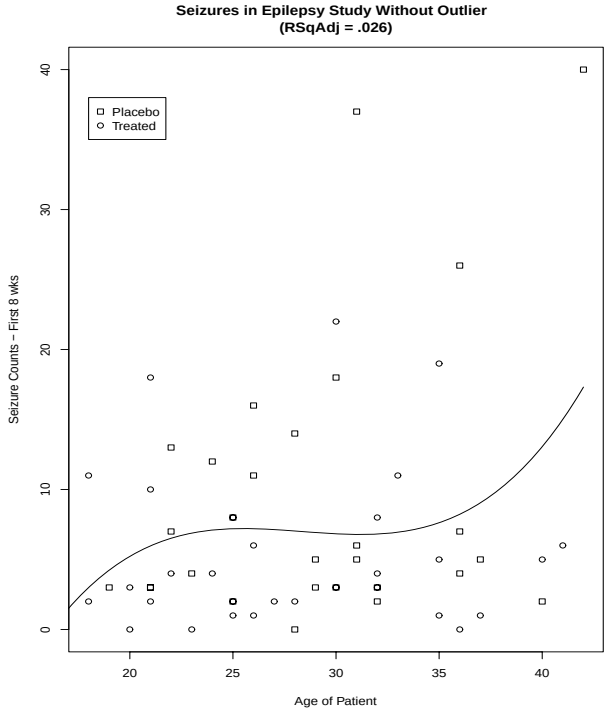

Y1
Y2
Y3
Y4
Trt
Base
Age

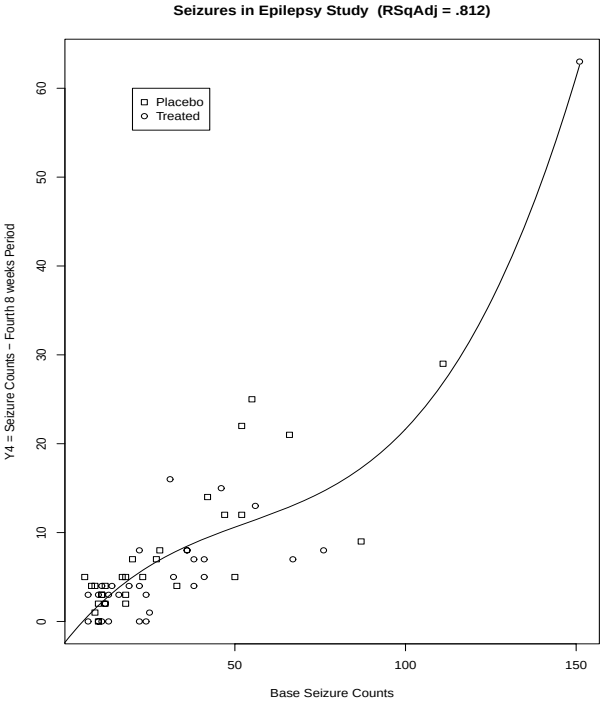

Y1
Y2
Y3
Y4
Base

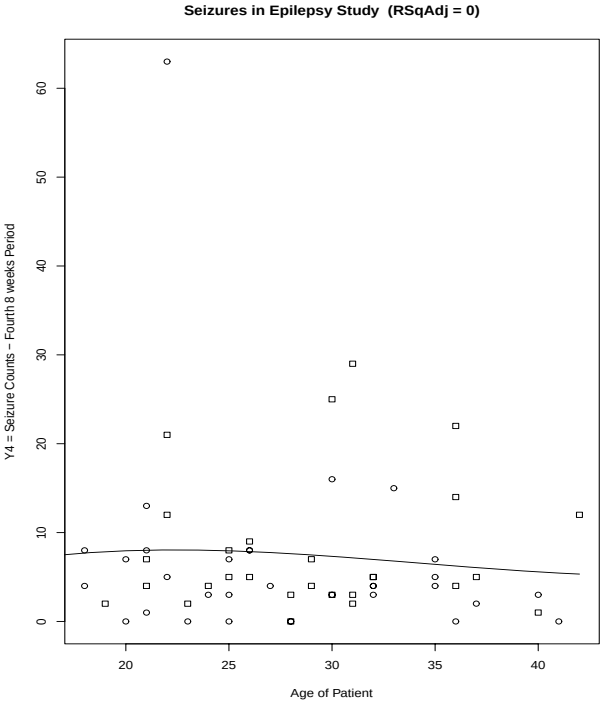

Seizures in Epilepsy Study (RSqAdj = .828)
Placebo
Treated
Y1 = Seizure Counts – First 8 weeks Period
Base Seizure Counts
0
20
40
60
80
100
50
100
150

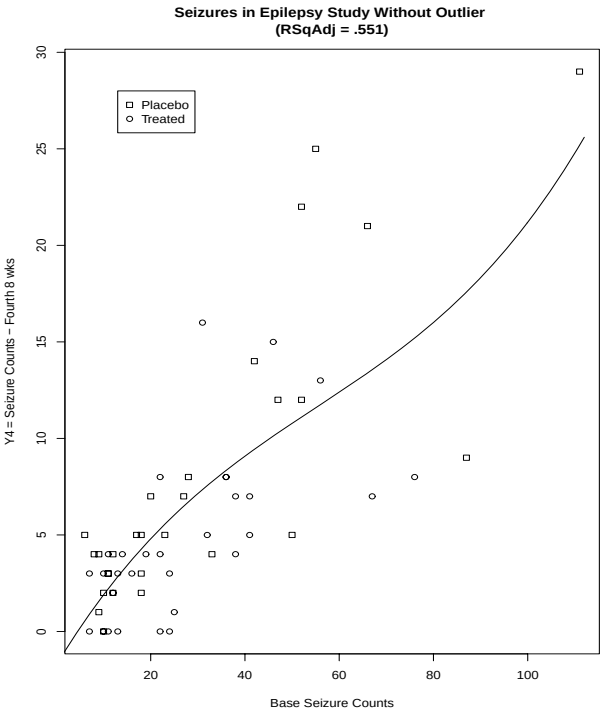

Seizures in Epilepsy Study (RSqAdj = 0)
Placebo
Treated
Y1 = Seizure Counts – First 8 weeks Period
Age of Patient
0
20
40
60
80
100
20
25
30
35
40

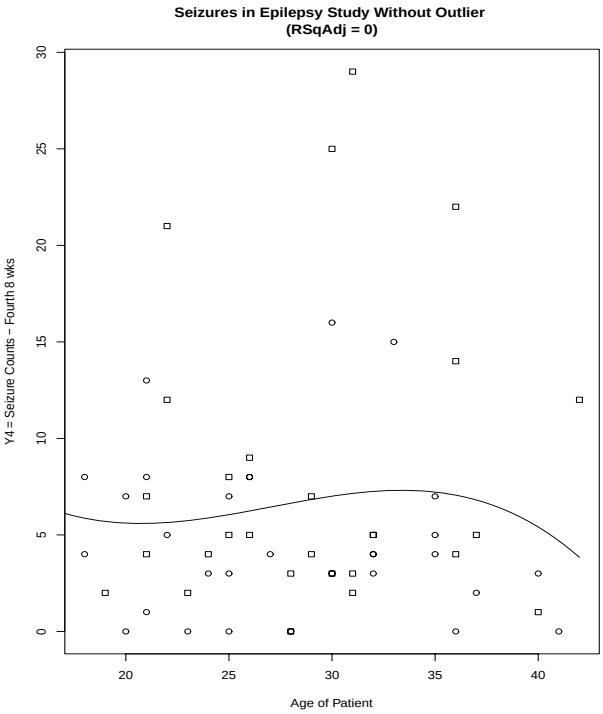

Seizures in Epilepsy Study Without Outlier
(RSqAdj = .449)
Placebo
Treated
Y1 = Seizure Counts – First 8 weeks Period
Base Seizure Counts
0
10
20
30
40
20
40
60
80
100


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


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


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


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


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

START: Wednesday 10/13/21

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

Supplemental Reading:

• Chapter 5, Sections 6.1, 14.6.2, and 15.1 in Tamhane/Dunlop Book

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?

Check conditions:

1. Is there a fixed number of trials? — yes
2. Does each trial result in a S or F? — yes
3. Are the trials independent? — yes

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

If violated, we can collapse categories with neighbor and make categories larger. In our example collapse 5 into 4 and make last category 4+

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.

A suggestion for a minimally acceptable n is to compute $p_1,\ p_2,\ \ldots,\ p_k$ and take $n = 5/min(p_1,\ p_2,\ \ldots,\ p_k)$

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.

usually on the neighboring cells.

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 provided tables 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?

No, one of the conditions is that none of the expected counts can be less than one.

BUT $E_5 = 0.0977 < 1$.

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

$$\begin{aligned}\hat{\theta} &= \frac{\text{Number of Defective Batteries}}{\text{Number of Batteries}} \\ &= \frac{(0)(26)+(1)(51)+(2)(47)+(3)(16)+(4)(10)}{(4)(150)} = \frac{233}{600} = 0.3883\end{aligned}$$

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 $i = 0, 1, 2, \ldots, 8$ and

$\hat{p}_{10} = P[X \geq 9] = 1 - P[X \leq 8] = \ 1 - ppois(8, 3.85)$ using R

$\hat{E}_i = 300\hat{p}_i$ for $i = 1, 2, \ldots, 10$

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	≥ 9	Total
$\hat{p}_i$	.021	.082	.158	.202	.195	.150	.096	.053	.026	.017	1.00
$\hat{E}_i$	6.4	24.6	47.3	60.7	58.4	45.0	28.9	15.9	7.6	5.2	300
O_i	7	24	42	59	62	44	41	14	5	2	300
$\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$	.06	.01	.60	.05	.22	.02	5.09	.22	.91	1.94	9.12

$$\hat{Q} = \sum_{i=1}^{10} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = 9.12, \quad df = 10 - 1 - 1 = 8 \Rightarrow$$

$$P[\chi^2_8 \geq 9.12] = 1 - F(9.12) = 1 - pchisq(9.12, 8) = 0.332,$$

where F is the cdf of a Chi-square distribution with $df = 8$.

With a p-value of 0.332, 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 that the researcher conjectures fits the observed data.

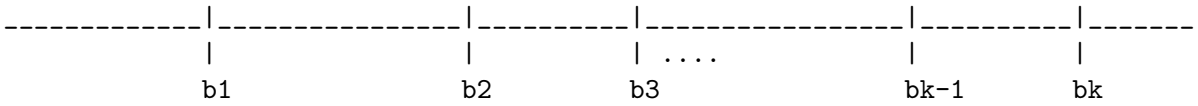
We will develop a measure of the degree to which $f_o(y; \boldsymbol{\theta})$ models the observed data.

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.

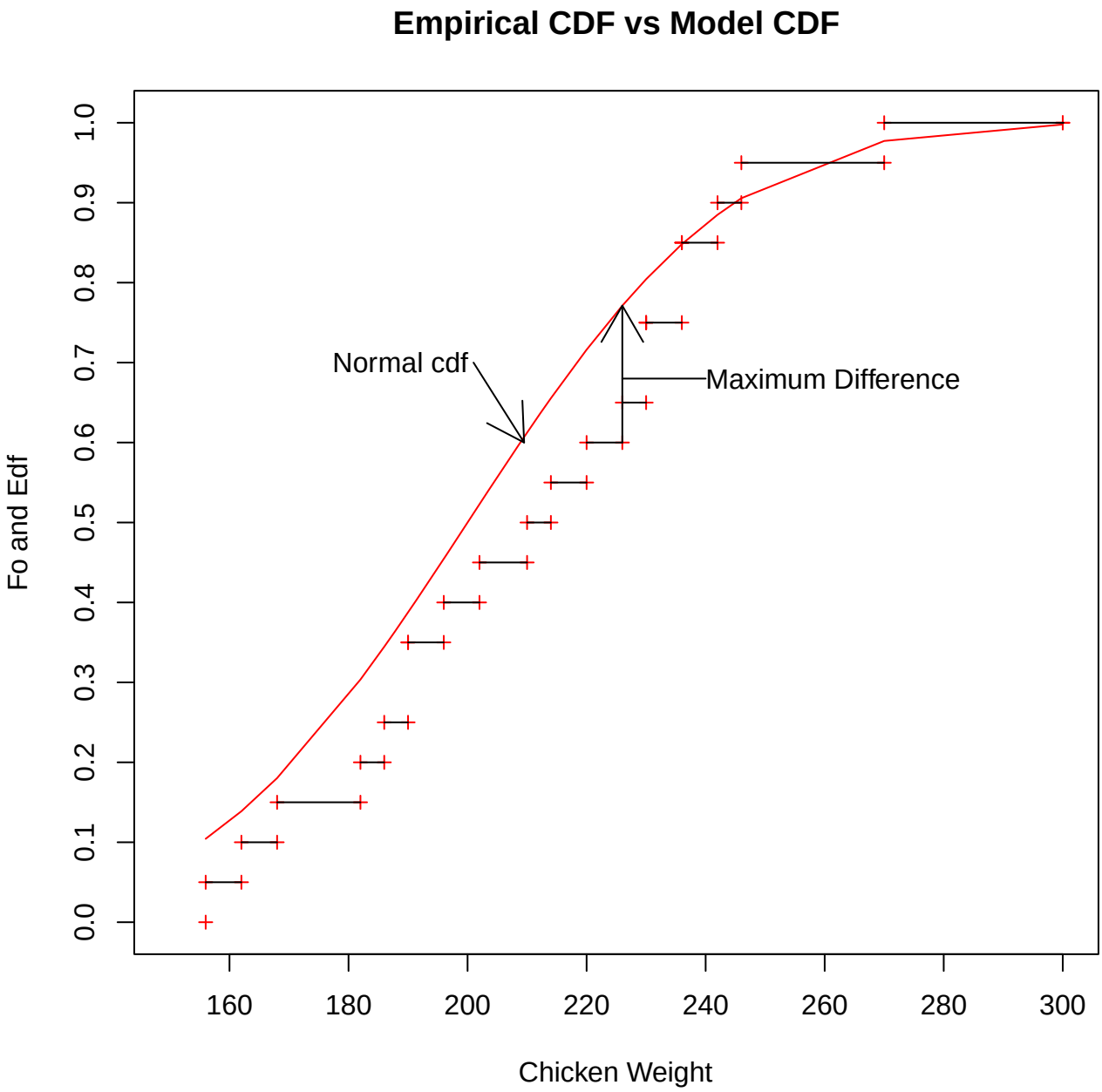
The following example will illustrate the above:

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). Recall, when F_o was the cdf of a discrete distribution, we used the $Q \sim \chi^2$ and the chi-square tables to determine the p-value.

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.

If F_o is NOT the correct model, cdf for Y, then $X = F_o(Y)$ will not have a Uniform on (0, 1) distribution. Also, $D_n = \max(D_n^-,\ D_n^+)$ will be large and D_n's distribution will depend on the form of F_o.

Tables for the distribution of D_n when $F = F_o$ are on page 17 of Handout 9. These tables can be used to compute the p-value for testing $F = F_o$, similar to computing the p-value in the case when F_o was a discrete distribution.

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.

END Class Notes 10/13/21 Wednesday

START: Class Date: Friday 10/15/21

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

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

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:

* look at book / papers mentioned in the beginning of slides to get more information about

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

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.

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

Is the distribution A_n^2 distribution-free when $F = F_o$, that is, is the distribution of A_n^2 the same for all selections of F_o provided F_o is the correct model for F?

Is this true, when F_o is not the correct selection of a cdf for F, that is, the cdf of the data, F is not F_o?

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 $n \geq 5$. The p-value is then obtained as

p-value $= 1 - G(z)$.

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.

Skip censoring stuff

GOF Measure - Model Completely Specified - Right 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 - Right 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 - Right 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
$Y_{(m+1)} \leq Y_{(m+2)} \leq \cdots \leq Y_{(n)}$ 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_m^2 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:

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

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 $X_i s$ and

the coefficients $a_i s$ are given in Table A28 on the following page.

When W is close to 1, the data appears to be from a normal distribution. 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

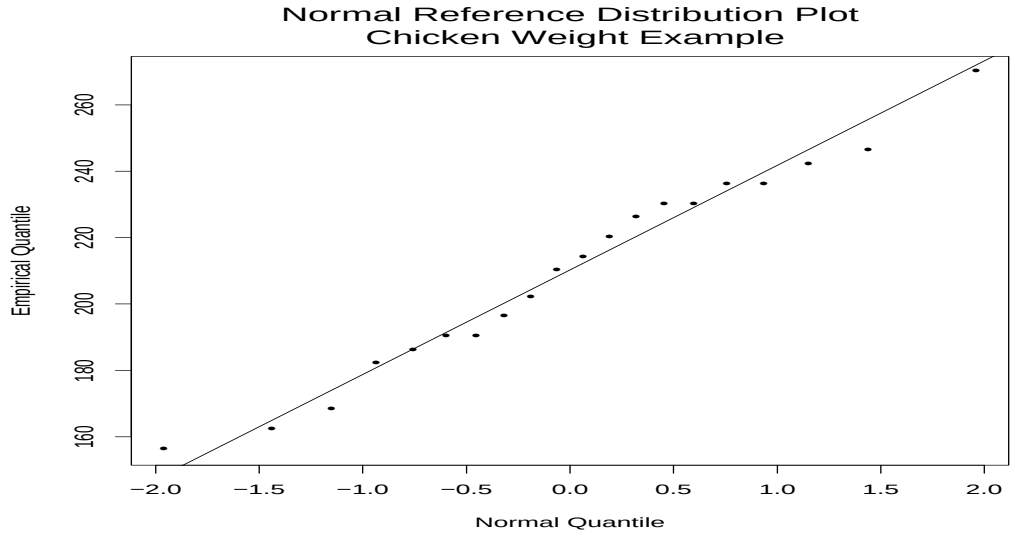
Suppose in the chicken weight example the researcher did not specify the values for μ and σ. Evaluate the degree of fit of the normal distribution to these 20 data values using the S-W statistic. The data in ordered form is given by:

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 which results in an approximation to the true distribution of the GOF statistics.

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*

	a_{n-i+1}											
i	$n=3$	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 (Q Q 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:

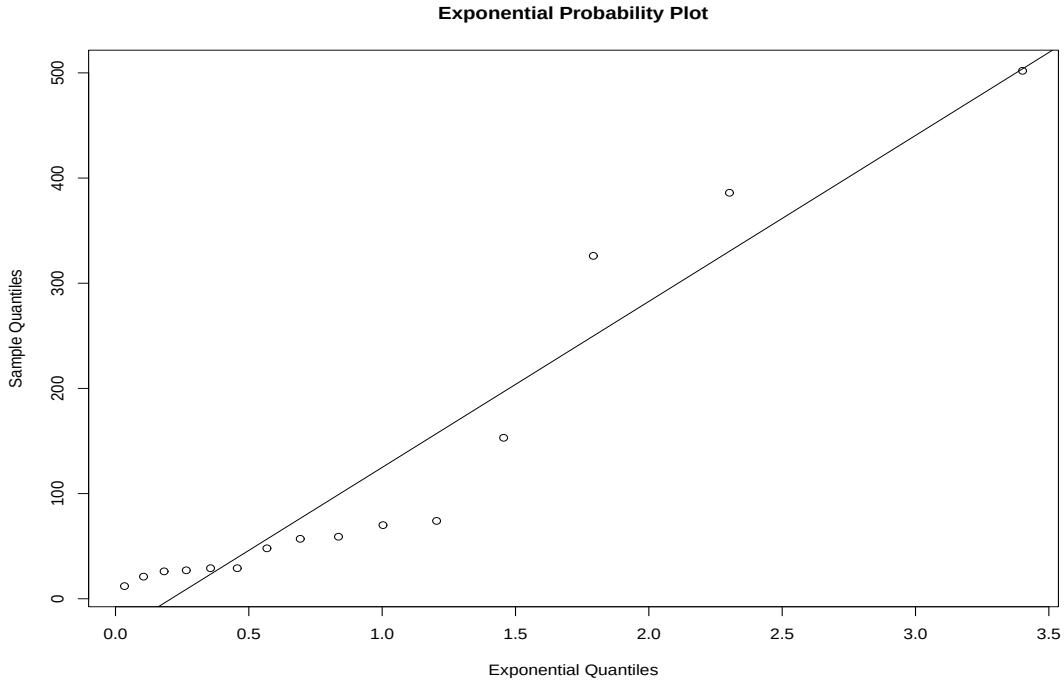
Table 4: Modifications and Percentiles for GOF Measures for Exponential Distribution with β Unknown

		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)
AD
1.1633

ADM = AD*(1+.6/n)
ADM
1.2098
```

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. However, if you run in R the function **ks.test(W,"pexp",1/121.2)**, you will obtain D = .2764 with p-value = .202 which would imply that the exponential distribution yielded a "Very Good Fit" to the data. The problem is that the p-value produced by ks.test does not take into account that an estimate, 1/121.2, was used for the scale parameter, β.

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:

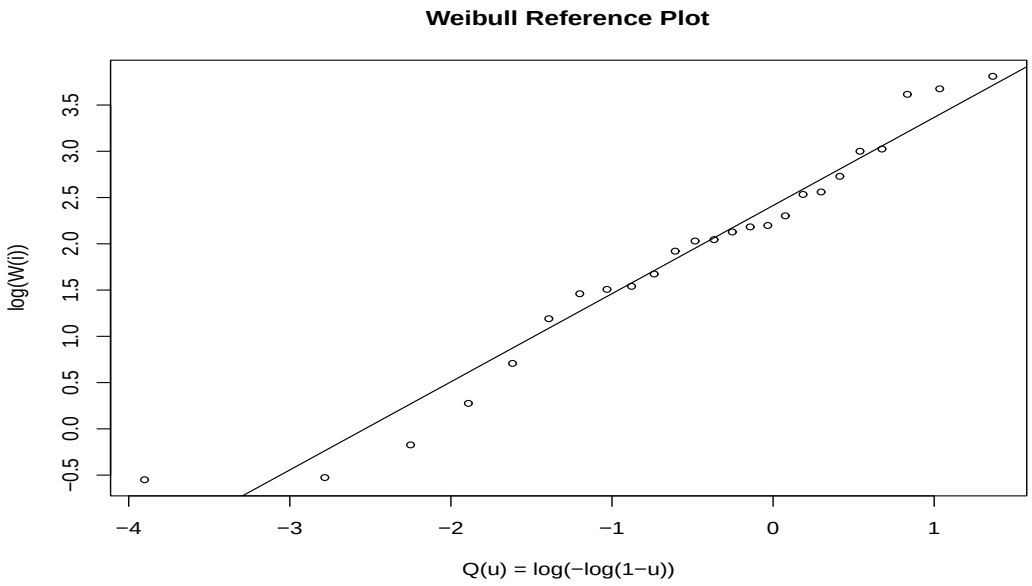
$$AD = AD\left[1 + \frac{.2}{\sqrt{n}}\right] = (.3277)\left[1 + \frac{.2}{\sqrt{23}}\right] = 0.3413$$

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))
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")

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

 AD  = 0.3276
 ADM = 0.3413
```

A Weibull Probability plot of the data is given here:

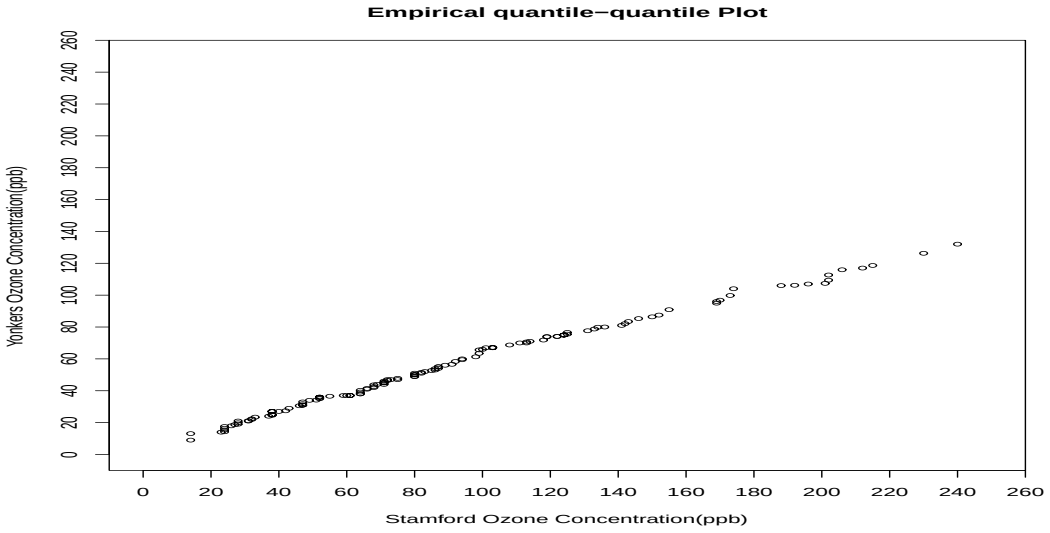


From the R output we have ADM = 0.3413 which implies from Table 5 that $p - value > 0.25$.

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.

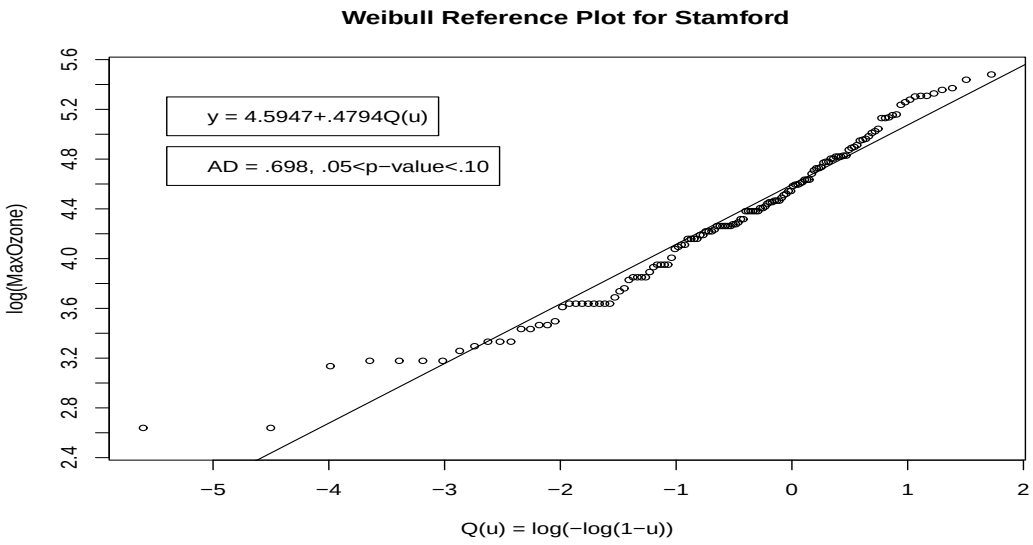
From the Anderson-Darling test we have AD=.698 with $.05 < p - value < .10$ for the Stamford ozone data and AD=.572 with $.10 < p - value < .25$ for the Yonkers ozone data. Thus, there is a moderately good fit of a Weibull model for the Stamford data and a good fit for the Yonkers data. This is depicted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

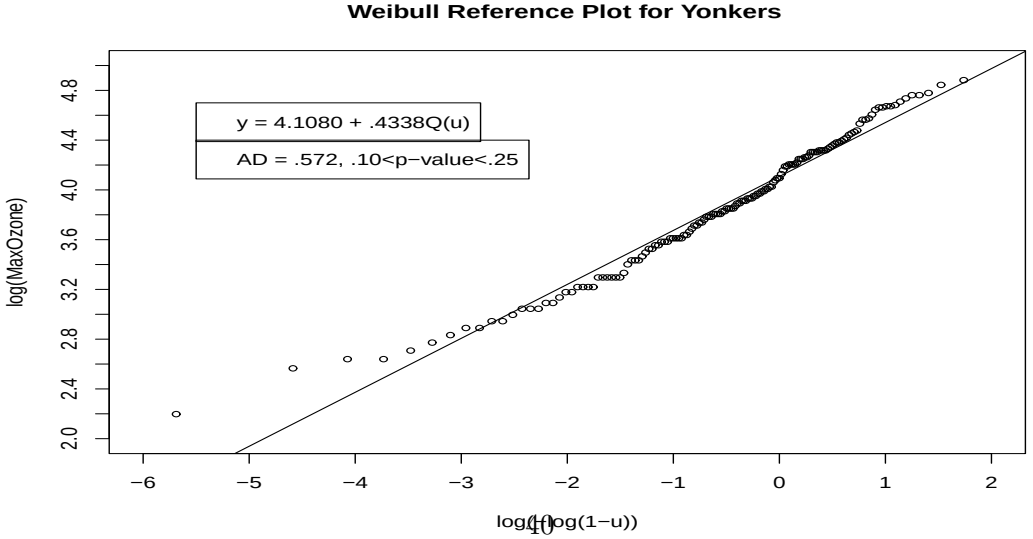
Stamford n=136 AD=0.698 .05 < p-value < .10

Yonkers n=148 AD=0.572 .10 < p-value < .25

Using KS test in R, we have ks.test(x1,"pweibull",1.8338,101.3680) yields a p-value=0.6011 and

ks.test(x2,"pweibull",2.0811,61.9644) yields a p-value=0.7833. These values are only approximations because estimates were used for the values of the scale and shape parameters.


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. The Box-Cox transformation of the data will often yield a data set which is approximately normally distributed. The Box-Cox transformation is for data sets which have only positive values. A modification of the their procedure for data which is both positive and negative is given in the article, Yeo and Johnson(2000), "A new family of power transformations to improve normality or symmetry". *Biometrika(2000)*, **87**, 4, pp. 954-959.

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

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

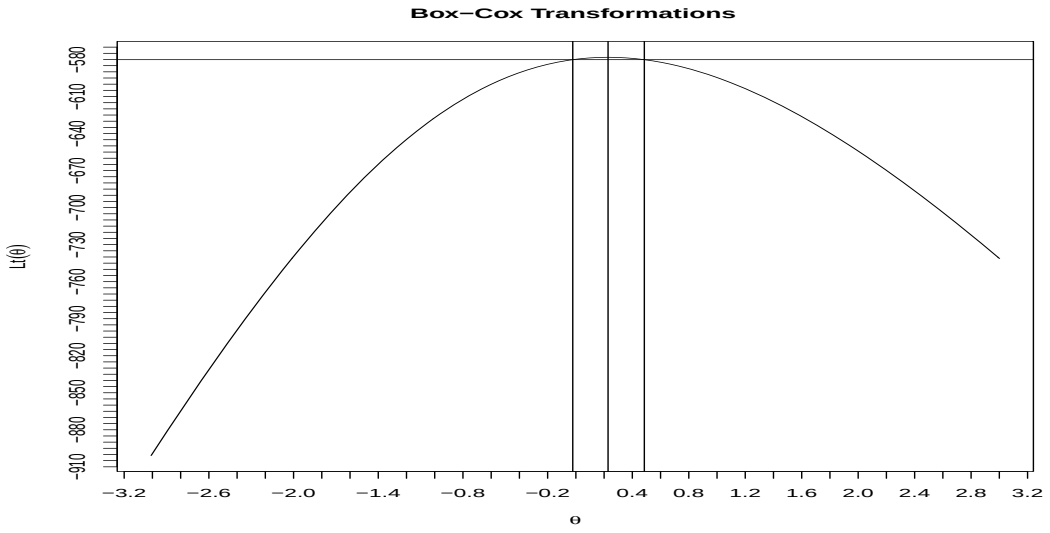
This line in "most cases" cut the curve at two values of θ and these values will be the endpoints of the approximate C.I.

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)
Ltmax
imax= which(Lt==max(Lt))
thmax= th[imax]
thmax

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:

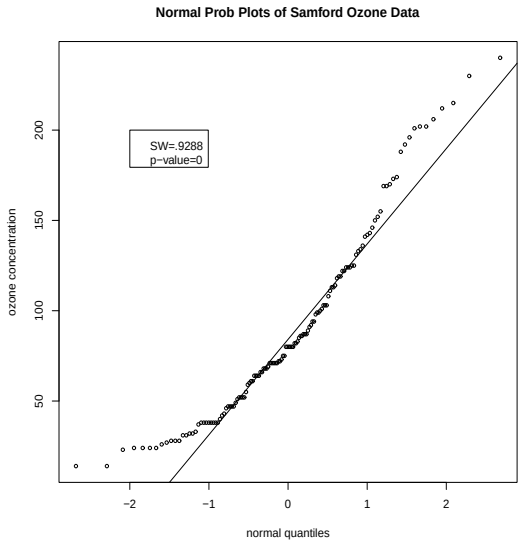
$$\theta = .23; \quad \theta = 0 \Rightarrow \quad \text{log transformation}; \quad \theta = .5 \Rightarrow \quad \text{square root transformation}$$

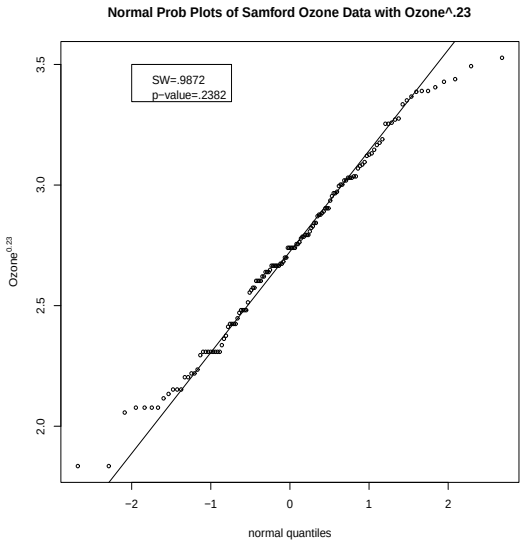
```
qqnorm(x,main="Normal Prob Plots of Samford Ozone Data",
                 xlab="normal quantiles",ylab="ozone concentration",cex=.65)
qqline(x)
text(-2,200,"SW=.9288")
text(-2,190,"p-value=0")

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)
```

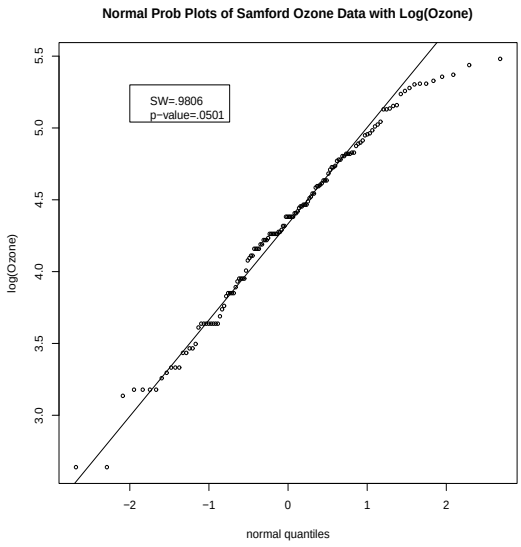
```
qqnorm(y2,main="Normal Prob Plots of Samford Ozone Data with (Ozone)^.23",
                 xlab="normal quantiles",ylab=expression(Ozone^.23),cex=.65)
          qqline(y2)
text(-2,3.5,"SW=.9872")
text(-2,3.4,"p-value=.2382")
```

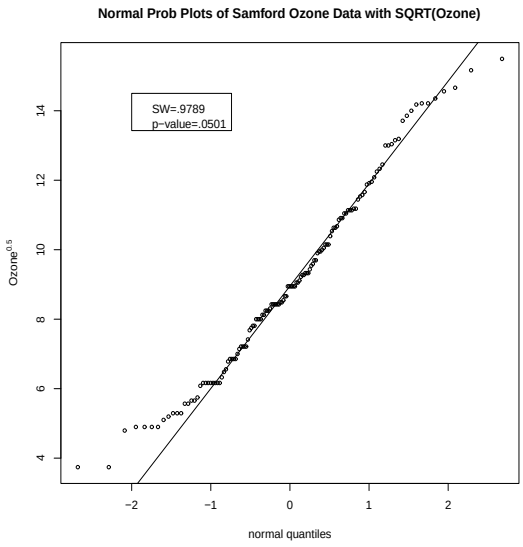
```
qqnorm(y1,main="Normal Prob Plots of Samford Ozone Data with Log(Ozone)",
                 xlab="normal quantiles",ylab="Log(Ozone)",cex=.65)
          qqline(y1)
text(-2,5.0,"SW=.9806")
text(-2,4.85,"p-value=.0501")
```

```
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")
```

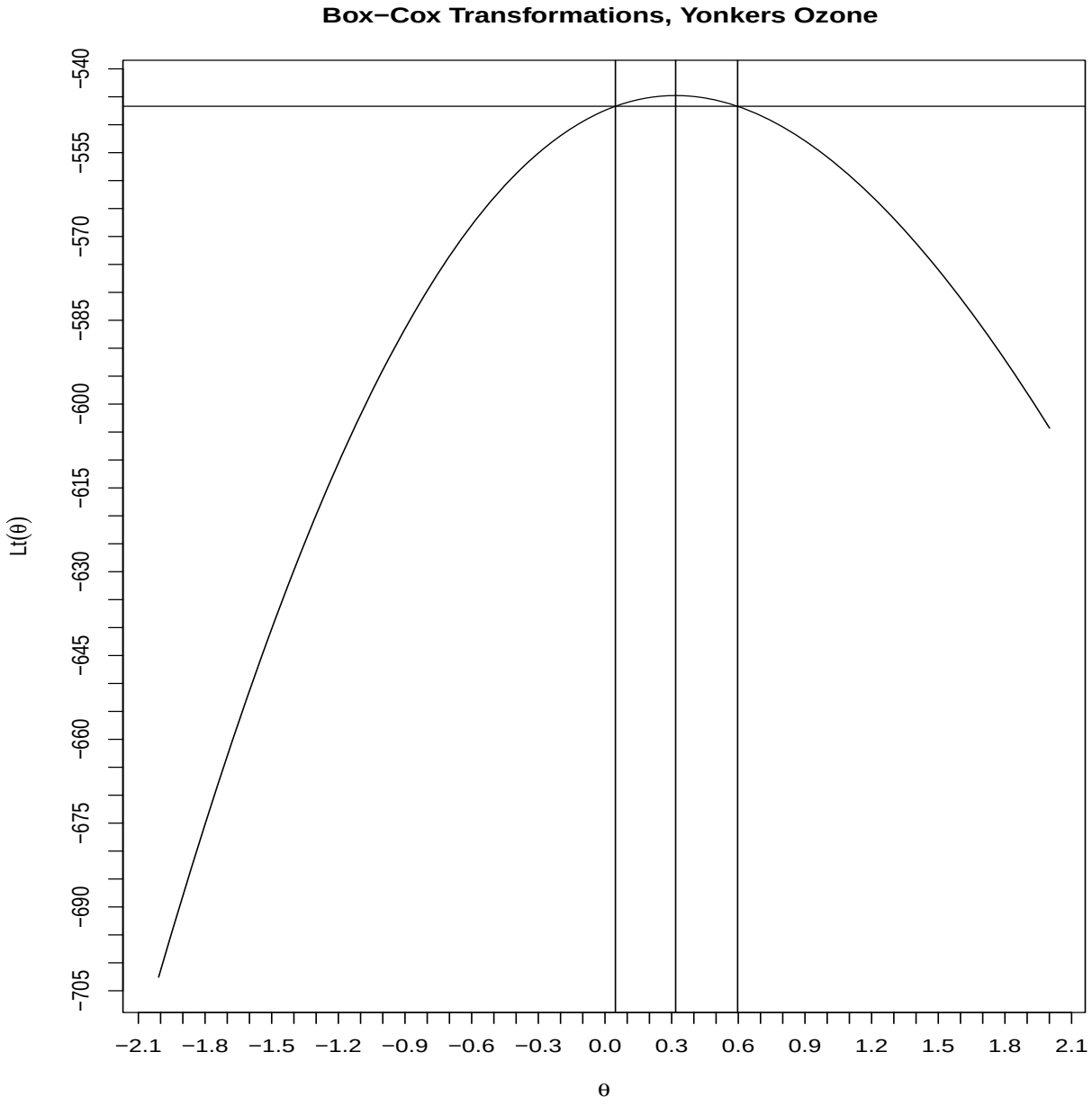

Normal Prob Plots of Samford Ozone Data
SW=.9288
p−value=0
ozone concentration
normal quantiles


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

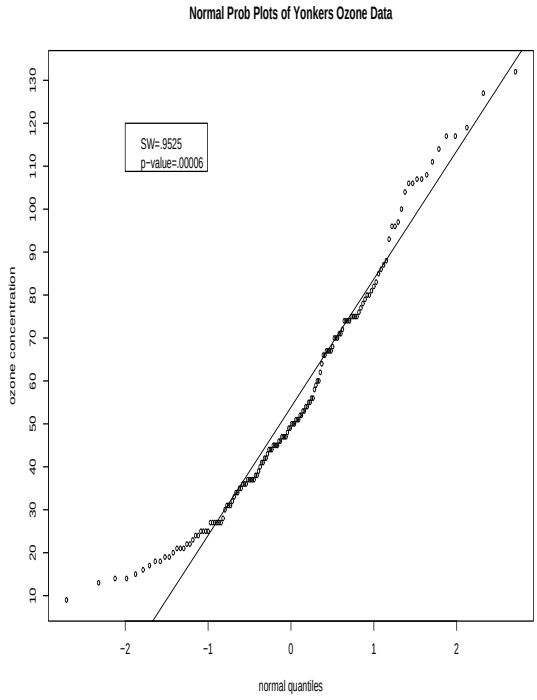

Normal Prob Plots of Samford Ozone Data with Log(Ozone)
SW=.9806
p−value=.0501
log(Ozone)
normal quantiles

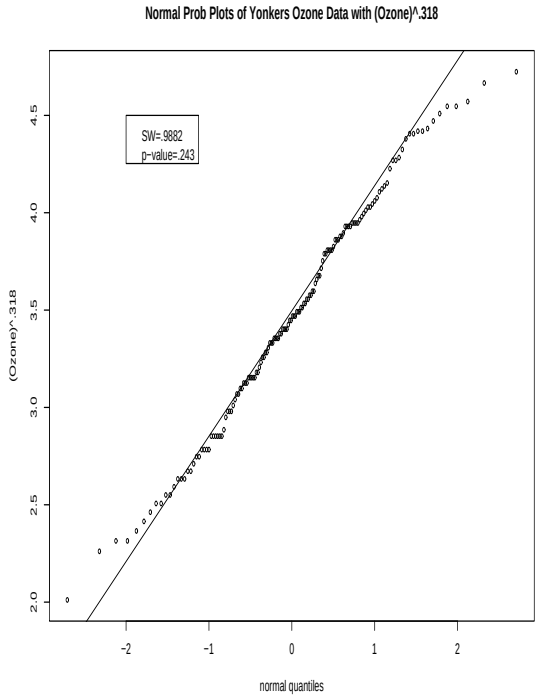

Normal Prob Plots of Samford Ozone Data with SQRT(Ozone)
SW=.9789
p−value=.0501
Ozone^.5
normal quantiles

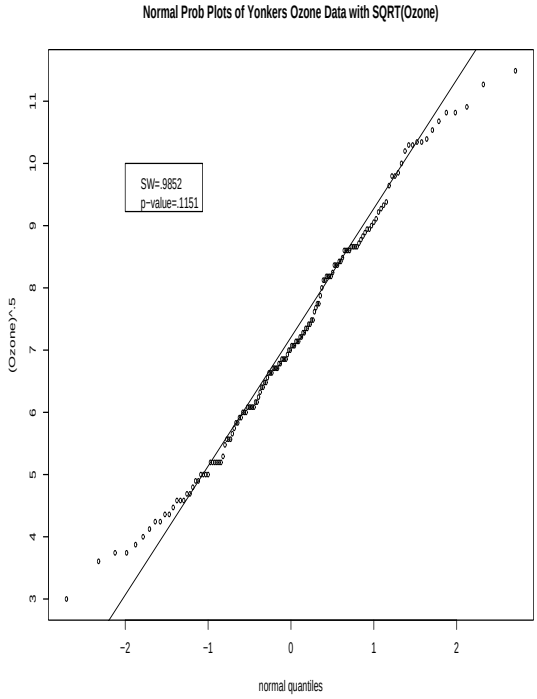
The plot of the likelihood function for the Yonkers ozone readings is given here with lines indicating a 95% confidence interval on values of θ which maximize the likelihood function. We have that $\hat{\theta} = .32$

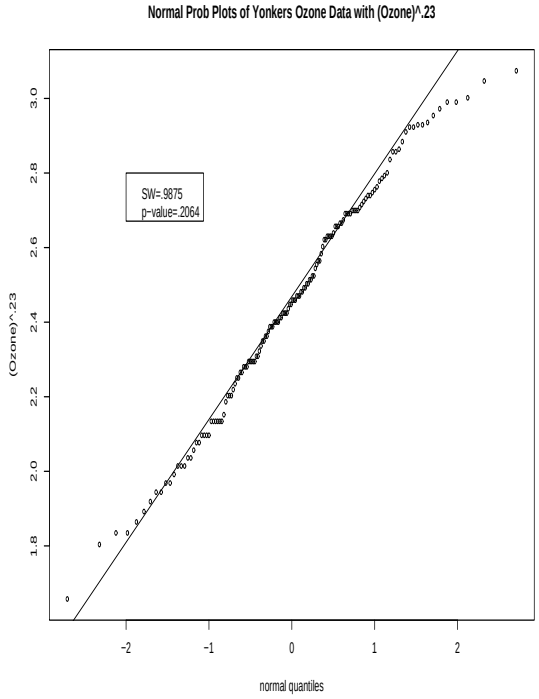


from the R output we obtain a 95% C.I. for θ (thLci,thUci)= $(.047, .60)$.


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
(Ozone)^.318
normal quantiles


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


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

$y^{23}_{Samford}$ has a p-value = .238 from SW test and $y^{23}_{Yonkers}$ has a p-value = .206 from SW test

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.

START Monday 10/18/21

HANDOUT #10: SAMPLING DISTRIBUTIONS

1. Sampling Distributions
   (a) Definition
   (b) Expected Value and Variance of Statistic
   (c) MSE and Bias

2. Methods for Determining Sampling Distribution:
   (a) Enumeration of All Possible Outcomes: Age of Coin Example
   (b) Theoretical Sampling Distribution: Mathematical Derivation
   (c) Simulation Results: Simulate Realizations from Specified Distributions
   (d) Asymptotic Results: Various Central Limit Theorems
   (e) Bootstrap Methods: Resampling From An Observed Sample

3. Sampling Distribution of
   (a) Sample Mean - $\bar{Y}$
   (b) Sample Median - $\widehat{Q}(.5) = \tilde{Y}$
   (c) Sample Standard Deviation - S
   (d) Sample Proportion - $\hat{p}$
   (e) Sample Quantiles - $\widehat{Q}(u)$
   (f) Maximum Likelihood Estimator - $\widehat{\theta}$

Supplemental Reading:

- Chapter 5, Sections 6.1, 14.6, 15.1 in Tamhane/Dunlop book

Sampling Distributions

Definition of Sampling Distribution of a Statistic:

After obtaining an estimator of a population parameter $(\mu, \sigma, Q(u))$ or a parameter associated with a particular family of distributions (β in an exponential family or β and γ in an Weibull family), the **Sampling Distribution of the estimator** is the distribution of the estimator as its possible realizations vary across all possible samples that may arise from a given population.

For example, let θ be a parameter for a population or for a family of distributions. Let $Y_1, \ldots, Y_n$ be iid random variables with cdf $F(\cdot, \theta)$, where θ is a vector of unknown parameters. Let $\hat{\theta} = \hat{\boldsymbol{\theta}}(Y_1, \ldots, Y_n)$ be an estimator of θ based on the observed data. We want to assess how well does $\hat{\theta}$ estimate θ. Some measures of this assessment are given here:

- Concentration of the values of $\hat{\theta}$ about θ:

  What is the chance that $\hat{\theta}$ will be close to θ? That is,

  Compute $P[|\hat{\theta} - \theta| < \epsilon]$ for small values of ϵ.

- On the average does $\hat{\theta}$ equal θ?

  Compute the **Bias** of using $\hat{\theta}$ as an estimator θ:

  Bias$=E[\hat{\theta}] - \theta$.

  If Bias=0, we state that $\hat{\theta}$ is an **unbiased estimator** of θ.

- On the average is $\hat{\theta}$ close to θ?

  Compute the average squared distance from $\hat{\theta}$ to θ, the **Mean Squared Error**:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias]^2$$

In order to calculate the above quantities, we need to know the distribution of $\hat{\theta}$ over all possible samples from the population. This would be nearly impossible for a large population but we can envision the procedure as follows:

- Take M samples of size n from the population

  Sample 1: $X_{11}, \ldots, X_{1n}$ then compute $\hat{\theta}_1$

  Sample 2: $X_{21}, \ldots, X_{2n}$ then compute $\hat{\theta}_2$

  ...

  Sample M: $X_{M1}, \ldots, X_{Mn}$ then compute $\hat{\theta}_M$

- Estimate the distribution of $\hat{\theta}$ using the M realizations of $\hat{\theta}$: $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$.

- We could then estimate the cdf of $\hat{\theta}$, its mean, variance, bias, MSE, etc.

In nearly all situations, this procedure is impossible because of cost, time, or the mere impossibility of every being able to take enough repeated samples from a fixed population. We can overcome this problem in a number of ways using one or a combination of the following procedures:

I. Mathematical Derivations: Consider the following examples

1. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then the distribution of $\hat{\mu} = \bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$

2. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then the distribution of $\hat{\sigma} = S$ can be obtained from the result

   $(n-1)S^2/\sigma^2$ is distributed chi-square with $df = n - 1$.

3. If $T_1, \ldots, T_n$ are iid $Exp(\beta)$, then then the distribution of $n\hat{\beta} = \sum_{i=1}^{n} T_i$ is distributed $Gamma(n, \beta)$

4. If $Y_1, \ldots, Y_n$ are iid Bernoulli with unknown value p, then the distribution of $\hat{p} = \bar{Y}$ is obtained from the result that $n\hat{p} = \sum_{i=1}^{n} Y_i$ is distributed $B(n, p)$

II. Asymptotic theory (large n) can also be applied in many situations.

1. Various versions of the central limit gives approximations to the sampling distributions of $\bar{Y}$, $\hat{p}$, $\hat{\sigma}$, $\hat{Q}(.5)$, etc.

2. If $\hat{\theta}$ is a MLE of θ then there are versions of the central limit theorem which describe the sampling distribution of $\hat{\theta}$.

3. The extreme value distributions can be used to approximate the sampling distributions of the sample minimum $Y_{(1)}$ and maximum $Y_{(n)}$.

III. Simulation studies provide some insight to the sampling distribution but are limited in that we must specify the population distribution exactly in order to conduct the simulation. Consider the following example:

Suppose we wanted to determine the sampling distribution of the estimators of (θ_1, θ_2) when sampling from a population having a Cauchy distribution. We can simulate observations from a Cauchy distribution using R but first we must design the simulation study.

1. How many different values of the sample size n will be needed?
2. How many different values of the location parameter θ_1 will be needed?
3. Which values of θ_1 should be selected?
4. How many different values of the scale parameter θ_2 will be needed?
5. Which values of θ_2 should be selected?
6. How many replications of the simulation are needed for each choice of (n, θ_1, θ_2)
7. How can we infer the sampling distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ for values of (n, θ_1, θ_2) not run in the simulation study.

IV. Suppose we have a random sample of n units from a population with n of a modest size. We want to determine the sampling distribution of the sample median. The population distribution is completely unknown hence a simulation study can not be utilized. The sample size is too small to have much confidence in applying the central limit theorem for the median. A possible method for obtaining an approximation to the sampling distribution of the statistics is to use a resampling procedure. This involves taking numerous samples of size n (with replacement) from the actual observed data and computing the value of the statistic from each of these samples. One such procedure is called the **bootstrap sampling procedure**.

There are many ways to demonstration the sampling distributions of various sample statistics. We will first have a physical demonstration involving the age of pennies.

Physical Demonstration of Sampling Distribution of $\bar{X}$

The following example will demonstrate the central limit theorem through the use of the ages of 500 pennies.

I have a collection of 500 pennies with X equal to the Age of penny: $X = 2019$ - (Date on penny).

What general shape do you think a histogram of age, X, would have? Why?

Stem and Leaf Plot for the Ages of 500 Pennies

37 - 0	1111111111111111111111111111111111111
85 - 0	22222222222222222222222222222222222222222222233333333333333333333333333333333333333333333
38 - 0	44444444444444444444445555555555555555
44 - 0	66666666666666666666677777777777777777777
37 - 0	88888888888888888888888899999999999
18 - 1	000000000000011111
38 - 1	22222222222222223333333333333333333
46 - 1	44444444444444444444455555555555555555555555
30 - 1	666666666677777777777777777
20 - 1	88888888888888999
21 - 2	000000000000000111111
18 - 2	222222333333333333
18 - 2	444444445555555555
17 - 2	66667777777777777
9 - 2	888889999
1 - 3	1
8 - 3	33333333
8 - 3	44444555
4 - 3	6777
2 - 3	88
1 - 4	0

This is a very right skewed distribution with five-number summary of the distribution:

$$\text{Minimum } = 1, \qquad Q_1 = 5, \qquad M = Q_2 = 10, \qquad Q_3, = 18, \qquad \text{Maximum } = 40$$

The mean and standard deviation for this population of all 500 pennies in the population were computed to be:

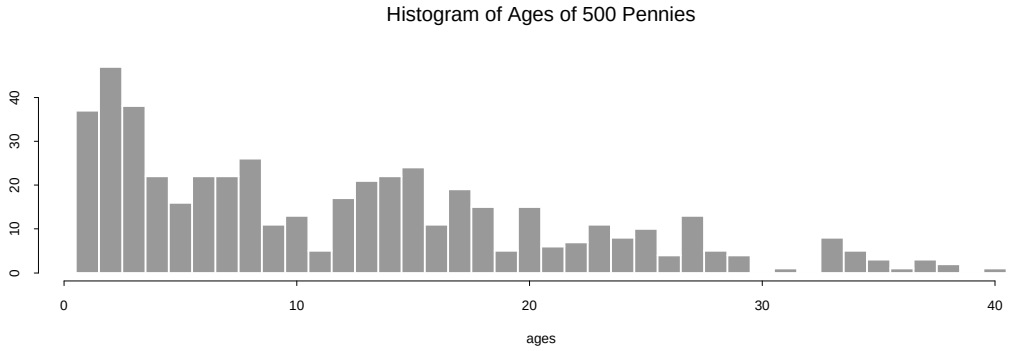
$$\mu_X = 12.038 \text{ and } \sigma_X = 9.281$$

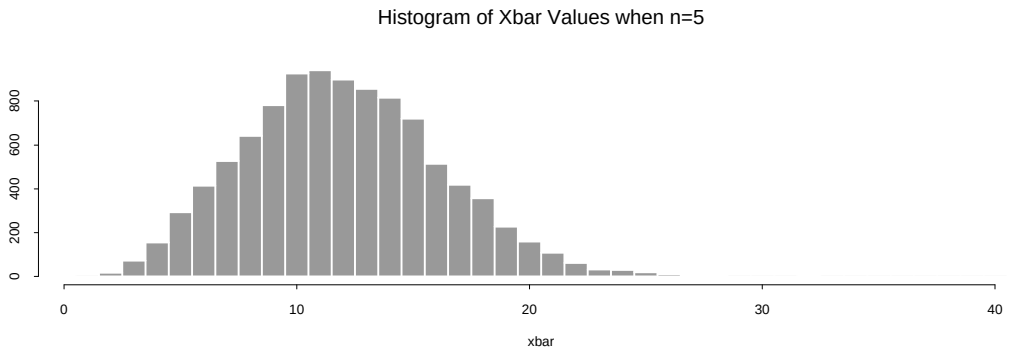
Sampling Distribution of Sample Mean, $\bar{X}$: To obtain the histogram of the sample means for *all* possible samples of a selected size n would be extremely difficult because of the enormous number of possible samples of size n:

n	Number of Possible Samples of Size n: $\binom{500}{n}$
5	255,244,687,600
10	2.458×10^{20}
20	2.667×10^{35}

Because of the enormous number of possible samples of size n needed to completely describe the sampling distribution of $\bar{X}$, approximations to the sampling distribution of $\bar{X}$ were generated by taking 10,000 random samples of size 5, 10, and 20 from the population of 500 coins.

Compare the shapes and spreads of the sampling distributions of $\bar{X}$ with the distribution of the original population.





The shape of the sampling distribution of $\bar{X}$ tends towards a symmetric distribution (normal distribution) as n increases from n=1 to 20 with its distribution concentrating near $\mu = 12$.

Sampling Distribution of $\bar{X}$, $\hat{Q}(.5)$, S, $\hat{p}$

Let $X_1, \ldots, X_n$ be iid random variables with cdf F having $\mu, \sigma < \infty$ for its mean and standard deviation and $\mu_3 < \infty$ and $\mu_4 < \infty$ as its 3rd and 4th central moments.

Sampling Distribution of the Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- For all n : If we sample with replacement,

  $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \frac{\sigma^2}{n}$

- For all n : If we sample without replacement,

  $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} < \frac{\sigma^2}{n}$,

  where N is the population size and $\left(\frac{N-n}{N-1}\right)$ is called the finite population correction factor (fpcf).

- For all values of n : If F is $N(\mu, \sigma^2)$, then the distribution of $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$.

- The Central Limit Theorem yields for large n : The distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges to a $N(0,1)$ as $n \to \infty$, provided $\sigma < \infty$. That is, for large n, the distribution of $\bar{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$.

- The accuracy of the approximate distribution of $\bar{X}$ depends on the size of n and the shape of F. The larger the value of n, the better the approximation. For a fixed value of n, the more symmetric with normal-like tails that F is, the greater the accuracy of the approximation.

- If the data is correlated, then $\bar{X}$ is still an unbiased estimator of μ but the variance of $\bar{X}$ may be larger or smaller than $\frac{\sigma^2}{n}$ depending on the type of correlation in the data.

- For example, suppose $corr(X_i, X_j) = \rho$ for all pairs $i \neq j$, (equi-correlated), where $\frac{-1}{n-1} < \rho < 1$. Then,

$$E(\bar{X}) = \mu \quad \text{but} \quad Var(\bar{X}) = \frac{\sigma^2}{n}[1 + (n-1)\rho] \Rightarrow \begin{cases} Var(\bar{X}) < \frac{\sigma^2}{n} & \text{if } \rho < 0 \\ Var(\bar{X}) > \frac{\sigma^2}{n} & \text{if } \rho > 0 \end{cases}$$

To verify the above, recall the following result for constants $c_1, c_2, \ldots, c_n$ and random variables $X_1, X_2, \ldots, X_n$:

$$Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 Var(X_i) + \sum_{i \neq j} c_i c_j Cov(X_i, X_j)$$

- Second example: Suppose that the $X_i s$ follow an AR(1) process: $X_t = \theta + \rho X_{t-1} + e_t$, where $X_t s$ are stationary and $e_t s$ are iid with $E[e_t] = 0, Var(e_t) = \sigma_e^2$, with $-1 < \rho < 1$. Then, we have

$$\mu_t = E[X_t] = \theta/(1-\rho), \qquad \sigma_t^2 = Var(X_t) = \sigma_e^2/(1-\rho^2), \qquad Corr(X_i, X_j) = \rho^{|i-j|} \text{ for } i \neq j,$$

The values of μ_t and σ_t are constant and do not depend on t. The amount of correlation decreases as time or distance between two units increases.

We have $E(\bar{X}) = \mu$ but

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left[1 + \frac{2}{n}\left(\frac{\rho}{1-\rho}\right)\left(n + \frac{1-\rho^n}{1-\rho}\right)\right] \approx \frac{\sigma^2}{n}\left[\frac{1+\rho}{1-\rho}\right] \Rightarrow$$

$$Var(\bar{X}) < \tfrac{\sigma^2}{n} \quad \text{if} \quad \rho < 0$$

$$Var(\bar{X}) > \tfrac{\sigma^2}{n} \quad \text{if} \quad \rho > 0$$

- For *iid* data, $\widehat{SE}(\bar{X}) = S/\sqrt{n}$

This estimate would underestimate $SE(\bar{X})$ if $X_i s$ are AR(1) with $\rho > 0$ which would result in a confidence interval for μ having coverage probability much less than the stated level, e.g., coverage probability of 80% for a stated 95% C. I.

A better estimator would be $\widehat{SE}(\bar{X}) = \frac{S}{\sqrt{n}}\sqrt{\frac{1+\hat{\rho}}{1-\hat{\rho}}}$

For example, consider the ozone data and suppose that an AR(1) model is adequate. Then, we have the following results

City	n	S	$\hat{\rho}$	$S/\sqrt{n}$	$\widehat{SE}(\bar{X})$	% $S/\sqrt{n}$ is less than $\widehat{SE}(\bar{X})$
Yonkers	148	28.11	0.4342	2.31	3.68	59.3%
Stamford	136	52.11	0.3342	4.47	6.33	41.6%

Sampling Distribution of the Sample Median: $\tilde{\mu} = \hat{Q}(.5)$

Let $\hat{Q}(.5)$ be the sample median given by

$$\hat{Q}(.5) = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ .5\left(X_{(\frac{n}{2})} + X_{(\frac{n+2}{2})}\right) & \text{if } n \text{ is even} \end{cases}$$

- The moments of $\hat{Q}(.5)$ depend on F and the joint distribution of the order statistics: For n odd

$$m_k = E\left[\left(\hat{Q}(.5)\right)^k\right] = E\left[X^k_{(\frac{n+1}{2})}\right] = n\binom{n-1}{\frac{n-1}{2}} \int_{-\infty}^{\infty} x^k \left[F(x)[1-F(x)]\right]^{\frac{n-1}{2}} f(x)dx$$

- The mean of $\hat{Q}(.5)$ is m_1 and the variance is $m_2 - m_1^2$,.

  Both of which depend on the population cdf F

  Similar results are obtained for n even.

- For large n : The distribution of $\hat{Q}(.5)$ is approximately $N\left(Q(.5), \frac{(.5)^2}{n(f(Q(.5))^2}\right)$, thus, the asymptotic mean and standard deviation for the sample median $\hat{Q}(.5)$ are given by

$$\mu_A = Q(.5) \text{ and } \sigma_A = \frac{.5}{f(Q(.5))\sqrt{n}} = \frac{.5/f(Q(.5))}{\sqrt{n}}.$$

  That is, $\hat{Q}(.5)$ is asymptotically unbiased but for small n it would be a biased estimator of Q(.5) in most situations.

- For estimating the location parameter θ of a symmetric distribution with $\sigma < \infty$,

  should we use $\bar{X}$ or $\hat{Q}(.5)$?

When to use $\bar{X}$ or $\widehat{Q}(0.5)$ — Estimator is more efficient when its variance is smaller

Definition: The Asymptotic Relative Efficiency (ARE) of $\widehat{Q}(.5)$ to $\bar{X}$ as an estimator of θ is given by

$$ARE(\hat{Q}(.5), \bar{X}) = \frac{\text{asymptotic variance of } \bar{X}}{\text{asymptotic variance of } \widehat{Q}(.5)} = \frac{\sigma^2/n}{\left(\frac{.5/f(Q(.5))}{\sqrt{n}}\right)^2} = 4\sigma^2 f^2(Q(.5))$$

When $ARE < 1$, the Sample Mean= $\bar{X}$ is a more efficient estimator of θ than the Sample Median=$\widehat{Q}(.5)$

1. F has a $N(\theta, \sigma^2)$, then $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{(x-\theta)^2}{2\sigma^2}}$, $Q(.5) = \theta$ and $f(Q(.5)) = f(\theta) = \frac{1}{\sigma\sqrt{2\pi}}$

   $\text{ARE} = 4\sigma^2 f^2(Q(.5)) = 4\sigma^2 \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 = \frac{2}{\pi} \approx .64 < 1$

- Sample Mean is more efficient estimator of θ than is the Sample Median

2. F has a uniform on $(\theta-a, \theta+a)$, then $f(x) = \frac{1}{2a} I(\theta-a < x < \theta+a)$, $Q(.5) = \theta$, $\sigma^2 = \frac{a^2}{3}$

   $f(Q(.5)) = f(\theta) = \frac{1}{2a}$ which yields

   $\text{ARE} = 4\sigma^2 f^2(Q(.5)) = 4\frac{a^2}{3}\frac{1}{4a^2} = \frac{1}{3} < 1$

- Sample Mean is more efficient estimator of θ than is the Sample Median

3. F is a logistic(θ_1, θ_2), then $f(x) = \frac{e^{-(x-\theta_1)/\theta_2}}{\theta_2(1+e^{-(x-\theta_1)/\theta_2})^2}$, $Q(.5) = \theta_1$, $\sigma^2 = \frac{\pi^2\theta_2^2}{3}$, and

   $f(Q(.5)) = f(\theta_1) = \frac{1}{4\theta_2}$ which yields

   $\text{ARE} = 4\sigma^2 f^2(Q(.5)) = 4\frac{\pi^2\theta_2^2}{3}\left(\frac{1}{4\theta_2}\right)^2 = \frac{\pi^2}{12} \approx .82 < 1$

- Sample Mean is more efficient estimator of θ than is the Sample Median

4. F has a shifted t distribution with two parameters: shift θ_1 and shape $\theta_2 = df > 2$,

   then $Q(.5) = \theta_1$, $\sigma^2 = \frac{\theta_2}{\theta_2-2}$, and $f(Q(.5)) = f(\theta_1) = \frac{\Gamma(\frac{\theta_2+1}{2})}{\Gamma(\frac{\theta_2}{2})\sqrt{\pi\theta_2}}$ which yields

$$\text{ARE}=4\sigma^2 f^2(Q(.5)) = 4\left(\frac{\theta_2}{\theta_2-2}\right)\left[\frac{\Gamma(\frac{\theta_2+1}{2})}{\Gamma(\frac{\theta_2}{2})\sqrt{\pi\theta_2}}\right]^2$$

θ_2	3	4	5	8	∞
ARE	1.62	1.12	0.96	0.8	0.64

For $2 < df \leq 4$, the Median is a more efficient estimator of the shift parameter, θ_1 than the Mean.

For $5 < df$, the Mean is a more efficient estimator of the shift parameter, θ_1 than the Median.

Thus, for a heavy-tailed, symmetric distribution, the Sample Median is a more efficient (less variable) estimator of the location parameter in comparison to the Sample Mean.

Sampling Distribution of the Sample Standard Deviation: S

Let $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ be the sample standard deviation.

- For all n : $E[S^2] = \sigma^2$
- For all n : $E[S] \neq \sigma$. Why?
- The $Var[S]$ depends on F
- For all n : If F is $N(\mu, \sigma^2)$, then the distribution of $\frac{(n-1)S^2}{\sigma^2}$ is chi-square with $df = n - 1$
- For all n : If F is $N(\mu, \sigma^2)$, then
  - $E[S] = \left[\sqrt{\frac{2}{n-1}}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)\right]\sigma = [c_n]\sigma, \quad \text{with} \quad c_n\sigma \to \sigma \quad \text{as} \quad n \to \infty$
  - $Var[S] = \sigma^2(1 - c_n^2) \quad \text{with} \quad \sigma^2(1 - c_n^2) \to 0 \quad \text{as} \quad n \to \infty$
- The bias in using S as an estimator of σ when the data is from a normally distributed population: Bias $= E[S] - \sigma = (c_n - 1)\sigma$

n	2	3	4	5	10	25	100	250
Bias	-.202σ	-.114σ	-.079σ	-.060σ	-.027σ	-.0104σ	-.00252σ	-.00100σ

- For large n : The distribution of S is approximately $N\left(\sigma, \frac{\mu_4 - \sigma^4}{4n\sigma^2}\right)$, thus, the asymptotic mean and standard deviation for the sample standard deviation S are given by

$$\mu_A = \sigma \quad \text{and} \quad \sigma_A = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}}.$$

- How accurate is this approximation? As is true for all asymptotic results, it depends on n and the population distribution F. A simulation study will demonstrate the accuracy of using the normal approximation in place of the true sampling distribution of $\bar{X}$, $\hat{Q}(.5)$, and S.
- How does either spatial or temporal correlation affect S as an estimator of σ?

  Let $Y_1, Y_2, \ldots, Y_n$ be stationary random variables with mean μ, variance σ^2, and covariance function $R(k) = E[(Y_t - \mu)(Y_{t+k} - \mu)]$. Then we have the following results.

  - $E[\bar{Y}] = \mu$ $\qquad$ $E[S^2] = \sigma^2 = R(0)$
  - $Var(\bar{Y}) = \frac{1}{n}\sum_{i=-n}^{n}\left(1 - \frac{i}{n}\right)R(i)$ $\qquad$ $Var(S^2) = \frac{2}{n}\sum_{i=-n}^{n}\left(1 - \frac{i}{n}\right)R^2(i)$

Sampling Distribution of the Sample Quantiles: $\hat{Q}(u)$

Let $\hat{Q}(u)$ be the sample quantile for values of u not too close to 0 or 1.

- The quantities $E[\widehat{Q}(u)]$ and $Var[\widehat{Q}(u)]$ depend on F through the distribution of the order statistics. Thus, we can compute these values using our smoothed definition of the sample quantile function:

  Let $\frac{1}{2n} \leq u \leq 1 - \frac{1}{2n}$ with $nu + .5 = k + r$ where

  $k = 1, \cdots, n-1$ and $0 < r < 1$ then we define

  $$\widehat{Q}(u) = Y_{(k)} + r[Y_{(k+1)} - Y_{(k)}]$$

  The pdf of the *kth* order statistic, $Y_{(k)}$ is given by

  $$f_{(k)}(y) = \frac{n!}{(k-1)!(n-k)!} [F(y)]^{k-1} [1 - F(y)]^{n-k} f(y)$$

  From which we can compute the mean of $Y_{(k)}$ which then yields

  $$E\left[\widehat{Q}(u)\right] = (1-r)E\left[Y_{(k)}\right] + E\left[Y_{(k+1)}\right]$$

  Similarly, using the joint pdf of two order statistics, we can obtain the covariance between any two order statistics. The $Var[\widehat{Q}(u)]$ can then be computed using the expressions for the variances and covariance of the order statistics.

- For large n : The distribution of $\hat{Q}(u)$ is approximately $N\left(Q(u), \frac{u(1-u)}{n(f(Q(u))^2}\right)$, thus, the asymptotic mean and standard deviation for the sample quantile $\hat{Q}(u)$ are given by

  $$\mu_A = Q(u) \qquad \text{and} \qquad \sigma_A = \frac{\sqrt{u(1-u)}}{f(Q(u))\sqrt{n}} = \frac{\sqrt{u(1-u)}/f(Q(u))}{\sqrt{n}}.$$

  For a $N(\mu, \sigma^2)$ distribution, $f(Q(u)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(Q(u)-\mu)^2/2\sigma^2}$

  For the median, $Q(.5) = \mu$ therefore $f(Q(.5)) = \frac{1}{\sigma\sqrt{2\pi}}$ thus

  $\sigma_A = \frac{\sqrt{.5(1-.5)}}{f(Q(.5))\sqrt{n}} = \frac{\sigma\sqrt{\frac{\pi}{2}}}{\sqrt{n}}$. This formula is only for the normal distribution.

  In practice, we would need to estimate $f(Q(u)))$ in order to obtain a value for σ_A.

Sampling Distribution of the Sample Minimum and Maximum

Let $X_1, \ldots, X_n$ iid with cdf F.

Define

$M_n = X_{(n)} = max[X_1, \ldots, X_n]$ as the sample maximum

$m_n = X_{(1)} = min[X_1, \ldots, X_n]$ as the sample minimum.

We then have the following results for m_n and M_n:

- The pdfs and cdfs of M_n and m_n are given by
  - For M_n: $G_n(y) = [F(y)]^n$ and $g_n(y) = nf(y)[F(y)]^{n-1}$
  - For m_n: $H_n(y) = 1 - [1 - F(y)]^n$ and $h_n(y) = nf(y)[1 - F(y)]^{n-1}$
- Thus, we can obtain the moments of M_n and m_n using the above expressions for the pdfs, although in many cases, these calculations are quite difficult.
- Bounds on Expectations: Suppose F has mean and variance μ and $\sigma^2 < \infty$.

  The following are bounds on the means of M_n and m_n:
  - $E(M_n) \leq \mu + \frac{(n-1)\sigma}{\sqrt{2n-1}}$
  - $E(m_n) \geq \mu - \frac{(n-1)\sigma}{\sqrt{2n-1}}$
- Asymptotic Results: Let $F_n(y)$ be the cdf of the standardized form of M_n: $\frac{M_n - a_n}{b_n}$,
  - $F_n(y) = P\left[\frac{M_n - a_n}{b_n} \leq y\right]$.
  - $F_n(y) \to G(y)$ as $n \to \infty$ where $G(y)$ is one of the following:
    - $G_{1,\gamma}(y) = e^{-y^{-\gamma}}$ for $y > 0$ and $G_{1,\gamma}(y) = 0$ for $y < 0$
    - $G_{2,\gamma}(y) = e^{-(-y)^{\gamma}}$ for $y < 0$ and $G_{2,\gamma}(y) = 0$ for $y > 0$
    - $G_3(y) = e^{-e^{-y}}$ for $-\infty < y < \infty$
  - If Y has a Weibull distribution, then W = -log(Y) has cdf $G_3(w) = e^{-e^{-(w-\theta_1)/\theta_2}}$ for $-\infty < y < \infty$
  - The sequences (a_n) and (b_n) and the form of $G(y)$ are determined by the form of the cdf $F(y)$ for the data: $Y_1, \ldots, Y_n$. See *A Course in Large Sample Theory*, by Thomas Ferguson for further details.
  - The asymptotic distributions for the sample minimum m_n can be obtained from the results of the sample maximum M_n by using $X_i = -Y_i$.

STOP Monday 10/18/21

START Wednesday 10/20/21

Sampling Distribution of the Sample Proportion, $\hat{p}$

Suppose we have a population in which all the units in the population are of one of two types: Type A Units or Type B Units. For example, people having a disease or not; or a warehouse containing good parts and defective parts. Let p be the proportion of Type A units in the population. Alternatively, p could be the probability of a particular outcome in a sequence of Bernoulli trials. For example, p is the probability experimental unit no longer has disease after receiving a drug. We apply the drug to n experimental units and observe whether or not they have the disease after a period of time.

Let $\hat{p} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = 1)$ be the sample proportion based on n iid Bernoulli trials or a random sample of n units taken from a population consisting of Type A and Type B units, where $Y_i = 1$ if the ith unit is a Type A unit and $Y_i = 0$ if the ith unit is not a Type A and $Y = \sum_{i=1}^{n} I(Y_i = 1)$ is the total number of Type A outcomes in the sample of n units.

- For sampling with replacement: $n\hat{p}$ has a binomial distribution.
- For sampling without replacement: $n\hat{p}$ has a hypergeometric distribution.
- For both sampling with and without replacement: $E[\hat{p}] = p$.
- For sampling with replacement: $Var[\hat{p}] = \frac{p(1-p)}{n}$
- For sampling without replacement: $Var[\hat{p}] = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$, where N is the number of units in the population.

  When $\frac{n}{N} < .05$, $\frac{N-n}{N-1} \approx 1 - \frac{n}{N} > .95 \quad \Rightarrow \quad Var[\hat{p}] \approx \frac{p(1-p)}{n}$
- For large n : The distribution of $\hat{p}$ is approximately $N\left(p, \frac{p(1-p)}{n}\right)$, thus, the asymptotic mean and standard deviation for the sample standard deviation $\hat{p}$ are given by

$$\mu_A = p \quad \text{and} \quad \sigma_A = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

  .
- The normal approximation to the binomial distribution should be used only when $min[np, n(1-p)] \geq 5$.

  Also, because we are approximating a discrete distribution with a continuous distribution, the following correction is generally suggested:

$$Pr[y_1 \leq \hat{p} \leq y_2] \approx \Phi\left(\frac{ny_2 + .5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{ny_1 - .5 - np}{\sqrt{np(1-p)}}\right),$$

  where Φ is the cdf of a standard normal distribution.

Where does the .5 come from in the approximation?

$$P[\hat{p} = y] = P[n\hat{p} = ny] \approx P[ny - .5 \leq Y \leq ny + .5]$$

where $n\hat{p}$ has a Binomial(n,p) distribution (discrete) and Y has a $N(np, np(1-p))$ distribution which is continuous

Example Let Y have a binomial distribution with n=100, p=0.2

$\mu_Y = E[Y] = np = 100(.2) = 20$ and $\sigma_Y = \sqrt{Var(Y)} = \sqrt{np(1-p)} = \sqrt{100(.2)(.8)} = 4$

Calculate $P(Y \leq 25)$:

- Exact calculation using B(100,.2) distribution:

  $P(Y \leq 25) = pbinom(25, 100, .2) = 0.9125246$

- Normal Approx to Binomial Without Correction:

  $$P(Y \leq 25) \approx P\left(Z \leq \frac{25-20}{\sqrt{100(.2)(.8)}}\right) = pnorn\left(\frac{25-20}{\sqrt{100(.2)(.8)}}\right) = 0.8943502$$

- Normal Approx to Binomial With Correction:

  $$P(Y \leq 25) \approx P\left(Z \leq \frac{25-20+.5}{\sqrt{100(.2)(.8)}}\right) = pnorm\left(\frac{25-20+.5}{\sqrt{100(.2)(.8)}}\right) = 0.9154343$$

- The following graphs illustrates that for the values of Y in the middle of the binomial distribution the normal approximation with the correction is more accurate than the approximation without the correction. However, in the tails of the distribution, there are regions where the calculation with the correction is less accurate than the calculation without the correction. Consider the following example.

- Exact calculation using B(100,.2) distribution:

  $P(Y \leq 12) = pbinom(12, 100, .2) = 0.02532875$

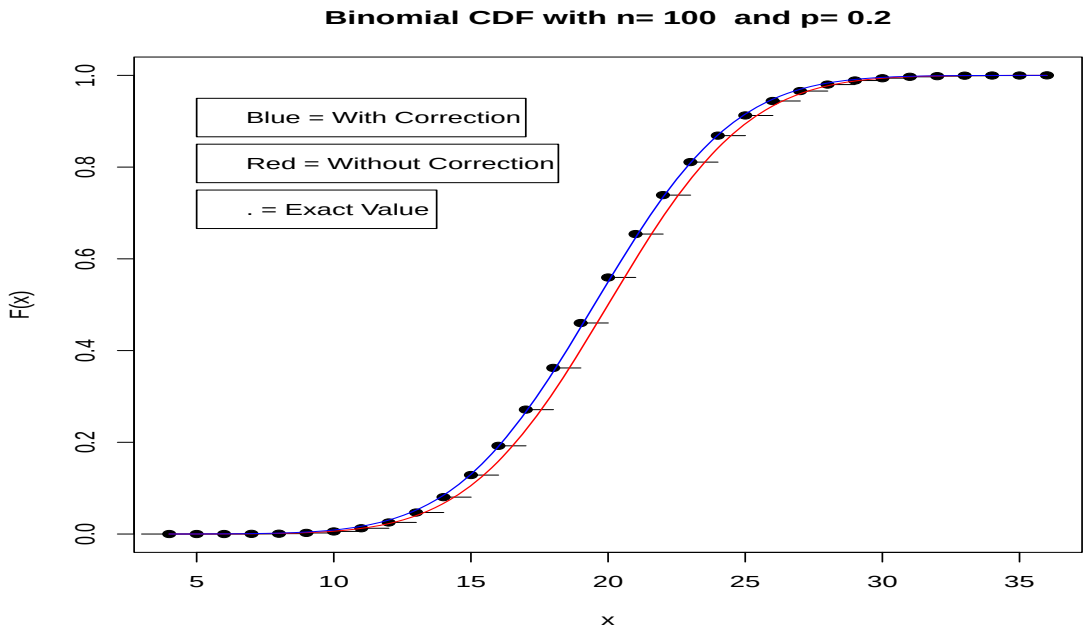
- Normal Approx to Binomial Without Correction:

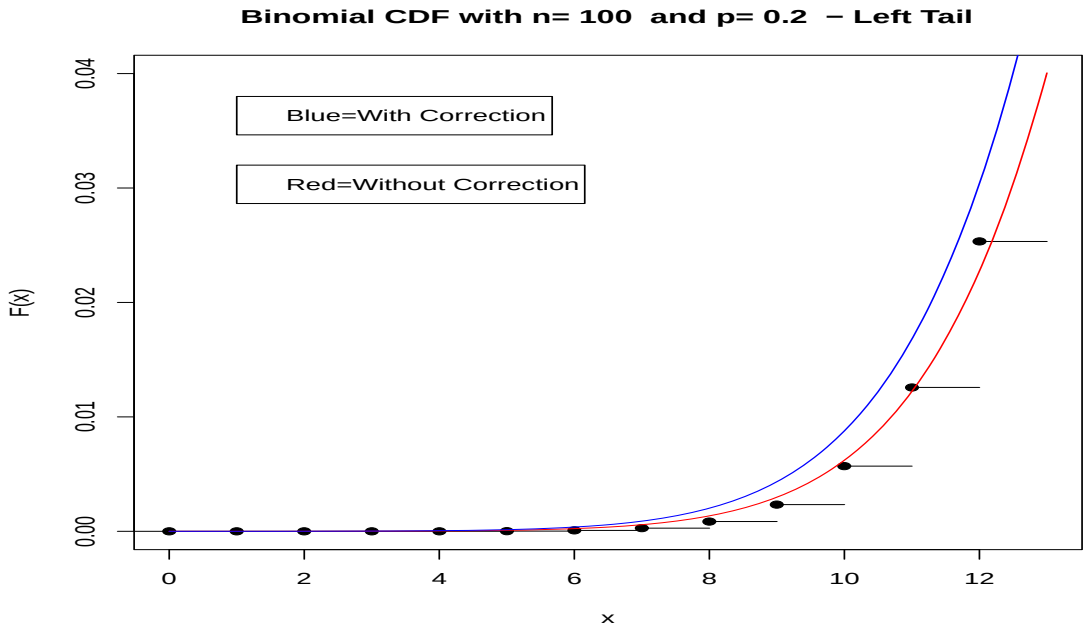
  $$P(Y \leq 12) \approx P\left(Z \leq \frac{12-20}{\sqrt{100(.2)(.8)}}\right) = pnorn\left(\frac{12-20}{\sqrt{100(.2)(.8)}}\right) = 0.02275013$$

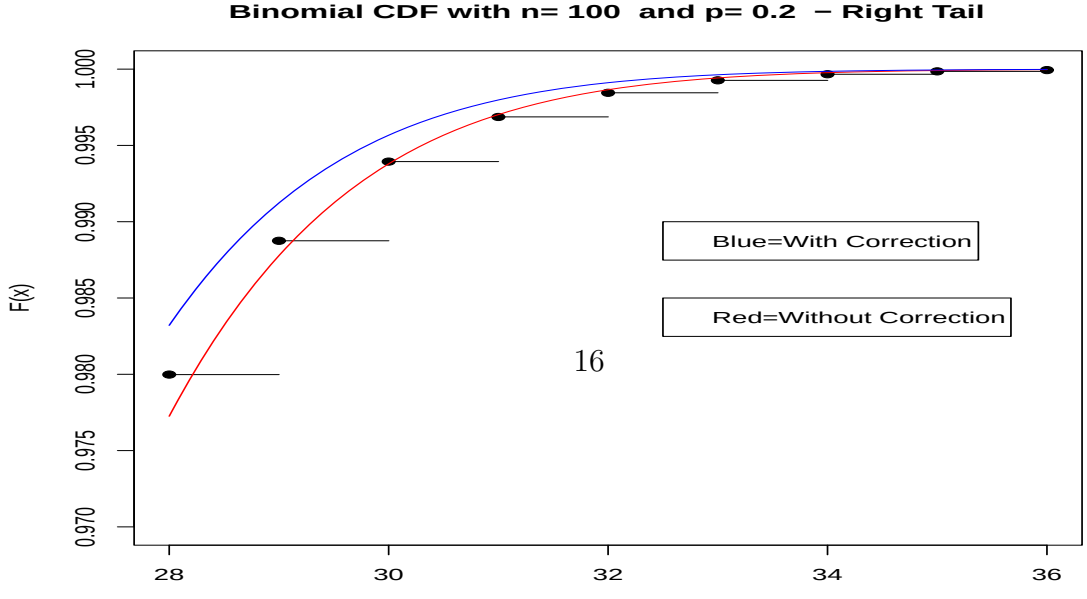
- Normal Approx to Binomial With Correction:

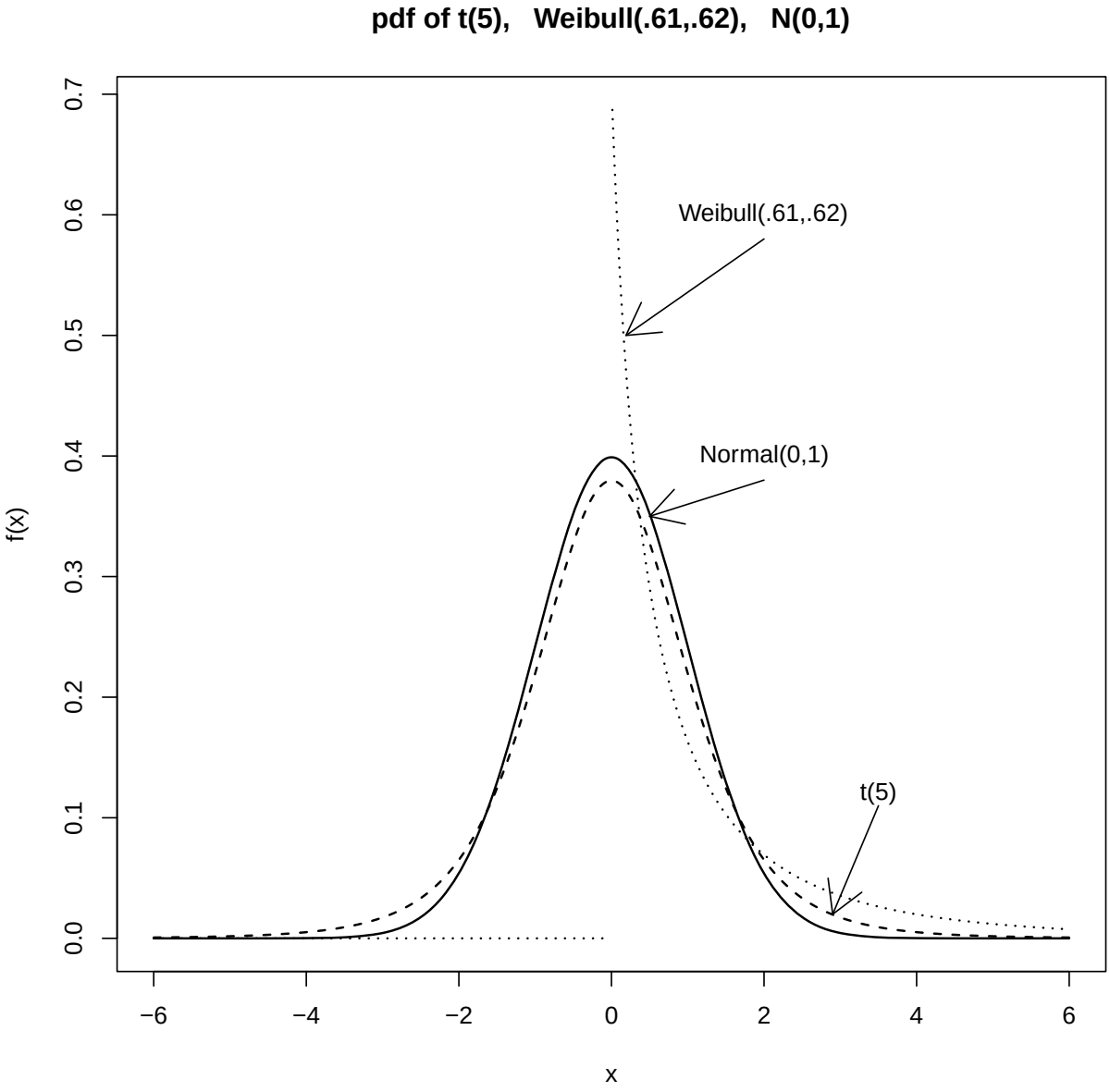
  $$P(Y \leq 12) \approx P\left(Z \leq \frac{12-20+.5}{\sqrt{100(.2)(.8)}}\right) = pnorm\left(\frac{12-20+.5}{\sqrt{100(.2)(.8)}}\right) = 0.03039636$$

  The error **with the correction** is -0.005067609 which is larger that the error **without the correction** .002578621.


Binomial CDF with n= 100 and p= 0.2
Blue = With Correction
Red = Without Correction
. = Exact Value
F(x)
x


Binomial CDF with n= 100 and p= 0.2 – Left Tail
Blue=With Correction
Red=Without Correction
F(x)
x


Binomial CDF with n= 100 and p= 0.2 – Right Tail
Blue=With Correction
Red=Without Correction
F(x)


pdf of t(5), Weibull(.61,.62), N(0,1)
Weibull(.61,.62)
Normal(0,1)
t(5)
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−6
−4
−2
0
2
4
6

Sampling Distribution of $\bar{X}$, $\hat{Q}(.5)$, S Using Simulation

Using the *R* program given on the next page, we will simulate data from the standard normal distribution for sample sizes of 10, 25, and 100. For each sample size, 10000 samples of that size are generated. The values of $\bar{X}$ $\hat{Q}(.5)$ S are computed. We thus have for each sample size, 10000 values of $\bar{X}$ $\hat{Q}(.5)$ S. The mean and standard deviation of these 10000 values are then computed. These means and standard deviations will then be compared to their corresponding asymptotic values based on the central limit theorem. A box plot of the 10000 values will also be provided to demonstrate the shape of the sampling distribution.

Sampling from N(0,1) Distribution

Population parameters:

$\mu = 0, \quad Q(.5) = 0, \quad f(Q(.5)) = f(0) = \frac{1}{\sqrt{2\pi}}, \quad \sigma = 1, \quad \mu_4 = 3\sigma^4 = 3.$

- Asymptotic Mean of $\bar{X}$ is $\mu_A = 0$;
- Asymptotic StDev of $\bar{X}$ is $\sigma_A = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}}$
- Asymptotic Mean of $\hat{Q}(.5)$ is $\mu_A = Q(.5) = 0$;
- Asymptotic StDev of $\hat{Q}(.5)$ is $\sigma_A = \frac{\sqrt{(.5)(.5)}}{\sqrt{n}f(Q(.5))} = \frac{(.5)}{\sqrt{n}f(0)} = \frac{(.5)}{\sqrt{n}(1/\sqrt{2\pi})} = \frac{1.2533}{\sqrt{n}}$
- Asymptotic Mean of S is $\mu_A = \sigma = 1$;
- Asymptotic StDev of S is $\sigma_A = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{2}\sigma^2}{2\sigma\sqrt{n}} = \frac{.707}{\sqrt{n}}$

Statistic	Sample Size	Asymp. Mean	Asymp. StDev	Sim. Mean	Sim. StDev
$\bar{X}$	10	0	0.3162	-0.00024	0.31419
	25	0	0.2000	-0.00009	0.20180
	100	0	0.1000	0.00062	0.10063
$\hat{Q}(.5)$	10	0	0.3963	-0.00242	0.37204
	25	0	0.2507	-0.00238	0.25143
	100	0	0.1253	0.00076	0.12647
S	10	1	0.2236	0.96878	0.23421
	25	1	0.1414	0.99065	0.14327
	100	1	0.0707	0.99683	0.07151

```
#Sampling Distribution of Mean, Median, Standard Deviation
#boxsamdistnorm.R also boxsamdistweib.R and boxsamdistt5.R are in R Files folder in eCampus
r=10000
x =  rep(0,10)
m10 =  rep(0,r)
q10 =  rep(0,r)
sq = rep(0,r)
s10 = rep(0,r)
for (i in 1:r){
  x =  rnorm(10)
   m10[i] =  mean(x)
   q10[i] =  median(x)
  s10[i] =  sd(x)}

x =  rep(0,25)
m25 =  rep(0,r)
q25 =  rep(0,r)
s2 = rep(0,r)
s25 = rep(0,r)
for (i in 1:r){
  x=  rnorm(25)
   m25[i] =  mean(x)
   q25[i] =  median(x)
  s25[i] =  sd(x)}

x= rep(0,100)
m100 =  rep(0,r)
q100 =  rep(0,r)
sq = rep(0,r)
s100 = rep(0,r)
for (i in 1:r){
  x=  rnorm(100)
   m100[i] =  mean(x)
   q100[i] =  median(x)
  s100[i] =  sd(x)}

outmean10 = c(mean(m10), mean(q10), mean(s10))
outmean25 = c(mean(m25), mean(q25), mean(s25))
outmean100 = c(mean(m100), mean(q100), mean(s100))
outmean = cbind(outmean10,outmean25,outmean100)
outmean
outsd10 = c(sd(m10), sd(q10), sd(s10))
outsd25 = c(sd(m25), sd(q25), sd(s25))
outsd100 = c(sd(m100), sd(q100), sd(s100))
outsd = cbind(outsd10,outsd25,outsd100)
outsd

boxplot(m10,m25,m100,xlab="Sample Size",yaxt="n",
                  ylab="Sample Mean",
          main="Boxplots of 10000 Sample Means from N(0,1)",
                  names=c("10","25","100"),cex=.75)
boxplot(q10,q25,q100,xlab="Sample Size",yaxt="n",
                  ylab="Sample Median",
          main="Boxplots of 10000 Sample Medians from N(0,1)",
                  names=c("10","25","100"),cex=.75)
boxplot(s10,s25,s100,xlab="Sample Size",yaxt="n",
                  ylab="Sample Standard Deviation",
          main="Boxplots of 10000 of Sample Std Dev from N(0,1)",
                  names=c("10","25","100"),cex=.75)
```

Sampling from Weibull($\gamma = .61, \alpha = .62$) Distribution

Simulate data from a Weibull distribution with $\alpha = .62, \gamma = .61$ for sample sizes of 10, 25, and 100. This Weibull distribution is highly right skewed. For each sample size, 10000 samples of that size are generated. The values of $\bar{X}$ $\hat{Q}(.5)$ S are computed. We thus have for each sample size, 10000 values of $\bar{X}$ $\hat{Q}(.5)$ S. The mean and standard deviation of these 10000 values are then computed. These means and standard deviations will then be compared to their corresponding asymptotic values based on the central limit theorem. A box plot of the 10000 values will also be provided to demonstrate the shape of the sampling distribution.

Population parameters: With $\alpha = .62$ and $\gamma = .61$, we obtain

$\mu = \alpha\Gamma\left[(\gamma+1)/\gamma)\right] = .9133,\ Q(.5) = \alpha(log(2))^{1/\gamma} = .3400,\ f(Q(.5)) = \alpha(log(2))^{1/\gamma} = .6218,$

$\sigma = \alpha\sqrt{(\Gamma\left[(\gamma+2)/\gamma)\right] - (\Gamma\left[(\gamma+1)/\gamma)\right])^2)} = 1.5730,\ \mu_4 = 309.2276,\ \text{skewness} = 4.38.$

- Mean of $\bar{X}$ is $\mu = .9133$; StDev of $\bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{1.5730}{\sqrt{n}}$
- Asymptotic Mean of $\hat{Q}(.5)$ is $Q(.5) = [-(.62)^{.61} log(.5)]^{1/.61} = .3400$;
- Asymptotic StDev of $\hat{Q}(.5) = \frac{\sqrt{(.5)(.5)}}{\sqrt{n} f(Q(.5))} = \frac{(.5)}{\sqrt{n} f(.34)} = \frac{(.5)}{\sqrt{n}(.6218)} = \frac{.8041}{\sqrt{n}}$
- Asymptotic Mean of S is $\sigma = 1.5730$;
- Asymptotic StDev of $S = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{309.2276 - (1.573)^4}}{2(1.573)\sqrt{n}} = \frac{5.534}{\sqrt{n}}$

Statistic	Sample Size	Asymp. Mean	Asymp. StDev	Sim. Mean	Sim. StDev
$\bar{X}$	10	0.9133	0.4974	0.91000	0.49061
	25	0.9133	0.3146	0.91166	0.31127
	100	0.9133	0.1573	0.91371	0.15548
$\hat{Q}(.5)$	10	0.3400	0.2543	0.42246	0.29338
	25	0.3400	0.1608	0.37020	0.17149
	100	0.3400	0.0804	0.34805	0.08185
S	10	1.5730	1.7500	1.27643	0.89891
	25	1.5730	1.1068	1.40861	0.67789
	100	1.5730	0.5534	1.52035	0.41083

Sampling from t with df=5 Distribution

Simulate data from a t with df=5 distribution for sample sizes of 10, 25, and 100. This t distribution is symmetric with heavier tables than N(0,1). For each sample size, 10000 samples of that size are generated. The values of $\bar{X}$, $\hat{Q}(.5)$, and S are computed. We thus have for each sample size, 10000 values of $\bar{X}$, $\hat{Q}(.5)$, and S. The mean and standard deviation of these 10000 values are then computed. These means and standard deviations will then be compared to their corresponding asymptotic values based on the central limit theorem. A box plot of the 10000 values will also be provided to demonstrate the shape of the sampling distribution.

Population parameters: with $\nu = df = 5$

$\mu = 0, \qquad Q(.5) = 0, \qquad f(Q(.5)) = f(0) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} = \frac{8}{3\pi\sqrt{5}} = .3796,$

$\sigma = \sqrt{\frac{\nu}{\nu-2}} = \sqrt{\frac{5}{3}} = 1.29099, \qquad \mu_4 = \frac{3\nu^2}{(\nu-2)(\nu-4)} = 25, \qquad \text{skewness} = 0.$

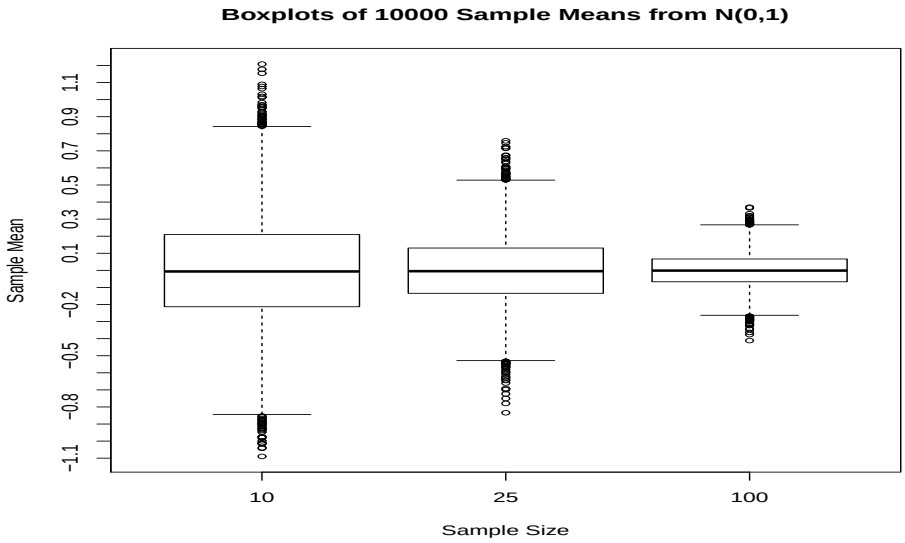
- Mean of $\bar{X}$ is $\mu = 0$;
- StDev of $\bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{1.291}{\sqrt{n}}$
- Asymptotic Mean of $\hat{Q}(.5)$ is $Q(.5) = 0$;
- Asymptotic StDev of $\hat{Q}(.5) = \frac{\sqrt{(.5)(.5)}}{\sqrt{n} f(Q(.5))} = \frac{(.5)}{\sqrt{n} f(0)} = \frac{(.5)}{\sqrt{n}(.3796)} = \frac{1.3172}{\sqrt{n}}$
- Asymptotic Mean of S is $\sigma = 1.291$;
- Asymptotic StDev of $S = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{25-(1.291)^4}}{2(1.291)\sqrt{n}} = \frac{1.8257}{\sqrt{n}}$

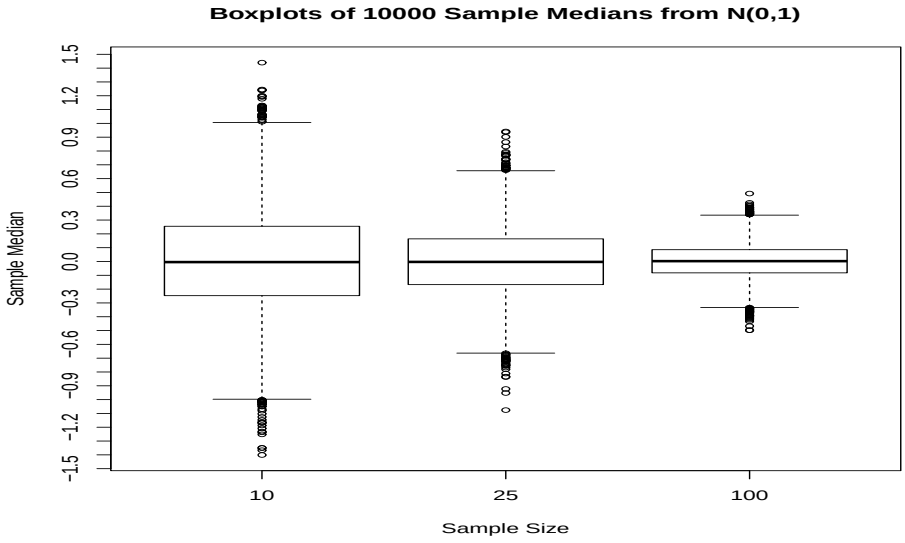
Statistic	Sample Size	Asymp. Mean	Asymp. StDev	Sim. Mean	Sim. StDev
$\bar{X}$	10	0	0.4083	-0.00275	0.41109
	25	0	0.2582	-0.00311	0.26146
	100	0	0.1291	0.00094	0.13002
$\hat{Q}(.5)$	10	0	0.4165	0.00123	0.40125
	25	0	0.2634	-0.00338	0.26451
	100	0	0.1317	0.00090	0.13037
S	10	1.291	0.5773	1.22537	0.42978
	25	1.291	0.3651	1.25945	0.28704
	100	1.291	0.1826	1.28177	0.15818

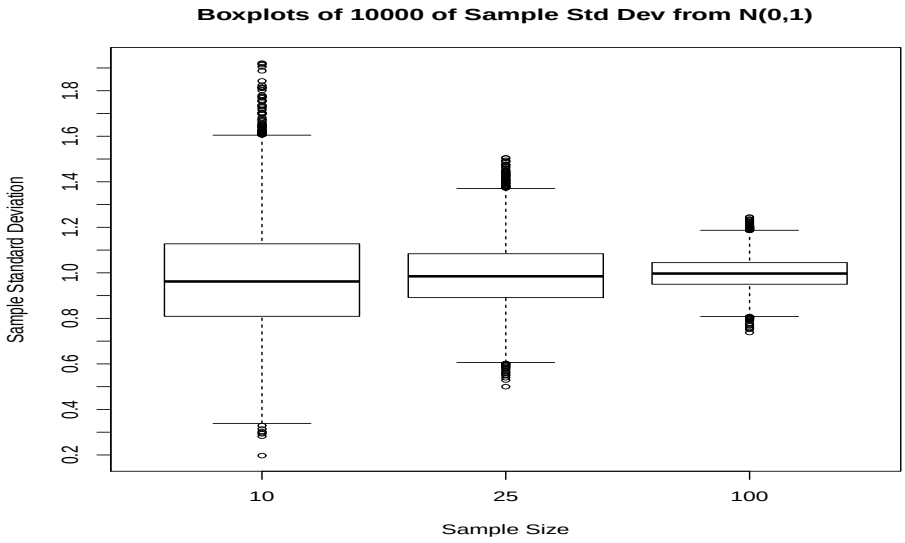
Number of Outliers in 10,000 Realizations of Three Sample Statistics

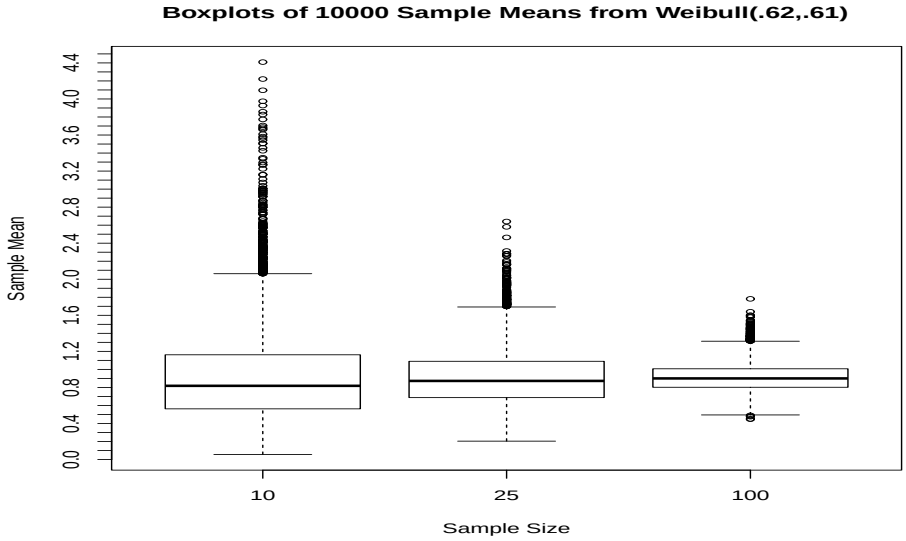
Distribution	Statistic	Sample Size	num lower outliers	num upper outliers	num outliers
N(0,1)	Mean	10	46	39	85
		25	36	35	71
		100	40	38	78
N(0,1)	Median	10	39	48	87
		25	34	26	60
		100	32	35	67
N(0,1)	S	10	4	66	70
		25	22	45	67
		100	27	33	60
Weibull	Mean	10	0	264	264
		25	0	231	231
		100	4	110	114
Weibull	Median	10	0	426	426
		25	0	251	251
		100	1	116	117
Weibull	S	10	0	466	466
		25	0	395	395
		100	0	274	274
t df = 5	Mean	10	67	59	126
		25	42	58	100
		100	53	42	95
t df = 5	Median	10	59	40	99
		25	31	44	75
		100	45	42	87
t df = 5	S	10	0	239	239
		25	0	249	249
		100	4	240	244
Expected number of outliers in a sample of 10,000 from a normal distribution is 70					

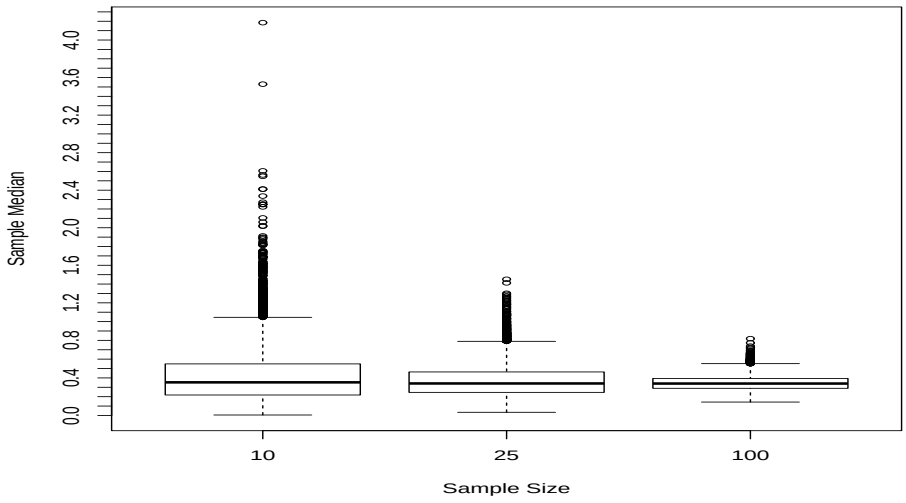
outliers.pdf

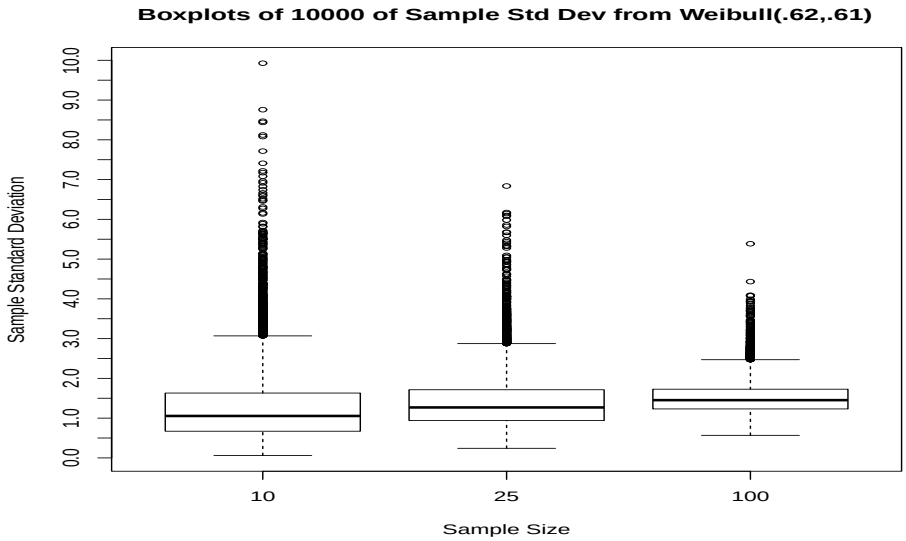

Boxplots of 10000 Sample Means from N(0,1)
Sample Mean
−1.1
−0.8
−0.5
−0.2
0.1
0.3
0.5
0.7
0.9
1.1
10
25
100
Sample Size

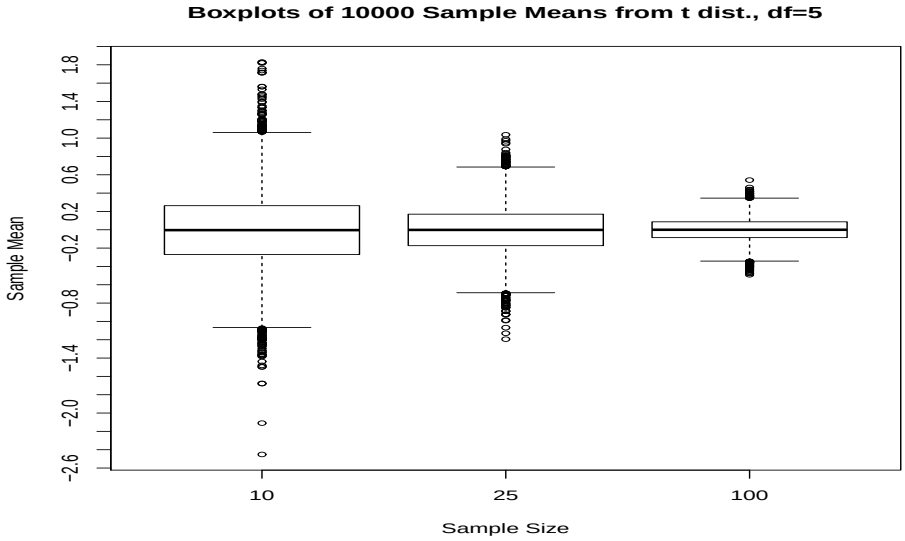

Boxplots of 10000 Sample Medians from N(0,1)
Sample Median
−1.5
−1.2
−0.9
−0.6
−0.3
0.0
0.3
0.6
0.9
1.2
1.5
10
25
100
Sample Size

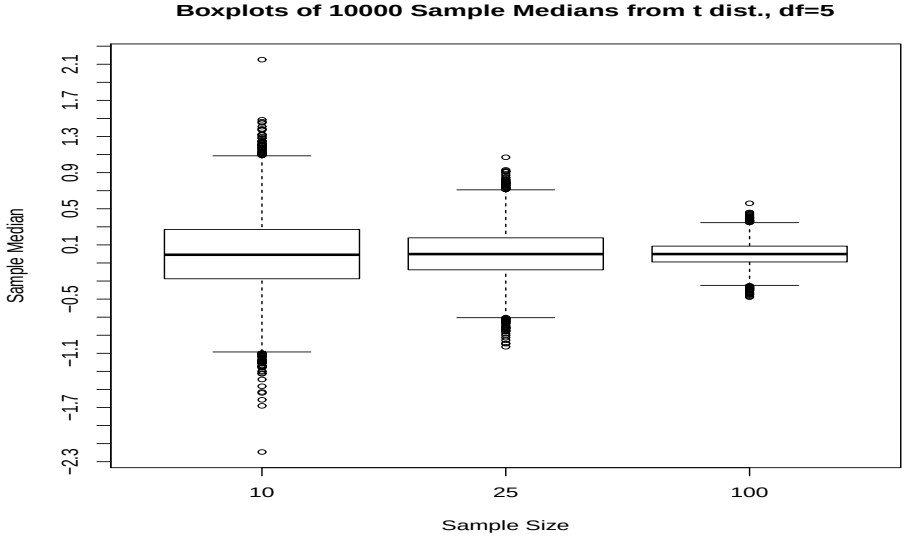

Boxplots of 10000 of Sample Std Dev from N(0,1)
Sample Standard Deviation
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
10
25
100
Sample Size

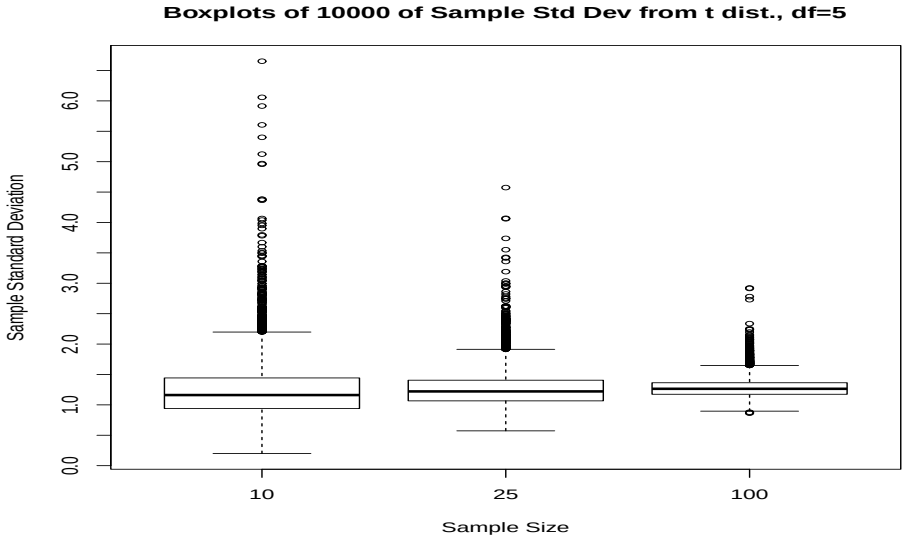

Boxplots of 10000 Sample Means from Weibull(.62,.61)
Sample Mean
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
10
25
100
Sample Size

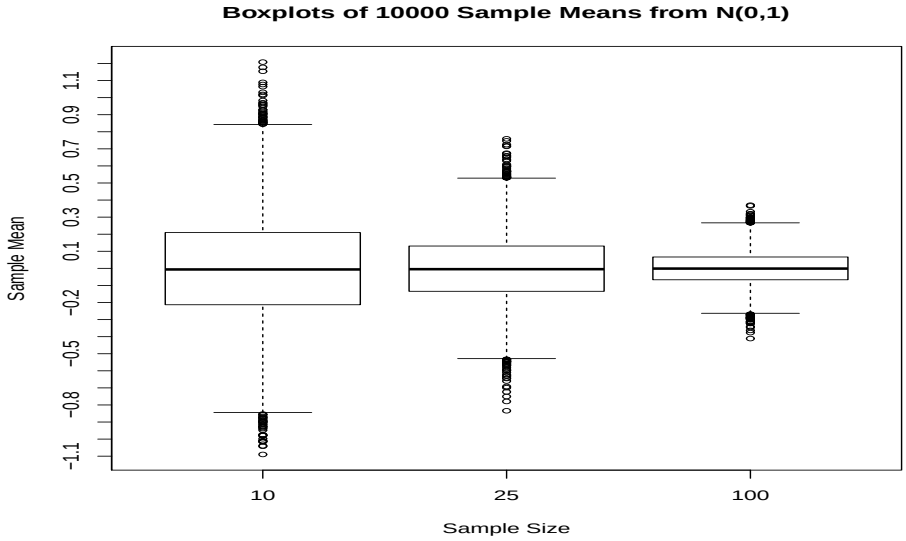

Boxplots of 10000 Sample Medians from Weibull(.62,.61)
Sample Median
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
10
25
100
Sample Size

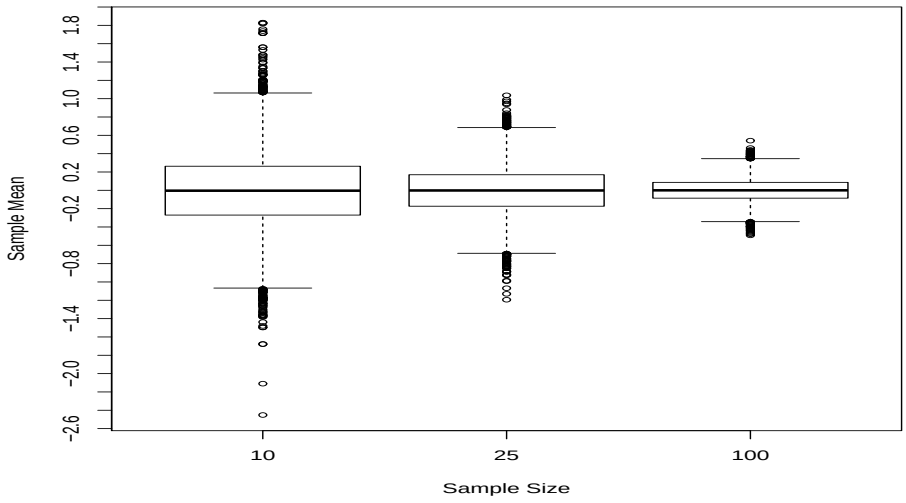

Boxplots of 10000 of Sample Std Dev from Weibull(.62,.61)
Sample Standard Deviation
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
10
25
100
Sample Size

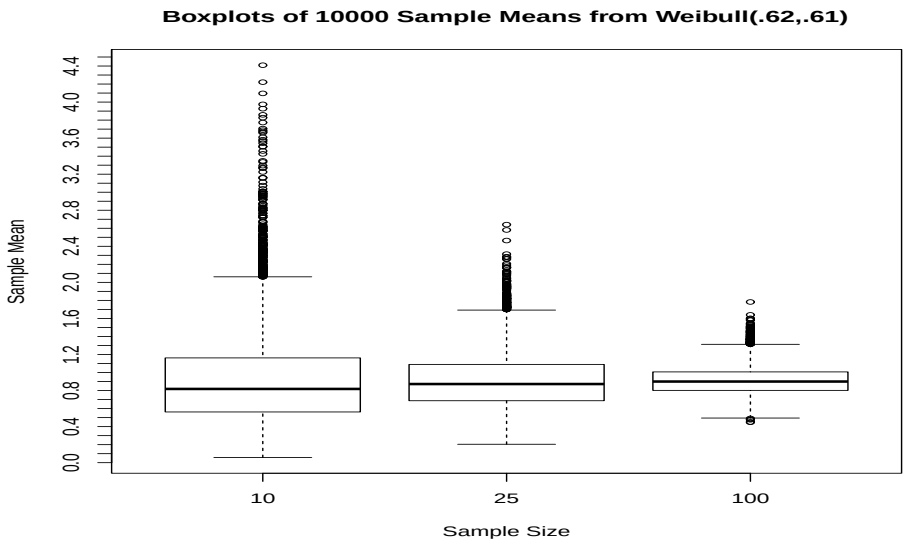

Boxplots of 10000 Sample Means from t dist., df=5
Sample Mean
1.8
1.4
1.0
0.6
0.2
−0.2
−0.8
−1.4
−2.0
−2.6
10
25
100
Sample Size

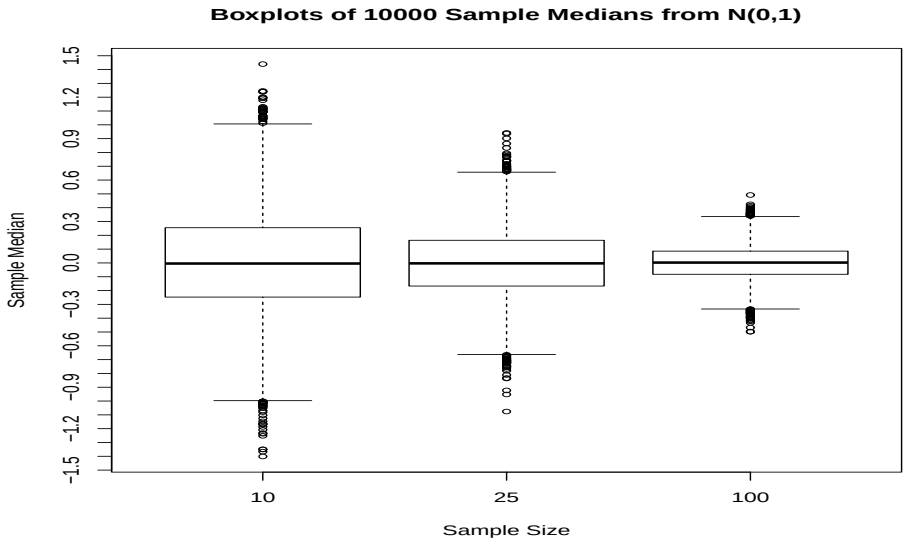

Boxplots of 10000 Sample Medians from t dist., df=5
Sample Median
2.1
1.7
1.3
0.9
0.5
0.1
−0.5
−1.1
−1.7
−2.3
10
25
100
Sample Size

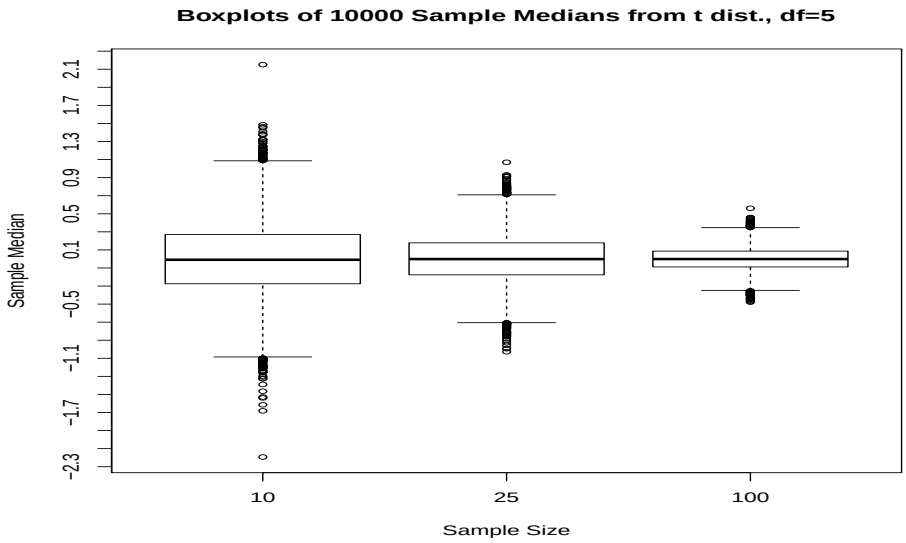

Boxplots of 10000 of Sample Std Dev from t dist., df=5
Sample Standard Deviation
6.0
5.0
4.0
3.0
2.0
1.0
0.0
10
25
100
Sample Size

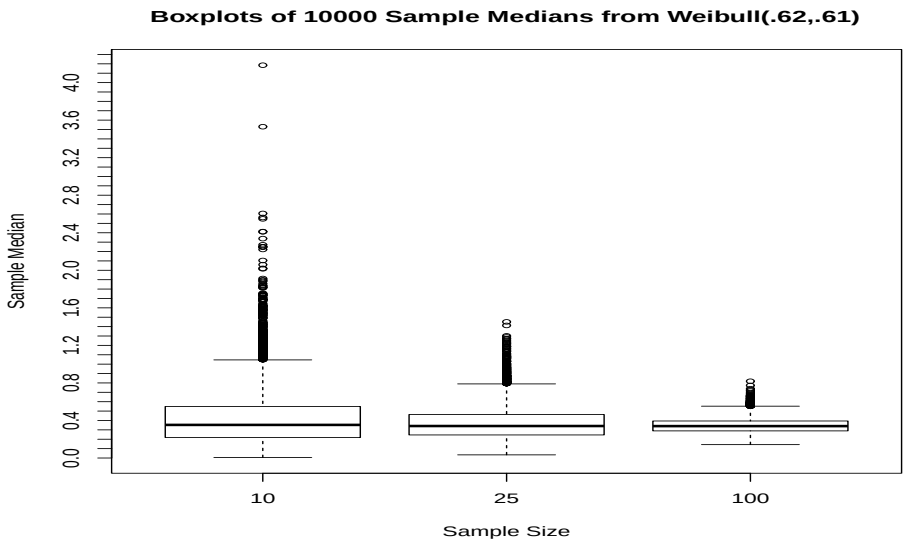

Boxplots of 10000 Sample Means from N(0,1)
Sample Mean
-1.1
-0.8
-0.5
-0.2
0.1
0.3
0.5
0.7
0.9
1.1
10
25
100
Sample Size

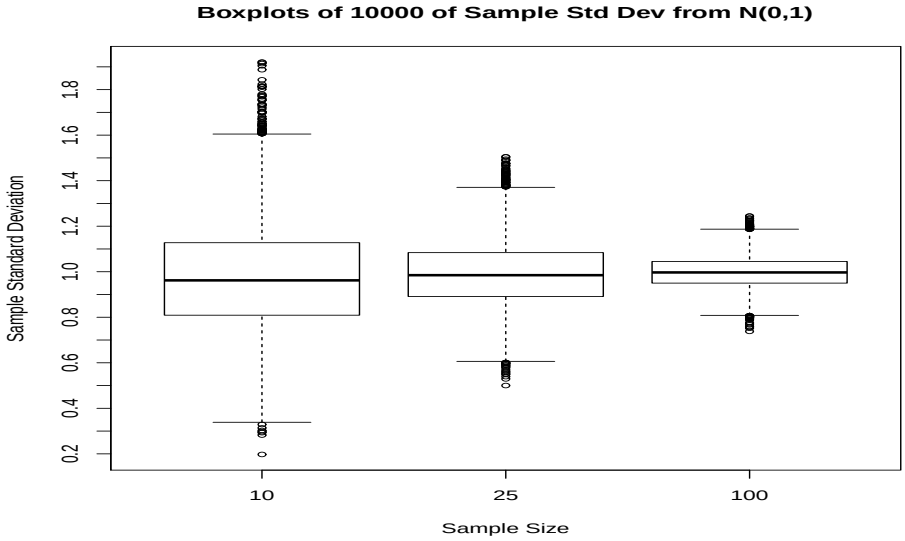

Boxplots of 10000 Sample Means from t dist., df=5
Sample Mean
-2.6
-2.0
-1.4
-0.8
-0.2
0.2
0.6
1.0
1.4
1.8
10
25
100
Sample Size

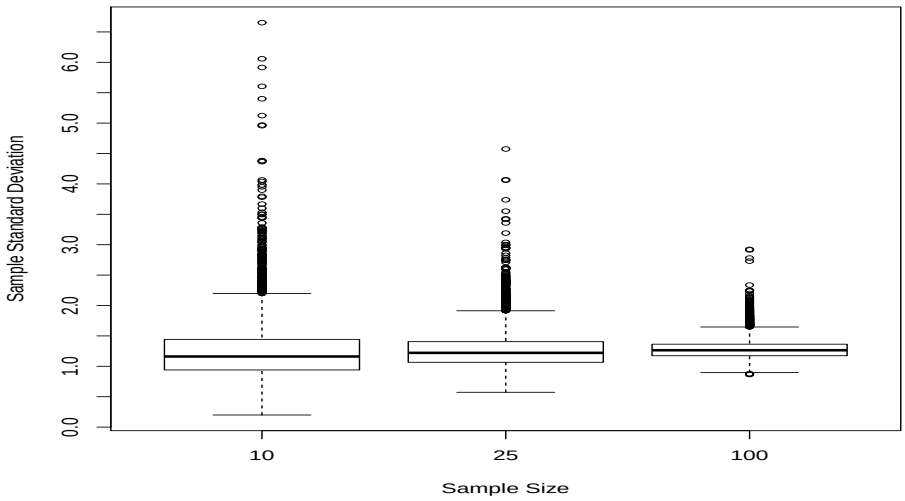

Boxplots of 10000 Sample Means from Weibull(.62,.61)
Sample Mean
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
10
25
100
Sample Size

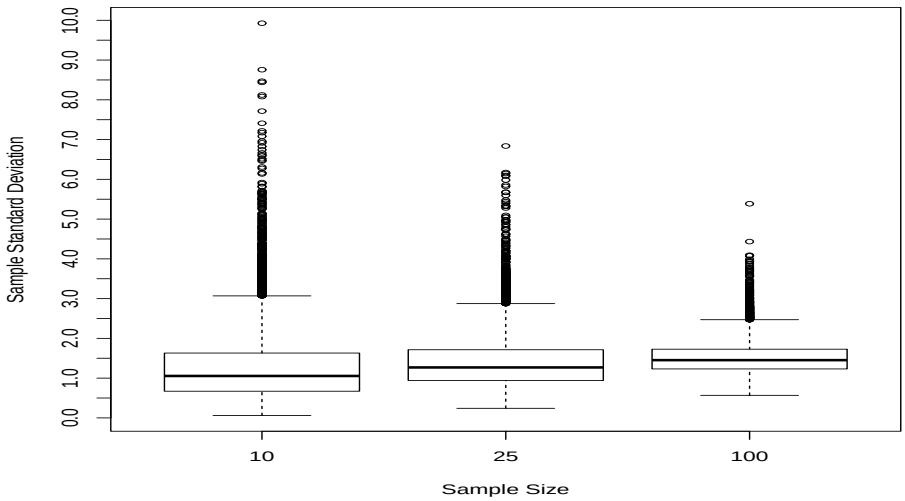

Boxplots of 10000 Sample Medians from N(0,1)
Sample Median
1.5
1.2
0.9
0.6
0.3
0.0
-0.3
-0.6
-0.9
-1.2
-1.5
10
25
100
Sample Size


Boxplots of 10000 Sample Medians from t dist., df=5
Sample Median
2.1
1.7
1.3
0.9
0.5
0.1
-0.5
-1.1
-1.7
-2.3
10
25
100
Sample Size


Boxplots of 10000 Sample Medians from Weibull(.62,.61)
Sample Median
4.0
3.6
3.2
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0.0
10
25
100
Sample Size


Boxplots of 10000 of Sample Std Dev from N(0,1)
Sample Standard Deviation
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
10
25
100
Sample Size


Boxplots of 10000 of Sample Std Dev from t dist., df=5
Sample Standard Deviation
0.0
1.0
2.0
3.0
4.0
5.0
6.0
10
25
100
Sample Size


Boxplots of 10000 of Sample Std Dev from Weibull(.62,.61)
Sample Standard Deviation
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
10
25
100
Sample Size

resampling technique that allows you to approximate a sampling distribution

Bootstrapping the Sampling Distribution of Various Statistics

Let $X_1, \ldots, X_n$ be iid random variables with a common cdf $F(\cdot)$. Let θ be a parameter which we wish to estimate using a function of the data $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

Suppose the cdf $F(\cdot)$ is unknown and the sample size n is small or that the asymptotic distribution of $\hat{\theta}$ is intractable.

We wish to determine the sampling distribution of $\hat{\theta}$ in order to be able to determine its bias as an estimator of θ or its variance or its percentiles in order to provide an assessment of how well $\hat{\theta}$ estimates θ. Suppose we can not mathematically derive the true sampling distribution of $\hat{\theta}$ because the cdf $F(\cdot)$ is unknown or the form of $\hat{\theta}$ may be too complex to obtain an exact result. The asymptotic distribution may not provide an adequate approximation of the true sampling distribution of $\hat{\theta}$ because n is too small. If F were known we could use simulation to determine the sampling distribution of $\hat{\theta}$.

For example, suppose we wanted to determine the sampling distribution of an estimate of the scale parameter, θ, in the Cauchy distribution when the location parameter is 0 and the sample size is $n = 100$. We would first select various values for θ. Next, for each value of θ, simulate $N = 10,000$ sets of 100 realizations from the Cauchy distribution, and compute a value for $\hat{\theta}$: $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$,. We could then use these N values of $\hat{\theta}$ to estimate the sampling distribution of $\hat{\theta}$ for the selected value of θ. This would then be repeated for the various choices of θ. However, in many cases, F will be unknown which would make simulation studies impossible.

An alternative to these approaches which can be used when F is unknown is the bootstrap procedure which will provide an approximation to the sampling distribution of $\hat{\theta}$ in the situation where we can write θ as a function of the cdf, that is, $\theta = g(F(\cdot))$. For example,

- the population mean $\mu = \int_{-\infty}^{\infty} x dF(x)$,
- the population variance, $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 dF(x)$,
- the population median M can be defined by $.5 = \int_{-\infty}^{M} dF(x)$.

To obtain the sample estimator, we simply replace the true cdf $F(\cdot)$ with the empirical (sample) cdf $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \frac{\#\{X_i \leq x\}}{n}$ in $\theta = g(F(\cdot))$ to obtain $\hat{\theta} = g(\hat{F}(\cdot))$.

In order to obtain the sampling distribution of $\hat{\theta}$, we simulate data from the edf $\hat{F}(\cdot)$ in place of the true cdf $F(\cdot)$. Recall, we used the true cdf in the simulations from the normal and Weibull distributions. That is, we will now consider the population to be the observed data having a cdf which places mass $\frac{1}{n}$ on each of the observed data values X_i. Thus, we select M random samples of size n (sampling with replacement) from this "new" population, compute $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$. We now have M realizations of $\hat{\theta}$ from which we can estimate the pdf (using a kernel density estimator), the quantile function, or specific parameters like its mean: $\mu_{\hat{\theta}} = E[\hat{\theta}] = \frac{1}{M}\sum_{i=1}^{M} \hat{\theta}_i$ and its variance $Var[\hat{\theta}] = \frac{1}{M}\sum_{i=1}^{M}(\hat{\theta}_i - \mu_{\hat{\theta}})^2$. Similarly, we can compute its median or any other percentiles.

When using a bootstrap procedure, we have two levels of approximation:

1. The estimation of the population cdf F using the edf $\hat{F}$

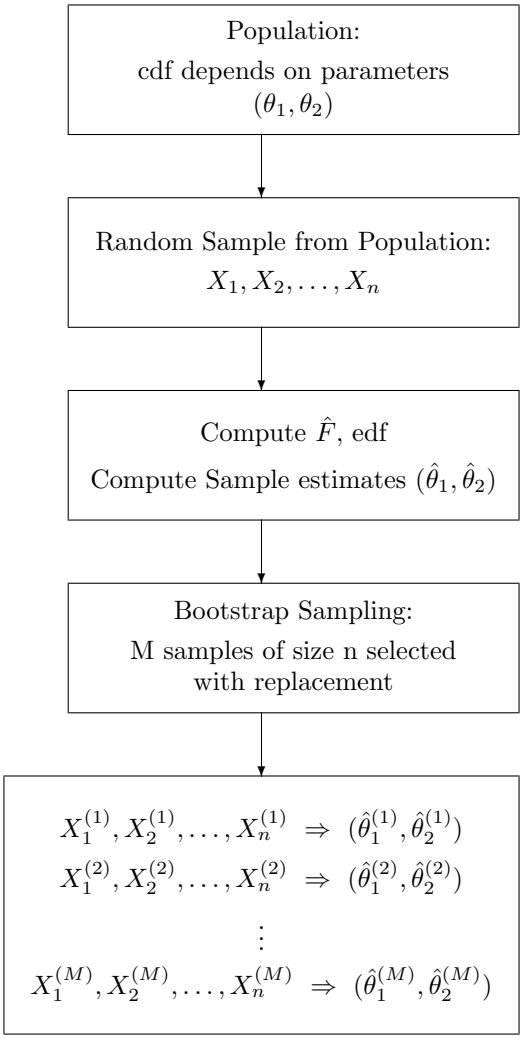
   Accuracy of approximation controlled by the size of n and the shape of F.

2. Repeated estimation of the edf $\hat{F}$ using the M bootstrap estimators $\tilde{F}$

   Accuracy of approximation limited by the value of n and how well $\hat{F}$ approximates F.

A flow chart of bootstraped procedure is given here:

1. Obtain data $X_1, \ldots, X_n$ iid with cdf $F(\cdot)$
2. Compute $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$
3. Select a random sample of size n with replacement from $X_1, \ldots, X_n$ (i.e., simulate n independent observations from the edf $\hat{F}(\cdot)$): Denote by $X_1^*, \ldots, X_n^*$
4. Compute $\hat{\theta}_i^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$
5. Repeat Step 3 and Step 4 M times yielding $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_M^*$
6. Use these M realizations of $\hat{\theta}$ to construct the sampling distribution of $\hat{\theta}$: Means, StDev, Percentiles, pdf, etc.



We will consider the following example to illustrate the application of the bootstrap procedure:

EXAMPLE: Suppose the life lengths of eleven engine parts are measured as

5700, 36300, 12400, 28000, 19300, 21500, 12900, 4100, 91400, 7600, 1600

We want to estimate the median life length θ of the engine part. From the data we compute $\hat{\theta} = X_{(6)} = 12900$. To study the variation in this estimator we need to know its sampling distribution. We will use the bootstrap to approximate this distribution.

We will first generate M = 200 bootstrap samples from $\hat{F}(\cdot)$ and then M = 20,000 bootstrap samples using the following R code: **bootsampMed.R**

```
y = c(1600,4100,5700,7600,12400,12900,19300,21500,28000,36300,91400)
mhat = median(y)
M =  20000
d = numeric(M)
for(i in 1:M)d[i] = median(sample(y,replace=T))
hist(d)
bootmean =  mean(d)
bootstd = sd(d)
bootquant = quantile(d)
probs = seq(0,1,.01)
Qd = quantile(d,probs)
boxplot(d,main="Empirical Quantile for Sample Median",
ylab="Median Life Lengths of Engine Part",plot=T)
plot(probs,Qd,type="l",ylab="Q(u) for Median",xlab="u",xlim=c(0,1),lab=c(10,11,7))
title("Empirical Quantile for Sample Median",cex=.75)
plot(density(d),type="l",xlab="Median Life Lengths",ylab="PDF of Sample Median")
title("Empirical pdf for Sample Median",cex=.75)
qqnorm(d,main="Normal Prob Plot of Sample Median",
xlab="normal quantiles",ylab="Sample Medians",
lab=c(7,7,7),cex=.75)
qqline(d)
```

In the following table the first five simulations are given with a "+" indicating which of the original data values was sampled. Note that some values will be sampled multiple times and some values may not be included:

First Five Bootstrap Samples

Original Data	Bootstrap Sample				
	1	2	3	4	5
1600			+		++
4100	+++	++	+	+	
5700	+	+	+	+++	+
7600				++	++
12400	+	+	+		+
12900	+		++		
19300	+	+	+	++	
21500		+++	+	+	+
28000	+	++	+	++	+
36300	+	+			+
91400	++		++		++
$\hat{\theta}^*$	12900	21500	12900	7600	12400

From the 200 realizations of $\hat{\theta}^*$, the following summary statistics were computed:

Average: $$E[\hat{\theta}^*] = \frac{1}{200}\sum_{i=1}^{200}\hat{\theta}_i^* = 14877.5$$

Standard Deviation: $$\sqrt{Var[\hat{\theta}^*]} = \sqrt{\frac{1}{200}\sum_{i=1}^{200}(\hat{\theta}_i^* - E[\hat{\theta}^*])^2} = 5552.6$$

Quantile				
0	.25	.50	.75	1.0
4100	12400	12900	19300	36300

If we extended the simulation to 20000 bootstrap samples, we obtain

Average: $$E[\hat{\theta}^*] = \frac{1}{20000}\sum_{i=1}^{20000}\hat{\theta}_i^* = 14924.1$$

Standard Deviation: $$\sqrt{Var[\hat{\theta}^*]} = \sqrt{\frac{1}{20000}\sum_{i=1}^{20000}(\hat{\theta}_i^* - E[\hat{\theta}^*])^2} = 5933.7$$

Quantile				
0	.25	.50	.75	1.0
1600	12400	12900	19300	91400

Thus, there was only minor differences in the mean and standard deviation for the sampling distribution of the median when comparing 200 bootstrap samples to 20000 bootstrap samples. However, note the big discrepancies between the quantiles. When generating 20000 samples of size 11 from the original data set, samples were obtained in which the median of the bootstrap sample was equal to the minimum value (1600) in the original data set. Because the bootstrap median equals $\hat{\theta}^* = X^*_{(6)}$, this result implies that, in the bootstrap samples having median=1600, at least 6 of the 11 data values must be equal to 1600. This seems very unlikely. However, if we calculate the expected number of samples in the 20000 samples having exactly 6 of their 11 values equal to 1600, we find that

$$\text{Expected Number} = 20000\ P[\text{exactly 6 of 11 values equal 1600}]$$

$$= (20000)\left(\binom{11}{6}(1)^6(10)^5\right)/(11)^{11} = 3.2$$

Therefore, on the average we would expect 3.2 occurrences of the event that exactly 6 of the 11 data values were equal to 1600.

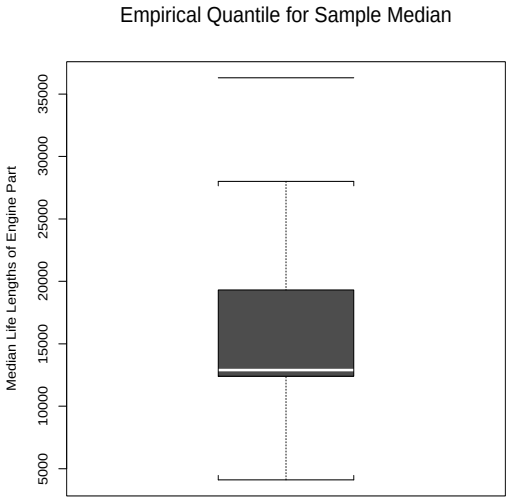
A good reference on bootstraping is *Bootstrap Methods and their Applications* by D.V. Hinkley and A.C. Davison.

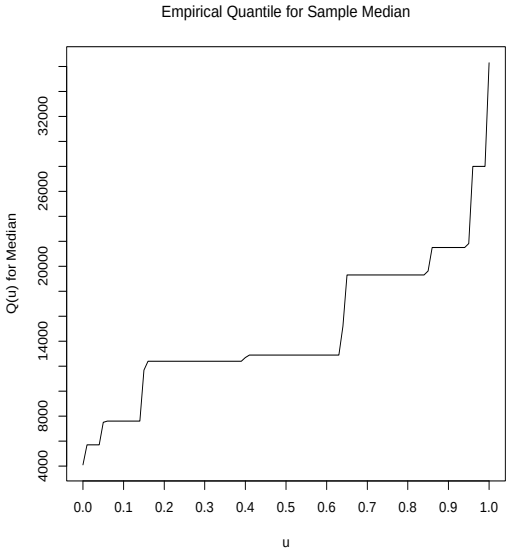
A plot of the quantile function, kernel density estimator of the pdf, a box plot, and a normal reference distribution plot for the sampling distribution of the sample quantile are given on the next pages for 200 and 20000 bootstrap samples. We note that there is considerable differences in the plots. The plots for 20000 bootstrap samples reveals the discreteness of the possible values for the median when the sample size ($n = 11$ in our case) is very small. Also, we note that $n = 11$ is too small for the sampling distribution for the median to achieve its asymptotic result (n large), an approximate normal distribution.

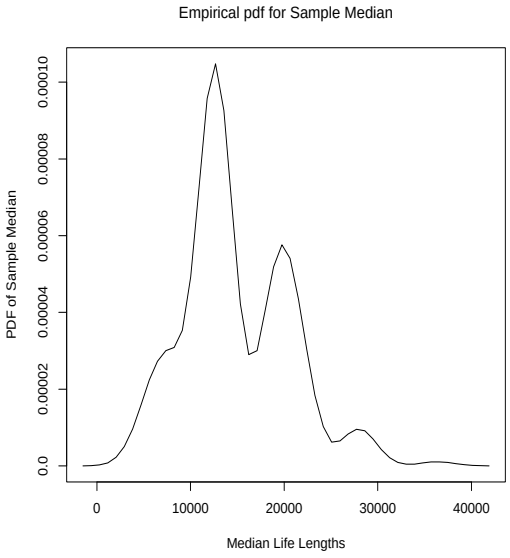
If you wanted to observe a few of the bootstrap samples, say the first K=20, then just use the following R code:

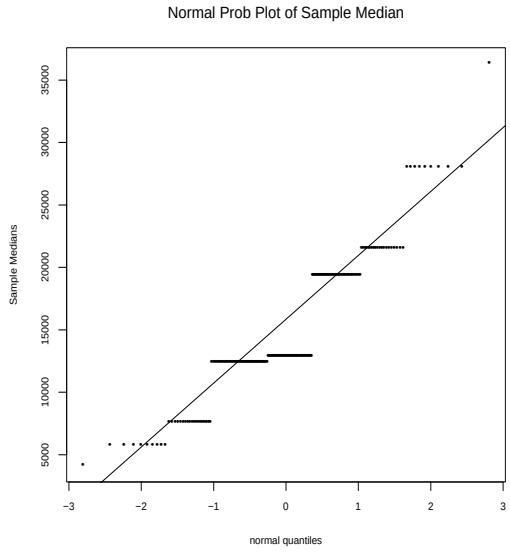
```
y = c(1600,4100,5700,7600,12400,12900,19300,21500,28000,36300,91400)
n = length(y)
K = 20
sam = matrix(0,K,n)
for(i in 1:K)
{
sam[i,] = sample(y,replace=T)
}
```

Plots of sampling distribution of sample median based on 200 resamples.

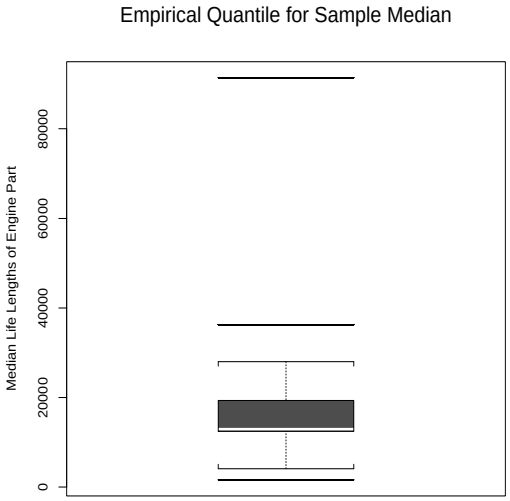


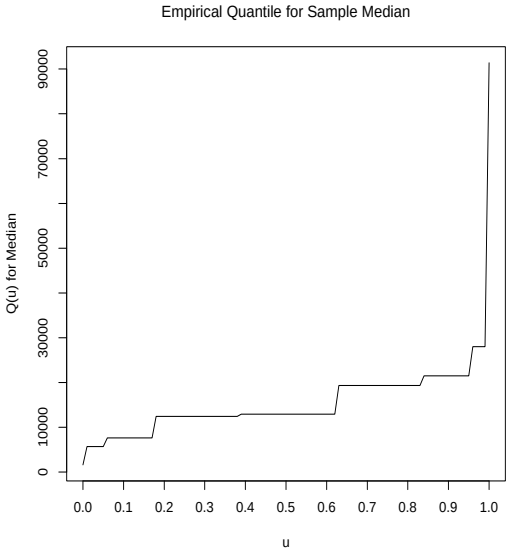


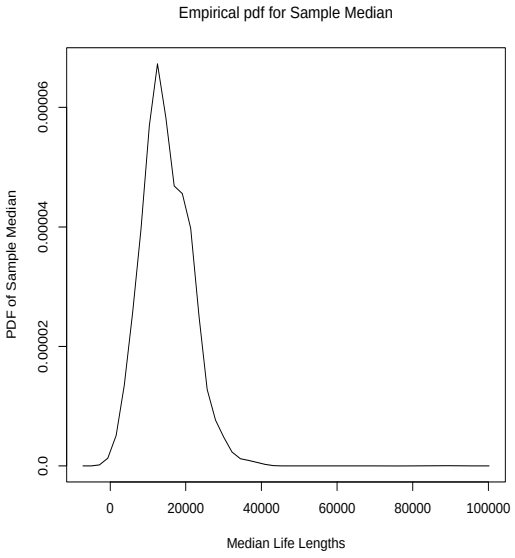


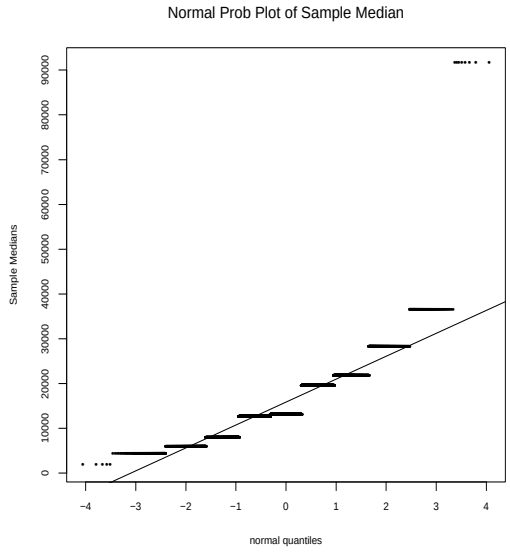


Plots of sampling distribution of sample median based on 20000 resamples.









Sampling Distribution of Maximum Likelihood Estimator

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (or iid observations) from a population/process having pdf $f(y)$ which depends on unknown parameters: θ_1, θ_2.

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

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

- Invariance Property of MLE:

  If $\hat{\theta}$ is the MLE of θ, then for any function $h(\theta)$, the MLE of $h(\theta)$ is $h(\hat{\theta})$

  Example: Suppose the random sample is from a Weibull population and $\hat{\alpha}$ and $\hat{\gamma}$ are the MLEs.

  Find the MLE of $S(t) = e^{-(t/\alpha)^{\gamma}}$

  Solution: By the Invariance Property of MLEs, the MLE of $S(t)$ is

  $\hat{S}(t) = e^{-(t/\hat{\alpha})^{\hat{\gamma}}}$

- The fixed n properties of MLEs depend on the population pdf $f(\cdot)$ and no general statement can be made about the distribution of $\widehat{\theta}$

- Asymptotic Properties of MLEs:

  Under some regularity conditions (see Casella-Berger,2nd Edition, p516), the central limit theorem for MLEs yields

  Let $\hat{\theta}$ denote the MLE of θ. Let $h(\theta)$ be any continuous function of θ,

  For large n : The distribution of $h(\hat{\theta})$ is approximately $N\left(h(\theta), \frac{(h'(\theta))^2}{I_n(\theta)}\right)$, where $I_n(\theta) = E_\theta\left(-\frac{\partial^2}{\partial \theta^2} log(L(\theta))\right)$.

  Thus, the asymptotic mean and standard deviation for the sampling distribution of $h(\hat{\theta})$ are given by

$$\mu_A = h(\theta) \qquad \sigma_A = \frac{h'(\theta)}{\sqrt{I_n(\theta)}},$$

an estimator of σ_A is given by

$$\widehat{\sigma}_A = \frac{h'(\hat{\theta})}{\sqrt{I_n(\hat{\theta})}}$$

In particular, if $h(\theta) = \theta$, then the MLE of θ, $\hat{\theta}$ is approximately (large n) normally distributed with asymptotic mean and standard deviation

$$\mu_A = \theta \qquad\qquad \sigma_A = \frac{1}{\sqrt{I_n(\theta)}}$$

- Example 1: Let $f(\cdot)$ be exponential(β). We derived in Handout 6 that the MLE was $\hat{\beta} = \bar{Y}$

  To find the asymptotic variance, we need to find the form of the Information number:

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-t_i/\beta} = \beta^{-n} e^{-\frac{1}{\beta}\sum_{i=1}^{n} t_i}$$

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

$$\frac{\partial l(\beta; y)}{\partial \beta^2} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} t_i \qquad\qquad \frac{\partial^2 l(\beta; y)}{\partial \beta^2} = \frac{n}{\beta^2} - \frac{2}{\beta^3}\sum_{i=1}^{n} t_i$$

$$I_n(\beta) = E\left[-\frac{\partial^2 l(\beta; y)}{\partial \beta^2}\right] = -\frac{n}{\beta^2} + \frac{2}{\beta^3} E\left[\sum_{i=1}^{n} t_i\right] \quad = -\frac{n}{\beta^2} + \frac{2n\beta}{\beta^3} \quad = \frac{n}{\beta^2}$$

  Thus, the asymptotic standard deviation is approximated by

$$\widehat{\sigma}_A = \frac{1}{\sqrt{I_n(\widehat{\beta})}} = \frac{\widehat{\beta}}{\sqrt{n}}$$

- Example 2: Let $f(\cdot)$ be Weibull(γ,α). We demonstrated in Handout 6 that the MLE estimators are obtained from solving the equations:

$$\widehat{\alpha} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i^{\widehat{\gamma}} \right)^{1/\widehat{\gamma}}$$

$$\frac{\sum_{i=1}^{n} t_i^{\widehat{\gamma}} log(t_i)}{\sum_{i=1}^{n} t_i^{\widehat{\gamma}}} - \frac{1}{\widehat{\gamma}} = \frac{1}{n} \sum_{i=1}^{n} log(t_i)$$

To find the asymptotic variances we would need to approximate numerically the second partial derivatives of the log-likelihood function at the values of $\hat{\gamma}$ and $\hat{\alpha}$.

These are the values displayed in the R output that were provided in Handout 6.

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

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


OUTPUT from R:

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

we can approximate these SEs with the bootstrap

We thus have

$\hat{\gamma} = 0.9739245$ with estimated standard error: $\widehat{SE}(\hat{\gamma}) = 0.1512936$

$\hat{\alpha} = 11.4852981$ with estimated standard error: $\widehat{SE}(\hat{\alpha}) = 2.4660607$

Parametric Bootstrap

To obtain the small sample sampling distribution of the MLE in those situations where the form of the MLE makes the exact derivation intractable or n is too small to invoke the asymptotic results, we can use parametric bootstrap techniques.

Let f be the pdf of the population or process that generates the data. Suppose f depends on unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$. To simplify the notation, let $k = 1$.

Let $Y_1, Y_2, \ldots, Y_n$ be iid with pdf $f(y, \theta)$.

Obtain the MLE of θ, $\hat{\theta}$

To obtain the sampling distribution of $\hat{\theta}$, make use of parametric bootstrap:

1. Generate M samples of size n from $f(y, \hat{\theta})$ and from each sample obtain a value for $\hat{\theta}$

   $y_{11}, y_{12}, \ldots, y_{1n} \Rightarrow \hat{\theta}_1$

   $y_{12}, y_{22}, \ldots, y_{2n} \Rightarrow \hat{\theta}_2$

   $y_{13}, y_{32}, \ldots, y_{3n} \Rightarrow \hat{\theta}_3$

   $\vdots$

   $y_{1M}, y_{M2}, \ldots, y_{Mn} \Rightarrow \hat{\theta}_M$

2. Use the values of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$ to estimate percentiles and moments of the sampling distribution of $\hat{\theta}$, for example,

$$\widehat{Q}(.05), \qquad \widehat{Q}(.95), \qquad \widehat{Q}(.5)$$

$$\hat{\mu}_{\hat{\theta}} = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i, \qquad \hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\hat{\theta}_i - \hat{\mu}_{\hat{\theta}})^2} \quad = \quad \widehat{SE}(\hat{\theta})$$

Example

A study is conducted to evaluate the time to failure of a new device. From previous studies, the reliability engineer is fairly certain that the time to failure, T, will have a Weibull distribution with parameters γ, and α, but the parameters are unknown because this a new device. The researcher wants to estimate $S(t) = e^{-(t/\alpha)^{\gamma}}$ for $t = 20$ and hence needs to know the sampling distribution of the MLE of $S(20) = P[T > 20]$, the proportion of the devices produced that will fail after 20 units of time.

$$\hat{S}(20) = e^{-(t/\hat{\alpha})^{\hat{\gamma}}} \text{ where } \hat{\alpha} \text{ and } \hat{\gamma} \text{ are the MLE's of } \alpha \text{ and } \gamma$$

The researcher was only able to test 25 of the devices and obtain their times to failure: $T_1, T_2, \ldots, T_{25}$ but $n = 25$ is too small to invoke the asymptotic distribution of $\hat{S}(20)_{MLE}$.

Therefore, a parametric bootstrap approximation is a possible alternative:

1. Use R to obtain values for the MLE's, $(\hat{\gamma}_{MLE}, \hat{\alpha}_{MLE})$ based on the 25 data values.
2. Use the computed values of ($\hat{\gamma}_{MLE}$ and $\hat{\alpha}_{MLE}$) to generate 10000 samples of size $n = 25$ from a Weibull distribution and then compute an estimate of $S(20)$ using $\hat{S}(20)$ from each of the 10000 samples yielding

$$\hat{S}(20)_1,\ \hat{S}(20)_2,\ \ldots,\ \hat{S}(20)_{10000}$$

- In R, iterate the function $rweibull(25, \hat{\gamma}_{MLE}, \hat{\alpha}_{MLE})$ $M = 10000$ times to obtain

Sample 1: $y_{1,1}, y_{1,2}, \ldots, y_{1,25} \ \Rightarrow\ (\hat{\gamma}_1, \hat{\alpha}_1) \ \Rightarrow\ \hat{S}(20)_1 = e^{-(20/\hat{\alpha}_1)^{\hat{\gamma}_1}}$

Sample 2: $y_{2,1}, y_{2,2}, \ldots, y_{2,25} \ \Rightarrow\ (\hat{\gamma}_2, \hat{\alpha}_2) \ \Rightarrow\ \hat{S}(20)_2 = e^{-(20/\hat{\alpha}_2)^{\hat{\gamma}_2}}$

$\vdots$

Sample 10000: $y_{10000,1}, y_{10000,2}, \ldots, y_{10000,25} \ \Rightarrow\ (\hat{\gamma}_{10000}, \hat{\alpha}_{10000}) \ \Rightarrow\ \hat{S}(20)_{10000} = e^{-(20/\hat{\alpha}_{10000})^{\hat{\gamma}_{10000}}}$

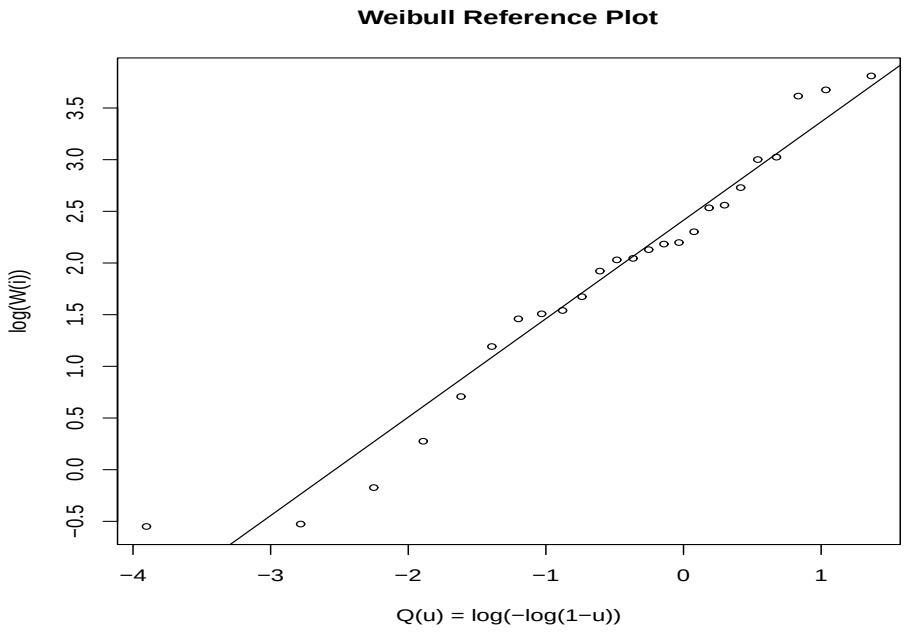
- Use the 10000 values of $\hat{S}(20)$: $\hat{S}(20)_1, \hat{S}(20)_2, \ldots, \hat{S}(20)_{10000}$ to estimate the standard error and necessary percentiles of $\hat{S}(20)$

The following are the times to failure of the 25 devices:

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

If we did not have a model for the data, a distribution-free estimate would be $\hat{S}(20) = P[T > 20] = 5/25 = .2$

Next, we evaluate a Weibull model for the failure times of the 25 devices. The following Weibull reference plot appears to confirm that a Weibull model would be appropriate.

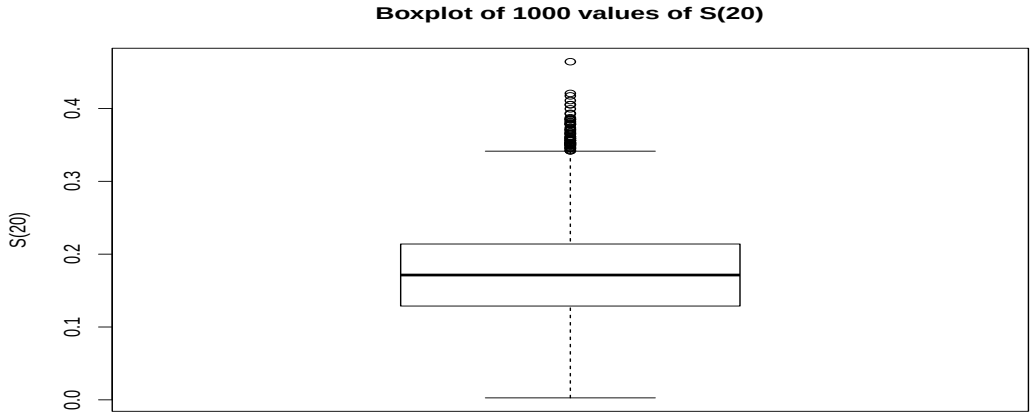


The Anderson-Darling test of a Weibull model has a modified value of 0.304 which yields a p-value > 0.25 using the table on page 34 of Handout 9.

Based on the Weibull reference plot and the A-D gof test, the Weibull model appears to be appropriate for the lifetime data.

R code - parabootweibull,seMLE.R - yields the parametric bootstrap estimates:

```
x =  c(15.321, 9.008, 20.104, 7.729, 45.154, 8.404, 5.332, 0.577, 4.305, 4.517,
  12.594, 6.829, 3.291, 37.175, 0.841, 1.317, 7.613, 20.582, 2.030, 10.001,
  4.666, 12.933, 0.591, 39.454,  8.875)
n =  length(x)
library(MASS)
fitdistr(x,"weibull")
#  OUTPUT from R:
#
#       shape        scale
#    0.9839245   11.4852981
#  ( 0.1512936) ( 2.4660607)
gamma = 0.9839245
alpha = 11.4852981
gamma = fitdistr(x,"weibull")$estimate[1]
alpha = fitdistr(x,"weibull")$estimate[2]
B  =  10000
W  =  matrix(0,B,n)
A =  numeric(B)
A =  rep(0,B)
G =  numeric(B)
G =  rep(0,B)
S = numeric(B)
S = rep(0,B)
{
for (i in 1:B)
W[i,]  =  rweibull(n,gamma,alpha)
}
{
for (i in 1:B)
G[i] = fitdistr(W[i,],"weibull")$estimate[1]
}
{
for (i in 1:B)
A[i] = fitdistr(W[i,],"weibull")$estimate[2]
}
{
for (i in 1:B)
S[i] = exp(-(20/A[i])^G[i])
}
summary(S)
sd(S)
boxplot(S)
out=c(mean(G),sd(G),mean(A),sd(A))
1.0437717  0.1747574 11.5987090  2.4618747
summary(S)
  Min.      1st Qu.    Median     Mean      3rd Qu.     Max.
0.0027      0.1288     0.1714    0.1727     0.2139     0.4643

sd(S)  0.06261848
```



The average value of the 10000 values of $\hat{S}(20)$ was 0.1747 which is nearly identical to our point estimator from the data using the MLE's:

$$\hat{\alpha} = 11.4852981, \quad \hat{\gamma} = .9839245 \quad \Rightarrow$$

$$\hat{S}(20) = e^{-(20/11.4852981)\cdot 9839245} = .178013$$

The standard error of the parameters using the asymptotic formulas are

$\widehat{se}(\hat{\gamma}) = .1512936$ and $\widehat{se}(\hat{\alpha}) = 2.4660607$

which are very close to our bootstrap estimates 0.1747574 and 2.4618747, even though n=25 is not very large.

Finally, we have the estimated standard error of $\hat{S}(20)$ is 0.06236332 which can be used to place a C.I. on S(20).

This would involve a fairly complicated computation to obtain the asymptotic standard error from MLE theory.

.

.

.

START Tuesday 10/22/24

HANDOUT #11: INTERVAL ESTIMATORS

I. Confidence Intervals (CI) for a Parameter θ

(a) Pivot Method
(b) Exact CI
(c) Asymptotic CI
(d) Bootstrap CI
(e) Example of Improper Use of CLThm
(f) Sample Size Determination
(g) Distribution-Free CI for Q(u)

II. Prediction Intervals (PI)

(a) PI for Normal Population Distribution
(b) PI for Exponential Population Distribution

III. Tolerance Intervals (TI)

(a) TI for Normal Population Distribution
(b) Lower Tolerance Bound for Exponential Population Distribution
(c) Distribution-Free Tolerance Bounds

IV. Alternative Approaches to CI, PI, TI

(a) Transformations
(b) Bootstraping

Supplemental Reading

- Sections 6.1, 6.2, 7.1-7.4, 9.1, 9.3, 14.6, and pages 565-566 in Tamhane/Dunlop book

Interval Estimators

Suppose we have a population or process having unknown parameters θ. For example, the population mean μ, standard deviation σ, population proportion p, or population median $Q(.5)$. Alternatively, we may know that the population cdf $F(\cdot;\theta)$ is a member of a family of distributions but θ is unknown. For example, Weibull(θ,γ) or Poisson(λ). We would like to estimate various components of the population. Three such interval estimators are

1. $100(1-\alpha)$ **Confidence Interval on the parameter θ:** Plausible set of values for θ

   a. Based on a random sample $Y_1,\ldots,Y_n$ from a population or process, obtain a point estimator $\hat{\boldsymbol{\theta}}$: MLE, MOM, Robust

   b. Construct interval of values $(\hat{\theta}_L,\hat{\theta}_U)$ such that $P[\hat{\theta}_L \le \theta \le \hat{\theta}_U] = 1-\alpha$.

   c. The $100(1-\alpha)$ C.I. $(\hat{\theta}_L,\hat{\theta}_U)$ reflects the uncertainty in using $\hat{\theta}$ as an estimator of θ.

2. $100(1-\alpha)$ **Prediction Interval on the R.V. Y:**

   a. Let $Y_1,\ldots,Y_n$ be a random sample from a population or process.

   b. Based on the data, predict the value of the next r.v. Y_{n+1} selected from the population or process: $\hat{Y}_{n+1}$.

   c. The $100(1-\alpha)$ P.I. is an interval of values $(\hat{Y}_{n+1,L},\hat{Y}_{n+1,U})$ such that $P[\hat{Y}_{n+1,L} \le Y_{n+1} \le \hat{Y}_{n+1,U}] = 1-\alpha$

3. (P,γ) **Tolerance Interval on the Population or Process:**

   Based on a random sample $Y_1,\ldots,Y_n$ from a population or process, construct an interval of values $(L_{p,\gamma},U_{p,\gamma})$ such that we are $100\gamma\%$ certain that the interval $(L_{p,\gamma},U_{p,\gamma})$ contains at least $100P\%$ of the population values.

   In $L_{p,\gamma}$, p is the proportion of the population values to be contained in the interval and γ is the level of confidence that the interval will in fact contain $100p\%$ of the population values.

Distinct differences in the three types of intervals:

1. The C.I. is making an inference about a fixed population parameter, μ, σ, or a parameter in a family of distributions, for example, β in an exponential family or λ in a Poisson family. For example, we are 99% certain that the mean time to appearance of a tumor is in the interval (40, 60) hours.

2. The P.I. is forecasting or predicting the value of a R.V., for example, we are 98% certain that the tensile strength of the next specimen of alloy tested is in the interval (1.23, 2.31) units. A meteorologist predicts with 80% confidence that the amount of rainfall tomorrow will be .5 to 1 cm.

3. The Tolerance Interval is estimating a region in a population that contains at least $100P\%$ of the population values. For example,

    a. Suppose based on the data we compute an interval of values such that we are 95% certain that 99% of 1 cm bearings produced next month by Company X will have diameters in the region $(.99, 1.02)$ cm. This is a $(P = .99, \gamma = .95)$ Tolerance Interval on distribution for the diameter of ball bearings.

    b. Suppose we compute from the data a lower bound such that we are 80% certain that at least 95% of all new Honda SUV's produced next year will have miles to failure of its transmission of at least 120000 miles. Then, $(120000, \infty)$ would be a $(P = .95, \gamma = .80)$ Lower Tolerance Bound on the miles to failure distribution.

Illustration of Level of Confidence In many samples 95% of intervals would contain μ

- 100 random samples of size $n = 25$ were generated from a population having a $N(27, (.8)^2)$ distribution:

$$Y_{i,1}, Y_{i,2}, \ldots, Y_{i,25}, \quad i = 1, 2, \ldots, 100$$

- From each of the 100 samples, a 95% C.I. for μ was constructed:

$$C.I._i = \bar{Y}_i \pm 1.96\frac{.8}{\sqrt{25}} \quad i = 1, 2, \ldots, 100$$

- There are now 100 estimates of μ: $C.I._{\cdot 1}, C.I._{\cdot 2}, \ldots, C.I._{\cdot 100}$.

- How many of the 100 intervals contain μ?

  The graph on the next page provides the answer to this question.

  This illustrates the relative frequency interpretation of a confidence interval.

- In repeated sampling (many more than 100), the proportion of $100(1-\alpha)\%$ confidence intervals that contain the population parameter is $(1-\alpha)$.

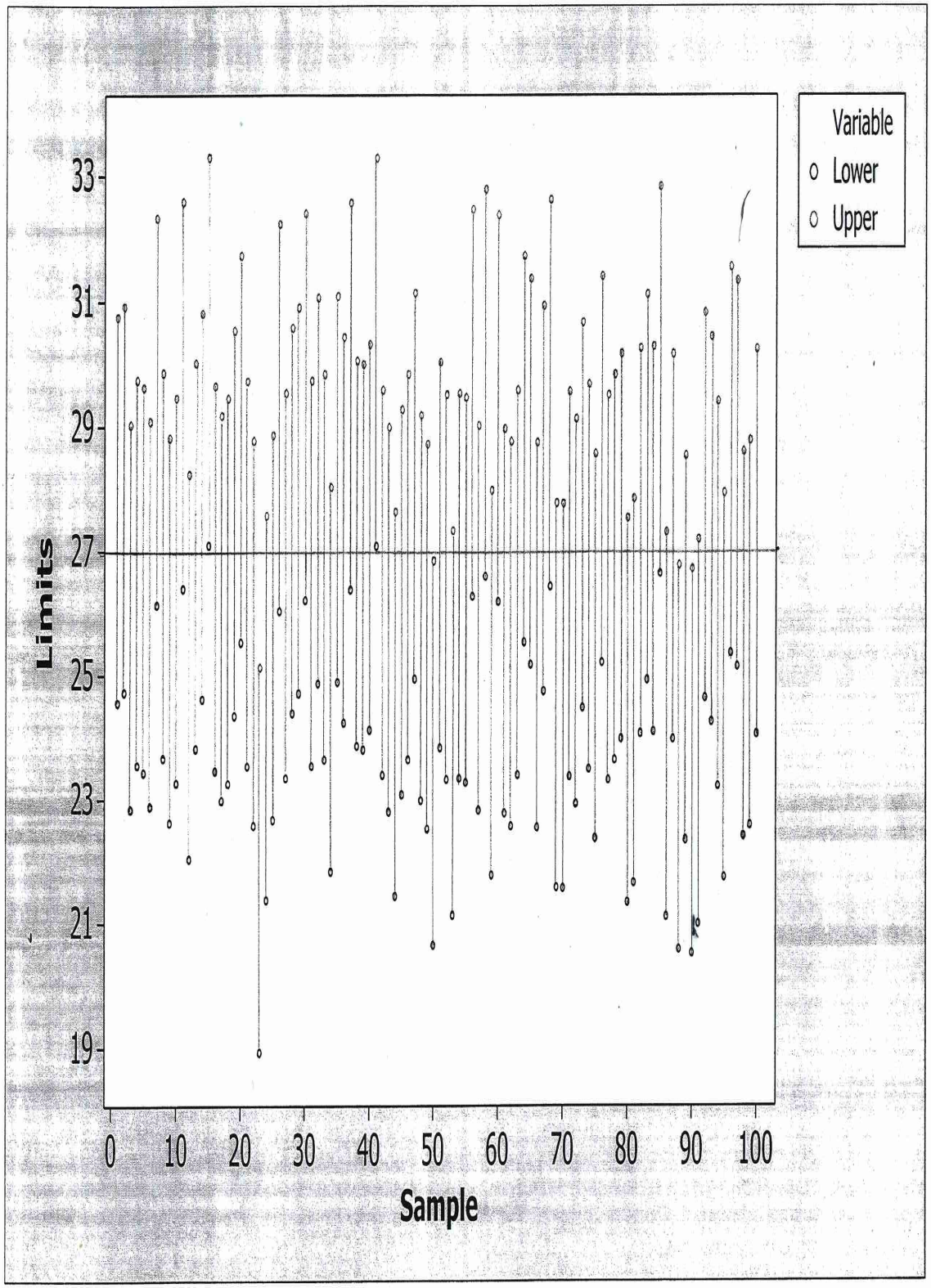
- The proportion of the intervals which **fail** to contain the population parameter is α.

  In the example on the next page, we have $6/100 = .06 \approx .05$ intervals fail to contain $\mu = 27$.

- However, the probability that a realized 95% C.I., (25.31, 25.95), for the population mean, actually contains μ is either 0 or 1.

  The value $100(1-\alpha)\%$ is relative to repeated sampling or relative to the proportion over a very large number of such intervals not to a particular interval.

- The level of confidence $100(1-\alpha)\%$ refers to the process of constructing the confidence interval and not to the actual confidence interval constructed from a given data set. That is, the process used to obtain the 95% C.I. (25.31, 25.95) for the population mean generates intervals of which 95% of the C.I.s produced will contain the population mean and 5% will not contain the population mean.


Variable
Lower
Upper
Limits
33
31
29
27
25
23
21
19
0
10
20
30
40
50
60
70
80
90
100
Sample

Construction of C.I.'s

The Pivot Method will often be used to construct C.I.'s for population or distribution parameters.

1. Find a function of the data $Y = (Y_1, \ldots, Y_n)$ and the parameter θ, $g(Y, \theta)$, such that the sampling distribution of $g(Y, \theta)$ does not depend on θ.

2. Use the cdf of $g(Y, \theta)$, $H(\cdot)$, to determine two percentiles $C_{\frac{\alpha}{2}}$ and $C_{1-\frac{\alpha}{2}}$ satisfying
   $P[C_{\frac{\alpha}{2}} \leq g(Y, \theta) \leq C_{1-\frac{\alpha}{2}}] = 1 - \alpha$, that is,
   $H(C_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $H(C_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$,
   where H is the cdf of the pivot.

3. Invert the inequality $C_{\frac{\alpha}{2}} \leq g(Y, \theta) \leq C_{1-\frac{\alpha}{2}}$ to obtain

   $L(C, g(Y, \theta)) \leq \theta \leq U(C, g(Y, \theta))$.

   Thus conclude that $(L(C, g(Y, \theta)), U(C, g(Y, \theta)))$ is a $100(1 - \alpha)\%$ C.I. for θ.

Determining the Distribution of Pivot

After we select the appropriate pivot, the cdf $H(\cdot)$ of the sampling distribution of the pivot must be determined in order to obtain $C_{\frac{\alpha}{2}}$ and $C_{1-\frac{\alpha}{2}}$. The methods for determining these distributions involve

1. Exact mathematical derivation of the distribution of the pivot

2. Using the asymptotic distribution of the pivot - Wald C.I.

3. Using a bootstrap approximation to the distribution of the pivot

1. Mathematical Determination of Distribution of Pivot

EXAMPLE 1. C.I. for μ for $N(\mu, \sigma^2)$ Distribution

Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$

Pivot is $g(Y, \mu) = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$

which is a function of the data $\bar{Y}$, S and the unknown parameter μ.

The sampling distribution of $g(Y, \mu)$ is a $t-$distribution with $df = n-1$ which does not depend on the unknown parameters $\boldsymbol{\theta} = (\mu, \sigma)$.

Next we obtain the percentiles

$C_{\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}} = -qt(1-\frac{\alpha}{2}, n-1)$ and $C_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} = qt(1-\frac{\alpha}{2}, n-1)$, where

$t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = n-1$. That is,

$$1-\alpha = P[-t_{\frac{\alpha}{2}} \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq t_{\frac{\alpha}{2}}] \Rightarrow \quad \bar{Y} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \Rightarrow$$

$$\bar{Y} \pm t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ C.I. for μ when population distribution is $N(\mu, \sigma^2)$.

EXAMPLE 2. C.I. for σ for $N(\mu, \sigma^2)$ Distribution

Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$

Pivot is $g(Y, \sigma) = \frac{(n-1)S^2}{\sigma^2}$

which is a function of the data through S and the unknown parameter σ.

The sampling distribution of $g(Y, \sigma)$ is a chi-square distribution with $df = n-1$ which does not depend on the unknown parameter, $\boldsymbol{\theta} = \sigma$.

Next we obtain the percentiles $C_{\frac{\alpha}{2}} = \chi^2_{\frac{\alpha}{2}} = qchisq(\frac{\alpha}{2}, n-1)$ and

$C_{1-\frac{\alpha}{2}} = \chi^2_{1-\frac{\alpha}{2}} = qchisq(1-\frac{\alpha}{2}, n-1)$,

which are the lower and the upper $\frac{\alpha}{2}$ percentile of the chi-square distribution with $df = n-1$. That is,

$$1-\alpha = P\left[\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\right] \Rightarrow \quad \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \Rightarrow$$

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}}\right)$$

is a $100(1-\alpha)\%$ C.I. for σ when population distribution is $N(\mu, \sigma^2)$.

EXAMPLE 3. C.I. for $\mu_2 - \mu_1$ for comparing two $N(\mu_i, \sigma_i^2)$ Distribution with $i = 1,2$

Let $X = (X_1, \ldots, X_{n_1})$ be iid $N(\mu_1, \sigma_1^2)$ and $Y = (Y_1, \ldots, Y_{n_2})$ be iid $N(\mu_2, \sigma_2^2)$, with X's and Y's independent.

Case 1: $\sigma_1 = \sigma_2 = \sigma$ (with σ unknown)

Pivot is $g(Y, X, \mu_1, \mu_2) = \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}} = \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$,

where $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_2+n_1-2}$ is the weighted (pooled) estimator of σ^2.

The sampling distribution of $g(Y, X, \mu_1, \mu_2)$ is a $t-$distribution with $df = n_1 + n_2 - 2$ which does not depend on the unknown parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2)$.

Next we obtain the percentiles $C_{\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}} = -qt(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)$ and $C_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} = qt(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)$,

where $t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = n_1 + n_2 - 2$. That is,

$$1 - \alpha = P\left[-t_{\frac{\alpha}{2}} \leq \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}} \leq t_{\frac{\alpha}{2}}\right] \Rightarrow$$

$$(\bar{Y} - \bar{X}) - t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \leq \mu_2 - \mu_1 \leq (\bar{Y} - \bar{X}) + t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \Rightarrow$$

$$\bar{Y} - \bar{X} \pm t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

is a $100(1 - \alpha)\%$ C.I. for $\mu_2 - \mu_1$ when $Y_i's$ and $X_i's$ are independent r.v's from population distributions $N(\mu_i, \sigma^2), i = 1, 2$.

Case 2: $\sigma_1 \neq \sigma_2$

Let $X = (X_1, \ldots, X_{n_1})$ be iid $N(\mu_1, \sigma_1^2)$ and $Y = (Y_1, \ldots, Y_{n_2})$ be iid $N(\mu_2, \sigma_2^2)$, with X's and Y's independent.

with $\sigma_1 \neq \sigma_2$, hence do not use pooled estimator

Pivot is $g(Y, X, \mu_1, \mu_2) = \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$,

where S_i^2 are the sample variances and estimators of σ_i^2.

The sampling distribution of $g(Y, X, \mu_1, \mu_2)$ is not exactly a t-distribution but Welch-Satterthwaite proved that the distribution is approximately a $t-$distribution with

$df_{WS} = \frac{(w_1+w_2)^2}{\frac{w_1^2}{n_1-1}+\frac{w_2^2}{n_2-1}}$,

where $w_i = \frac{S_i^2}{n_i}, i = 1, 2.$

Next we obtain the percentiles

$C_{\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}} = -qt(1-\frac{\alpha}{2}, df_{WS})$ and $C_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} = qt(1-\frac{\alpha}{2}, df_{WS})$,

where $t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = df_{WS}$.
That is,

$$1-\alpha \approx P\left[-t_{\frac{\alpha}{2}} \leq \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \leq t_{\frac{\alpha}{2}}\right] \quad \Rightarrow$$

$$(\bar{Y}-\bar{X}) - t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \leq \mu_2-\mu_1 \leq (\bar{Y}-\bar{X}) + t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \quad \Rightarrow$$

$$\bar{Y}-\bar{X} \pm t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

is an approximate $100(1-\alpha)\%$ C.I. for $\mu_2 - \mu_1$ when $Y_i's$ and $X_i's$ are independent r.v's from population distributions $N(\mu_i, \sigma_i^2), i = 1, 2.$

Case 3: Y_i and X_i paired

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be iid pairs of observations.

The goal is to construct a C.I. on $\mu_Y - \mu_X$.

The problem with using the pivot for the case were the X's and Y's were independent is displayed as follows:

When using a paired analysis, the pairing reduces the variability of the individual differences if the $(X, Y)-$pairs are positively correlated:

$$Var(\bar{Y} - \bar{X}) = Var(\bar{Y}) + Var(\bar{X}) - 2\sigma_X \sigma_Y Corr(\bar{Y}, \bar{X}) < Var(\bar{Y}) + Var(\bar{X})$$

provided $Corr(\bar{Y}, \bar{X}) > 0$

In many experiments involving paired observations, the sample size n is relatively small and hence an estimate of $Corr(\bar{Y}, \bar{X})$ is not feasible. Thus, the data is reduced to the differences in the n pairs:

Let $D_i = Y_i - X_i,\ i = 1, \ldots, n$ with $D_i's$ iid $N(\mu_D, \sigma_D^2)$, and $\mu_D = \mu_2 - \mu_1$.

Pivot is

$$g(D, \mu_D) = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

which is a function of the data $\bar{D}$, S_D and the unknown parameter μ_D.

The sampling distribution of $g(Y, \mu)$ is a $t-$distribution with $df = n - 1$ which does not depend on the unknown parameters $\boldsymbol{\theta} = (\mu_D, \sigma_D)$. We have converted the problem to finding a C.I. for the mean of a normal population. Therefore,

$$\bar{D} \pm t_{\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

is a $100(1 - \alpha)\%$ C.I. for $\mu_2 - \mu_1$ when $D_i's$ are normally distributed.

Note: $t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = n - 1$, not $df = 2(n - 1)$.

When using a paired analysis, the pairing reduces the variability of the individual differences if the $(X, Y)-$pairs are positively correlated:

$$Var(\bar{D}) = Var(\bar{Y} - \bar{X}) = Var(\bar{Y}) + Var(\bar{X}) - 2Cov(\bar{Y}, \bar{X}) < Var(\bar{Y}) + Var(\bar{X})$$

The reduction of the variance would lead to a C.I. of narrow width, and hence a more precise estimation of $\mu_2 - \mu_1$, if a paired data experiment was conducted rather than having two independent data sets. However, the reduction in variance is obtained only by a concurrent reduction in the *df* for the $t-$percentiles. That is,

$$t_{\frac{\alpha}{2}, n-1} > t_{\frac{\alpha}{2}, 2(n-1)}$$

Thus, in attempting to decide between using a paired analysis or two independent samples, this trade-off between reduced variability and reduced *df* must be taken into consideration.

EXAMPLE 4. C.I. for β in an Exponential Distribution 

Let $Y_1, \ldots, Y_n$ be iid Exp(β)

The MLE of β is $\hat{\beta} = \bar{Y}$ which will be used to form the Pivot:

Pivot is $g(Y, \beta) = \frac{2n\bar{Y}}{\beta}$.

Verify that the sampling distribution of $g(Y, \beta)$ does not depend on β.

Recall, $n\bar{Y} = \sum_{i=1}^{n} Y_i$ has a $Gamma(\alpha = n, \beta)$ distribution.

Further recall, that if W has a $Gamma(\alpha = n, \beta)$ distribution then $X = 2W/\beta$ has a chi-square distribution with $df = 2n$.

Therefore, the sampling distribution of the pivot $g(Y, \beta) = \frac{2n\bar{Y}}{\beta}$ is chi-square with df=2n.

It then follows that

$1 - \alpha = P\left[\chi^2_{\frac{\alpha}{2}} \leq \frac{2n\bar{Y}}{\beta} \leq \chi^2_{1-\frac{\alpha}{2}}\right]$,

where $\chi^2_{\frac{\alpha}{2}} = qchisq(\frac{\alpha}{2}, 2n)$ and $\chi^2_{1-\frac{\alpha}{2}} = qchisq(1 - \frac{\alpha}{2}, 2n)$

are respectively the lower and upper $\frac{\alpha}{2}$-percentiles of a chi-square distribution with $df = 2n$.

Inverting the inequalities yields a $100(1-\alpha)\%$ C.I. for β :

$$\left(\frac{2n\bar{Y}}{\chi^2_{1-\frac{\alpha}{2}}}, \frac{2n\bar{Y}}{\chi^2_{\frac{\alpha}{2}}}\right)$$

The tables in most textbooks provide the upper percentiles of the chi-square distribution.

For example, if we wanted a 95% C.I. with n=10, then we would select the following values from the table:

$\chi^2_{\frac{\alpha}{2}} = \chi^2_{.025,20} = 9.591$ using $\alpha = 1 - .025 = .975$ in Chisquared table or use **qchisq(.025,20** in R

$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{1-.025,20} = 34.170$ using $\alpha = .025$ in Chisquared table or use **qchisq(.975,20** in R

EXAMPLE 5. C.I. for $\theta = \beta_1/\beta_2$ for comparing Two Exponential cdf's

Let $Y = (Y_1, \ldots, Y_{n_2})$ be iid Exp(β_2) and $X = (X_1, \ldots, X_{n_1})$ be iid Exp(β_1) with $Y_i's$ and $X_i's$ independent.

Pivot is $g(Y, X, \beta_1, \beta_2) = \frac{\bar{Y}/\beta_2}{\bar{X}/\beta_1} = \frac{\left(2n_2\bar{Y}/\beta_2\right)/(2n_2)}{\left(2n_1\bar{X}/\beta_1\right)/(2n_1)}$

The sampling distribution of $g(Y, X, \beta_1, \beta_2)$ is an F-distribution with $df = (2n_2, 2n_1)$.

This follows from the results in EXAMPLE 4 and the result that the ratio of independent chisquare r.v.s/df has an $F-$distribution. It then follows that

$$1 - \alpha = P\left[F_{\frac{\alpha}{2}} \leq \frac{\left(2n_2\bar{Y}/\beta_2\right)/(2n_2)}{\left(2n_1\bar{X}/\beta_1\right)/(2n_1)} \leq F_{1-\frac{\alpha}{2}}\right],$$

where $F_{\frac{\alpha}{2}} = qf(\frac{\alpha}{2}, 2n_2, 2n_1)$ and $F_{1-\frac{\alpha}{2}} = qf(1-\frac{\alpha}{2}, 2n_2, 2n_1)$ are the lower and upper $\frac{\alpha}{2}$-percentiles of a F-distribution with $df = (2n_2, 2n_1)$.

Inverting the inequalities yields a $100(1-\alpha)\%$ C.I. for β_1/β_2 :

$$\left(\frac{\bar{X}}{\bar{Y}}F_{\frac{\alpha}{2}}, \frac{\bar{X}}{\bar{Y}}F_{1-\frac{\alpha}{2}}\right)$$

Note that in using the F-tables, the lower percentile is related to the upper percentile through $F_{\alpha,n_2,n_1} = 1/F_{1-\alpha,n_1,n_2}$,

$$\begin{aligned} 1 - \alpha = P[F_{n_2,n_1} \geq F_{\alpha,n_2,n_1}] &= P\left[\frac{1}{F_{n_2,n_1}} \leq \frac{1}{F_{\alpha,n_2,n_1}}\right] \\ &= \\ &= P\left[F_{n_1,n_2} \leq \frac{1}{F_{\alpha,n_2,n_1}}\right] \\ &= 1 - \alpha \end{aligned}$$

Therefore, we have that

$$F_{1-\alpha,n_1,n_2} = \frac{1}{F_{\alpha,n_2,n_1}}$$

EXAMPLE 6. C.I. for Population Proportion p

Suppose we have a population consisting of two types of units A or B with p being the proportion of Type A units. Alternatively, we may have a process that produces one of two types of units A or B with p being the probability of a Type A unit occurring. Let Y be the number of Type A outcomes in n *iid* trials or the number of Type A units occurring in a random sample taken with replacement from a population. In either case, $\hat{p} = Y/n$ and Y has a $Bin(n, p)$ distribution. A C.I. for p can be constructed in a number of ways.

The Clopper-Pearson C.I. for p:

See page 434 and Exercise 9.21 in Cassela-Berger for mathematical details.

The Clopper-Pearson $100(1-\alpha)\%$ C.I. for p can be expressed as

$$\{p \mid P[B(n;p) \leq y] \geq \alpha/2\} \cap \{p \mid P[B(n;p) \geq y] \geq \alpha/2\}$$

where the observed value of Y is y.

The interval can alternatively be expressed as $(P_L,\ P_U)$ where the values of P_L and P_U are obtained from the following equations: Suppose we observe Y=y in the study, then solve for P_L and P_U

$$(1) \quad \sum_{k=y}^{n} \binom{n}{k} P_L^k (1-P_L)^{n-k} = \frac{\alpha}{2}$$

$$(2) \quad \sum_{k=0}^{y} \binom{n}{k} P_U^k (1-P_U)^{n-k} = \frac{\alpha}{2}$$

Determining Limits for Clopper-Pearson C.I.

Case 1: If $y = 0$ then

$P_L = 0$ and $P_U = 1 - (\frac{\alpha}{2})^{1/n}$

The justification of these solutions is as follows:

If $y = 0$ then (1) $\Rightarrow \sum_{k=0}^{n} \binom{n}{k} P_L^k (1-P_L)^{n-k} = 1$ unless $P_L = 0$

If $y = 0$ then (2) $\Rightarrow \sum_{k=0}^{0} \binom{n}{k} P_U^k (1-P_U)^{n-k} = (1-P_U)^n = \frac{\alpha}{2} \Rightarrow P_U = 1 - (\frac{\alpha}{2})^{1/n}$

Case 2: If $y = n$

then $P_L = (\frac{\alpha}{2})^{1/n}$ and $P_U = 1$.

The justification of these solutions is as follows:

If $y = n$ then (2) $\Rightarrow \sum_{k=0}^{n} \binom{n}{k} P_U^k (1-P_U)^{n-k} = 1$ unless $P_U = 1$

If $y = n$ then (1) $\Rightarrow \sum_{k=n}^{n} \binom{n}{k} P_L^k (1-P_L)^{n-k} = P_L^n = \frac{\alpha}{2} \Rightarrow P_L = (\frac{\alpha}{2})^{1/n}$

Case 3: For $y = 1, 2, \cdots, n-1$; (more general case)

Using the relationship between the binomial distribution, beta distribution, and the Fisher - F distribution, the solution to the equations can be expressed by using the upper $\frac{\alpha}{2}$ percentiles from the

F distribution: $F_{df_1, df_2, \frac{\alpha}{2}} = qf(1 - \frac{\alpha}{2}, df_1, df_2)$:

$$P_L = \frac{1}{1 + \left(\frac{n-y+1}{y}\right) F_{2(n-y+1),2y,\frac{\alpha}{2}}}; \qquad P_U = \frac{\left(\frac{y+1}{n-y}\right) F_{2(y+1),2(n-y),\frac{\alpha}{2}}}{1 + \left(\frac{y+1}{n-y}\right) F_{2(y+1),2(n-y),\frac{\alpha}{2}}}$$

Examples when $y = 0$:

With $n = 20$, $y = 0$ then $\hat{p} = 0$ and 95% C.I. for p is

$P_L = 0$ and $P_U = 1 - (.025)^{1/20} = .168 \quad \Rightarrow (0, .168)$

With $n = 100$, $y = 0$ then $\hat{p} = 0$ and 95% C.I. for p is

$P_L = 0$ and $P_U = 1 - (.025)^{1/100} = .0362 \quad \Rightarrow (0, .0362)$

Examples when $y = n$:

With $n = 20$, $y = 20$ then $\hat{p} = 1$ and 95% C.I.. for p is

$P_L = (.025)^{1/20} = .832$ and $P_U = 1 \quad \Rightarrow (.832, 1)$

With $n = 100$, $y = 100$ then $\hat{p} = 1$ and 95% C.I. for p is

$P_L = (.025)^{1/100} = .9638$ and $P_U = 1 \quad \Rightarrow (.9638, 1)$

Example when $0 < y < n$ Suppose $n = 20$ and $y = 5$. Obtain a 95% C.I. for p.

$$F_{2(n-y+1),2y,\frac{\alpha}{2}} = F_{32,10,.025} = qf(1 - .025, 32, 10) = 3.297234$$

$$F_{2(y+1),2(n-y),\frac{\alpha}{2}} = F_{12,30,.025} = qf(1 - .025, 12, 30) = 2.412034$$

$$P_L = \frac{1}{1 + \left(\frac{16}{5}\right) F_{32,10,.025}} = \frac{1}{1 + \left(\frac{16}{5}\right)(3.297234)} = .087$$

$$P_U = \frac{\left(\frac{6}{15}\right) F_{12,30,.025}}{1 + \left(\frac{6}{15}\right) F_{12,30,.025}} = \frac{\left(\frac{6}{15}\right)(2.412034)}{1 + \left(\frac{6}{15}\right)(2.412034)} = .491$$

Therefore, the 95% C.I. for p is (.087, .491).

Comments

- What is the connection between the binomial distribution and the F-distribution?

  Suppose X is distributed $Bin(n, p)$, then

  $P[X \geq x] = P[Y \leq p]$ where Y is distributed $Beta(\alpha = x, \beta = n - x + 1)$.

  Furthermore, suppose W has an $F-$distribution with $df_1 = \nu_1, df_2 = \nu_2$ then

  $$\frac{\left(\frac{\nu_1}{\nu_2}\right) W}{1 + \left(\frac{\nu_1}{\nu_2}\right) W} \quad \text{has a} \quad Beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) \quad \text{distribution}$$

- One problem with the Clopper-Pearson C.I. is that it is necessarily conservative, that is, it has coverage probability greater than $1 - \alpha$. This is due to that the binomial distribution being discrete and hence it is impossible to obtain bounds having exactly a $1 - \alpha$ probability.

  Therefore, the bounds are set so that the coverage probability is at least $1-\alpha$. A problem with having a slightly higher coverage probability is that the width of the C.I. may be somewhat wider than a corresponding interval having a coverage probability of exactly $1 - \alpha$.

STOP Friday 10/22/21

START Monday 10/25/21

Asymptotic (large n) Results for the Sampling Distribution of Pivot

When we are unable to derive the exact sampling distribution of the pivot it may be able to obtain asymptotic (large n) approximations. For example, in EX 1, EX 2, and EX 3, even if the population distributions were non-normal but the sample sizes were large it would be possible to construct the C.I.'s using the central limit theorem results with the sampling distribution of the sample mean having approximately a normal distribution and S being a consistent estimator of σ for large n. Because n is large the t-based percentiles would be essentially the same as the standard normal percentiles. Therefore, the C.I.'s would have approximately the correct level of confidence. However, the sample size necessary to invoke the central limit theorem results varies depending on the true population distribution as was seen in Handout 10, sampling distribution handout. The following example will illustrate the problems that may result when n is too small.

Improper Use of Central Limit Theorem

If we have a simple random sample of size $n = 20$ from a $N(\mu, \sigma^2)$ distribution and construct 100 95% C.I.'s for μ using the t-distribution based pivot,

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}$$

we would expect approximately 5 of the 100 intervals to fail to contain μ. The fact that the t-distribution is the valid sampling distribution for the pivot,

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

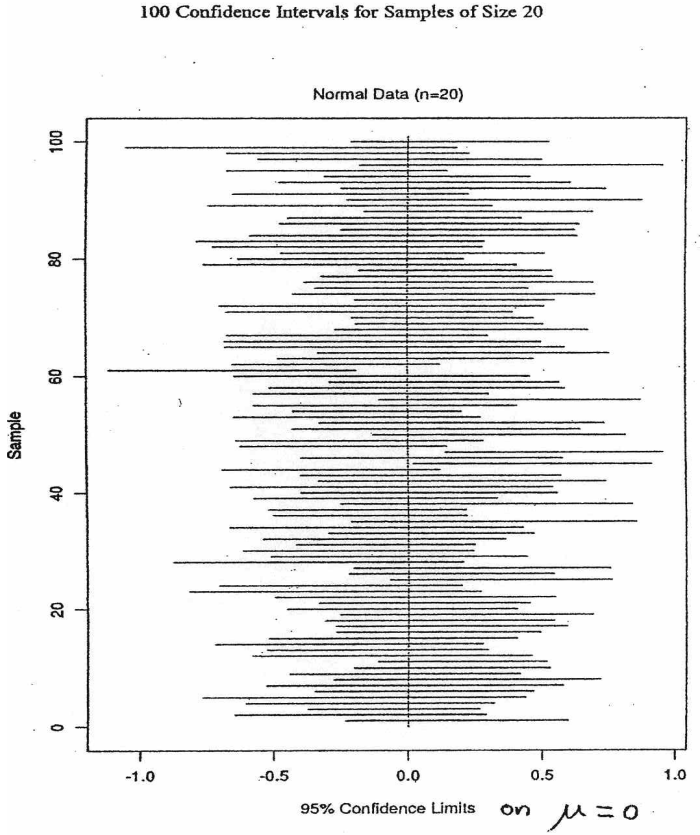
requires that

1. $\bar{X}$ have a normal distribution
2. $\frac{(n-1)S^2}{\sigma^2}$ have a chi-square distribution
3. $\bar{X}$ and S be independent.

These conditions are met when sampling from a $N(\mu, \sigma^2)$ distribution. This is illustrated in the graphs on the next page. Note that only 3 of the 100 C.I.'s fail to contain μ, the sampling distribution of $\bar{X}$ is nearly normal in shape and the scatterplot indicates that $\bar{X}$ and S are uncorrelated.

Next, suppose we have a simple random sample of size $n = 20$ from a $Exp(\beta)$ distribution and construct 100 95% C.I.'s for $\mu = \beta$ using the t-distribution. What is the impact of the skewness of the exponential distribution on the level of the C.I.'s. Note that now instead of having approximately 5 of the 100 C.I.'s fail to contain μ we now have 9 of the 100 C.I.'s fail to contain μ. Why is the coverage probability so much less than the stated 95%? Examining the graphs, we note that the sampling distribution of $\bar{X}$ appears normal in shape and the sampling distribution of S is similar to the shape from sampling from a normal distribution. However, the scatterplot of S vs $\bar{X}$ show a very strong positive correlation which would violate our independence requirement. Why are $\bar{X}$ and S correlated when sampling from an exponential distribution? A heuristic explanation is that in the exponential distribution both $\bar{X}$ and S are estimating the same parameter β because $\mu = \beta$ and $\sigma = \beta$. However, a similar sort of behavior can be seen in many other right skewed distributions.

Population Distribution is $N(0,1)$



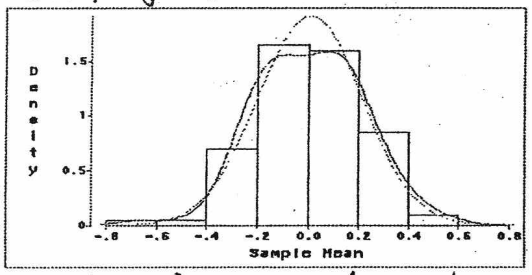
Based on $X_1, \cdots, X_{20}$ iid $N(0,1)$

$\bar{X} \pm t_{.025,19} \frac{S}{\sqrt{20}}$

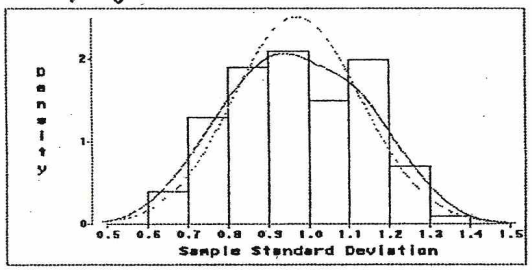
(100 Reps of the above Process)

Population Has a $N(0,1)$ Distribution

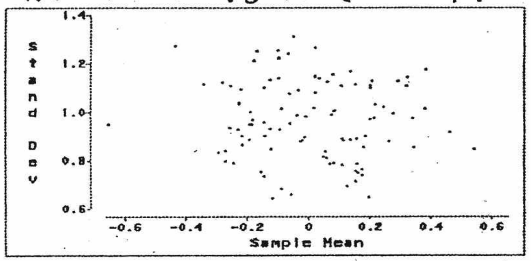
Sampling Distribution of $\bar{X}$ with $n=20$



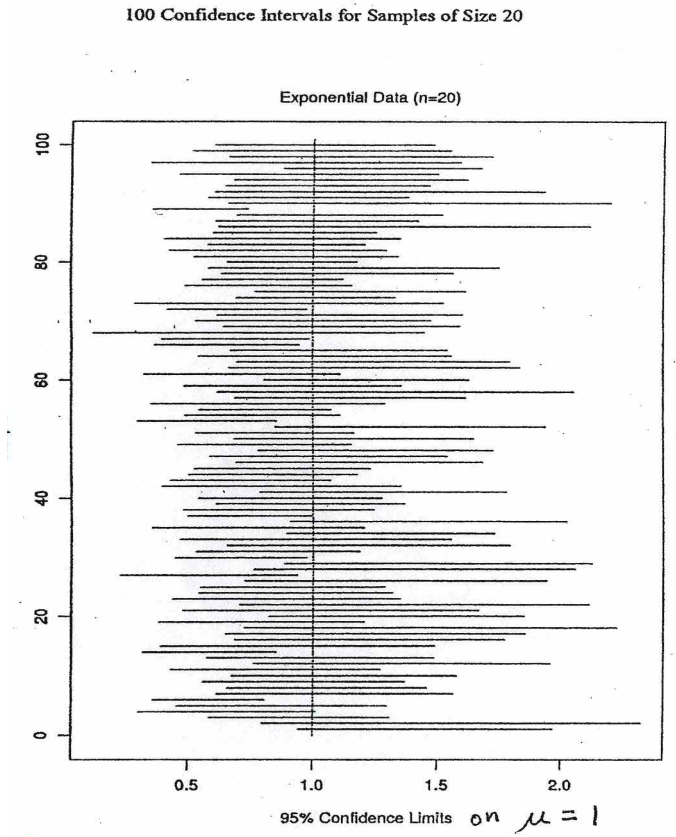
Sampling Distribution of S with $n=20$



Plot of S vs $\bar{X}$ (100 reps)



Population Distribution is $Exp(\lambda=1)$

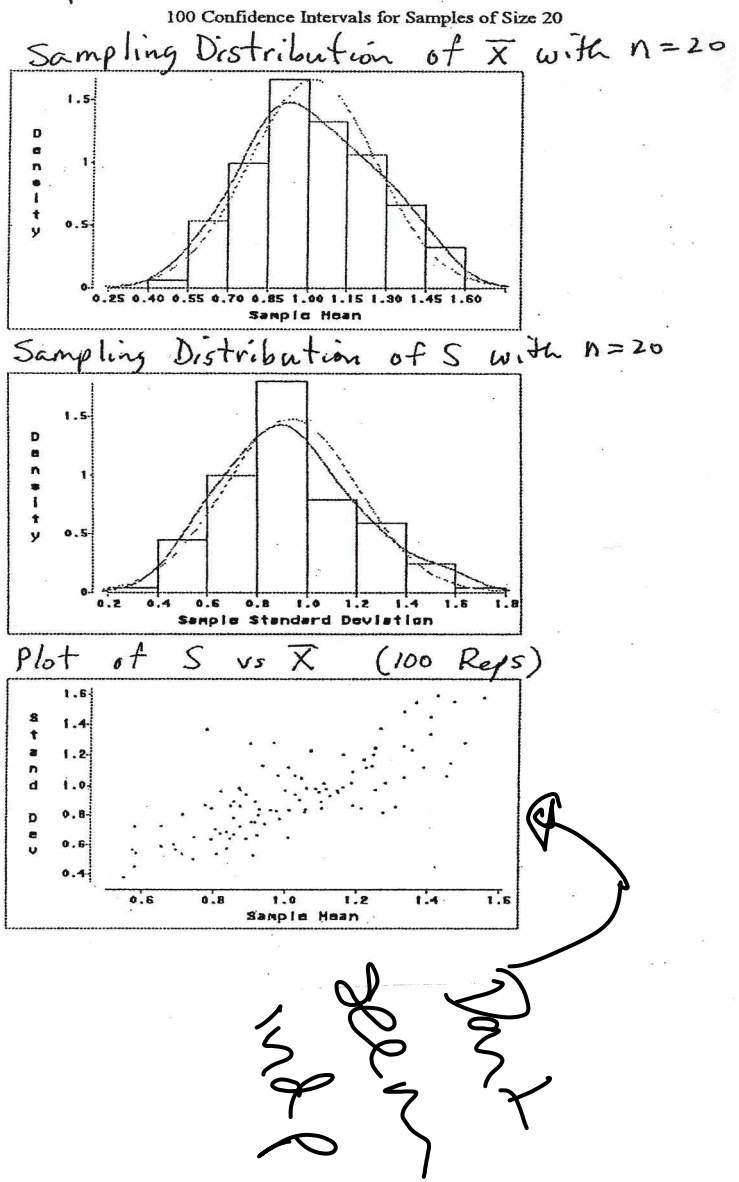


Based on $X_1, \cdots, X_{20}$ iid $Exp(\lambda=1)$

$\bar{X} \pm t_{.025} \frac{S}{\sqrt{20}}$

(100 Reps of the Above Prouss)

Population Distribution is $Exp(\lambda=1)$



Don't Memorize Me!

With very heavy tailed distributions, the sample standard deviation S has a much more skewed distribution than its distribution when sampling from a normal distribution. Thus the values of S are unusually small too often with a few very large values. With highly skewed distributions, the values of $\bar{Y}$ and S may be positively correlated when using a relatively small value for n. The following simulation results demonstrate the effects of these two results if the t-based 95% C.I. is constructed for a population mean μ: $\bar{Y} \pm t_{.025}S/\sqrt{n}$. Five symmetric distributions were used and their corresponding "half" distribution was obtained by folding the distribution over its point of symmetry. For each of the three samples sizes $n = 10, 20, 30$ and ten distributions, 10000 samples of size n were generated. The table provides the Coverage Probability of these 10000 95% C.I.'s for μ and the correlation between $\bar{Y}$ and S. (These results are from Dr. Cline).

n=10		
Distribution	Coverage Probability	Corr(Y, S)
Parabola	0.942	0.019
Half-Parabola	0.944	0.331
Normal	0.950	-0.003
Half-Normal	0.937	0.595
Double-Exponential	0.937	-0.006
Exponential	0.900	0.759
Symmetric Lognormal	0.960	0.038
Lognormal	0.839	0.866
Symmetric Pareto	0.839	-0.012
Pareto	0.839	0.853
n=20		
Distribution	Coverage Probability	Corr(Y, S)
Parabola	0.950	0.018
Half-Parabola	0.948	0.342
Normal	0.948	0.001
Half-Normal	0.946	0.590
Double-Exponential	0.946	-0.012
Exponential	0.918	0.750
Symmetric Lognormal	0.959	-0.027
Lognormal	0.868	0.853
Symmetric Pareto	0.868	-0.020
Pareto	0.868	0.817
n=30		
Distribution	Coverage Probability	Corr$(\bar{Y}, S)$
Parabola	0.951	0.020
Half-Parabola	0.951	0.349
Normal	0.948	-0.008
Half-Normal	0.941	0.600
Double-Exponential	0.941	0.009
Exponential	0.921	0.739
Symmetric Lognormal	0.960	0.007
Lognormal	0.884	0.825
Symmetric Pareto	0.884	-0.044
Pareto	0.884	0.789

If data are not normal 95% CIs do not actually contain 0 95% of the time.

Asymptotic C.I. for Population Proportion p - Wald Confidence Intervals

Let Y be the number of Type A outcomes in n *iid* trials or the number of Type A units occurring in a random sample taken with replacement from a population.

In either case, $\hat{p} = Y/n$ and Y has a $Bin(n, p)$ distribution.

Using the C.L.Th. for the binomial distribution we obtain the following results:

- The sampling distribution of $\hat{p} = Y/n$ has

  asymptotic mean and standard deviation: $\mu_A = p$ and $\sigma_A = \sqrt{p(1-p)}/\sqrt{n}$

  Therefore, the appropriate pivot is given by

- Pivot=$g(Y, p) = \frac{\hat{p}-p}{\sqrt{p(1-p)}/\sqrt{n}}$

  which has approximately a $N(0, 1)$ distribution for large n by the Central Limit Theorem.

Approach 1: Wald C.I. for p:

Replace p with $\hat{p}$ in the denominator of the pivot

$$g(Y, p) = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}$$

and require that $min(n\hat{p}, n(1-\hat{p})) \geq 5$.

This results in the Wald $100(1-\alpha)\%$ C.I. for p:

$$CI_S = \quad \hat{p} \pm Z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

The Wald C.I. does not provide adequate coverage for small n due to the estimation of the standard deviation of $\hat{p}$.

Also, the C.I. is meaningless if $\hat{p} = 0$ or $\hat{p} = 1$ because in both cases C.I. $= \hat{p} \pm 0$

Therefore, an alternative approach was given by Wilson.

Approach 2, Wilson C.I. for p

Wilson(1927) used the original pivot, with p not $\hat{p}$ to obtain a C.I. for p by inverting the following inequalities:

$$1-\alpha \approx P\left[\frac{|\hat{p}-p|}{\sqrt{p(1-p)}/\sqrt{n}} \leq Z_{\frac{\alpha}{2}}\right] \Rightarrow \text{ C.I. is } \frac{|\hat{p}-p|}{\sqrt{p(1-p)}/\sqrt{n}} \leq Z_{\frac{\alpha}{2}}$$

Squaring this interval yields

$$(\hat{p}-p)^2 \leq Z^2_{\frac{\alpha}{2}} \frac{p(1-p)}{n} \Rightarrow$$

$$h(p) = (1+C)p^2 - (C+2\hat{p})p + \hat{p}^2 \leq 0 \text{ where } C = \frac{1}{n} Z^2_{\frac{\alpha}{2}}$$

Next, need to solve inequality $h(p) \leq 0$ for p, that is, find the region $\{p : h(p) \leq 0\}$.

This yields the following region:

$$\frac{n\hat{p} + .5Z^2_{\frac{\alpha}{2}}}{n + Z^2_{\frac{\alpha}{2}}} \pm \frac{\sqrt{n} Z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p}) + \frac{1}{4n} Z^2_{\frac{\alpha}{2}}}}{n + Z^2_{\frac{\alpha}{2}}}$$

Set $\tilde{Y} = Y + .5Z^2_{\frac{\alpha}{2}}, \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}}, \quad \tilde{p} = \tilde{Y}/\tilde{n}$ yields the Wilson C.I. for p:

$$CI_W = \quad \tilde{p} \pm \frac{\sqrt{n} Z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p}) + \frac{1}{4n} Z^2_{\frac{\alpha}{2}}}}{\tilde{n}}$$

Note that for large n ($n \geq 40$),

$$\begin{aligned} \tilde{p} &= \frac{Y + .5Z^2_{\alpha/2}}{n + Z^2_{\alpha/2}} \\ &= \frac{Y}{n + Z^2_{\alpha/2}} + \frac{.5Z^2_{\alpha/2}}{n + Z^2_{\alpha/2}} \\ &\approx \frac{Y}{n} \\ &= \hat{p} \end{aligned}$$

This leads to the Agresti-Coull C.I. for p.

Approach 3, Agresti-Coull C.I. for p

Agresti-Coull C.I. for p:

$$CI_{AC} = \quad \tilde{p} \pm Z_{\frac{\alpha}{2}} \frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{\tilde{n}}}$$

Special Case of Agresti-Coull for 95% C.I.

For a 95% C.I. we have

$$Z_{\frac{\alpha}{2}} = Z_{.025} = 1.96 \approx 2 \Rightarrow$$

$$\tilde{Y} = Y + .5Z^2_{\frac{\alpha}{2}} \approx Y + 2 \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}} \approx n + 4 \quad \Rightarrow$$

$$\tilde{p} = \frac{Y+2}{n+4} \quad \text{Requires a level of 95\%}$$

This is the formula seen in many textbooks.

There are many other approaches but we will only consider four of these confidence intervals for p:

1. the confidence interval using binomial distribution, Clopper-Pearson interval: CI_{CP},
2. the Wald confidence interval using the asymptotic distribution: CI_{WALD},
3. the Wilson confidence interval: CI_{WILS}
4. the Agresti-Coull confidence interval: CI_{AC}.

The following table computes all four C.I.'s for a variety of values for n and Y to illustrate how the width of the intervals vary considerably from the Binomial based C.I.

When $Y = n\hat{p} < 5$, the Wald C.I.'s are not recommended.

The C.I.'s are presented to just illustrate their inaccuracies in these situations relative to the Binomial based C.I.'s.

From the formulas for the Wilson and Agresti-Coull C.I.'s we know that the Agresti-Coull C.I. is always wider than the Wilson C.I.

From the table, for n relatively small both Wilson and Agresti-Coull C.I.'s may be too narrow and hence have coverage probability less than 95%.

Comparison of Various 95% C.I.'s for Proportion

n	$\hat{p}$	y	Wald	Wilson	Agresti-Coull	Clopper-Pearson
10	.10	1	(.0000, .2859)	(.0179, .4042)	(.0000, .4260)	(.003, .445)
10	.20	2	(.0000, .4479)	(.0567, .5098)	(.0459, .5206)	(.025, .556)
10	.50	5	(.1901, .8099)	(.2366, .7634)	(.2366, .7634)	(.187, .813)
25	.04	1	(.0000, .1168)	(.0071, .1954)	(.0000, .2114)	(.001, .204)
25	.20	5	(.0432, .3568)	(.0886, .3913)	(.0841, .3958)	(.068, .407)
50	.02	1	(.0000, .0588)	(.0035, .1050)	(.0000, .1148)	(.001, .107)
50	.04	2	(.0000, .0943)	(.0110, .1346)	(.0034, .1422)	(.005, .137)
50	.10	5	(.0178, .1832)	(.0435, .2136)	(.0391, .2179)	(.003, .218)
50	.20	10	(.0891, .3109)	(.1124, .3304)	(.1105, .3323)	(.100, .338)
50	.50	25	(.3614, .6386)	(.3664, .6336)	(.3664, .6336)	(.355, .645)
100	.02	2	(.0000, .0474)	(.0055, .0700)	(.0011, .0744)	(.002, .070)
100	.04	4	(.0016, .0784)	(.0157, .0984)	(.0124, .1016)	(.011, .099)
100	.10	10	(.0412, .1588)	(.0552, .1744)	(.0535, .1761)	(.049, .176)
100	.20	20	(.1216, .2784)	(.1334, .2888)	(.1326, .2896)	(.127, .292)
100	.50	50	(.4020, .5980)	(.4038, .5962)	(.4038, .5962)	(.398, .602)
250	.02	5	(.0026, .0374)	(.0085, .0460)	(.0072, .0473)	(.007, .046)
250	.04	10	(.0157, .0643)	(.0218, .0721)	(.0209, .0730)	(.019, .072)
250	.10	25	(.0628, .1372)	(.0686, .1435)	(.0682, .1439)	(.066, .144)
250	.20	50	(.1504, .2496)	(.1551, .2540)	(.1549, .2542)	(.152, .255)
250	.50	125	(.4380, .5620)	(.4384, .5615)	(.4385, .5615)	(.436, .564)
500	.02	10	(.0077, .0323)	(.0108, .0364)	(.0104, .0369)	(.010, .036)
500	.04	20	(.0228, .0572)	(.0260, .0610)	(.0257, .0613)	(.025, .061)
500	.10	50	(.0737, .1263)	(.0766, .1294)	(.0765, .1296)	(.075, .130)
500	.20	100	(.1649, .2351)	(.1672, .2373)	(.1672, .2374)	(.161, .238)
500	.50	250	(.4562, .5438)	(.4563, .5437)	(.4563, .5437)	(.455, .545)
1000	.02	20	(.0113, .0287)	(.0129, .0307)	(.0128, .0309)	(.012, .031)
1000	.04	40	(.0279, .0521)	(.0295, .0540)	(.0294, .0541)	(.029, .054)
1000	.10	100	(.0814, .1186)	(.0829, .1201)	(.0828, .1202)	(.082, .120)
1000	.20	200	(.1752, .2248)	(.1763, .2259)	(.1764, .2259)	(.176, .226)
1000	.50	250	(.4690, .5310)	(.4690, .5309)	(.4691, .5309)	(.469, .531)

Note: When $y = n\hat{p} < 5$, the Asymptotic C.I.'s are not recommended.

The C.I.'s are given to illustrate their inaccuracies in these situations relative to the Clopper-Pearson C.I.

The Clopper-Pearson C.I.'s tend to produce intervals having coverage probabilities somewhat larger than the nominal value. This is due to the discreteness of the binomial distribution.

=> Clopper CI's tend to be conservative estimates.

- would prefer to be conservative

Comparison of Performance of C.I.'s

Given an interval estimator of a parameter θ: $100(1-\alpha)$ C.I. $= (\hat{\theta}_L, \hat{\theta}_U)$.

There a number of methods for assessing the performance of an $100(1-\alpha)$ C.I. as an estimator of θ:

1. Accuracy of C.I. is measured by Coverage Probability: $C(\theta, n) = P[\theta \epsilon (\hat{\theta}_L, \hat{\theta}_U)]$

   Compare $C(\theta, n)$ to $100(1-\alpha)$ to determine how close the true level of confidence is to the nominal (stated) level.

2. Precision of C.I. is measured by Expected Width of C.I.: $E[W(\theta, n)]$

   Because $(\hat{\theta}_L, \hat{\theta}_U)$ is a r.v., we need to compute its average width.

   That is, let $W(\theta, n) = \hat{\theta}_U - \hat{\theta}_L$ and then compute $E[W(\theta, n)]$.

   In comparing two C.I.'s for the same parameter having the same **coverage probability**, the C.I. having shortest expected width is the better C.I.

The article **Interval Estimation for a Binomial Proportion** in *Statistical Science*, Vol. 16, pp. 101-133, by L. Brown, T. Cai and A. DasGupta contains a discussion of the performance of various confidence intervals for p using the above measures and several others. I have included a few of the graphs from their article. The following recommendations are given in the article:

- For larger n, the Wilson and A-C are comparable.
- The A-C CI is recommended for $n \geq 40$ due to its ease of calculation.
- For $n < 40$, there are several alternatives:

  1. Jeffreys Confidence Interval: This is a Bayesian confidence interval involving a $B(n, p)$ data distribution with a beta prior distribution on p, $Beta(a_1, a_2)$ (See STAT 638 for details).

     Observe Y distributed $B(n, p)$ then a $100(1-\alpha)\%$ Bayesian confidence interval on p is

     $$[Beta(\alpha/2;\ Y + a_1,\ n - Y + a_2),\ Beta(1-\alpha/2;\ Y + a_1,\ n - Y + a_2)]$$

     where $Beta(\alpha;\ m_1,\ m_2)$ denotes the α quantile of a $Beta(m_1, m_2)$ distribution.

     The values of a_1 and a_2 depend on the researcher's prior knowledge of the value of p.

  2. Several other confidence intervals are discussed in the article by Agresti and Coull (1998), "Approximate is Better than Exact for Interval Estimation of Binomial Proportions", *The American Statistican*, **Vol. 52**.

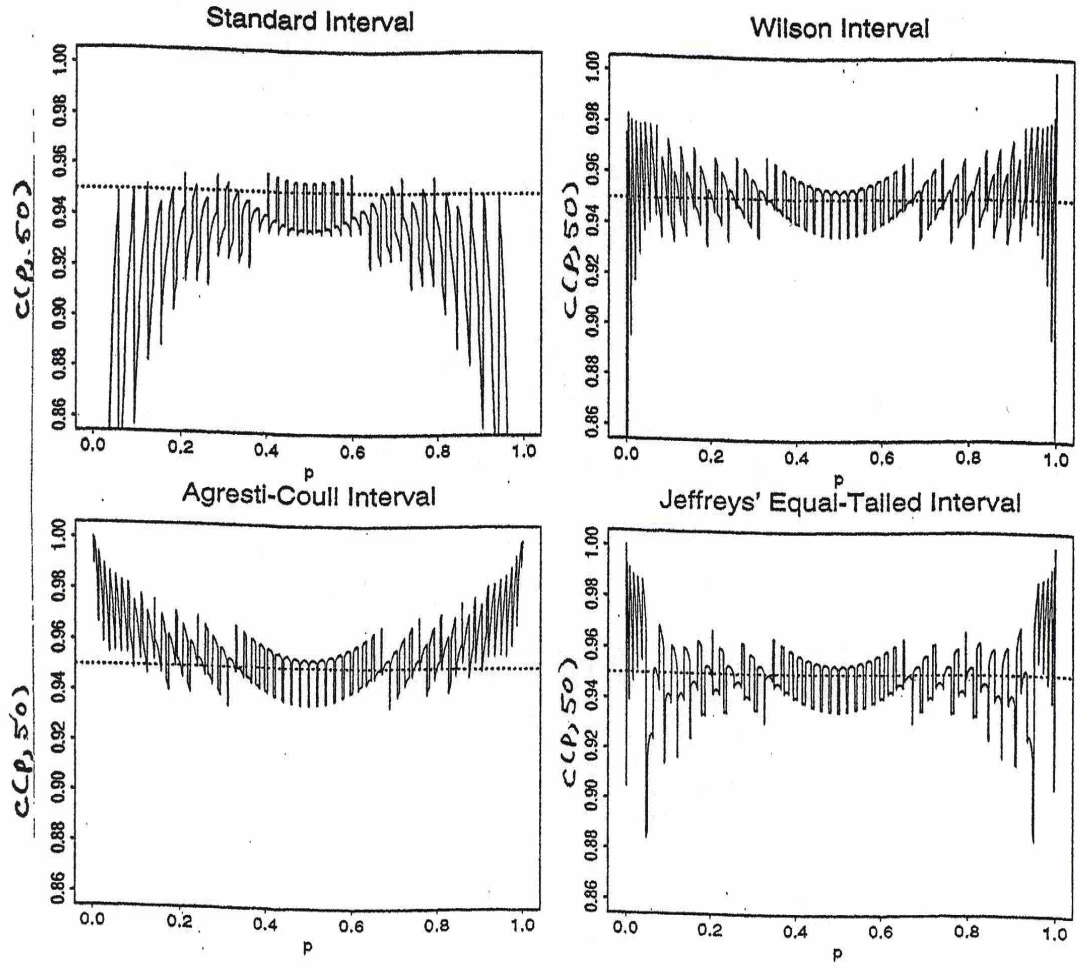


FIG. 5. *Coverage probability for* $n = 50$. (nominal = 0.95)

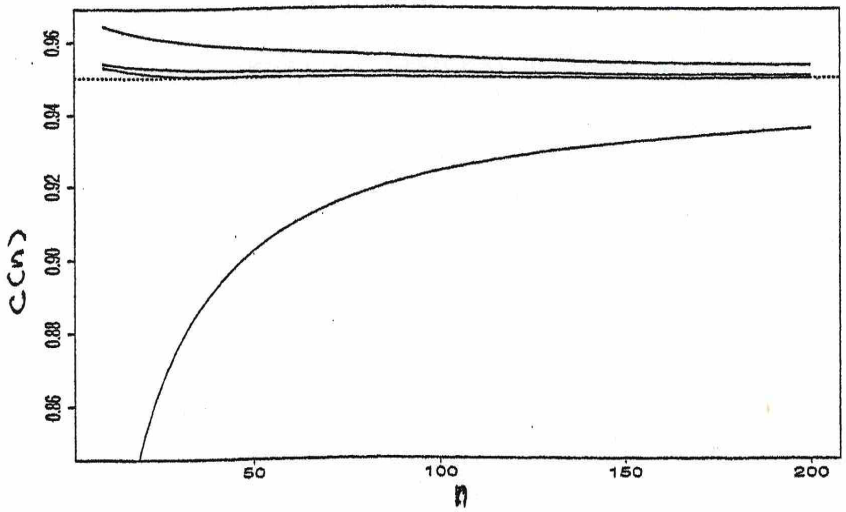


FIG. 6. *Comparison of the average coverage probabilities. From top to bottom: the Agresti–Coull interval* CI_{AC}, *the Wilson interval* CI_W, *the Jeffreys prior interval* CI_J *and the standard interval* CI_s. *The nominal confidence level is* 0.95. (Averaged over p)

3.3 Expected Length

Besides coverage, length is also very important in evaluation of a confidence interval. We compare both the expected length and the average expected length of the intervals. By definition,

$$\begin{aligned}&\text{Expected length}\\&\quad = E_{n,p}(\text{length}(CI))\\&\quad = \sum_{x=0}^{n}(U(x,n)-L(x,n))\binom{n}{x}p^x(1-p)^{n-x},\end{aligned}$$

where U and L are the upper and lower limits of the confidence interval CI, respectively. The average expected length is just the integral $\int_0^1 E_{n,p}(\text{length}(CI))\,dp$.

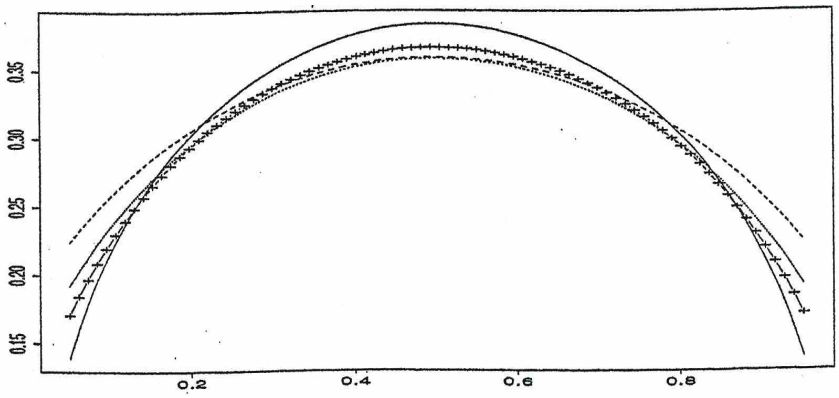


FIG. 8. *The expected lengths of the standard (solid), the Wilson (dotted), the Agresti–Coull (dashed) and the Jeffreys (+) intervals for $n = 25$ and $\alpha = 0.05$.*

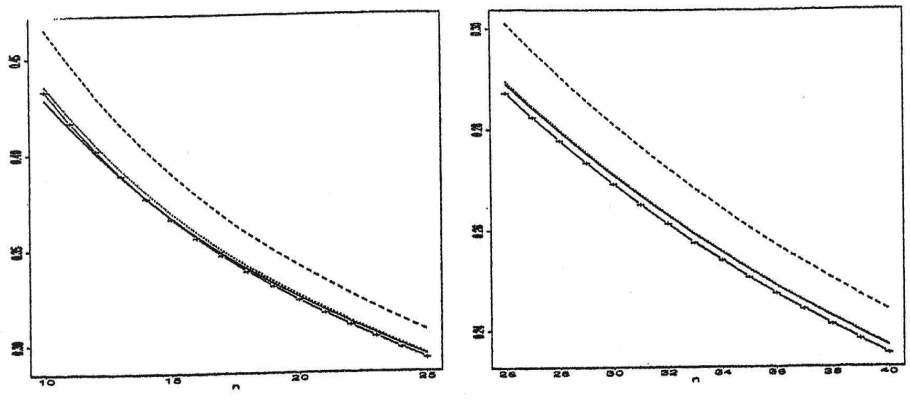


FIG. 9. *The average expected lengths of the standard (solid), the Wilson (dotted), the Agresti–Coull (dashed) and the Jeffreys (+) intervals for $n = 10$ to 25 and $n = 26$ to 40.*

Sample Size Determination

Statisticians are often asked the question "How much data do I need to collect?" One technique to answer this question is by way of C.I.'s.

Find the sample size n such that the estimator $\hat{\theta}$ of the parameter θ is within Δ units of the true value of θ with $100(1-\alpha)\%$ confidence.

We can then set up an equation and solve for the sample size n once the client provides the values of $100(1-\alpha)\%$ and Δ.

Consider the following two situations:

1. Sample Size for Estimating Population Mean μ

   Find n such that we are $100(1-\alpha)\%$ confident that $\bar{Y}$ is within Δ units of μ. The asymptotic sampling distribution for $\bar{Y}$ yields

   $$P\left[|\bar{Y}-\mu| \leq \Delta\right] = P\left[\left|\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right| \leq \Delta\frac{\sqrt{n}}{\sigma}\right] \approx 1-\alpha.$$

   Thus, set $\Delta\frac{\sqrt{n}}{\sigma} = Z_{\frac{\alpha}{2}}$ and solve for n yielding

   $$n = \frac{Z^2_{\frac{\alpha}{2}}\sigma^2}{\Delta^2}.$$

   Note that σ is often unknown so the client will have to have at least a rough guess of this value.

   This can be obtained from previous studies, literature, pilot study, or by using the crude estimator

   $$\hat{\sigma} \approx \frac{Range}{4}.$$

2. Sample Size for Estimating Population Proportion p

   Find n such that we are $100(1-\alpha)\%$ confident that $\hat{p}$ is within Δ units of p. The asymptotic sampling distribution for $\hat{p}$ yields

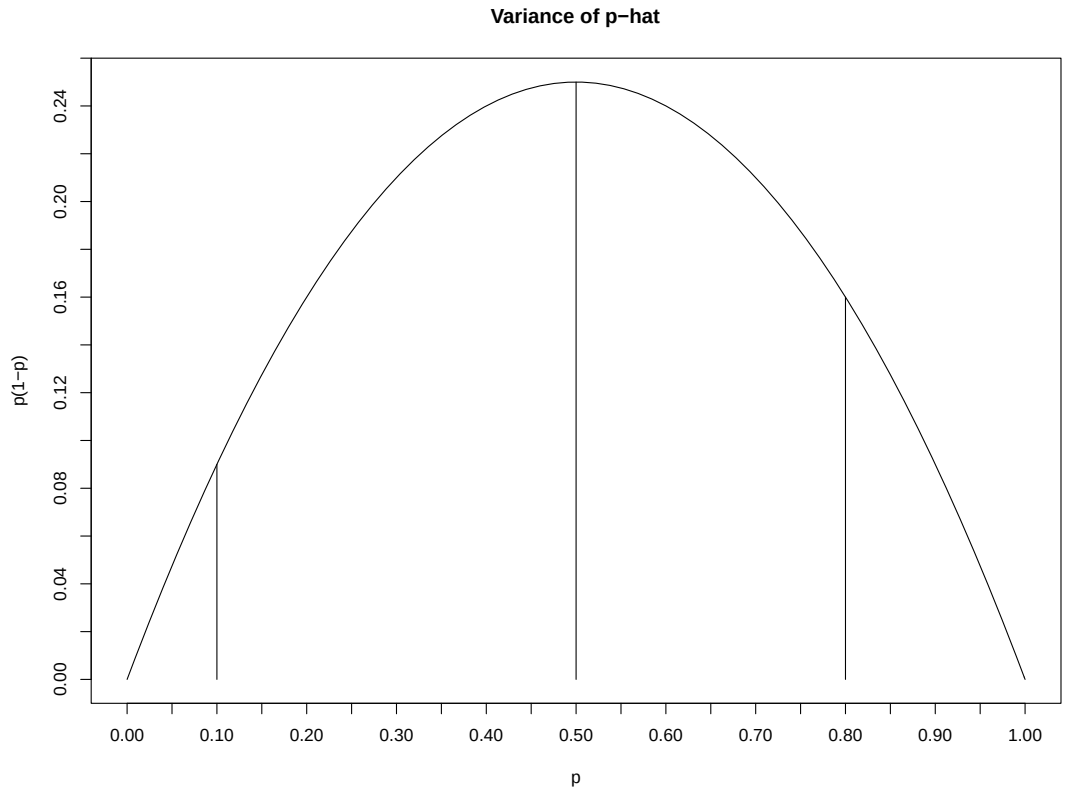
   $$P\left[|\hat{p}-p| \leq \Delta\right] = P\left[\left|\frac{\hat{p}-p}{\sqrt{p(1-p)}/\sqrt{n}}\right| \leq \Delta\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right] \approx 1-\alpha.$$

   Thus, set $\Delta\frac{\sqrt{n}}{\sqrt{p(1-p}} = Z_{\frac{\alpha}{2}}$ and solve for n yielding

   $$n = \frac{Z^2_{\frac{\alpha}{2}}p(1-p)}{\Delta^2}.$$

   Note that p, an unknown value, is in the formula so the client will have to have at least a rough guess of the value of p.

When the client can provide a bound on p by knowing that $p \leq p_L$ or $p \geq p_U$.



Then replace p with either p_L or p_U.

If you cannot bound p away from 0.5 then just replace p with 0.5

Replacing p with 0.5 generally will produce very large values for n.

Example Find n such that 99% confident that $\hat{p}$ is within 0.01 units of p.

a. Suppose we know that $0 \leq p \leq 0.10$

   Then, $n = \frac{(2.576)^2(.1)(.9)}{(.01)^2} = 5973$.

b. Suppose we know that $.8 \leq p \leq 1.0$

   Then, $n = \frac{(2.576)^2(.8)(.2)}{(.01)^2} = 10618$.

c. If we were not able to bound p away from 0.5 then

   $n = \frac{(2.576)^2(.5)(.5)}{(.01)^2} = 16,590$.

conservative way to choose n is to choose p close to 0.5

3. More accurate calculations can be obtained using the formula for the Wilson CI:

$$n = \frac{2z^2_{\frac{\alpha}{2}}\hat{p}\hat{q} - 4z^2_{\frac{\alpha}{2}}\Delta^2 + \sqrt{4z^4_{\frac{\alpha}{2}}\hat{p}\hat{q}(\hat{p}\hat{q} - 4\Delta^2) + 4\Delta^2 z^4_{\frac{\alpha}{2}}}}{4\Delta^2}$$

For example, find n such that 99% confident that $\hat{p}$ is within 0.01 units of p.

Suppose we know that $0 \leq p \leq 0.10$. Then,

$$n = \frac{2(2.576)^2(.1)(.9) - 4(2.576)^2(.01)^2 + \sqrt{4(2.576)^4(.1)(.9)((.1)(.9) - 4(.01)^2) + 4(.01)^2(2.576)^4}}{4(.01)^2}$$

$$n = 5977.3 \Rightarrow$$

$$n = 5978$$

Distribution-Free C.I. for Population Quantile Q(u)

Suppose $Y_1, \ldots, Y_n$ are iid with strictly increasing continuous cdf $F(\cdot)$ and quantile function $Q(\cdot) = F^{-1}(\cdot)$.

A $100(1-\alpha)$ C.I. for $Q(u)$ is

$$\left(Y_{(r)},\ Y_{(s)}\right)$$

where $Y_{(1)} < \cdots < Y_{(n)}$, and r is the largest integer and s is the smallest integer such that

$$1 \le r < s \le n \ \text{ and } \ P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] \ge 1-\alpha$$

The values of r and s are selected such that

$$1-\alpha = \sum_{j=r}^{s-1} \binom{n}{j} u^j (1-u)^{n-j} = P\left[r \le B \le s-1\right]$$

where B is $Bin(n,u)$.

This result is obtained from

$$\begin{aligned} P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] &= P\left[F(Y_{(r)}) \le F(Q(u)) \le F(Y_{(s)})\right] \\ &= P\left[U_{(r)} \le u \le U_{(s)}\right] \end{aligned}$$

$F(Y) \sim Unif(0,1)$

where $U_{(1)} < \cdots < U_{(n)}$ are order statistics from iid Uniform on (0,1) distribution.

Next, we observe that

$$P\left[U_{(r)} \le u \le U_{(s)}\right] = P\left[\text{at least } \ n-s+1 \ U_{i's} \ge u \ \text{ and at least } \ r \ U_{i's} \le u\right] \ \Rightarrow$$

$$P\left[U_{(r)} \le u \le U_{(s)}\right] = P\left[\text{at most } \ s-1 \ U_{i's} \le u \ \text{ and at least } \ r \ U_{i's} \le u\right] \ \Rightarrow$$

$$\begin{aligned} P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] &= P\left[U_{(r)} \le u \le U_{(s)}\right] \\ &= \sum_{j=r}^{s-1} \binom{n}{j} (P(U_i \le u)^j (1-P(U_i \le u))^{n-j} \\ &= \sum_{j=r}^{s-1} \binom{n}{j} u^j (1-u)^{n-j} \\ &= P\left[r \le B \le s-1\right] = pbinom(s-1,n,u) - pbinom(r-1,n,u) \end{aligned}$$

where B is $Bin(n,u)$.

Finally, select the values of r and s such that

$$P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] = P\left[r \le B \le s-1\right] = 1-\alpha$$

In most cases, we cannot obtain exactly, $1-\alpha$, in the above expressions so the coverage probability is always taken to be as close to $1-\alpha$ as possible but never less that $1-\alpha$.

STOP Monday 10/25/2021

EXAMPLE Suppose we wanted to find a 95% C.I. for the upper quartile,Q(.75), based on n=50 iid observations on the cdf F. The following Rcode will determine the values of r, s, and the true coverage:

```
n=50
L=.95
P=.75
s=ceiling(n*P)-1
r=floor(n*P)+1
cov=0
while(s<n-1 && r>1 && cov<L)
{s=s+1
cov=pbinom(s-1,n,P)-pbinom(r-1,n,P)
if(cov>=L) break;
r=r-1
cov=pbinom(s-1,n,P)-pbinom(r-1,n,P)
}
r
s
cov
> r
[1] 32
> s
[1] 44
> cov
[1] 0.951876
```

The 95% C.I. on Q(.75) would be $(Y_{(32)},\ Y_{(44)})$ with coverage probability of 95.2%.

Special Case: C.I. for Median

A C.I. for the population median is obtained by just setting

$u = .5$ and requiring that the C.I. to be symmetric, that is, $s = n - r + 1$.

This then yields a $100(1-\alpha)\%$ C.I. for the median $Q(.5)$:

$$\left(Y_{(r)},\ Y_{(n-r+1)}\right)$$

where r is the largest integer such that

$$1 - \alpha \leq P\left[r \leq B \leq n - r\right]$$

and B is $Bin(n, .5)$.

The following example will illustrate how to use R code to select the value of r and determine the true coverage probability.

Example Suppose we want a 95% C.I. on $Q(.5)$ based on $n = 50$.

Find the largest r such that $.95 \leq P\left[r \leq B \leq 50 - r\right]$, where B is $Bin(50, .5)$.

```
n =  50
cov =  .95
r =  0
imin =  0
```

```
i =  0
ans =  0
anst =  0
m =  1:n
ans =  pbinom(n-m,n,.5)-pbinom(m-1,n,.5)
while(i<n)
{
i =  i+1
if(ans[i]<cov)  anst[i] =  2
if(ans[i]>=cov) anst[i] =  ans[i]
}
ansmin =  min(anst)
imin =  which(anst==ansmin)
r =  imin
coverage =  ans[r]
```

From the above R-code we have $r = 18$ with coverage probability 0.96716.
Therefore, the 95% C.I. for the median is

$\left(Y_{(r)}, Y_{(n-r+1)}\right) = \left(Y_{(18)}, Y_{(33)}\right)$.

The true coverage probability is computed using the $B(50, .5)$ distribution.
Coverage Probability = (see previous page) $P\left[r \leq B \leq n-r\right] = P\left[18 \leq B \leq 32\right] = .96716$
So the true coverage probability is a little higher than 0.95.

The following table from *CRC Handbook of Tables for Probability and Statistics* provides r for a variety of values for n and levels of confidence.

START Wednesday 10/27/21

CRC Handbook of Tables for Probability and Statistics

VII.3 CONFIDENCE INTERVALS FOR MEDIANS

If the observations $x_1, x_2, \ldots, x_n$ are arranged in ascending order $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, a $100(1-\alpha)\%$ confidence interval on the median of the population can be found. This table gives values of k and α such that one can be $100(1-\alpha)\%$ confident that the population median is between $x_{(k)}$ and $x_{(n-k+1)}$.

A—compute using whichever one is in this table.

CONFIDENCE INTERVALS FOR THE MEDIAN

n	Largest k	Actual $\alpha \leq 0.05$	Largest k	Actual $\alpha \leq 0.01$	N	Largest k	Actual $\alpha \leq 0.05$	Largest k	Actual $\alpha \leq 0.01$
6	1	0.031			36	12	0.029	10	0.004
7	1	0.016			37	13	0.047	11	0.008
8	1	0.008	1	0.008	38	13	0.034	11	0.005
9	2	0.039	1	0.004	39	13	0.024	12	0.009
10	2	0.021	1	0.002	40	14	0.038	12	0.006
11	2	0.012	1	0.001	41	14	0.028	12	0.004
12	3	0.039	2	0.006	42	15	0.044	13	0.008
13	3	0.022	2	0.003	43	15	0.032	13	0.005
14	3	0.013	2	0.002	44	16	0.049	14	0.010
15	4	0.035	3	0.007	45	16	0.036	14	0.007
16	4	0.021	3	0.004	46	16	0.026	14	0.005
17	5	0.049	3	0.002	47	17	0.040	15	0.008
18	5	0.031	4	0.008	48	17	0.029	15	0.006
19	5	0.019	4	0.004	49	18	0.044	16	0.009
20	6	0.041	4	0.003	50	18	0.033	16	0.007
21	6	0.027	5	0.007	51	19	0.049	16	0.005
22	6	0.017	5	0.004	52	19	0.036	17	0.008
23	7	0.035	5	0.003	53	19	0.027	17	0.005
24	7	0.023	6	0.007	54	20	0.040	18	0.009
25	8	0.043	6	0.004	55	20	0.030	18	0.006
26	8	0.029	7	0.009	56	21	0.044	18	0.005
27	8	0.019	7	0.006	57	21	0.033	19	0.008
28	9	0.036	7	0.004	58	22	0.048	19	0.005
29	9	0.024	8	0.008	59	22	0.036	20	0.009
30	10	0.043	8	0.005	60	22	0.027	20	0.006
31	10	0.029	8	0.003	61	23	0.040	21	0.010
32	10	0.020	9	0.007	62	23	0.030	21	0.007
33	11	0.035	9	0.005	63	24	0.043	21	0.005
34	11	0.024	10	0.009	64	24	0.033	22	0.008
35	12	0.041	10	0.006	65	25	0.046	22	0.006

The following table provides C.I.s for a variety of situations and parameters.

Parameter	Population Conditions	Endpoints of Confidence Intervals
Q(p)	$X_1, \cdots, X_n$ iid cont. cdf	$(X_{(r)}, X_{(s)})$,
		where r,s selected using Binomial(n,p) tables
μ	$X_1, \cdots, X_n$ iid $N(\mu, \sigma^2)$ σ unknown	$\bar{X} \pm t_{(\frac{\alpha}{2}, df)} \frac{S}{\sqrt{n}}$ where t has d.f.= n-1
$\mu_1 - \mu_2$	$X_1, \cdots, X_{n1}$ iid $N(\mu_1, \sigma_1^2)$ $Y_1, \cdots, Y_{n2}$ iid $N(\mu_2, \sigma_2^2)$ X's,Y's ind., $\sigma_1 = \sigma_2$	$(\bar{X} - \bar{Y}) \pm t_{(\frac{\alpha}{2}, df)} S_p \sqrt{\frac{1}{n1} + \frac{1}{n2}}$ where t has d.f.= $n1 + n2 - 2$
$\mu_1 - \mu_2$	$X_1, \cdots, X_{n1}$ iid $N(\mu_1, \sigma_1^2)$ $Y_1, \cdots, Y_{n2}$ iid $N(\mu_2, \sigma_2^2)$ X's,Y's ind., $\sigma_1 \neq \sigma_2$	$(\bar{X} - \bar{Y}) \pm t_{(\frac{\alpha}{2}, df)} \sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}$ where t has d.f.= $\frac{(C+1)^2 (n1-1)(n2-1)}{C^2(n2-1)+(n1-1)}$, and $C = \frac{S_1^2/n_1}{S_2^2/n_2}$
$\mu_1 - \mu_2$	$(X_1, Y_1), \cdots, (X_n, Y_n)$ iid with $D_i = X_i - Y_i$ $N(\mu_D, \sigma_D^2)$	$\bar{D} \pm t_{(\frac{\alpha}{2}, df)} S_D / \sqrt{n}$ where t has d.f. = n-1
p	Y is Bin(n,p)	$\tilde{p} \pm \frac{Z_{(\frac{\alpha}{2})} \sqrt{n} \sqrt{\hat{p}(1-\hat{p}) + \frac{1}{4n} Z^2_{(\frac{\alpha}{2})}}}{n + Z^2_{\frac{\alpha}{2}}}$
	$min(n\hat{p}, n(1-\hat{p}) \geq 5$	$Z_{(\frac{\alpha}{2})}$ upper N(0,1) percentile
	and $n \leq 40$	$\tilde{Y} = Y + Z^2_{\frac{\alpha}{2}}/2, \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}}, \quad \tilde{p} = \frac{\tilde{Y}}{\tilde{n}}$
p	Y is Bin(n,p)	$\tilde{p} \pm Z_{(\frac{\alpha}{2})} \sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$
	$min(n\hat{p}, n(1-\hat{p}) \geq 5$	$Z_{(\frac{\alpha}{2})}$ upper N(0,1) percentile
	and $n > 40$	$\tilde{Y} = Y + Z^2_{\frac{\alpha}{2}}/2, \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}}, \quad \tilde{p} = \frac{\tilde{Y}}{\tilde{n}}$
p	Y is Bin(n,p)	Clopper-Pearson CI: $(P_L, \ P_U)$, where $P_L = \frac{1}{1+\left(\frac{n-y+1}{y}\right) F_{2(n-y+1), 2y, \frac{\alpha}{2}}}$; $\quad P_U = \frac{\left(\frac{y+1}{n-y}\right) F_{2(y+1), 2(n-y), \frac{\alpha}{2}}}{1+\left(\frac{y+1}{n-y}\right) F_{2(y+1), 2(n-y), \frac{\alpha}{2}}}$ Upper $F-$ quantiles: $F_{df_1, df_2, \frac{\alpha}{2}} = qf(1 - \frac{\alpha}{2}, df_1, df_2)$

Parameter	Population Conditions	Endpoints of Confidence Intervals
$p_1 - p_2$	Count Data $min(n\hat{p}_i, n(1-\hat{p}_i) \geq 5$	$\hat{p}_1 - \hat{p}_2 \pm Z_{(\frac{\alpha}{2})}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ $Z_{(\frac{\alpha}{2})}$ upper N(0,1) percentile
σ	Normal Data	$\left(\frac{\sqrt{n-1}\ \ S}{\sqrt{\chi^2_{(\frac{\alpha}{2},n-1)}}},\ \frac{\sqrt{n-1}\ \ S}{\sqrt{\chi^2_{(1-\frac{\alpha}{2},n-1)}}}\right)$, $\chi^2_{(\frac{\alpha}{2},n-1)}$ and $\chi^2_{(1-\frac{\alpha}{2},n-1)}$ upper percentiles- Chi-square tables
$\frac{\sigma_1}{\sigma_2}$	Normal Data	$\left(\frac{S_1}{S_2}\sqrt{\frac{1}{F_{(\frac{\alpha}{2},n_1-1,n_2-1)}}},\ \frac{S_1}{S_2}\sqrt{F_{(\frac{\alpha}{2},n_2-1,n_1-1}}\right)$, $F_{(\frac{\alpha}{2},n_1-1,n_2-1)}$ and $F_{(\frac{\alpha}{2},n_2-1,n_1-1)}$ upper percentiles- F-tables
β	Exponential Data	$\left(\frac{2n\bar{Y}}{\chi^2_{(1-\frac{\alpha}{2},2n)}},\ \frac{2n\bar{Y}}{\chi^2_{(\frac{\alpha}{2},2n)}}\right)$, $\chi^2_{(\frac{\alpha}{2},2n)}$ and $\chi^2_{(1-\frac{\alpha}{2},2n)}$ upper percentiles- Chi-square tables
$\frac{\beta_1}{\beta_2}$	Exponential Data	$\left(\frac{\bar{Y}_1}{\bar{Y}_2}\frac{1}{F_{(\frac{\alpha}{2},2n_1,2n_2)}},\ \frac{\bar{Y}_1}{\bar{Y}_2}F_{(\frac{\alpha}{2},2n_2,2n_1)}\right)$, $F_{(\frac{\alpha}{2},2n_1,2n_2)}$ and $F_{(\frac{\alpha}{2},2n_2,2n_1)}$ upper percentiles- F-tables
θ	parameter in pdf $f(y,\theta)$	$\hat{\theta} \pm Z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\theta})$ where $\hat{\theta}$ is MLE

Prediction Intervals

In some studies or experiments, the researcher will want to predict the next outcome of the process or experiment.

For example, suppose your company is interested in purchasing a new airplane engine from supplier X. Your company is not interested in the average time to failure of all engines of that model but rather what is the predicted time to failure of the randomly selected engine that your company will be receiving.

Other examples, predicting the amount of rainfall in the next month, predicting demand for a product in the next quarter, and predicting the enrollment in a large undergraduate class for the next semester. Many of these types of predictions will use time series modeling and explanatory variables in regression models which you will study in STAT 626 and STAT 608.

An interval estimator of this predicted value is called a $100(1-\alpha)\%$ Prediction Interval P.I.

Case 1: Prediction Interval for a $N(\mu, \sigma^2)$ Population Distribution

Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$.

We can write the $Y_i's$ in the following model:

$$Y_i = \mu + \sigma Z_i \quad \text{where} \quad Z_i's \text{ are } iid \quad N(0,1)$$

The next unit selected from the population will have measured value:

$$Y_{n+1} = \mu + \sigma Z_{n+1}$$

Our estimator of Y_{n+1} is obtained by

1. Replacing μ and σ with their MLE's and
2. Replacing Z_{n+1} with $E[Z_{n+1}] = 0$, yielding

$$\hat{Y}_{n+1} = \hat{\mu} + \hat{\sigma}\hat{Z}_{n+1} = \hat{\mu} = \bar{Y}$$

To obtain, a P.I. for Y_{n+1} we need a pivot which involves the quantities

$$Y_{n+1}, \quad \hat{Y}_{n+1}, \quad \text{and} \quad \hat{\sigma}:$$

A possible candidate is

$$\text{Pivot } = g(Y, Y_{n+1}, \sigma) = \frac{Y_{n+1} - \bar{Y}}{S\sqrt{\frac{n+1}{n}}}$$

This will be a good candidate for a pivot only if we can determine its distribution.

From previous results, we know that the pivot has a t-Distribution with $df = n - 1$. Why?

1. $Y_{n+1} = \mu + \sigma Z_{n+1} \Rightarrow Y_{n+1}$ is distributed $N(\mu, \sigma^2)$
2. $Y_{n+1}, \bar{Y}$ are independent $\Rightarrow Y_{n+1} - \bar{Y}$ is distributed $N\left(\mu - \mu, \sigma^2 + \frac{\sigma^2}{n}\right) = N\left(0, \sigma^2\left(\frac{n+1}{n}\right)\right)$
3. $Y_{n+1} - \bar{Y}$ is distributed independent of S which has $\frac{(n-1)S^2}{\sigma^2}$ distributed as Chi-square with df=n-1
4. $t = \frac{N(0,1)}{\sqrt{\text{Chi-square/df}}} \Rightarrow \frac{(Y_{n+1}-\bar{Y})/\sigma\sqrt{\left(\frac{n+1}{n}\right)}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{(Y_{n+1}-\bar{Y})}{S\sqrt{\left(\frac{n+1}{n}\right)}}$

Thus, we have the following results:

$$\begin{aligned} 1 - \alpha &= P\left[-t_{\frac{\alpha}{2}} \leq \frac{Y_{n+1} - \bar{Y}}{S\sqrt{\frac{n+1}{n}}} \leq t_{\frac{\alpha}{2}}\right] \\ &= P\left[\bar{Y} - t_{\frac{\alpha}{2}} S\sqrt{\frac{n+1}{n}} \leq Y_{n+1} \leq \bar{Y} + t_{\frac{\alpha}{2}} S\sqrt{\frac{n+1}{n}}\right] \end{aligned}$$

The $100(1-\alpha)\%$ Prediction Interval for Y_{n+1} is

$$\bar{Y} \pm t_{\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n}}$$

Recall that the $100(1-\alpha)\%$ C.I. for μ in a $N(\mu, \sigma^2)$ population is

$$\bar{Y} \pm t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n}}$$

Thus, the added width in the P.I. reflects the additional uncertainty in attempting to predict a realization from a $N(\mu, \sigma^2)$ population in comparison to estimating the population mean.

Case 2: Prediction Interval for a $Exponential(\beta)$ Distribution

Suppose the data $W_1, \ldots, W_n$ is iid $Exponential(\beta)$ r.v.'s with β unknown

A prediction is needed for the next realization from this population.

For example, the time to the next hurricane in the Gulf of Mexico or the life length of the transmission placed in the Porsche 911 Turbo that you will purchase upon graduation.

The MLE point estimator of β is $\hat{\beta} = \bar{W}$.

A pivot should involve

- the point estimator, $\hat{\beta} = \bar{W}$
- the next realization W_{n+1}, and
- the unknown parameter β.

  Recall that if $Y_1, \ldots, Y_k$ are iid $Exp(\beta)$ then

- $k\bar{Y} = \sum_{i=1}^{k} Y_i$ has a $Gamma(k, \beta)$ distribution.
- If X has a $Gamma(k, \beta)$ distribution, then $\frac{2X}{\beta}$ has a chi-square distribution with $df = 2k$.

  Therefore, we know have the following results:

1. $\frac{2W_{n+1}}{\beta}$ has a chi-square distribution with $df = 2$,
2. $\frac{2n\bar{W}}{\beta}$ has a chi-square distribution with $df = 2n$, and
3. $\frac{2W_{n+1}}{\beta}$ and $\frac{2n\bar{W}}{\beta}$ are independent.

   The pivot for this model will be:

$$\text{Pivot} = \frac{W_{n+1}}{\bar{W}} = \frac{\left(\frac{2W_{n+1}}{\beta}\right)/2}{\left(\frac{2n\bar{W}}{\beta}\right)/2n}, \quad \text{which is distributed F-distribution with } \ df = 2, 2n$$

The $100(1-\alpha)\%$ P.I. for W_{n+1} is

$$\left(\bar{W}F_{1-\frac{\alpha}{2}}, \bar{W}F_{\frac{\alpha}{2}}\right)$$

which is obtained from

$$P\left[\bar{W}F_{1-\frac{\alpha}{2}} \leq W_{n+1} \leq \bar{W}F_{\frac{\alpha}{2}}\right] = P\left[F_{1-\frac{\alpha}{2}} \leq \frac{W_{n+1}}{\bar{W}} \leq F_{\frac{\alpha}{2}}\right] = 1-\alpha$$

Tolerance Intervals

Confidence intervals are interval estimators of **population parameters** and reflect our uncertainty in estimating these parameters based on a random sample of observations from the population.

Prediction intervals are forecasts or predictions of the value of a **random variable**. The prediction interval reflects our uncertainty in predicting the measured value on a randomly selected unit from the population.

There are many other types of intervals used in production processes and laboratories:

1. **Engineering Tolerance Interval**

   A set of specification limits (S_L, S_U) placed on a product which define an acceptable range of values for the product.

   For example,

   a. A 1.2 Kg box of cereal has $(S_L, S_U) = 1.2 \pm .1$ Kg

   b. A 3 cm diameter piston ring has $(S_L, S_U) = 3 \pm .01$ cm

   c. A 100 ohm resister has $(S_L, S_U) = 100 \pm 5$ ohm

   If the product measurement falls outside of the range (S_L, S_U) then the product is not acceptable for its intended use. The only statistical question is what proportion, p, of the process's output is Acceptable.

   Case 1: Population pdf f is known

$$p = P[Acceptable] = P[Y \epsilon (S_L, S_U)] = \int_{S_L}^{S_U} f(y)dy = p$$

   Case 2: Population pdf f is unknown

   Estimate $p = P[Acceptable]$ using a $100(1-\alpha)$ C.I. based on inspecting n units selected from the process output.

   Let $\hat{p}$ be the proportion of the n units which are acceptable, and the just use $\hat{p}$ to construct a 95% C.I. on p. For example, use the Agresti-Coull C.I. for a proportion.

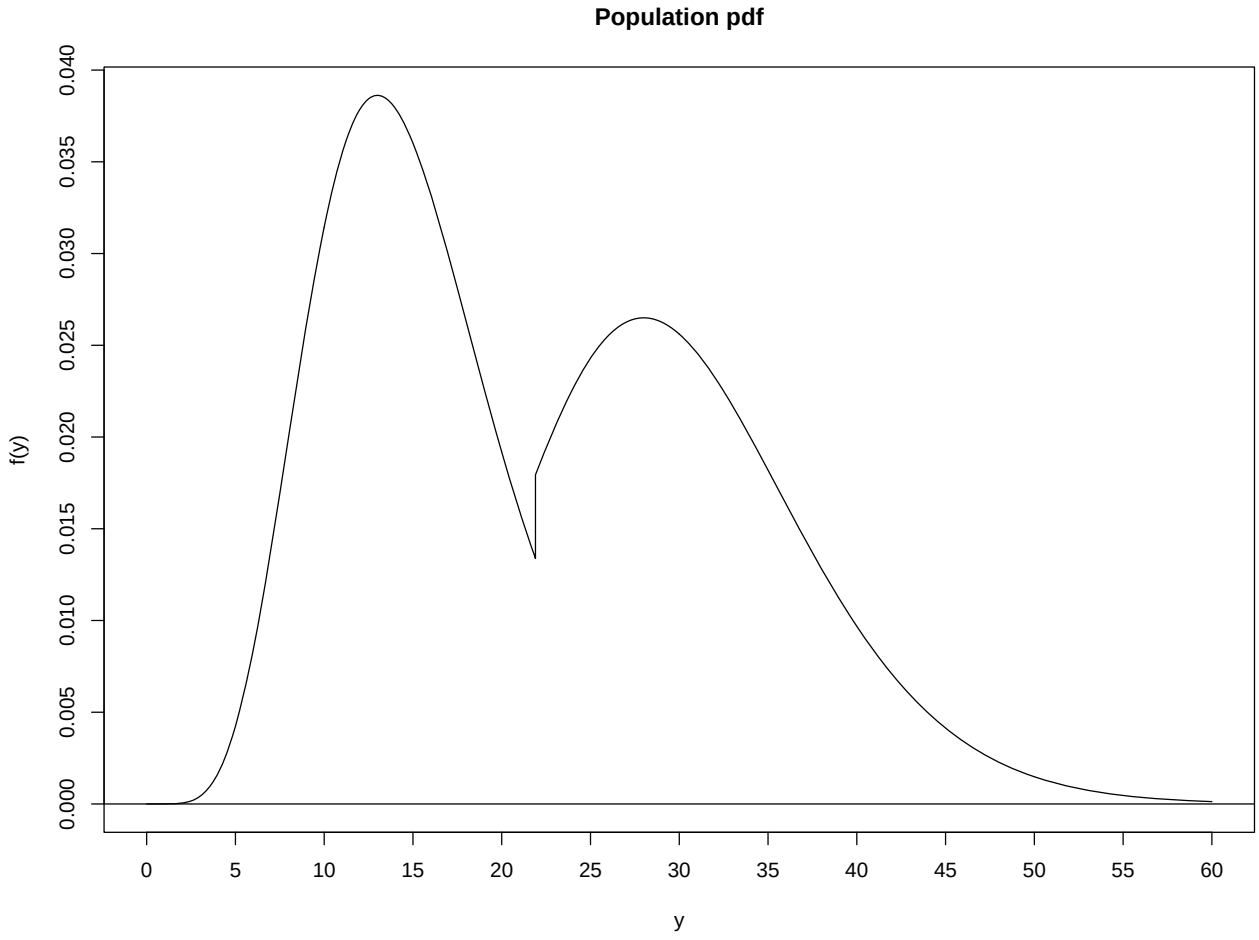
2. **Natural Tolerance of size** $100(1-\alpha)$

When the cdf F for the measurements from a process is completely specified, no unknown parameters, then an interval of values, (T_L, T_U), can we constructed such that

$P[Y\epsilon(T_L, T_U)] = 1 - \alpha$

In fact, one such interval would $T_L = Q(u_1)$ and $T_U = Q(u_2)$, where

$Q(u)$ is the quantile function associated with F and $u_1 + 1 - u_2 = \alpha$.

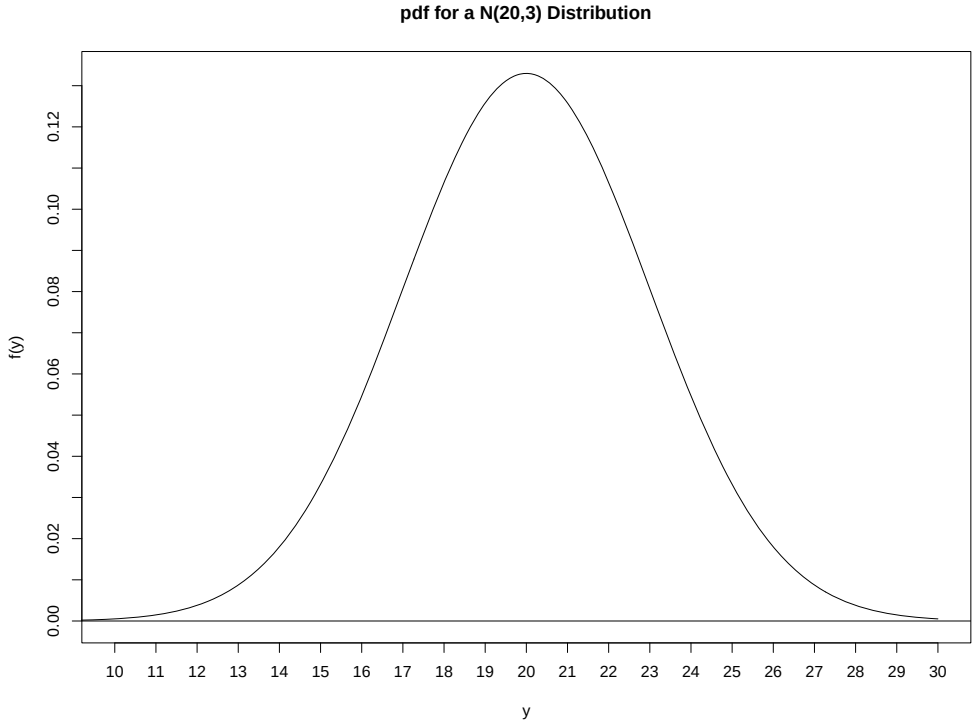


In particular, for a process in which the product characteristic in distributed $N(\mu, \sigma^2)$, with μ and σ known, we have

$Q(u) = \mu + \sigma Z_u$, where Z_u is the $100u$ percentile from a $N(0,1)$ distribution. Thus, the $100(1-\alpha)\%$ natural tolerance would be

$$(T_L, T_U) = \mu \pm \sigma Z_{\frac{\alpha}{2}}, \text{ where we take } u_1 = 1 - u_2 = \frac{\alpha}{2}.$$

For example, with $\mu = 20$ and $\sigma = 3$ we would have a 99% Natural Tolerance Interval would be $20 \pm (3)(Z_{.005}) = 20 \pm (3)(2.58) = (12.26, 27.74)$



3. **Statistical Tolerance Interval** (what you use in practice)

Statistical tolerance intervals are natural tolerance intervals when the population distribution is unknown or contains unknown parameters.

That is, a $100(P,\gamma)\%$ Statistical Tolerance Intervals, T.I., establish limits, $(L_{P,\gamma}, U_{P,\gamma})$ that include $100P\%$ of the responses in a population or from a process with $100\gamma\%$ confidence.

These intervals are used in evaluating production, quality, and service characteristics in many manufacturing and service industries. The T.I. reflects the actual variability of the product or service. T.I.'s are not intervals about a population or process parameter, they are intervals that include a specified portion of the observations from a population or process.

For example, what is the typical systolic blood pressure of health adult females of age 25-50? What would be more useful: a value of mean pressure, μ, or a C.I. on μ or a range of values such that we are 99% confident that 90% of all health adult females of age 25-50 fall into this range?

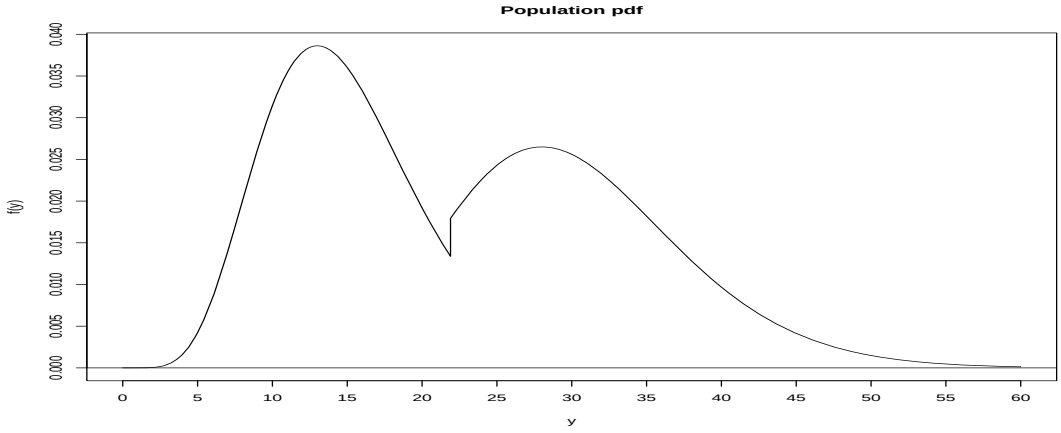
We will consider two examples of parametric T.I.'s and a more general distribution-free T.I.

General expression for a $100(P,\gamma)\%$ T.I. :

A $100(P,\gamma)\%$ T.I $(L_{P,\gamma}, U_{P,\gamma})$ for a population having cdf $F(\cdot)$ and pdf $f(\cdot)$ satifies

$$P\left[A(P,\gamma) \geq P\right] = \gamma \quad \text{where} \quad A(P,\gamma) = \int_{L_{P,\gamma}}^{U_{P,\gamma}} f(y)dy = F(U_{P,\gamma}) - F(L_{P,\gamma}).$$

That is, there is a $100\gamma\%$ probability that at least $100P\%$ of the population falls in the interval $(L_{P,\gamma}, U_{P,\gamma})$.



If f is known completely, no unknown parameters, then just take

$$L_{P,\gamma} = Q(P_1), \quad U_{P,\gamma} = Q(P_2) \quad \text{where} \quad P_2 - P_1 = P$$

Tolerance Intervals for $N(\mu, \sigma^2)$

Case 1: μ and σ Known

If μ and σ are known, then a $\gamma = 1.0$, 100% certainty interval, (L_P, U_P) exists. Namely,

$$A_p = \int_{L_P}^{U_P} \phi(y)dy = \Phi(U_{P,\gamma}) - \Phi(L_{P,\gamma}) = P$$

.

In fact, just let

$$L_P = Q_Y(P_1) = \mu + \sigma Z_{1-P_1}, \quad U_P = Q_Y(P_2) = \mu + \sigma Z_{1-P_2}$$

where $P_2 - P_1 = P$ and

Z_{1-P_1} and Z_{1-P_2} are the upper percentiles from a $N(0,1)$.

If we put equal weight in both tails, that is, $P_1 = 1 - P_2$, then we obtain

$$P_1 = .5(1-P), \quad P_2 = .5(1+P)$$

Case 2: μ and σ Unknown

If μ and σ are unknown, then we must use data to estimate them.

Let $Y_1, \ldots, Y_n$ be $iidN(\mu, \sigma^2)$ r.v.'s.

Then use the following estimators of μ and σ:

$$\hat{\mu} = \bar{Y}, \quad \hat{\sigma} = S$$

Need to find constant $K_{P,\gamma}$ to reflect the uncertainty in using $\hat{\mu}$ and $\hat{\sigma}$ in the intervals:

$$L_{P,\gamma} = \hat{\mu} - K_{P,\gamma}S \quad U_{P,\gamma} = \hat{\mu} + K_{P,\gamma}S$$

where

$$P\left[A(P,\gamma) \geq P\right] = \gamma \text{ and } A(P,\gamma) = \int_{L_{P,\gamma}}^{U_{P,\gamma}} f(y)dy, \text{ with } f(y) \text{ a } N(\mu,\sigma^2) \text{ pdf.}$$

Numerical approximations are used to determine $K_{P,\gamma}$ and a table of values is given on the following page. These tables are from the book, *Statistical Design and Analysis of Experiments*, by Mason, Gunst, and Hess. The values in these tables are slightly different from the values given in the textbook.

One-sided Tolerance Intervals

1-sided tolerance intervals are obtained by just placing all the probability in one tail of the normal distribution. That is,

lower tolerance bound: $L^*_{P,\gamma} = \hat{\mu} - K^*_{P,\gamma}S$

upper tolerance bound: $U^*_{P,\gamma} = \hat{\mu} + K^*_{P,\gamma}S$

Values of $K^*_{P,\gamma}$ for these intervals are also given in the tables on the following pages.

The lower tolerance bound is such that we are $100\gamma\%$ confident that at least $100P\%$ of the population values are greater than $L^*_{P,\gamma}$, yielding $(L^*_{P,\gamma},\ \infty)$

For example, this value could be used as a Warranty. Suppose we are placing on a package of light bulbs the life length of the light bulbs. We are 95% confident that 99% of the company's light bulbs will last at least $L^*_{.99,.95}$ hours: $(L^*_{.99,.95},\ \infty)$

The upper tolerance bound is such that we are $100\gamma\%$ confident that at least $100P\%$ of the population values are less than $U^*_{P,\gamma}$, yielding $(0,\ U^*_{P,\gamma})$

For example, suppose we are producing paint which contains a small amount of lead. We could state that we are 99% confident that 99.5% of our containers of paint will have at most $U^*_{.995,.99}$ units of lead in the container of paint, yielding $(0,\ U^*_{.995,.99})$,

The following R code yields approximations to the exact coefficients.

However, only use when all three of the following conditions hold:

1. $P \geq .95$
2. $\gamma \geq .95$
3. $n > 50$.

```
#Coefficients for One and Two Sided Tolerance Intervals
n =  100
G =  .90
P =  .99
Chi =  qchisq(1-G,n-1)
z =  qnorm((1+P)/2)
K2Side =  sqrt(((n-1)*(n+1)*z^2)/(n*Chi))

za =  qnorm(G)
zb =  qnorm(P)
a =  1-za^2/(2*(n-1))
b =  zb^2-za^2/n
K1Side =  (zb+sqrt(zb^2-a*b))/a
```

Factors for Determining Two-sided Tolerance Limits

	$\gamma = 0.90$			$\gamma = 0.95$			$\gamma = 0.99$		
	p			p			p		
n	0.900	0.950	0.990	0.900	0.950	0.990	0.900	0.950	0.990
2	15.512	18.221	23.423	31.092	36.519	46.944	155.569	182.720	234.877
3	5.788	6.823	8.819	8.306	9.789	12.647	18.782	22.131	28.586
4	4.157	4.913	6.372	5.368	6.341	8.221	9.416	11.118	14.405
5	3.499	4.142	5.387	4.291	5.077	6.598	6.655	7.870	10.220
10	2.546	3.026	3.958	2.856	3.393	4.437	3.617	4.294	5.610
15	2.285	2.720	3.565	2.492	2.965	3.885	2.967	3.529	4.621
20	2.158	2.570	3.372	2.319	2.760	3.621	2.675	3.184	4.175
25	2.081	2.479	3.254	2.215	2.638	3.462	2.506	2.984	3.915
30	2.029	2.417	3.173	2.145	2.555	3.355	2.394	2.851	3.742
35	1.991	2.371	3.114	2.094	2.495	3.276	2.314	2.756	3.618
40	1.961	2.336	3.069	2.055	2.448	3.216	2.253	2.684	3.524
45	1.938	2.308	3.032	2.024	2.412	3.168	2.205	2.627	3.450
50	1.918	2.285	3.003	1.999	2.382	3.129	2.166	2.580	3.390
60	1.888	2.250	2.956	1.960	2.335	3.068	2.106	2.509	3.297
70	1.866	2.224	2.922	1.931	2.300	3.023	2.062	2.457	3.228
80	1.849	2.203	2.895	1.908	2.274	2.988	2.028	2.416	3.175
90	1.835	2.186	2.873	1.890	2.252	2.959	2.001	2.384	3.133
100	1.823	2.172	2.855	1.875	2.234	2.936	1.978	2.357	3.098
150	1.786	2.128	2.796	1.826	2.176	2.859	1.905	2.271	2.985
200	1.764	2.102	2.763	1.798	2.143	2.816	1.866	2.223	2.921
250	1.750	2.085	2.741	1.780	2.121	2.788	1.839	2.191	2.880
300	1.740	2.073	2.725	1.767	2.106	2.767	1.820	2.169	2.850
350	1.732	2.064	2.713	1.757	2.094	2.752	1.806	2.152	2.828
400	1.726	2.057	2.703	1.749	2.084	2.739	1.794	2.138	2.810
450	1.721	2.051	2.695	1.743	2.077	2.729	1.785	2.127	2.795
500	1.717	2.046	2.689	1.737	2.070	2.721	1.777	2.117	2.783
550	1.713	2.041	2.683	1.733	2.065	2.713	1.770	2.109	2.772
600	1.710	2.038	2.678	1.729	2.060	2.707	1.765	2.103	2.763
650	1.707	2.034	2.674	1.725	2.056	2.702	1.759	2.097	2.755
700	1.705	2.032	2.670	1.722	2.052	2.697	1.755	2.091	2.748
750	1.703	2.029	2.667	1.719	2.049	2.692	1.751	2.086	2.742
800	1.701	2.027	2.664	1.717	2.046	2.688	1.747	2.082	2.736
850	1.699	2.025	2.661	1.715	2.043	2.685	1.744	2.078	2.731
900	1.697	2.023	2.658	1.712	2.040	2.682	1.741	2.075	2.727
950	1.696	2.021	2.656	1.711	2.038	2.679	1.738	2.071	2.722
1000	1.695	2.019	2.654	1.709	2.036	2.676	1.736	2.068	2.718
∞	1.645	1.960	2.576	1.645	1.960	2.576	1.645	1.960	2.576

Factors for Determining One-sided Tolerance Limits

	$\gamma = 0.90$			$\gamma = 0.95$			$\gamma = 0.99$		
	p			p			p		
n	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
2	10.253	13.090	18.500	20.581	26.260	37.094	103.029	131.426	185.617
3	4.258	5.311	7.340	6.155	7.656	10.553	13.995	17.370	23.896
4	3.188	3.957	5.438	4.162	5.144	7.042	7.380	9.083	12.387
5	2.742	3.400	4.666	3.407	4.203	5.741	5.362	6.578	8.939
10	2.066	2.568	3.532	2.355	2.911	3.981	3.048	3.738	5.074
15	1.867	2.329	3.212	2.068	2.566	3.520	2.521	3.102	4.222
20	1.765	2.208	3.052	1.926	2.396	3.295	2.276	2.808	3.832
25	1.702	2.132	2.952	1.838	2.292	3.158	2.129	2.633	3.601
30	1.657	2.080	2.884	1.777	2.220	3.064	2.030	2.515	3.447
35	1.624	2.041	2.833	1.732	2.167	2.995	1.957	2.430	3.334
40	1.598	2.010	2.793	1.697	2.125	2.941	1.902	2.364	3.249
45	1.577	1.986	2.761	1.669	2.092	2.898	1.857	2.312	3.180
50	1.559	1.965	2.735	1.646	2.065	2.862	1.821	2.269	3.125
60	1.532	1.933	2.694	1.609	2.022	2.807	1.764	2.202	3.038
70	1.511	1.909	2.662	1.581	1.990	2.765	1.722	2.153	2.974
80	1.495	1.890	2.638	1.559	1.964	2.733	1.688	2.114	2.924
90	1.481	1.874	2.618	1.542	1.944	2.706	1.661	2.082	2.883
100	1.470	1.861	2.601	1.527	1.927	2.684	1.639	2.056	2.850
150	1.433	1.818	2.546	1.478	1.870	2.611	1.566	1.971	2.740
200	1.411	1.793	2.514	1.450	1.837	2.570	1.524	1.923	2.679
250	1.397	1.777	2.493	1.431	1.815	2.542	1.496	1.891	2.638
300	1.386	1.765	2.477	1.417	1.800	2.522	1.475	1.868	2.608
350	1.378	1.755	2.466	1.406	1.787	2.506	1.461	1.850	2.585
400	1.372	1.748	2.456	1.398	1.778	2.494	1.448	1.836	2.567
450	1.366	1.742	2.448	1.391	1.770	2.484	1.438	1.824	2.553
500	1.362	1.736	2.442	1.385	1.763	2.475	1.430	1.814	2.540
550	1.358	1.732	2.436	1.380	1.757	2.468	1.422	1.806	2.530
600	1.355	1.728	2.431	1.376	1.752	2.462	1.416	1.799	2.520
650	1.352	1.725	2.427	1.372	1.748	2.456	1.411	1.792	2.512
700	1.349	1.722	2.423	1.368	1.744	2.451	1.406	1.787	2.505
750	1.347	1.719	2.420	1.365	1.741	2.447	1.401	1.782	2.499
800	1.344	1.717	2.417	1.363	1.737	2.443	1.397	1.777	2.493
850	1.343	1.714	2.414	1.360	1.734	2.439	1.394	1.773	2.488
900	1.341	1.712	2.411	1.358	1.732	2.436	1.390	1.769	2.483
950	1.339	1.711	2.409	1.356	1.729	2.433	1.387	1.766	2.479
1000	1.338	1.709	2.407	1.354	1.727	2.430	1.385	1.762	2.475
∞	1.282	1.645	2.326	1.282	1.645	2.326	1.282	1.645	2.326

Lower Tolerance Bound for Exponential Distribution

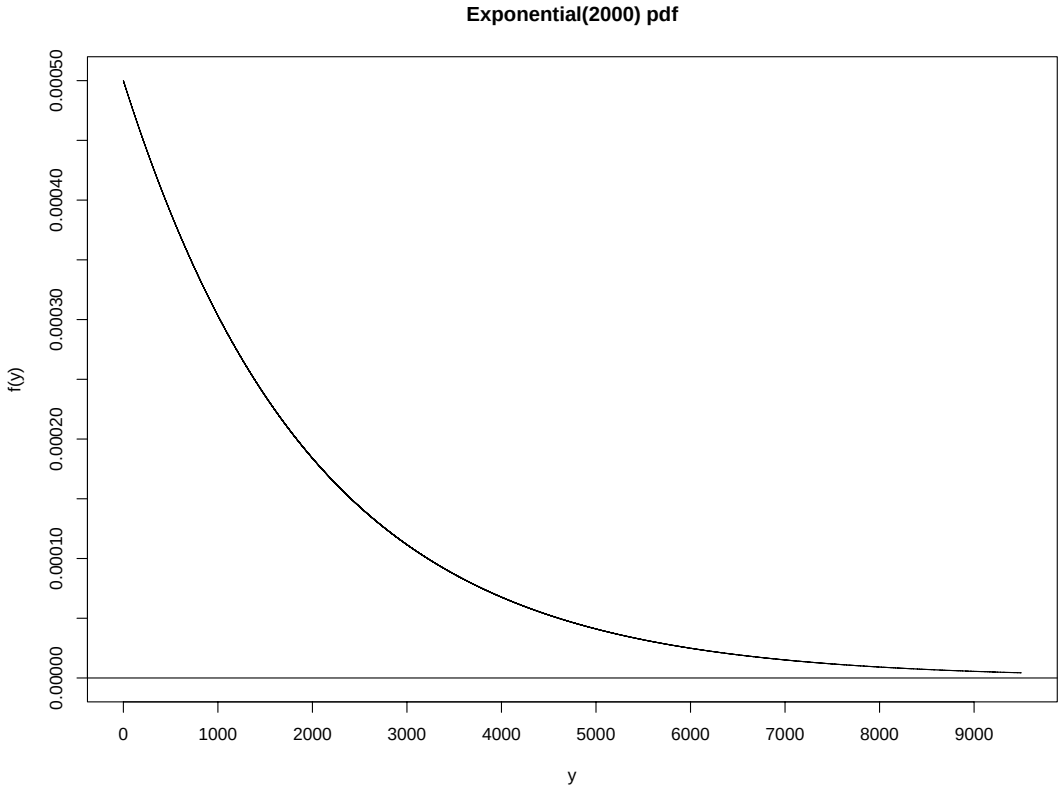
Suppose a company is producing an electronic device which has time to failure modelled by T which has an exponential distribution with average time to failure β. The company wants to set a warranty for the product W such that if the device fails prior to time W they will repair the device without cost to the customer. They want to determine the value for W such that at most 2% of all devices sold will need to be repaired under the warranty. Thus, we need to find W such that $P[T \geq W] \geq .98$.

Case 1: β known

Based on previous repair history, the failure times are known to have an exponential distribution with average time to failure for the device β known with near certainty. Therefore, we have in general the following specification (P=.98 in our example)

$$P[T \geq W_P] = P \Rightarrow \int_{W_P}^{\infty} \frac{1}{\beta} e^{-t/\beta} dt = P \quad \Rightarrow \quad e^{-W_p/\beta} = P \quad \Rightarrow \; W_P = -\beta log(P)$$

If $\beta = 2000$ hours and $P = .98$, we have $W_{.98} = -2000log(.98) = 40.4$ hours.



Case 2: β unknown

The company has made some modifications to the device and thus needs to possibly change the value of W_P. They are certain that the distribution of T is still exponential but the average time to failure for the device β hopefully will have been increased because of improvements in product design.

A survival analysis is conducted on n of the new devices yielding times to failure $T_1, \ldots, T_n$ which are *iid* *Exponential*(β) with β to be estimated from the n data values.

We now need to construct a new lower bound to reflect our uncertainty in the value of β.

Construct $W_{P,\gamma}$ such that we are $100\gamma\%$ confident that $100P\%$ of the population of devices will have times to failure exceeding $W_{P,\gamma}$.

That is, construct $W_{P,\gamma}$ such that with

$$A(W_{P,\gamma}) = \int_{W_{P,\gamma}}^{\infty} \frac{1}{\beta} e^{-t/\beta} dt \quad \text{implies} \quad P\left[A(W_{P,\gamma}) \geq P\right] = \gamma$$

With β known we determined $W_{P,\gamma} = -\beta log(P)$ for all possible values of γ.

To generalize to the case of β unknown we will estimate β using its MLE $\hat{\beta} = \bar{T}$ and

then determine the constant $K_{P,\gamma}$ such that we have

$P\left[A(W_{P,\gamma}) \geq P\right]$, where $W_{P,\gamma} = -\hat{\beta} K_{P,\gamma} log(P)$,

Claim:

$$K_{P,\gamma} = \frac{2n}{\chi^2_{1-\gamma,2n}}$$

where $\chi^2_{1-\gamma,2n}$ is the upper $100(1-\gamma)\%$ percentile from a chi-square distribution with $df = 2n$.

Proof of Claim:

$$A(W_{P,\gamma}) \geq P \iff \int_{W_{P,\gamma}}^{\infty} \frac{1}{\beta} e^{-t/\beta} dt \geq P \iff$$

$$e^{-W_{P,\gamma}/\beta} \geq P \iff -\frac{1}{\beta} W_{P,\gamma} \geq log(P) \iff$$

$$\frac{\hat{\beta}}{\beta} K_{P,\gamma}\ log(P) \geq log(P) \iff \frac{\hat{\beta}}{\beta} K_{P,\gamma} \leq 1 \iff \frac{2n\bar{T}}{\beta} \leq \frac{2n}{K_{P,\gamma}}$$

where $\dfrac{2n\bar{T}}{\beta}$ is distributed chi-square with df=2n

$$\gamma = P\left[A(W_{P,\gamma}) \geq P\right] \iff \gamma = P\left[\frac{2n\bar{T}}{\beta} \leq \frac{2n}{K_{P,\gamma}}\right] \iff \frac{2n}{K_{P,\gamma}} = \chi^2_{1-\gamma,2n} = qchisq(\gamma, 2n)$$

We thus conclude that a $100(P,\gamma)\%$ Lower Tolerance Bound for an Exponential Distribution is

$$W_{P,\gamma} = -\hat{\beta}\left[\frac{2n}{\chi^2_{1-\gamma,2n}}\right] log(P)$$

END Wednesday 10/27/21

START Friday 10/29/21

Example: Determine a Lower Bound on the time to failure of a device having Exponential failure times such that we are 95% confident that at least 90% of the devices will have failure times greater than the Lower Bound, that is,

Determine a $100(.9, .95)\%$ Lower Tolerance Bound for the device.

Suppose we have $n = 10$ failure times and compute $\hat{\beta} = \bar{T}$.

Then the $100(.9, .95)\%$ Lower Tolerance Bound is given by

$W_{P,\gamma} = -\hat{\beta}\left[\frac{2n}{\chi^2_{1-\gamma,2n}}\right] log(P) = -\bar{T}\left[\frac{20}{\chi^2_{.05,20}}\right] log(.9) = -\bar{T}\left[\frac{20}{31.41043}\right] log(.9) = 0.067\bar{T}.$

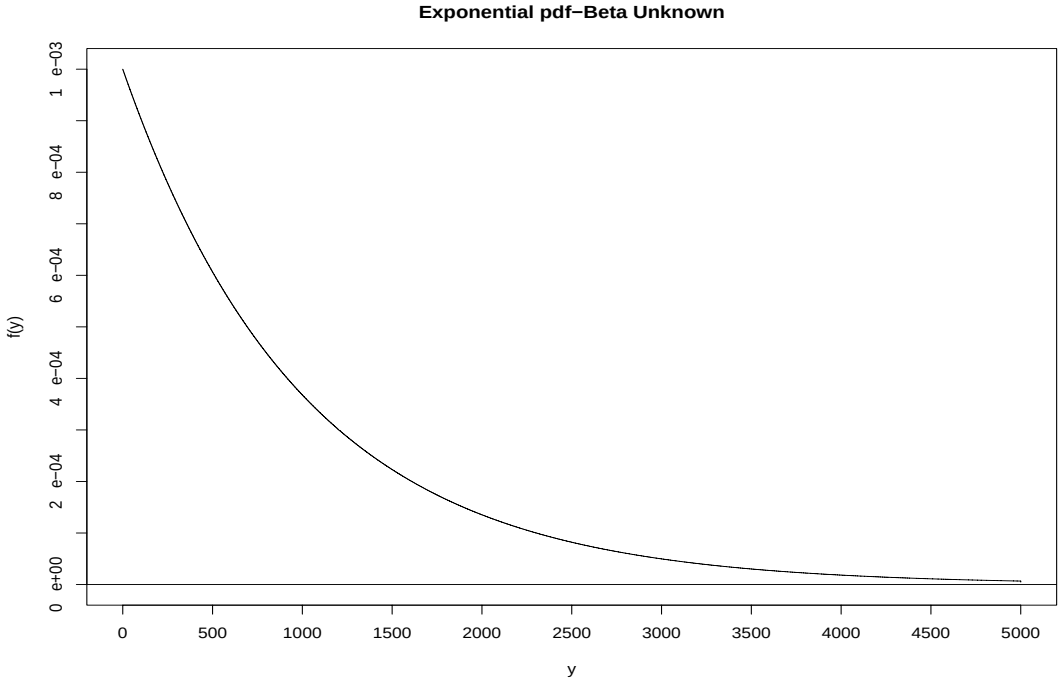
If we would have known β, the Lower Bound would have been

$W_P = -\beta log(P) = -\beta log(.9) = .105\beta.$

Note that for unknown β we have

$E[W_{.9,.95}] = 0.067\beta < .105\beta = W_{.9}.$

Does this relationship seem logical?



Distribution-Free Tolerance Intervals (Bounds) for a Population/Process

In many experiments or studies, the form of the population distribution function F is completely unknown or intractable.

The following procedure yields a distribution-free tolerance interval for a population or process.

Let $Y_1, \ldots, Y_n$ be *iid* with continuous cdf $F(\cdot)$ and pdf $f(\cdot)$.

Let $Y_{(1)} < \ldots < Y_{(n)}$ be the corresponding order statistics for the $Y_i's$.

A $100(P,\gamma)\%$ tolerance interval for the population is given by

$$\left(Y_{(r)},\ Y_{(n-s+1)}\right)$$

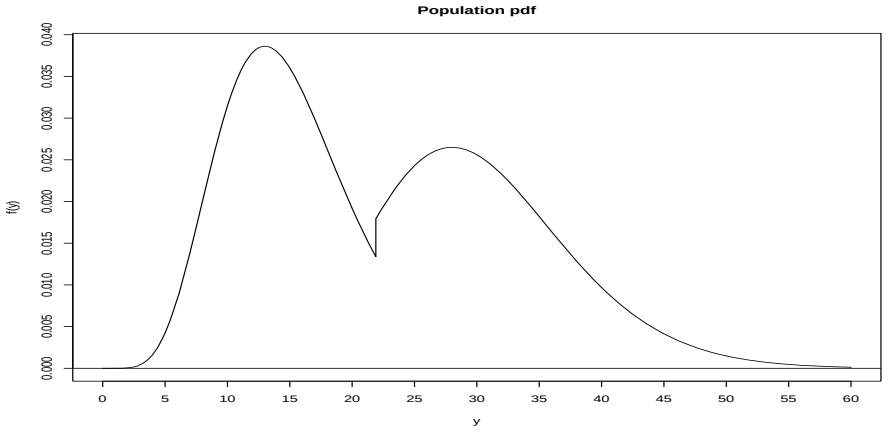
```
   |           |                          |           |
   |--- r ---->|                          |<-- s ----|
---|-----------|------------.......-------|----------|---
 Y(1)        Y(r)                      Y(n-s+1)     Y(n)
```

where r and s are integers satisfying

$$P\left[A(P,\gamma) \geq P\right] = \gamma \quad \text{with}$$

$$A(P,\gamma) = \int_{Y_{(r)}}^{Y_{(n-s+1)}} f(y)dy = F(Y_{(n-s+1)}) - F(Y_{(r)}) = U_{(n-s+1)} - U_{(r)},$$

where $U_{(1)} < \ldots < U_{(n)}$ are the order statistics from n *iid* Uniform on (0,1) r.v.'s.



The tolerance interval is distribution-free by using the probability integral transform theorem, $Y \sim F \Rightarrow F(Y) \sim U_{(0,\ 1)}$ has a uniform on (0, 1) distribution.

The remaining problem is to find the largest integers r and s which yields the narrowest interval $\left(Y_{(r)},\ \ Y_{(n-s+1)}\right)$ such that

$$P\left[Y_{(n-s+1)} - Y_{(r)} \geq P\right] = P\left[U_{(n-s+1)} - U_{(r)} \geq P\right] = \gamma$$

Using properties of the order statistics from an Uniform on (0,1) distribution,

$$P\left[U_{(n-s+1)} - U_{(r)} \geq P\right] = 1 - I_P(n-r-s+1, r+s)$$

where

$$I_P(a,b) = \frac{1}{I_1(a,b)} \int_0^P t^{a-1}(1-t)^{b-1}dt$$

and

$$I_1(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

are incomplete Beta functions.

A few facts about this function will be useful:

1. $I_p(a,b) = 1 - I_{1-p}(b,a)$
2. For $m < n$ integers, $I_p(m, n-m+1) = \sum_{j=m}^{n} \binom{n}{j} p^j (1-p)^{n-j}$
3. $1 - I_p(n-r-s+1, r+s) \geq \gamma$ iff $P[Y \leq m-1] \leq 1-\gamma$

   where Y is distributed $Bin(n, 1-p)$. Thus, we can use the binomial distribution to determine r and s.

- Therefore, $P\left[U_{(n-s+1)} - U_{(r)} \geq P\right] \geq \gamma$ iff $P\left[Y \leq m-1\right] \leq 1-\gamma$

  where $m = r+s$

  Tables for the values of (r,s) are given in Somerville(1958) *Annals of Mathematical Statistics* 29, pp. 599-601. I have included these tables on the next page to obtain $m = r+s$ in the expression: $\left(Y_{(r)},\ \ Y_{(n-s+1)}\right)$.

  To obtain an Upper Tolerance Bound take $r = 0$ yielding $\left(Y_{(0)},\ \ Y_{(n-m+1)}\right)$

  to obtain an Lower Tolerance Bound take $s = 0$ yielding $\left(Y_{(m)},\ \ Y_{(n+1)}\right)$

  where $Y_{(0)}$ is the **population minimum** and $Y_{(n+1)}$ is the **population maximum.**

  When population values are $(0,\ \ \infty)$,

  Upper Tolerance Bound is $\left(0,\ \ Y_{(n-m+1)}\right)$ and lower tolerance bound is $\left(Y_{(m)},\ \ \infty\right)$

Example Based on a random sample of $n = 130$ data values: $Y_1, \ldots, Y_{130}$,

find an interval of values

$[L_{(P,\gamma)}, U_{(P,\gamma)}]$

such that we are 95% certain that at least 90% of the population values are in this interval,

$[L_{(P,\gamma)}, U_{P,\gamma}]$.

That is, find a $100(.9, .95)\%$ tolerance interval for the population.

From the table with $P = .90$, $\gamma = .95$, and $n = 130$, we have $r + s = m = 8$.

Thus, one choice would be to take $r = 4$ and $s = 4$.

Using this choice, the $100(.9, .95)\%$ tolerance interval would be

$[Y_{(4)}, Y_{(130-4+1)}] = [Y_{(4)}, Y_{(127)}]$.

That is, we are 95% certain that at least 90% of the population values are between

$$Y_{(4)} \text{ and } Y_{(127)}.$$

The following R code will compute the value of m given in the tables on the next page along with the actual value of γ:

```
n= 130
G= .95
P= .90
m= 0
imin= 0
i= 0
ans= 0
anst= 0
r= 1:n
ans= pbinom(r-1,n,1-P)
while(i<n)
{
i= i+1
if(ans[i]<=1-G)  anst[i]= ans[i]
if(ans[i]>1-G) anst[i]= -1
}
ansmax= max(anst)
imax= which(anst==ansmax)
m= imax
coverage= 1-ans[m]
out = cbind(m, coverage)
out
# m coverage
  8 .9544028
```

Annals of Mathematical Statistics Vol 29 (1958)

Tolerance Intervals (P, γ): $[X_{(r)}, X_{(n-s+1)}]$

Values of $m = r + s$ such that we may assert with confidence at least γ that 100 P percent of a population lies between the rth smallest and the sth largest of a random sample of n from that population (continuous distribution function assumed)

n	P																								
	$\gamma = 0.50$					$\gamma = 0.75$					$\gamma = 0.90$					$\gamma = 0.95$					$\gamma = 0.99$				
	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99
50	25	12	5	2	0	22	10	3	1	—	20	9	2	1	—	19	8	2	—	—	16	6	1	—	—
55	28	14	5	3	0	25	12	4	2	—	23	10	3	1	—	21	9	2	—	—	19	7	1	—	—
60	30	15	6	3	0	27	13	4	2	—	25	11	3	1	—	24	10	2	1	—	21	8	1	—	—
65	33	16	6	3	0	30	14	5	2	—	27	12	4	1	—	26	11	3	1	—	23	9	2	—	—
70	35	17	7	3	1	32	15	5	2	—	30	13	4	1	—	28	12	3	1	—	25	10	2	—	—
75	38	19	7	4	1	35	16	6	2	—	32	14	4	1	—	30	13	3	1	—	27	10	2	—	—
80	40	20	8	4	1	37	17	6	3	—	34	15	5	2	—	33	14	4	1	—	30	11	2	—	—
85	43	21	8	4	1	39	19	7	3	—	37	16	5	2	—	35	15	4	1	—	32	12	3	—	—
90	45	22	9	4	1	42	20	7	3	—	39	17	5	2	—	37	16	5	1	—	34	13	3	1	—
95	48	24	9	5	1	44	21	7	3	—	41	18	6	2	—	39	17	5	2	—	36	14	3	1	—
100	50	25	10	5	1	47	22	8	3	—	44	20	6	2	—	42	18	5	2	—	38	15	4	1	—
110	55	27	11	5	1	51	24	9	4	—	48	22	7	3	—	46	20	6	2	—	43	17	4	1	—
120	60	30	12	6	1	56	27	10	4	—	53	24	8	3	—	51	22	7	2	—	47	19	5	1	—
130	65	32	13	6	1	61	29	11	5	—	58	26	9	3	—	56	25	8	3	—	52	21	6	2	—
140	70	35	14	7	1	66	31	12	5	1	62	28	10	4	—	60	27	8	3	—	56	23	6	2	—
150	75	37	15	7	1	71	34	12	6	1	67	31	10	4	—	65	29	9	3	—	61	26	7	2	—
170	85	42	17	8	2	81	39	14	7	1	77	35	12	5	—	74	33	11	4	—	70	30	9	3	—
200	100	50	20	10	2	95	46	17	8	1	91	42	15	6	—	88	40	13	5	—	84	36	11	4	—
300	150	75	30	15	3	144	70	26	12	2	139	65	23	10	1	136	63	22	9	1	130	58	19	7	—
400	200	100	40	20	4	193	94	35	17	3	187	89	32	15	2	184	86	30	13	1	177	80	27	11	—
500	250	125	50	25	5	242	118	45	22	3	236	113	41	19	2	232	109	39	17	2	224	103	35	14	1
600	300	150	60	30	6	292	143	55	26	4	284	136	51	23	3	280	133	48	21	2	272	126	44	18	1
700	350	175	70	35	7	341	167	65	31	5	333	160	60	28	4	328	156	57	26	3	319	149	52	22	2
800	400	200	80	40	8	390	192	74	36	6	382	184	69	32	5	377	180	66	30	4	367	172	61	26	2
900	450	225	90	45	9	440	216	84	41	7	431	208	79	37	5	425	204	75	35	4	415	195	70	30	3
1000	500	250	100	50	10	489	241	94	45	8	480	233	88	41	6	474	228	85	39	5	463	219	79	35	3

Tolerance Interval (P, γ): $[X_{(1)}, X_{(n)}]$

Confidence γ with which we may assert that 100 P percent of the population lies between the largest and smallest of a random sample of n from that population (continuous distribution assumed)

n	$P = .50$	$P = .75$	$P = .90$	$P = .95$	$P = .99$	n	$P = .75$	$P = .90$	$P = .95$	$P = .99$
3	.50	.16	.03	.01	.00	17	.95	.52	.21	.01
4	.69	.26	.05	.01	.00	18	.96	.55	.22	.01
5	.81	.37	.08	.02	.00	19	.97	.58	.25	.02
6	.89	.47	.11	.03	.00	20	.98	.61	.26	.02
7	.94	.56	.15	.04	.00	25	.99	.73	.36	.03
8	.96	.63	.19	.06	.00	30	1.00−	.82	.45	.04
9	.98	.70	.23	.07	.00	40		.92	.60	.06
10	.99	.76	.26	.09	.00	50		.97	.73	.09
11	.99	.80	.30	.10	.01	60		.99	.81	.12
12	1.00−	.84	.34	.12	.01	70		.99	.87	.16
13		.87	.38	.14	.01	80		1.00−	.91	.19
14		.90	.42	.15	.01	90			.94	.23
15		.92	.45	.17	.01	100			.96	.26
		.94	.49	.19	.01					

Comparison of Distribution-Free to Parametric Tolerance Intervals

Because there is a Distribution-Free method to obtain Tolerance Intervals, why bother with using the specific Tolerance Intervals for given families of distributions?

The major reason is that the Distribution-Free Tolerance Intervals may be considerably wider than the corresponding interval for a particular family because the Distribution-Free Tolerance Intervals must be applicable to a wide-variety of distributions.

Suppose the population or process cdf F is from a $N(\mu, \sigma^2)$ family of distributions.

Normal based Tolerance Interval: $\bar{Y} \pm K_{p,\gamma} S$

Distribution-free Tolerance Interval: $\left(Y_{(r)},\ Y_{(n-s+1)}\right)$

For $100(.9, .95)\%$ Tolerance Intervals, compare the widths of the normal-based tolerance interval to the distribution-free tolerance interval. Because the widths are random variables we will consider the expected widths for comparison under the assumption that the population distribution is $N(\mu, \sigma^2)$.

$W_N = 2K_{P,\gamma} S$ $\qquad$ $W_{DF} = Y_{(n-s+1)} - Y_{(r)}$ which yields

$E[W_N] = 2K_{P,\gamma} E[S] = 2K_{P,\gamma} \frac{\sigma\sqrt{2}}{\sqrt{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$

$E[W_{DF}] = E[Y_{(n-s+1)}] - E[Y_{(r)}]$

with $E[Y_{(r)}] = \mu + \sigma \frac{n!}{(n-r)!(r-1)!} \int_{-\infty}^{\infty} y(.5 - \Phi(y))^{r-1}(.5 + \Phi(y))^{n-r}\phi(y)dy,$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the $N(0,1)$ cdf and pdf, respectively.

Evaluating these expressions yields the following results:

n	$E[W_{DF}]$	$E[W_N]$	$E[W_{DF}]/E[W_N]$
50	4.498σ	3.978σ	1.15
100	4.094σ	3.741σ	1.10
200	3.753σ	3.591σ	1.05

Thus, even for a relatively large value of n, $n = 100$ the Distribution-Free Tolerance Interval is 10% wider than the normal-based Tolerance Interval.

Alternative Approaches to Constructing C.I.'s, P.I.'s, and T.I.'s

Two other approaches to constructing C.I.'s, P.I.'s, and T.I.'s are to transform the data to normality and then use a normal-based approach or use bootstrap procedures on the untransformed data.

Box-Cox Transformations to Normality

Suppose the population distributions is non-normal, that is, Y has cdf $F_Y(\cdot)$ which is non-normal but the Box-Cox transformation $X = g(Y)$ results in X having a cdf $F_X(\cdot)$ which is approximately a normal distribution. How can Intervals generated for X yield a corresponding interval for Y?

1. **Tolerance Interval for Distribution of Y based on $X = g(Y)$**

   Because X is approximately normally distributed we obtain a $100(P,\gamma)\%$ T.I. for the distribution for X:

   $(L^*_{P,\gamma}, U^*_{P,\gamma}) \Rightarrow \quad \gamma = P[F_X(U^*_{P,\gamma}) - F_X(L^*_{P,\gamma}) \geq P].$

   Case 1: Increasing Function: Ex $\theta = 2$

   Let $g^{-1}(\cdot)$ be the inverse of g where g is a strictly increasing function, a $100(P,\gamma)\%$ T.I. for the distribution of Y is

   $(L_{P,\gamma}, U_{P,\gamma}) = \left(g^{-1}(L^*_{P,\gamma}), g^{-1}(U^*_{P,\gamma})\right).$

   This result follows from:

   $F_Y(y) = P[Y \leq y] = P[g^{-1}(X) \leq y] = P[g(g^{-1}(X)) \leq g(y)] = P[X \leq g(y)] = F_X(g(y)) \Rightarrow$

   $$\begin{aligned} P[F_Y(U_{P,\gamma}) - F_Y(L_{P,\gamma}) \geq P] &= P[F_Y(g^{-1}(U^*_{P,\gamma})) - F_Y(g^{-1}(L^*_{P,\gamma})) \geq P] \\ &= P[F_X(g(g^{-1}(U^*_{P,\gamma}))) - F_X(g(g^{-1}(L^*_{P,\gamma}))) \geq P] \\ &= P[F_X(U^*_{P,\gamma}) - F_X(L^*_{P,\gamma}) \geq P] \\ &= \gamma \end{aligned}$$

   Case 2: Decreasing Function: Ex: $\theta = -0.5$

   Let g be a strictly decreasing function, a $100(P,\gamma)\%$ T.I. for the distribution of Y is

   $(L_{P,\gamma}, U_{P,\gamma}) = \left(g^{-1}(U^*_{P,\gamma}), g^{-1}(L^*_{P,\gamma})\right).$

   This result follows from:

   $F_Y(y) = P[Y \leq y] = P[g^{-1}(X) \leq y] = P[g(g^{-1}(X)) \geq g(y)] = P[X \geq g(y)] = 1 - F_X(g(y))$

   $$\begin{aligned} P[F_Y(U_{p,\gamma}) - F_Y(L_{p,\gamma}) \geq P] &= P[F_Y(g^{-1}(L^*_{p,\gamma})) - F_Y(g^{-1}(U^*_{p,\gamma})) \geq P] \\ &= P[1 - F_X(g(g^{-1}(L^*_{P,\gamma}))) - 1 + F_X(g(g^{-1}(U^*_{P,\gamma}))) \geq P] \\ &= P[F_X(U^*_{P,\gamma}) - F_X(L^*_{P,\gamma}) \geq P] \\ &= \gamma \end{aligned}$$

2. **Prediction Interval for Distribution of Y based on $X = g(Y)$**

The same results from above would apply to prediction intervals.

Suppose $Y_1, \ldots, Y_n$ are *iid* but have a non-normal distribution but $g(Y_i) = X_i$ is approximately normally distributed.

Case 1: g a strictly increasing function

Let $\bar{X}$ and S_X be the sample mean and standard deviation of the $X_i s$

Let $\bar{X} \pm t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}$ be a $100(1-\alpha)\%$ P.I. for X_{n+1}, that is,

$$\begin{aligned} 1-\alpha &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq X_{n+1} \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq g(Y_{n+1}) \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right) \leq Y_{n+1} \leq g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right)\right] \Rightarrow \end{aligned}$$

$$\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right), g^{-1}\left(X + t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right)\right)$$

is a $100(1-\alpha)\%$ P.I. for Y_{n+1}

Case 2: g a strictly decreasing function

$$\begin{aligned} 1-\alpha &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq X_{n+1} \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq g(Y_{n+1}) \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right) \geq Y_{n+1} \geq g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right)\right] \Rightarrow \end{aligned}$$

$$\left(g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right), g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right)\right)$$

is a $100(1-\alpha)\%$ P.I. for Y_{n+1}

3. Confidence Intervals for Population μ and σ Using Transformations

The results from above **DO NOT** in general apply to confidence intervals for μ, σ, and most other parameters related to population moments. However, transformations are valid for Quantiles, such as the population median, $Q(.5)$.

Suppose $Y_1, \ldots, Y_n$ are *iid* but have a non-normal distribution but $g(Y_i) = X_i$ is approximately normally distributed with g a strictly increasing function. We want to obtain a $100(1-\alpha)$ C.I. for μ_Y.

Let $\bar{X} \pm t_{\frac{\alpha}{2}} S/\sqrt{n}$ be a $100(1-\alpha)\%$ C.I. for μ_X.

Is $\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right), g^{-1}\left(X + t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right)\right)$ a $100(1-\alpha)\%$ C.I. for μ_Y?

We cannot conclude this because, in general, $\mu_X \neq g(\mu_Y)$ and hence $g^{-1}(\mu_X) \neq \mu_Y$.

This results from the following:

$\mu_X = E[X] = E[g(Y)] \neq g(E[Y]) = g(\mu_Y)$, thus, $g^{-1}(\mu_X) \neq \mu_Y$.

Using a Taylor series expansion of $g(y)$ about μ_Y we obtain

$X = g(Y) = g(\mu_Y) + g'(\mu_Y)(Y - \mu_Y) + \frac{1}{2} g''(\mu_Y)(Y - \mu_Y)^2 + R \Rightarrow$

$$\begin{aligned} \mu_X = E[X] = E[g(Y)] &= g(\mu_Y) + g'(\mu_Y)E[(Y - \mu_Y)] + \frac{1}{2} g''(\mu_Y)E[(Y - \mu_Y)^2] + E[R] \\ &= g(\mu_Y) + \frac{1}{2} g''(\mu_Y)Var(Y) + E[R] \\ &\approx g(\mu_Y) \end{aligned}$$

provided $Var(Y)$ is small and/or $g''(\mu_Y)$ is small, and $E[R]$ is small.

In many cases, neither $Var(Y)$ is small nor $E[R]$ is small.

This leads to a very bad approximation of μ_X using $g(\mu_Y)$ and hence the C.I. for μ_Y obtained from $\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right), g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right)\right)$ will not be an appropriate C.I. for μ_Y.

An alternative approach in such situations is to attempt to find a C.I. for μ_Y directly from the distribution of Y.

If this is not possible then the **bootstrap C.I.** may be an alternative methodology for obtaining the confidence interval.

Example Let Y have a lognormal distribution, that is, $Y = e^W$, where W is $N(\mu, \sigma^2)$.

Let

$$X = g(Y) = log(Y)$$

Then, X has a $N(\mu, \sigma^2)$ distribution with

$\mu_X = E[X] = E[log(Y)] = \mu$ and

$\mu_Y = E[Y] = e^{\mu + \frac{1}{2}\sigma^2}$.

How close is μ_X to $g(\mu_Y) = log(\mu_Y)$?

$log(\mu_Y) = \mu + \frac{1}{2}\sigma^2 = \mu_X + \frac{1}{2}\sigma^2 \quad \Rightarrow \quad \mu_X = log(\mu_Y) - \frac{1}{2}\sigma^2$.

Therefore, if σ^2 is large relative to μ_X, then

μ_X will not be very close to $g(\mu_Y) = log(\mu_Y)$.

A $100(1-\alpha)\%$ C.I. for μ_X is $\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n} \;\; , \;\; \bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n} \;\right)$

Inverting the end points of the C.I. for μ_X, we obtain:

$$\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}\;\right), \;\; g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n}\;\right)\right) = \left(e^{\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}}, \;\; e^{\bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n}}\right)$$

This interval would be a C.I. for the parameter

$$e^{\mu_X} = e^{log(\mu_Y) - \frac{1}{2}\sigma^2} = \mu_Y e^{-\frac{1}{2}\sigma^2}$$

and not a C.I. for μ_Y.

To obtain an approximate C.I. for μ_Y would involve simultaneous C.I.'s for μ and σ^2

4. **Confidence Intervals for Population Quantiles: $Q(u)$ Using Transformations**

Suppose we want a C.I. for $Q_Y(u)$ for the distribution of the r.v. Y but standard techniques are not feasible.

However, there exists a transformation, of Y, $X = g(Y)$ such that a $100(1-\alpha)\%$ C.I. can be constructed for $Q_X(u)$

C.I. on $Q_X(u) = (L_x,\ U_x)$

A. If g is an increasing function then recall that $Q_X(u) = g(Q_Y(u))$ thus a $100(1-\alpha)\%$ C.I. on $Q_Y(u)$ is given by

$(g^{-1}(L_x),\ g^{-1}(U_x))$

B. If g is a decreasing function then recall that $Q_X(u) = g(Q_Y(1-u))$ thus a $100(1-\alpha)$ C.I. on $Q_Y(u)$ is given by

$(g^{-1}(U_x^*),\ g^{-1}(L_x^*))$, where

$(L_x^*,\ U_x^*)$ is a $100(1-\alpha)\%$ on $Q_X(1-u)$

Bootstrap Confidence Intervals for Parameters Related to cdf

Let $X_1, \ldots, X_n$ be iid random variables with a common cdf $F(\cdot)$.

Let θ be a parameter which we wish to estimate using a function of the data $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$.

Furthermore, we want to construct a $100(1-\alpha)\%$ C.I. for θ based on the pivot $R_n = \hat{\theta}_n - \theta$.

Suppose the cdf $F(\cdot)$ is unknown and the sample size n is small or that the asymptotic distribution of $\hat{\theta}_n$ is intractable.

We need to determine the sampling distribution of the pivot $R_n = \hat{\theta}_n - \theta$ in order to be able to determine its percentiles, R_α, which are necessary in order to obtain a $100(1-\alpha)\%$ C.I. for θ.

We cannot use exact mathematical derivations to derive the true sampling distribution of $\hat{\theta}_n - \theta$ because the cdf $F(\cdot)$ is unknown. Also, the asymptotic distribution may not provide an adequate approximation of the true sampling distribution of $\hat{\theta}_n - \theta$ when n is small to moderate in size.

An alternative to these two approaches is the **bootstrap procedure** which will provide an approximation to the sampling distribution of the pivot, $\hat{\theta}_n - \theta$, in those situations where we can write θ as a function of the cdf, that is, $\theta = g(F(\cdot))$.

To obtain the sample estimator, we simply replace the true cdf $F(\cdot)$ with the empirical (sample) cdf $\hat{F}(x)$ in $\theta = g(F(\cdot))$ to obtain $\hat{\theta}_n = g(\hat{F}(\cdot))$.

We want to obtain the sampling distribution of $\hat{\theta}_n - \theta$ using simulated data from the edf $\hat{F}(\cdot)$ in place of the true cdf $F(\cdot)$.

Let $\hat{\theta}_D$ be the value of $\hat{\theta}_n$ computed from the edf, $\hat{F}$, that is, from the data, $X_1,\ X_2,\ \ldots,\ X_n$.

The sampling distribution of $\hat{\theta}_n - \theta$, will be approximated by the sampling distribution of $\hat{\theta}^*_n - \hat{\theta}_D$, where $\hat{\theta}^*_n$ is the value of $\hat{\theta}_n$ from a bootstrap sample.

Thus, we have replaced $\hat{\theta}_n$ with $\hat{\theta}^*_n$, its value from the bootstrap sample and θ with its estimator from the edf, $\hat{F}$, $\hat{\theta}_D$.

More formally, if $G(\cdot)$ is the sampling cdf of $\hat{\theta}_n - \theta$, that is,

$G(y) = P[\hat{\theta}_n - \theta \leq y]$,

then the bootstrap simulation estimator of $G(y)$ based on B bootstrap samples is given by

$$\widehat{G}_B(y) = \frac{1}{B}\sum_{i=1}^{B} I[\hat{\theta}^*_i - \hat{\theta}_D \leq y]$$

the proportion of the B samples in which $\hat{\theta}^*_i - \hat{\theta}_D$ are less than or equal to y and

where $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$

are B estimates of θ based on the B bootstrap samples of size n taken from the original data set.

STOP Friday 10/29/21

The approximation of $\widehat{G}_B(u)$ to $G(u)$ contains two sources of error:

1. the difference between $\widehat{G}(u)$ and $G(u)$ due to data variability from the original population
2. the difference between $\widehat{G}_B(u)$ and $\widehat{G}(u)$ due to a finite number of bootstrap samples.

We will approximate the percentiles of the sampling distribution of the pivot $R_n = \hat{\theta}_n - \theta$ using the sample percentiles from the

B values of $R_n^* = \hat{\theta}^* - \hat{\theta}_D$: $R_{n,1}^*, R_{n,2}^*, \cdots, R_{n,B}^*$

This results from the fact that if $X_1, \ldots, X_M$ are iid with cdf $K(\cdot)$ and if $X_{(i)}$ denotes the ith ordered value of the $X_i's$, then

$E[X_{(i)}] \approx K^{-1}(\frac{i}{M+1})$.

Thus, a sensible estimator of $Q(p) = K^{-1}(p)$ is $X_{((M+1)p)}$, provided $(M+1)p$ is an integer.

So we will estimate the $100pth$ quantile of the pivot $R_n = \hat{\theta}_n - \theta$ by

the $(B+1)pth$ ordered value in $R_{(1)}^*, R_{(2)}^*, \cdots, R_{(B)}^*$,

that is, $R_{((B+1)p)}^* = \hat{\theta}_{((B+1)p)}^* - \hat{\theta}_D$,

where $\hat{\theta}_{((B+1)p)}^*$ is the $(B+1)pth$ ordered value of $\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \cdots, \hat{\theta}_{(B)}^*$,.

In our development we will select B such that $(B+1)p$ is an integer.

For example, suppose $n = 20$ and we want to obtain

the $100(.025) = 2.5$th percentile and $100(.975) = 97.5$th percentile.

With B=9999 then we would use

the $(B+1)(.025) = 250$th ordered value of R_n^*, $R_{(250)}^*$ to estimate the 2.5th percentile and

$(B+1)(.975) = 9750$th ordered value of R_n^*, $R_{(9750)}^*$ to estimate the 97.5th percentile

START Monday 7/11/21

Basic Bootstrap Confidence Limits

For a given confidence level $1-\alpha$, and using the Pivot $= \hat{\theta}_n - \theta$,

replace w/ sample estimate i.e. MLE.

we need to find values $L_{\frac{\alpha}{2}}$ and $U_{\frac{\alpha}{2}}$ such that

replace w/ bootstrap estimate

$P(L_{\frac{\alpha}{2}} \leq \hat{\theta}_n - \theta \leq U_{\frac{\alpha}{2}}) \approx 1-\alpha$ which would yield

$\left(\hat{\theta}_n - U_{\frac{\alpha}{2}},\ \hat{\theta}_n - L_{\frac{\alpha}{2}}\right)$ as the $100(1-\alpha)$ C.I. for θ.

Because the sampling distribution of the Pivot $= \hat{\theta}_n - \theta$ is unknown, we will convert the problem to the following with

$\hat{\theta}_n^*$ in place of $\hat{\theta}_n$ and → $\hat{\theta}_n^*$ is $\hat{\theta}_n$ estimate from bootstrap

$\hat{\theta}_D$ in place of θ in the Pivot and use the bootstrap samples to obtain the necessary percentiles. $\hat{\theta}_n$ MLE $\hat{\theta}_D$ estimate

Let $R^* = \hat{\theta^*}_{ni} - \hat{\theta}_D$.

Generate B bootstrap values for $R^*:\ R_1^*,\ R_2^*,\ \ldots, R_B^*$ and then order these values:

$$R_{(1)}^* \leq R_{(2)}^* \leq \ldots \leq R_{(B)}^*$$

For a given confidence level $1-\alpha$, find values $L_{\frac{\alpha}{2}}^*$ and $U_{\frac{\alpha}{2}}^*$ such that

$P(L_{\frac{\alpha}{2}}^* \leq \hat{\theta}_n^* - \hat{\theta}_D \leq U_{\frac{\alpha}{2}}^*) \approx 1-\alpha \Rightarrow$ (using the results on the previous page)

$L_{\frac{\alpha}{2}}^* = R_{((B+1)(\frac{\alpha}{2}))}^* = \theta_{((B+1)(\frac{\alpha}{2}))}^* - \hat{\theta}_D$ and

$U_{\frac{\alpha}{2}}^* = R_{((B+1)(1-\frac{\alpha}{2}))}^* = \hat{\theta}_{((B+1)(1-\frac{\alpha}{2}))}^* - \hat{\theta}_D$.

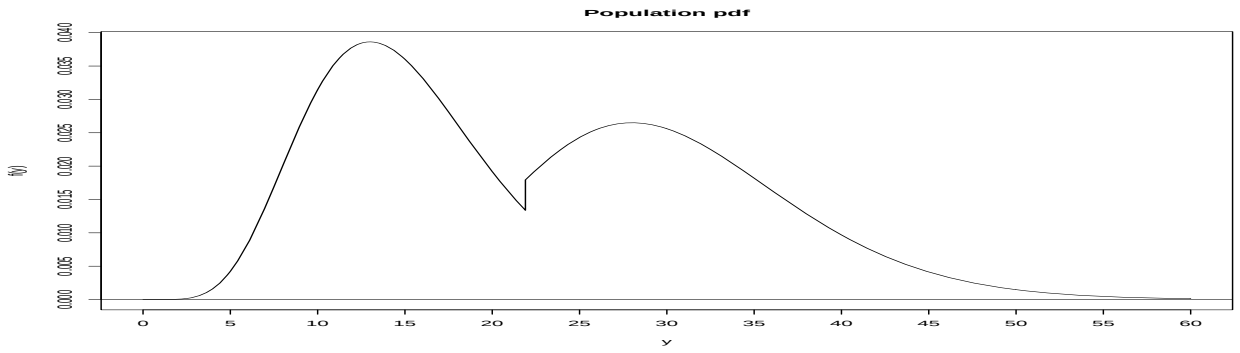
Therefore, an approximate $100(1-\alpha)\%$ C.I. for θ is given by

$$\left(\hat{\theta}_D - R_{((B+1)(1-\frac{\alpha}{2}))}^*,\quad \hat{\theta}_D - R_{((B+1)(\frac{\alpha}{2}))}^*\right) = \left(2\hat{\theta}_D - \hat{\theta}_{((B+1)(1-\frac{\alpha}{2}))}^*,\quad 2\hat{\theta}_D - \hat{\theta}_{((B+1)(\frac{\alpha}{2}))}^*\right)$$

The accuracy of the approximation depends on the size of B, thus one typically takes $B \geq 10000$.

Accuracy also depends upon the extent to which the distribution of $\hat{\theta}_n^* - \hat{\theta}_D$ agrees with that of $\hat{\theta}_n - \theta$.

If the distribution of $\hat{\theta}_n - \theta$ does depend on unknown parameters, then alternative formulations of the C.I.'s must be invoked. The reference *Bootstrap Methods and their Applications*, Davison and Hinkley, discuss this situation in detail, including diagnostics to determine if the pivot depends on θ. We will consider this situation following the next example.



Example The following example is from *Computer Intensive Statistical Methods*, by Hjorth.

In a large installation of electric bulbs, all the bulbs are planned to be replaced regularly after 1200 hours. In order to form an opinion about this strategy, the probability that a bulb will survive 1200 hours is of interest. Let T be the time to failure of this type of bulb. The parameter of interest is

$\theta = P[T \geq 1200] = 1 - F(1200) \Rightarrow \theta = g(F) = 1 - F(1200)$, where $F(\cdot)$ is the cdf of T.

A limited test of the bulbs is conducted and the following 20 life times are observed:

1354	1552	1766	1325	2183	1354	1299	627	695	2586
2420	71	2195	1825	159	1577	3725	884	1014	965

From the data, we compute $\hat{\theta}_n = 1 - \hat{F}(1200) = \frac{13}{20} = .65 = \theta_D$

because 13 of the 20 bulbs failed after 1200 hours.

We will compute an approximate 95% C.I. for θ using the following bootstrap program: **boot1,ci.R**:

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,2420,71,2195,1825,159,1577,3725,884,1014,965)

n= length(x)

m= sum(x>1200)

thest = m/n

B = 9999

thestS = numeric(B)

thestS = rep(0,times =B)

for (i in 1:B)

thestS[i] = sum(sample(x,replace=T)>1200)/20

RS = sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS
```

The approximate 95% C.I. for θ is (thL, thU) = (.45, .85).

Suppose we also want to compute an approximate 95% confidence interval for the median time to failure:
$\theta = Median$.

We would proceed similarly as above
using the following bootstrap program: **boot2,ci.R**:

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,2420,71,2195,1825,159,1577,3725,884,1014,965)

n= length(x)

thest = median(x)

B = 9999

thestS = numeric(B)

thestS = rep(0,times =B)

for (i in 1:B)

thestS[i] = median(sample(x,replace=T))

RS= sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS
```

The point estimate is $\hat{\theta}_n = 1354$.

The approximate 95% C.I. for θ is

$$(\hat{\theta}_L,\ \hat{\theta}_U) = (thL,\ thU) = (912.5,\ 1718.5)$$

Using our nonparametric C.I. for the median, an approximated 95% c.i. for the median is

$$\left(\hat{\theta}_L,\ \hat{\theta}_U\right) \ = \ \left(Y_{(k)},\ Y_{(n-k+1)}\right) = \left(Y_{(6)},\ Y_{(15)}\right) = (965,\ 1825)$$

The two intervals are reasonable close considering that we only have $n = 20$ data values.

Studentized Bootstrap C.I. (Improvement from above when we have additional information.)

If the distribution of $\hat{\theta}_n - \theta$ depends on unknown parameters, then the basic bootstrap procedure may not be very accurate.

For example, suppose $\theta = Q(u)$ for some value of u, $0 < u < 1$.

$\hat{\theta}_n = \hat{Q}(u)$ and $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$

has asymptotic standard error:

$$SE_{Asy}(\hat{\theta}) = \frac{\sqrt{u(1-u)}}{f(Q(u))} = \frac{\sqrt{u(1-u)}}{f(\theta)}$$

which is a function of θ

A method which regains some of the lost accuracy is to use the *Studentized* version of $\hat{\theta}_n - \theta$:

$$Z = \frac{\hat{\theta}_n - \theta}{\sqrt{\widehat{V}}}$$; Pivot in studentized bootstrap

where $\widehat{V}$ is an estimate of $Var(\hat{\theta}_n)$.

The **Studentized Bootstrap C.I. for** θ is given by

$$\left(\hat{\theta}_D - \sqrt{\widehat{V}_D}\ Z^*_{((B+1)(1-\frac{\alpha}{2})},\ \ \hat{\theta}_D - \sqrt{\widehat{V}_D}\ Z^*_{((B+1)(\frac{\alpha}{2})}\ \right)$$

where $\hat{\theta}_D$ is the sample estimate of θ and $\widehat{V}_D$ is the sample estimate of V.

$Z^*_{((B+1)(1-\frac{\alpha}{2})}$ and $Z^*_{((B+1)(\frac{\alpha}{2})}$

are the $(B+1)(\frac{\alpha}{2})$ and $(B+1)(1-\frac{\alpha}{2})$ ordered values obtained from

B bootstrap samples of Z: $Z^*_1, \ldots, Z^*_B$.

The difficulty is that we need to obtain a form for $\hat{V}$.

In some cases, the form will be known or we can use the asymptotic variance for $\hat{V}$.

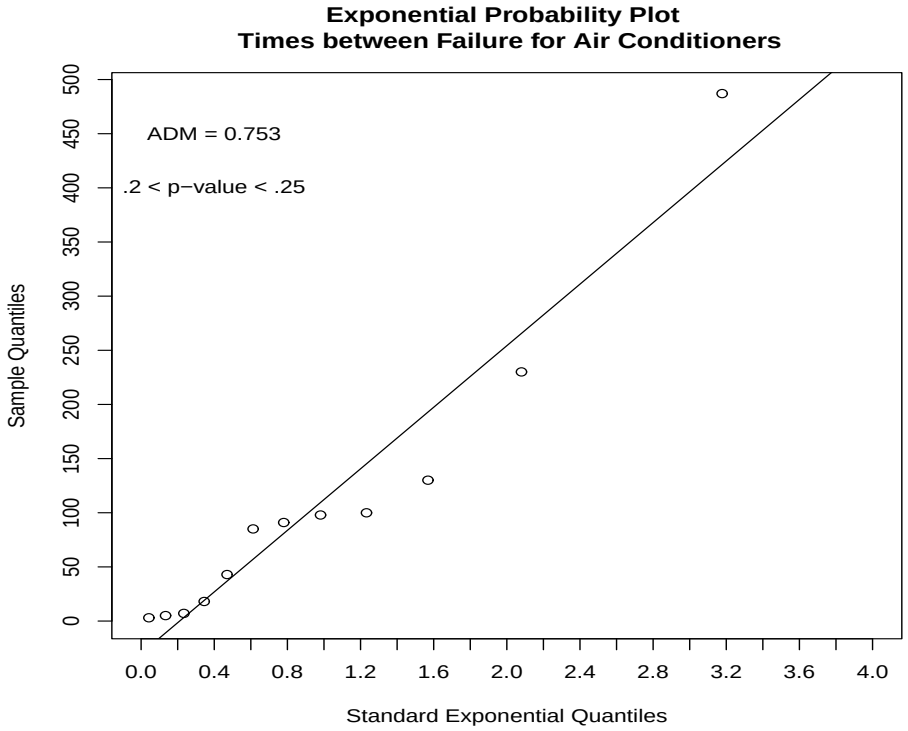
Davison and Hinkley provide a general form for $\hat{V}$ involving influence functions and the nonparametric delta method.

Example A study yielded the n=12 times between failure of air-conditioning units, for which we wish to estimate the average time between failures:

3	5	7	18	43	85
91	98	100	130	230	487

Suppose we model the times to failure as an exponential distribution (see Exponential Reference Distribution Plot) and compute

AD=0.753 with $.20 < p - value < .25$.



For the exponential model, the unknown parameter $\beta = E[T]$ and $Var[T] = \beta^2 \ \Rightarrow \ SE(\hat{\beta}) = \frac{\beta}{\sqrt{n}}$.

If we did have prior information that $T_1,\ T_2, \ldots,\ T_n$ were iid Exp(β) r.v.s, and used $\bar{T} - \mu$ as the pivot for the C.I. for μ, then the sampling distribution of $\bar{T} - \mu$ would depend on the unknown parameter because $\text{Var}(\bar{T} - \mu) = \beta/\sqrt{n}$.

Using the knowledge that the data is from an exponential distribution, the MLE's are $\hat{\beta} = \bar{T}$ and $\hat{V} = \hat{\beta}^2/n$.

Thus, let $\hat{V} = \bar{T}^2/n$ which yields

$$Z = \frac{\hat{\theta}_D - \theta}{\sqrt{\hat{V}}} = \frac{\bar{T} - \beta}{\bar{T}/\sqrt{n}} = \frac{\sqrt{n}(\bar{T} - \beta)}{\bar{T}}$$

An exact C.I.'s for β can be obtained using the fact that the sampling distribution of $2n\bar{T}/\beta$ has a chi-squared distribution with df=2n.

To illustrate the studentized bootstrap technique, the following R code was used **bootexp,ci.R:**

bootexp,ci.R:

```
x= c(3,5,7,18,43,85,91,98,100,130,230,487)

n= length(x)

thest = mean(x)

V = thest**2/n

B = 9999

W = numeric(B)

W = rep(0,times =B)

for (i in 1:B)

W[i] = mean(sample(x,replace=T))

Z = sqrt(n)*(W-thest)/W

Z = sort(Z)

LZ = Z[250]

UZ = Z[9750]

thL = thest-UZ*sqrt(V)

thU = thest-LZ*sqrt(V)
```

We obtain $Z^*_{((B+1)(1-\frac{\alpha}{2})} = Z^*_{(9750)} = 1.494$ and $Z^*_{((B+1)(\frac{\alpha}{2}))} = Z^*_{(250)} = -4.474$.

Thus, an approximate 95% C.I. for β is

$$\left(\hat{\theta}_n - \sqrt{\hat{V}} Z^*_{((B+1)(1-\frac{\alpha}{2})}, \quad \hat{\theta} - \sqrt{\hat{V}} Z^*_{((B+1)(\frac{\alpha}{2})} \right) =$$

$$\left(108.08 - \sqrt{973.5006}(1.494), \quad 108.08 - \sqrt{973.5006}(-4.474) \right) = (61.5, 247.7)$$

The C.I. for β using the fact we are sampling from an exponential distribution and using the pivot $2n\bar{T}/\beta$ is given by

$$\left(\frac{2n\bar{T}}{\chi^2_{.025}}, \frac{2n\bar{T}}{\chi^2_{.975}} \right) = \left(\frac{24(108.08)}{39.364}, \frac{24(108.08)}{12.401} \right) = (65.9, 209.2)$$

The bootstrap C.I. and Exact C.I. are surprising close considering that the C.I.'s are based on only $n = 12$ data values.

— Basic Bootstrap example -> NOTE: Not Studentized

Example In the light bulb example on page 61, we wanted to estimate the value of bootstrap b/c

$\theta = P[T \geq 1200] = g(F) = 1 - F(1200)$, where $F(\cdot)$ is the cdf of T.

If we take $F(t) = 1 - e^{-t/\beta}$, exponential model, then the parameter of interest is we don't know the

$\theta = 1 - F(1200) = e^{-1200/\beta}$, with $\hat{\beta} = \bar{T} = 1478.8$, standard error of $\hat{\theta}$.

we obtain $\hat{\theta}_n = e^{-1200/1478.8} = 0.444$

To obtain a basic bootstrap C.I. for θ we will use the following program **boot3,ci.R**:

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,2420,71,2195,1825,159,1577,3725,884,1014,965)
n = length(x)

mn= mean(x)

thest = exp(-1200/mn)

R = 9999
thestS = numeric(R)
thestS = rep(0,times =R)

for (i in 1:R)
thestS[i] = exp(-1200/mean(sample(x,replace=T)))

RS = sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS
```

From the R output we have:

$R^*_{(1000)(.025)} = RS[250] = -.100, \quad R^*_{(1000)(.975)} = RS[9750] = .080, \quad \hat{\theta}_D = .444$

Then, our approximate 95% C.I. for θ is

$$\begin{aligned} \left(\hat{\theta}_D - R^*_{.975}, \;\; \hat{\theta}_D - R^*_{.025}\right) &= (.444 - (.080), \;\; .444 - (-.100)) \\ &= (.364, \;\; .544) \end{aligned}$$

Example of Studentized Bootstrap C.I. for θ

Assuming that an exponential model for the light bulb data is appropriate, a studentized bootstrap C.I. for the parameter

$\theta = P[T \geq 1200] = e^{-1200/\beta}$

could be calculated provided we are able to obtain an approximation to the variance of $\hat{\theta} = e^{-1200/\hat{\beta}}$.

If the standard deviation of $\bar{T}$ is small, then a 1-term Taylor expansion of $g(y) = e^{-1200/y}$ about μ_T would be appropriate. This yields

$$\hat{\theta} = g(\bar{T}) \approx g(\mu_T) + g^{'}(\mu_T)(\bar{T} - \mu_T) = e^{-1200/\mu_T} + \left(\frac{1200}{\mu_T^2}\right) e^{-1200/\mu_T} (\bar{T} - \mu_T)$$

From which we obtain

$$V = Var(\hat{\theta}) \approx \left(\frac{1200}{\mu_T^2}\right)^2 e^{-2400/\mu_T} Var\left(\bar{T} - \mu_T\right) \Rightarrow \hat{V} = \left(\frac{1200}{\bar{T}^2}\right)^2 e^{-\frac{2400}{\bar{T}}} \frac{S^2}{n}$$

where $\hat{\mu}_T = \bar{T}$ and S^2 is the sample variance of the $T_i's$.

We can now bootstrap $Z = (\hat{\theta}_n - \hat{\theta}_D)/\sqrt{V}$.

From each bootstrap sample, compute $\bar{T}^*, S^*, V^*, Z^*$ and

then obtain the upper and lower sample $\frac{\alpha}{2}$ percentiles of Z^*.

Using the following program **boot4,ci.s**, we can obtain a 95% C.I. for θ

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,
      2420,71,2195,1825,159,1577,3725,884,1014,965)
n = length(x)
mn= mean(x)
thhat = exp(-1200/mn)

S2 = var(x)

Vest= ((1200/(mn**2))**2)*(exp(-2400/mn))*(S2/n)

R = 9999
z = numeric(R)
z = rep(0,times =R)
for (i in 1:R)
{t= sample(x,replace=T)

V= ((1200/(mean(t)**2))**2)*(exp(-2400/mean(t)))*(var(t)/n)

z[i] = (exp(-1200/mean(t))-thhat)/sqrt(V)}

z = sort(z)

L = z[250]

U = z[9750]

thL = thhat-U*sqrt(Vest)

thU = thhat-L*sqrt(Vest)
```

From the output, we have

$\bar{T} = 1478.8; \quad \hat{V}_D = .002246486; \quad \hat{\theta} = 1 - \hat{F}(1200) = e^{-1200/\bar{T}} = e^{-1200/1478.8} = .444$

$Z^*_{((R+1)(1-\frac{\alpha}{2})} = Z^*_{250} = -2.030352$

$Z^*_{((R+1)(\frac{\alpha}{2})} = Z^*_{9750} = 2.193155$

Then, our approximate 95% C.I. for θ is

$$\left(\hat{\theta}_n - \sqrt{\widehat{V}_D} Z^*_{((R+1)(1-\frac{\alpha}{2})}, \quad \hat{\theta} - \sqrt{\widehat{V}_D} Z^*_{((R+1)(\frac{\alpha}{2})} \quad \right) =$$

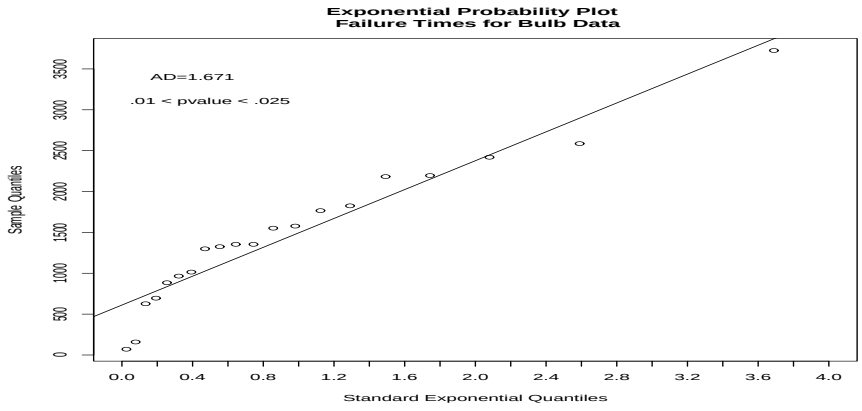
$$\left(.444 - (\sqrt{.002246486})(2.193155), \quad .444 - (\sqrt{.002246486})(-2.030352)\right) = (.340, \quad .540)$$

We have calculated 5 different confidence intervals for the parameter $\theta = P[T > 1200]$:

Method	Confidence Interval
Clapper-Pearson(with θ a population proportion)	(.408, .846)
Wilson C.I.(with θ a population proportion)	(.433, .819)
Basic Bootstrap (with θ a population proportion)	(.450, .850)
Basic Bootstrap (with $\theta = 1 - e^{-1200/\beta}$)	(.361, .547)
Studentized Bootstrap (with $\theta = 1 - e^{-1200/\beta}$)	(.340, .540)

Why do you think there is such a difference between the intervals obtained by treating the parameter θ as a proportion in comparison to the intervals obtained by treating the parameter θ as a function of β?

For the exponential distribution, the two representation of θ are identical, why are the two C.I.'s for θ so different?



Parametric Bootstrap

In some cases a parametric model may be known for the data but the sampling distribution of some statistics is of such a complex nature that the sampling distribution cannot be determined mathematically and/or the sample size n is too small to invoke asymptotic properties of the sampling distribution. In these types of situations, a **parametric bootstrap** procedure can be implemented.

Example Suppose we are studying the effectiveness of a new insecticide for controlling the damage on various crops due to infestation of fire ants. The treatments are applied to one acre plots of land planted with the crops of interest. The response variable is the amount of useful crop harvested from the field.

The parameter of interest is the coefficient of variation: $\theta = \frac{\sigma}{\mu}$ since the crops are very different in terms of the mean yield.

From many previous studies it is known that the crop yields follow a log-normal distribution but the parameters, (μ, σ), will be unknown.

The maximum likelihood estimator of θ is denoted as $\hat{\theta} = \frac{\hat{\sigma}}{\hat{\mu}}$.

The sampling distribution of $\hat{\theta}$ is known asymptotically but there is no explicit expression for the sampling distribution for small sample sizes. A C.I. for θ can be estimated using the parametric bootstrap.

In our previous discussion the bootstrap samples were obtained by random sampling with replacement from the n original data values, that is, sampling from a population having cdf $\widehat{F}$, the edf obtained from the original data.

In the case of the parametric bootstrap, the unknown parameters in the cdf are estimated using the MLEs computed from the original data set. Samples are then simulated from the parametric cdf with the unknown parameters replaced with their MLEs. We thus run the simulation B times yielding

B bootstrap samples $(Y_1^*, Y_2^*, \ldots, Y_n^*)_i \quad \text{for} \quad i = 1, 2, \ldots, B.$

The modification from the procedures used previously to determine a bootstrap C.I. for θ is that when a parametric model was unknown we were generating the B bootstrap samples of n observations from the original data $Y_1, \ldots, Y_n$. In the parametric bootstrap we are generating the B bootstrap samples from a log-normal distribution with the values of the parameters μ and σ determined by the MLEs computed from the original data.

START Wednes 11/3/21

Example of Parametric Bootstrap

Suppose we have data from a logistic distribution with (location, scale) parameters $(\beta,\ \gamma)$ unknown. A $100(1-\alpha)\%$ is desired on the coefficient of variation, $\theta = \frac{\sigma}{\mu} = \frac{\pi\gamma/\sqrt{3}}{\beta}$.

The sampling distribution of the MLE for θ is unknown for small to moderate sized samples. In particular, n=25 would be too small to apply the asymptotic results for the functions of MLE's.

Let $X_1, X_2, \ldots, X_{25}$ be the data from which we compute the MLE's $\hat{\beta}$ and $\hat{\gamma}$.

Next we use R to generate $B = 10,000$ samples of size $n = 25$ from a logistic distribution with parameters $\hat{\beta}$ and $\hat{\gamma}$.

$$\text{rlogis}(25,\hat{\beta},\hat{\gamma}) \Rightarrow\ (X_1^*, X_2^*, \ldots, X_{25}^*)_1\ \Rightarrow\ (\hat{\beta}_1^*,\ \hat{\gamma}_1^*)\ \Rightarrow\ \hat{\theta}_1^* = \frac{\pi\hat{\gamma}_1^*/\sqrt{3}}{\hat{\beta}_1^*}$$

$$\text{rlogis}(25,\hat{\beta},\hat{\gamma}) \Rightarrow\ (X_2^*, X_2^*, \ldots, X_{25}^*)_2\ \Rightarrow\ (\hat{\beta}_2^*,\ \hat{\gamma}_2^*)\ \Rightarrow\ \hat{\theta}_1^* = \frac{\pi\hat{\gamma}_2^*/\sqrt{3}}{\hat{\beta}_2^*}$$

$$\vdots$$

$$\text{rlogis}(25,\hat{\beta},\hat{\gamma}) \Rightarrow\ (X_1^*, X_2^*, \ldots, X_{25}^*)_B\ \Rightarrow\ (\hat{\beta}_B^*,\ \hat{\gamma}_B^*)\ \Rightarrow\ \hat{\theta}_1^* = \frac{\pi\hat{\gamma}_B^*/\sqrt{3}}{\hat{\beta}_B^*}$$

From the B estimates of θ: $\hat{\theta}_1,\ \hat{\theta}_2,\ \ldots, \hat{\theta}_B$ we could then construct the C.I. on θ using a bootstrap confidence interval.

Suppose we have 25 observations from a logistic distribution with (Location, Scale) = $(\beta,\ \gamma)$ unknown

```
Data:  16.20  9.37 25.80  9.55 12.86 15.34 18.08 10.76 14.92  9.75 17.10 13.97
       15.08  9.24 11.99 13.60  8.16 12.82 12.89 13.59 16.23 14.19  9.03  9.58  13.68
```

Using the following Rcode: **parambootlogistic,ci.R**, we will obtain a 95% c.i. for the coefficient of determination, CV = $\frac{\sigma}{\mu}$

Because CV is a scale type parameter, an appropriate pivot would be $R = \frac{\widehat{CV}}{CV}$.

First find $R_{\alpha/2}$ and $R_{1-\alpha/2}$ such that $P[R_{\alpha/2} \le R \le R_{1-\alpha/2}] \approx 1-\alpha$.

Obtain estimates by simulating B bootstrap samples of size n and computing

$R_1^*,\ R_2^*,\ \ldots,\ R_B^*$, where $R_i^* = \frac{\widehat{CV}_i^*}{\widehat{CV}_D}$ and

$\widehat{CV}_D$ is the value of CV computed from the original data.

The approximate $100(1-\alpha)$ C.I. for CV would be

$$\left(\frac{\widehat{CV}_D}{R_{((B+1)(1-\alpha/2))}},\ \frac{\widehat{CV}_D}{R_{((B+1)(\alpha/2))}}\right)$$

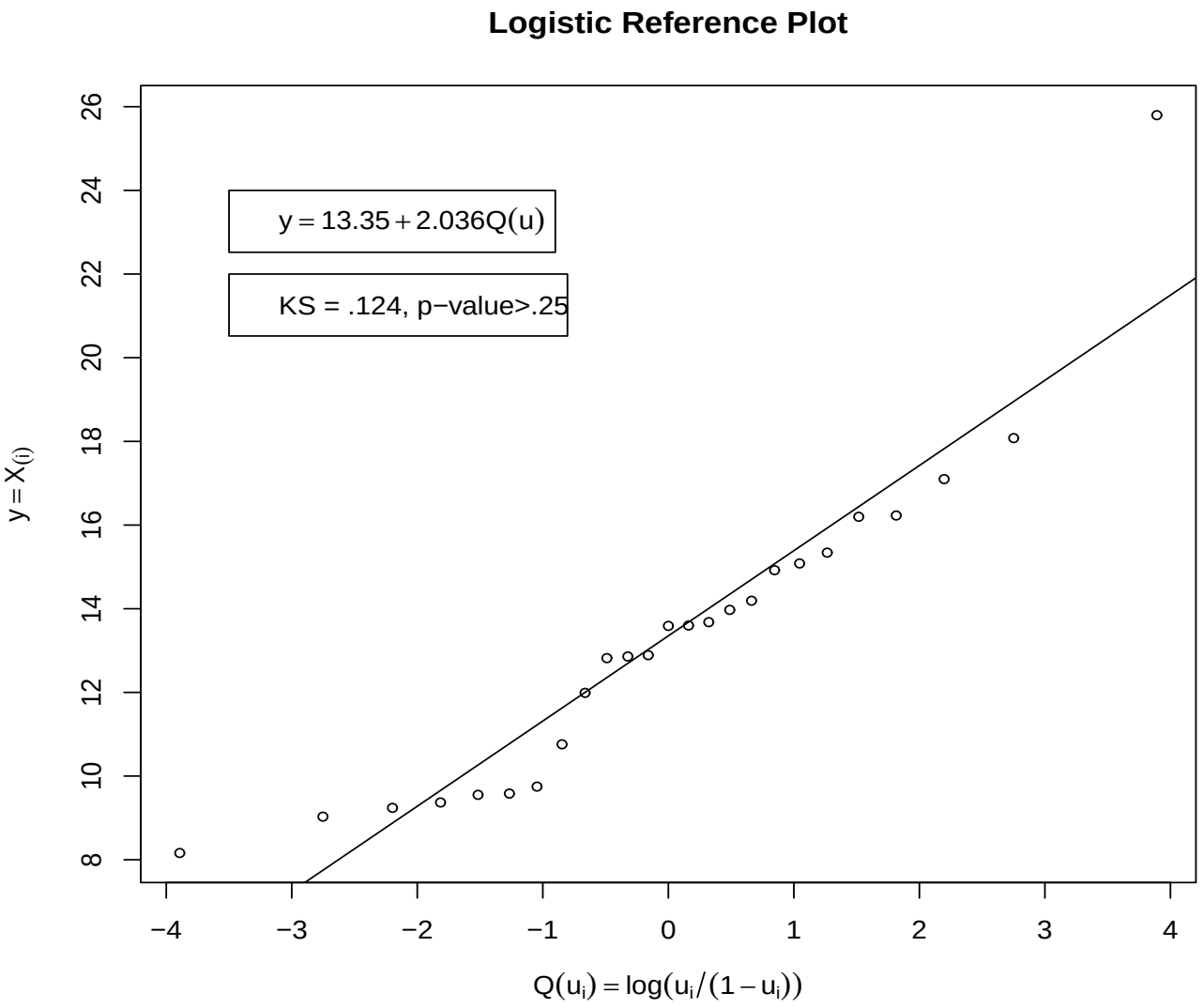
The logistic distribution has location and scale parameters $(\beta,\ \gamma)$ and quantile function:

$$Q(u) = \beta + \gamma log(u/(1-u))$$

thus the quantile function for the standard member is $(\beta = 0,\ \gamma = 1)$: $Q(u) = log(u/(1-u))$

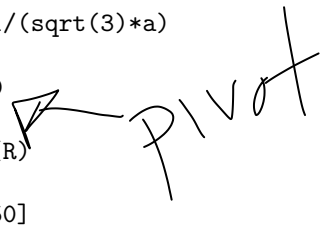
The Logistic Reference Plot for the 25 data values is given below along with the approximate K-S test yielding a p-value from the R-code

```
library(MASS)
mleestD = coef(fitdistr(x,"logistic"))
aD = mleestD[1]
bD = mleestD[2]
ks.test(x,"plogis",aD,bD)
D = 0.12448, p-value = 0.7888
```



parambootlogistic,ci.R

```
library(MASS)

x = c(16.20,9.37,25.80,9.55,12.86,15.34,18.08,10.76,14.92,9.75,17.10,13.97,
15.08,9.24,11.99,13.60,8.16,12.82,12.89,13.59,16.23,14.19,9.03,9.58,13.68)

#obtain MLE of the a=location and b=scale parameters in logistic model

mleestD = coef(fitdistr(x,"logistic"))
aD = mleestD[1]
bD = mleestD[2]
cvD = bD*pi/(sqrt(3)*aD)

n =  length(x)
B  =  9999
W  =  matrix(0,B,n)
cv =  numeric(B)
cv =  rep(0,B)
a =  numeric(B)
a =  rep(0,B)
b =  numeric(B)
b =  rep(0,B)
mleest =  matrix(0,B,2)

{
for (i in 1:B)
W[i,]  =  rlogis(n,aD,bD)
}

{
for (i in 1:B)
mleest[i,] = coef(fitdistr(W[i,],"logistic"))
}

a = mleest[,1]

b = mleest[,2]

cv =  b*pi/(sqrt(3)*a)

R = cv/cvD

R =  sort(R)

L  =  R[250]

U  =  R[9750]

ci = c(cvD/U, cvD/L)

ci
```



The point estimator for the CV is 0.2742 and a 95% C.I. for the CV is (0.202, 0.408)

Skip rest of HD. End @ 12:27- cn 11/5 Lecture

Problems Encountered in Using Bootstrap Procedures

Great care must be taken in using bootstrap procedures, both nonparametric and parametric bootstrap methods. Many inappropriate applications of bootstrap procedures are conducted due to the attitude that there are no assumptions necessary in order to apply a bootstrap procedure. This is far from the truth.

An excellent discussion of some of these problems can be found in Hinkley and Davison's book, *Bootstrap Methods and Their Applications*

A quote from their book, "The error in resampling methods is generally a combination of statistical error and simulation error."

The basic bootstrap idea is to approximate a quantity $c(F)$, where F is the population/process cdf, by the estimate $c(\widehat{F})$, where $\widehat{F}$ is either a parametric or nonparametric estimate of F based on the observed data, $Y_1, \ldots, Y_n$.

The statistical error in the bootstrap procedure is the difference between $c(\widehat{F})$ and $c(F)$. The problem that may arise is that the sampling distribution of the elected pivot may depend on unknown parameters. They describe procedures for checking empirically if this problem exists and remedies when it does. The more basic consideration is that the observed data does not adequately represent the population either distributionally or with respect to outliers.

The simulation errors can generally removed by selecting a large value for B. The only question is the size of B to guarantee a specified degree of accuracy. They discuss the issue of how to select the value of B in a number of situations.

Various conditions are given under which the bootstrap procedure will produce a mathematically justifiable estimate of the sampling distribution of the statistic of interest.

The following example is provided to demonstrate that bootstrap procedures are not universally applicable.

Example Suppose a random sample $Y_1, \ldots, Y_n$ is selected from a uniform on $(0, \theta)$ distribution. The MLE of θ is $\hat{\theta} = Y_{(n)}$.

The pivot for obtaining a C.I. on θ is $PV_n = \frac{n(\theta - \hat{\theta})}{\theta}$.

It is easily shown that the asymptotic distribution of PV_n is a standard exponential distribution, $F(y) = 1 - e^{-y}$

This would suggest that in the bootstrap procedure the pivot should be computed as $PV_n^* = \frac{n(\hat{\theta}_D - \hat{\theta}^*)}{\hat{\theta}_D}$, where $\hat{\theta}_D$ is $\hat{\theta}$ computed from the original data and $\hat{\theta}^*$ is $\hat{\theta}$ computed from each bootstrap sample.

We can easily determine: $P[PV_n^* = 0|\widehat{F}] = P[\hat{\theta}^* = \hat{\theta}_D|\widehat{F}] = 1 - (1 - \frac{1}{n})^n$

Thus, $lim_{n\to\infty} Pr[PV_n^* = 0|\widehat{F}] = 1 - e^{-1}$

Therefore, the limiting distribution PV_n^* cannot be a standard exponential distribution.

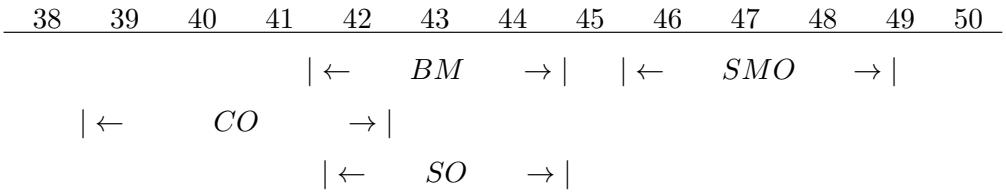
The above illustrates care must be taken in making sure the required regularity conditions prior to using a bootstrap procedure.

Simultaneous Confidence Intervals

In many studies or experiments, there will be more than one population/process of interest to the researcher. Consider the following example. A large percentage of the dietary energy in the bodies of infants is provided by lipids. Lipids are a class of hydrocarbon-containing organic compounds. The following data on total polyunsaturated fats(%) was reported for infants who were randomized to four different feeding regimens: BM (breast milk), CO corn-oil-based formula, SO (soy-oil-based formula), or SMO (soy-marine-oil based formula).

Regimen	n	$\bar{X}$	S	95% C.I. on μ_i
BM	18	43.0	3.5	(41.27, 44.73)
CO	13	40.4	3.3	(38.42, 42.38)
SO	17	43.1	3.2	(41.46, 44.74)
SMO	14	47.1	3.2	(45.27, 48.93)

The researcher determined that there was significant evidence that the data from the four regimen was from a normal distribution. He then proceeded to construct 95% confidence intervals on the mean percentage of polyunsaturated fat in the infants for each of the four regimens, as reported in the above table.



Based on the above confidence intervals, the researcher concluded that there was significant evidence at the 95% level that SMO produced a mean percentage of polyunsaturated fat in the infants that was significantly higher than the other regimens.

A statistician reviewed the results and told the researcher that he may be incorrect in his conclusions. She stated that the above confidence intervals are **individual** confidence intervals and that the researcher needed **simultaneous** confidence intervals in order to reach his stated conclusions.

The researcher was attempting to make a statement about all k (k=4, in the example) parameters: $\theta_1, \theta_2, \ldots, \theta_k$ simultaneously not as separate estimates of the individual parameters. That is, the above intervals were derived using an inference procedure based on

$$\text{Individual } 100(1-\alpha)\% \text{ C.I.'s:} \qquad P\left[\theta_i \epsilon \left(L_i, U_i\right)\right] = 1-\alpha \qquad \text{for all } \; i = 1, \ldots, k$$

In order to reach the conclusion stated by the researcher it is necessary to construct confidence intervals based on the following probability statement:

$$\text{Simultaneous } 100(1-\alpha)\% \text{ C.I.'s:} \qquad P\left[\theta_i \epsilon \left(L_i, U_i\right) \quad \text{for all } \; i = 1, \ldots, k\right] = 1-\alpha$$

Suppose $100(1-\alpha)\%$ individual C.I.s are constructed on k parameters: $\theta_1, \ldots, \theta_k$.

What is the simultaneous coverage probability, γ, of these k intervals? That is, compute

$$\gamma = P\left[\theta_i \epsilon\, (L_i, U_i) \quad \text{for} \quad i = 1, \ldots, k\right]$$

Let A_i be the event $\{\theta_i \epsilon\, (L_i, U_i)\}$

If the intervals (L_i, U_i) are independent random variables for $i = 1, \ldots, k$ with coverage probability $1-\alpha_i$ then

$$\gamma = P\left[A_1 \cap A_2 \cap \cdots \cap A_k\right] = \prod_{i=1}^{k} P\left[A_i\right] = \prod_{i=1}^{k} (1-\alpha_i) = (1-\alpha)^k \quad \text{if} \quad \alpha_i \equiv \alpha$$

The following table will demonstrate that the simultaneous coverage probability can be considerably smaller than the individual coverage probabilities.

k	1	2	3	4	5	$\cdots$	13	14
$1-\alpha$	0.95	0.95	0.95	0.95	0.95	$\cdots$	0.95	0.95
γ	0.95	0.903	0.857	0.815	0.774	$\cdots$	0.513	0.488

After examining the above table, the researcher then asked the statistician how he could construct simultaneous C.I.s for the parameters. The statistician suggested the following procedure but emphasized that it would only be **valid if the individual C.I.s are independent.**

Determining α to Obtain Specified Value for γ

To obtain a $100\gamma\%$ simultaneous C.I. for k parameters using a procedure which generates $100(1-\alpha)\%$ independent individual C.I.s for k parameters, just set the value of $\alpha/2$ at

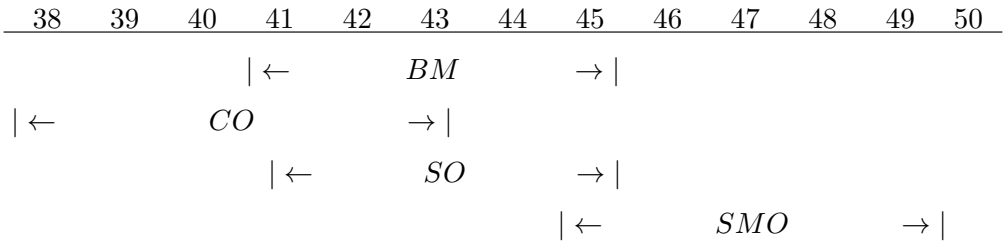
$$1-\alpha = \gamma^{1/k} \;\Rightarrow\; \alpha/2 = .5\left(1-\gamma^{1/k}\right)$$

k	1	2	3	4	5	$\cdots$	13	14
γ	0.9500	0.9500	0.9500	0.9500	0.9500	$\cdots$	0.9500	0.9500
$1-\alpha$	0.9500	0.9747	0.9830	0.9873	0.9898	$\cdots$	0.9960	0.9963
$\alpha/2$	.0250	0.01266	0.008476	0.006371	0.005103	$\cdots$	0.001969	0.001829

Example:

In the lipid example, the researcher wanted a 95% simultaneous C.I. on the $k = 4$ means. Therefore, he would need to use $t_{.00635}$ in the C.I. construction: $\bar{X} \pm t_{\frac{\alpha}{2}} S/\sqrt{n}$ in place of $t_{.025}$, as was done in constructing 4 individual C.I.s. The impact of this replacement can be seen in the following table:

Regimen	n	$\bar{X}$	S	95% Indiv. C.I. on μ_i	95% Simul. C.I. on $\mu_i s$	$t_{.025}$	$t_{.00635}$
BM	18	43.0	3.5	(41.27, 44.73)	(40.72, 45.28)	2.110	2.785
CO	13	40.4	3.3	(38.42, 42.38)	(37.76, 43.04)	2.179	2.926
SO	17	43.1	3.2	(41.46, 44.74)	(40.94, 45.26)	2.120	2.805
SMO	14	47.1	3.2	(45.27, 48.93)	(44.66, 49.54)	2.160	2.888



The widths of the individual C.I.s are considerably wider than the standard 95% C.I. in order to achieve a 95% simultaneous C.I. for the k C.I.s.

The four C.I.'s now overlap and thus it would not appear that the means for the four regimens are different.

In STAT 642, we will learn procedures to perform the above inferences in a more efficient manner using multiple comparison and ANOVA techniques.

In many instances, the individual C.I.s are not independent. How do we construct simultaneous C.I.s?

In some settings there will be more efficient methods for deriving simultaneous C.I.s. than the method described below.

However, the Bonferroni procedure is widely used in those cases where more efficient procedures do not exist.

Bonferroni Simultaneous C.I.s for k parameters

Suppose we have $100(1-\alpha)\%$ C.I.s on k parameters $\theta_1, \ldots, \theta_k$. However, the C.I.s are not necessarily independent. A lower bound on the simultaneous coverage probability of the k C.I.s is obtained using the Bonferroni inequality:

Let A_i be the event $\{\theta_i \epsilon\, (L_i, U_i)\}$

$$\begin{aligned}
\gamma &= P\left[\theta_i \epsilon\, (L_i, U_i) \quad \text{for} \quad i = 1, \ldots, k\right] \\
&= P\left[\cap_{i=1}^{k} A_k\right] \\
&= 1 - P\left[\cup_{i=1}^{k} A_k^c\right] \\
&\geq 1 - \sum_{i=1}^{k} P\left[A_i^c\right] = 1 - \sum_{i=1}^{k} \alpha_i = 1 - k\alpha \quad \text{if} \quad \alpha_i \equiv \alpha
\end{aligned}$$

Thus, to set the level of the k individual C.I.s, let

$$\alpha = \frac{1-\gamma}{k}$$

Simultaneous inferences and C.I.s obtained using the Bonferroni inequality are often inefficient because the actual coverage probability may be much larger than the nominal coverage γ.

The following table illustrates the coverage probability needed on k individual C.I.s to obtain a $100\gamma\%$ simultaneous C.I.s for k parameters using the Bonferroni inequality:

$$1 - \alpha = 1 - \frac{1-\gamma}{k} \;\Rightarrow\; \alpha/2 = .5(1-\gamma)/k$$

k	1	2	3	4	5	$\cdots$	13	14
γ	0.9500	0.9500	0.9500	0.9500	0.9500	$\cdots$	0.9500	0.9500
$1-\alpha$	0.9500	0.9750	0.9833	0.9875	0.9900	$\cdots$	0.9962	0.9964
$\alpha/2$	0.02500	0.0125	0.008333	0.006250	0.005000	$\cdots$	0.001923	0.001786

Note, that the width of the individual C.I.s will be somewhat wider than a 0.95 C.I. in order to achieve a 95% simultaneous coverage probability.

For example, if $k = 4$, then to obtain four C.I.s having simultaneous Coverage Probability of .95, it would be necessary to construct four individual CIs have coverage probability of .9875, that is, we would use $\alpha/2 = .00625$ in place of $\alpha/2 = .025$ if we were not concerned about simultaneous coverage. Also, note that $\alpha/2$ has decreased from .006371 to .00625 due to the lack of independence between the four C.I.'s. This will result in slightly wider intervals.

Wald C.I. based on MLE's

When we obtain MLE of parameters, the R or SAS output will provide asymptotic estimates of the standard errors of the MLE's.

For example, if $\hat{\theta}$ is the MLE of θ, then we can also obtain the asymptotic standard error of $\hat{\theta}$, $\widehat{SE}(\hat{\theta})$. Using these values, an asymptotic approximate $100(1-\alpha)$ for θ is given by

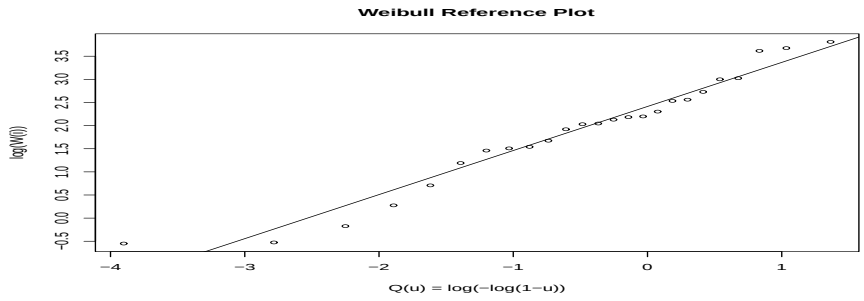
$$\hat{\theta} \pm Z_{\alpha/2} \cdot \widehat{SE}(\hat{\theta}) \;=\; (\hat{\theta} - Z_{\alpha/2} \cdot \widehat{SE}(\hat{\theta}),\;\; \hat{\theta} + Z_{\alpha/2} \cdot \widehat{SE}(\hat{\theta}))$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the $N(0,1)$ distribution.

Example: Suppose we have a random sample of n=23 ball bearings and observe the number of revolutions to failure for the 23 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
```

The researcher states that from previous studies that the Weibull distribution was a good approximation to cdf of the r.v. R, Revolutions to Failure. A Weibull Probability plot and the Anderson-Darling GOF measure were applied to the data. The results are given here:



From the R output we have ADM = 0.3413 which implies from Table 5 in Handout 9 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. Next, we obtain the MLE's of the parameters, γ and α:

```
*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 $\hat{\gamma} = 2.1012$ with estimated standard error $\widehat{SE}(\hat{\gamma}) = .3286$

Estimate of α is $\hat{\alpha} = 81.8324$ with estimated standard error $\widehat{SE}(\hat{\alpha}) = 8.5971$

We can then construct approximate 95% C.I. for the two parameters:

$$95\% \text{ C.I. for } \gamma: \quad \hat{\gamma} \pm 1.96 \cdot \widehat{SE}(\hat{\gamma}) = 2.1012 \pm 1.96 \cdot .3286 = (1.457,\ 2.745)$$

$$95\% \text{ C.I. for } \alpha: \quad \hat{\gamma} \pm 1.96 \cdot \widehat{SE}(\hat{\alpha}) = 81.8324 \pm 1.96 \cdot 8.5971 = (73.235,\ 90.430)$$

The two intervals are fairly wide reflecting the small sample size, n=23.

We can construct similar intervals for the mean and median which are outputted from R when we use the Kaplan-Meier estimates of the survival function.

Recall the example from Handout 7:

EXAMPLE: In an experiment to determine the strength of a braided cord after weathering, the strengths of 48 pieces of cord that had been weathered for a specified length of time were investigated. The company wanted to estimate the probability that the cord would have strength of at least 53, that is, estimate S(53). Seven cords were damaged during the study which resulted in a decrease in their strength. Therefore, the study produced right censored strength values. The strengths of the remaining 41 cords were determined as shown below:

```
36.3  52.4  54.8  57.1  60.7  41.7  52.6  54.8  57.3
43.9  52.7  55.1  57.7  49.4  53.1  55.4  57.8
50.1  53.6  55.9  58.1  50.8  53.6  56.0  58.9
51.9  53.9  56.1  59.0  52.1  53.9  56.5  59.1
52.3  54.1  56.9  59.6  52.3  54.6  57.1  60.4
```

The 7 censored strength values from the damaged cords are given next:

```
26.8  29.6  33.4  35.0  40.0  41.9 42.5
```

The true strength values of the 7 cords, T_i are unobservable but we know $T_i > Y_i$, where Y_i are the observed values.

An analysis of the above data will be conducted using the following R code:

```
 library(MASS)
 library(survival)


st = c(36.3,  52.4,  54.8,  57.1,  60.7,  41.7,  52.6,  54.8, 57.3,
       43.9,  52.7,  55.1,  57.7,  49.4,  53.1,  55.4,  57.8, 50.1,
       53.6,  55.9,  58.1,  50.8,  53.6,  56.0,  58.9,  51.9, 53.9,
       56.1,  59.0,  52.1,  53.9,  56.5,  59.1,  52.3,  54.1, 56.9,
       59.6,  52.3,  54.6,  57.1,  60.4,
       26.8,  29.6,  33.4,  35.0,  40.0,  41.9,  42.5)

stcens = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           0,0,0,0,0,0,0)

Surv(st, stcens)

cords.surv <- survfit(Surv(st, stcens) ~ 1,conf.type="log-log")
summary(cords.surv)
print(cords.surv,print.rmean=TRUE)
```

Estimators from R Code:

```
records   n.max    n.start  events  *rmean  *se(rmean)  median  0.95LCL    0.95UCL
 48.000  48.000    48.000   41.000   54.182   0.723     54.800   53.100     56.100
* restricted mean with upper limit =  60.7
```

The above output from R displays the estimated value of the mean

$\hat{\mu} = 54.182$ with estimated standard error $\widehat{SE}(\hat{\mu}) = .723$.

From these values, we can compute an approximate 99% C.I. for the mean, $Z_{.005} = 2.576$:

$$\hat{\mu} \pm 2.576 \cdot \widehat{SE}(\hat{\mu}) \quad = \quad 54.182 \pm 2.576 \cdot .723 \quad = \quad (52.32,\ 56.04)$$

A 99% confidence interval for the median can computed in a similar fashion once we know $\widehat{SE}(\hat{median})$

The width of the 95% C.I. is $(56.1 - 53.1) = 2(1.96)\widehat{SE}(median) \ \Rightarrow \ \widehat{SE}(median) = 0.7653$

The 99% C.I. for the median is

$$54.8 \pm 2.576 \cdot .7653 \quad = \quad (52.83,\ 56.77)$$

.

.

.