MIKE LEDERLE - STATISTICS 641 - ASSIGNMENT 5

• Read Handouts 8 & 9

I. (6 points) Suppose 50 iid observations $Y_1, Y_2, \ldots, Y_{50}$ from a process yield the following normal probability plot.

Using the above plot, describe the shape of the process distribution.

I DON'T HAVE THE FIGS, SO I COULDN"T GENERATE THE GRAPHS

Looks like heavy-tails compared to a normal, like a t with low df.

II. (15 points) A researcher is studying the relative brain weights (brain weight divided by body weight) for 51 species of mammal whose average litter size is less than 2 and for 45 species of mamma whose average litter size is greater than or equal to 2. The researcher was interested in determining what evidence that brain sizes tend to be different for the two groups. (Data from *The Statistical Sleuth* by Fred Ramsey and Daniel Schafer).

RELATIVE BRAIN WEIGHTS - SMALL LITTER SIZE

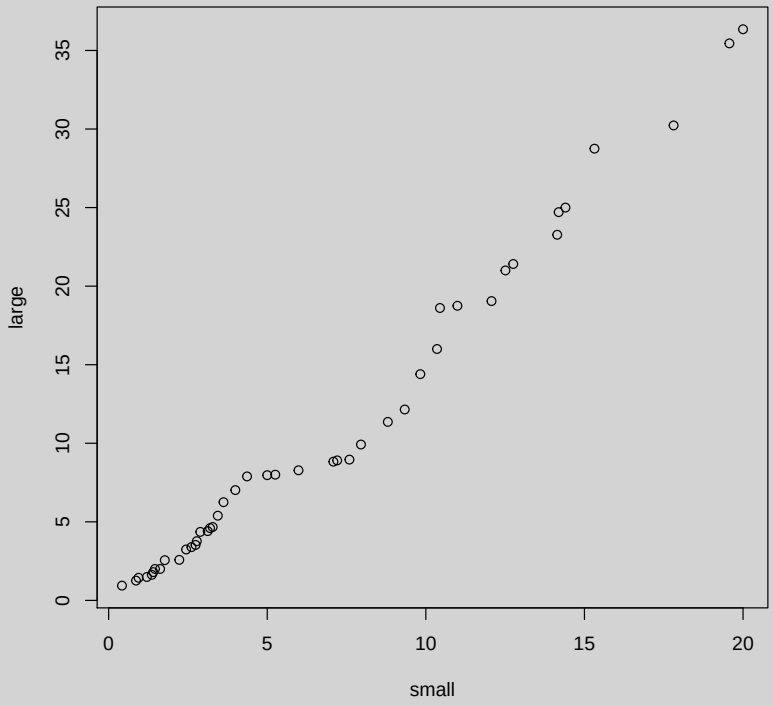
0.42	0.86	0.88	1.11	1.34	1.38	1.42	1.47	1.63
1.73	2.17	2.42	2.48	2.74	2.74	2.79	2.90	3.12
3.18	3.27	3.30	3.61	3.63	4.13	4.40	5.00	5.20
5.59	7.04	7.15	7.25	7.75	8.00	8.84	9.30	9.68
10.32	10.41	10.48	11.29	12.30	12.53	12.69	14.14	14.15
14.27	14.56	15.84	18.55	19.73	20.00			

RELATIVE BRAIN WEIGHTS - LARGE LITTER SIZE

0.94	1.26	1.44	1.49	1.63	1.80	2.00	2.00	2.56
2.58	3.24	3.39	3.53	3.77	4.36	4.41	4.60	4.67
5.39	6.25	7.02	7.89	7.97	8.00	8.28	8.83	8.91
8.96	9.92	11.36	12.15	14.40	16.00	18.61	18.75	19.05
21.00	21.41	23.27	24.71	25.00	28.75	30.23	35.45	36.35

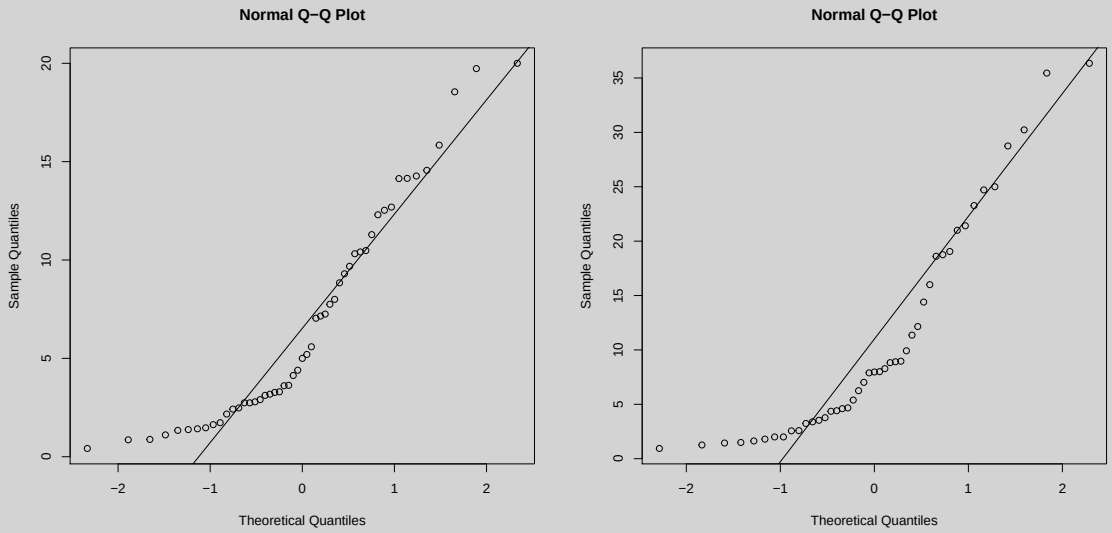
1. Produce a q-q plot of the two data sets.

```
small <- c(0.42, 0.86, 0.88, 1.11, 1.34, 1.38, 1.42, 1.47, 1.63, 1.73, 2.17, 2.42, 2.48, 2.74,
    2.74, 2.79, 2.9, 3.12, 3.18, 3.27, 3.3, 3.61, 3.63, 4.13, 4.4, 5, 5.2, 5.59, 7.04, 7.15,
    7.25, 7.75, 8, 8.84, 9.3, 9.68, 10.32, 10.41, 10.48, 11.29, 12.3, 12.53, 12.69, 14.14,
    14.15, 14.27, 14.56, 15.84, 18.55, 19.73, 20)

large <- c(0.94, 1.26, 1.44, 1.49, 1.63, 1.8, 2, 2, 2.56, 2.58, 3.24, 3.39, 3.53, 3.77, 4.36,
    4.41, 4.6, 4.67, 5.39, 6.25, 7.02, 7.89, 7.97, 8, 8.28, 8.83, 8.91, 8.96, 9.92, 11.36,
    12.15, 14.4, 16, 18.61, 18.75, 19.05, 21, 21.41, 23.27, 24.71, 25, 28.75, 30.23, 35.45,
    36.35)

qqplot(small, large)
```



2. Produce separate normal reference distribution plot.

```
qqnorm(small)
qqline(small)
qqnorm(large)
qqline(large)
```



3. What conclusions can you reach about the effect of the Litter size on Relative Brain weights?

I don't see a straight line or functional relationship in the qqplot, so I don't think we can make any conclusion. Both distributions are right skewed, from the qqnorm plots.

III. (15 Points) A data value Y is an outlier if ˇ $Y < Q(.25) - 1.5 * IQR$ or $Y > Q(.75) + 1.5 * IQR$ Calculate the probability that a randomly selected observation is a outlier if the distribution of Y is

1. exponential

Of course, this is using a single "standard" distribution for each case, when we should leave the parameters in to see the functional dependency; this seemed like fun though.

```
prob.of.outlier <- function(dist) {
    d <- paste("p", dist, sep = "")
    q <- paste("q", dist, sep = "")
    left <- eval(call(q, 0.25))
    right <- eval(call(q, 0.75))
    IQR <- right - left
    lower <- eval(call(d, left - 1.5 * IQR, lower.tail = TRUE))
    upper <- eval(call(d, right + 1.5 * IQR, lower.tail = FALSE))
    list(left.endpoint = lower, right.endpoint = upper)
}
prob.of.outlier("exp")

## $left.endpoint
## [1] 0
##
## $right.endpoint
## [1] 0.04811
```

2. Weibull

Have to modify `prob.of.outlier` since *shape* parameter isn't given a default. Let *shape* be 1; i.e., an exponential, so that we should get answer above.

```
W_prob.of.outlier_W <- function(dist, shape.param = 1) {
    d <- paste("p", dist, sep = "")
    q <- paste("q", dist, sep = "")
    left <- eval(call(q, 0.25, shape = shape.param))
    right <- eval(call(q, 0.75, shape = shape.param))
    IQR <- right - left
    lower <- eval(call(d, left - 1.5 * IQR, shape = shape.param, lower.tail = TRUE))
    upper <- eval(call(d, right + 1.5 * IQR, shape = shape.param, lower.tail = FALSE))
    list(left.endpoint = lower, right.endpoint = upper)
}
W_prob.of.outlier_W("weibull")

## $left.endpoint
## [1] 0
##
## $right.endpoint
## [1] 0.04811
```

Of course, any shape parameter can be substituted:

```
W_prob.of.outlier_W("weibull", 1:5)

## $left.endpoint
## [1] 0.000000 0.000000 0.000000 0.001706 0.005043
##
## $right.endpoint
## [1] 0.0481125 0.0103037 0.0030085 0.0011264 0.0005086
```

3. uniform on (0,1)

```
prob.of.outlier("unif")

## $left.endpoint
## [1] 0
##
## $right.endpoint
## [1] 0
```

4. normal

```
prob.of.outlier("norm")

## $left.endpoint
## [1] 0.003488
##
## $right.endpoint
## [1] 0.003488
```

5. t with df=2

```
T_prob.of.outlier_T <- function(dist, df.param) {
    d <- paste("p", dist, sep = "")
    q <- paste("q", dist, sep = "")
    left <- eval(call(q, 0.25, df = df.param))
    right <- eval(call(q, 0.75, df = df.param))
    IQR <- right - left
    lower <- eval(call(d, left - 1.5 * IQR, df = df.param, lower.tail = TRUE))
    upper <- eval(call(d, right + 1.5 * IQR, df = df.param, lower.tail = FALSE))
    list(left.endpoint = lower, right.endpoint = upper)
}
T_prob.of.outlier_T("t", 2:5)

## $left.endpoint
## [1] 0.04117 0.02751 0.02072 0.01676
##
## $right.endpoint
## [1] 0.04117 0.02751 0.02072 0.01676
```

IV. (10 points) Nylon bars were tested for brittleness. Each of 280 bars was molded under similar conditions and was tested by placing a specified stress at 5 locations on the bar. Assuming that each bar has uniform composition, the number of breaks on a given bar should be binomially distributed with an unknown probability p of breaking. The following table summarizes the outcome of the experiment:

Breaks/Bar	0	1	2	3	4	5	total
Frequency	121	110	38	7	3	1	280

Use a GOF test to evaluate whether the data appears to be from a binomial model.

We need an estimate of θ:

```
(theta.hat <- as.numeric(crossprod(c(0, 1, 2, 3, 4, 5), c(121, 110, 38, 7, 3, 1))/(280 * 5)))

## [1] 0.16
```

Now calculate $\hat{p}_i$, E and the ratio:

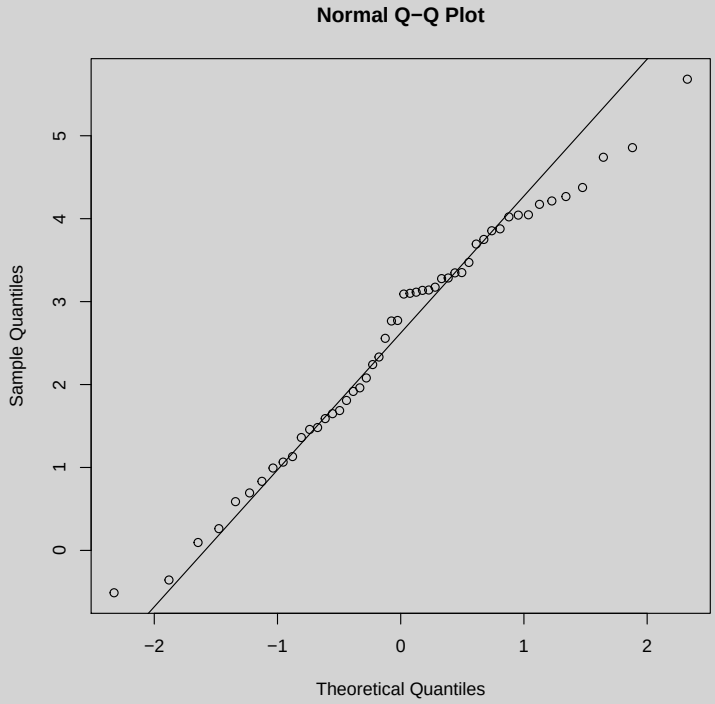
```
(p.hat <- dbinom(0:5, size = 5, prob = theta.hat))

## [1] 0.4182119 0.3982971 0.1517322 0.0289014 0.0027525 0.0001049

(E <- 280 * p.hat)

## [1] 117.09934 111.52318  42.48502   8.09239   0.77070   0.02936

O <- c(121, 110, 38, 7, 3, 1)
(ratio <- sum((O - E)^2/E))

## [1] 39.31

df <- 6 - 1 - 1
(p.val <- 1 - pchisq(ratio, df))

## [1] 6.014e-08
```

V. (16 points) A major problem in the Gulf of Mexico is the excessive capture of game fish by shrimpers. A random sample of the catch of 50 shrimpers yield the following data concerning the catch per unit effort (CPUE) of Red Snappers, a highly sought game fish. Let C_i be the CPUE for the ith shrimper. The data, $C_1, C_2, \ldots, C_{50}$ is given next.

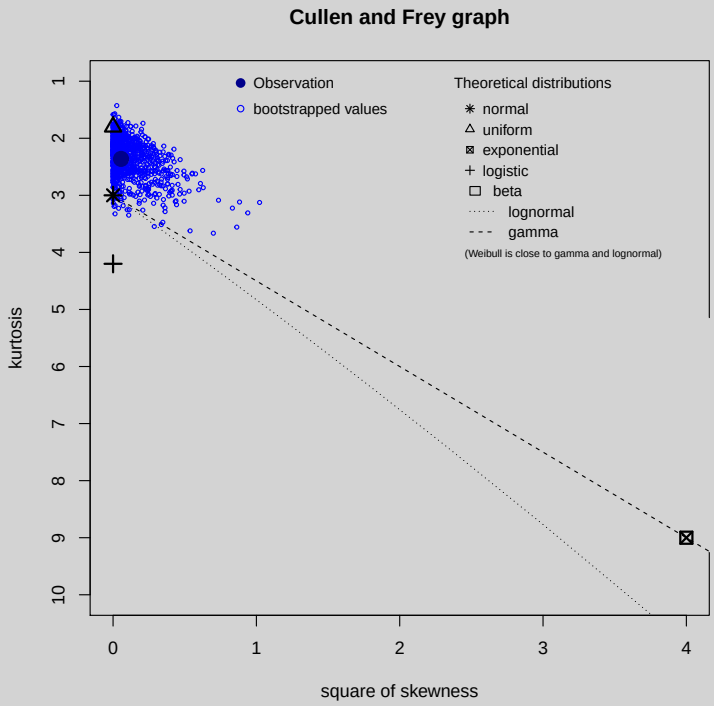
0.6	0.7	1.1	1.3	1.8	2.0	2.3	2.7	2.9	3.1
3.9	4.3	4.4	4.9	5.2	5.4	6.1	6.8	7.1	8.0
9.4	10.3	12.9	15.9	16.0	22.0	22.2	22.5	23.0	23.1
23.9	26.5	26.7	28.4	28.5	32.2	40.2	42.5	47.2	48.3
55.8	57.0	57.2	64.9	67.6	71.3	79.5	114.5	128.6	293.5

1. CPUE data is often modelled using a Log-Normal distribution. Does the above data appear to be from a Log-Normal distribution? Explain your answer with both a normal reference distribution plot and a GOF test.

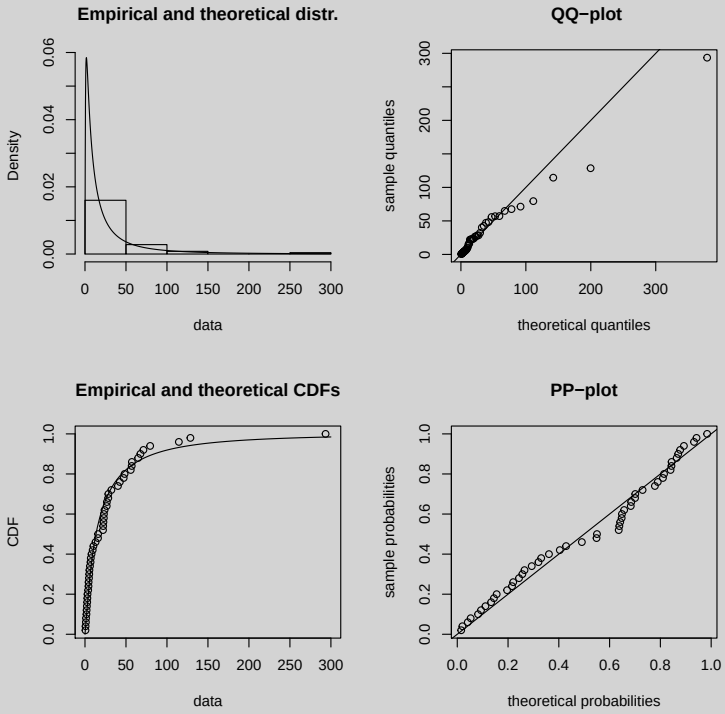
```
CPUE <- c(0.6, 0.7, 1.1, 1.3, 1.8, 2, 2.3, 2.7, 2.9, 3.1, 3.9, 4.3, 4.4, 4.9, 5.2, 5.4, 6.1,
    6.8, 7.1, 8, 9.4, 10.3, 12.9, 15.9, 16, 22, 22.2, 22.5, 23, 23.1, 23.9, 26.5, 26.7, 28.4,
    28.5, 32.2, 40.2, 42.5, 47.2, 48.3, 55.8, 57, 57.2, 64.9, 67.6, 71.3, 79.5, 114.5, 128.6,
    293.5)
qqnorm(log(CPUE))
qqline(log(CPUE))
```




```
library(fitdistrplus)
descdist(log(CPUE), boot = 1000)

## summary statistics
## ------
## min:  -0.5108   max:  5.682
## median:  2.932
## mean:  2.591
## estimated sd:  1.453
## estimated skewness:  -0.236
## estimated kurtosis:  2.362
```




```
f <- fitdist(CPUE, "lnorm")
plot(f)
```



```
gofstat(f, print.test = TRUE)

## Kolmogorov-Smirnov statistic:  0.136
## Kolmogorov-Smirnov test:  not rejected
##    The result of this test may be too conservative as it
##    assumes that the distribution parameters are known
## Cramer-von Mises statistic:  0.09077
## Cramer-von Mises test:  not rejected
## Anderson-Darling statistic:  0.4774
## Anderson-Darling test:  not rejected
```

2. Use the Box-Cox transformation of the CPUE data to determine the most appropriate power transformation to transform the CPUE distribution to Normality. How does the fit from the Box-Cox transformation compare to the fit for the log transformation?

```
bc <- function(y, theta) {
    if (theta == 0)
        return(log(y)) else return((y^theta - 1)/theta)
}
theta <- seq(0, 3, by = 0.25)
orig.mat <- matrix(rep(CPUE, length(theta)), nrow = length(theta), byrow = TRUE)
trans.mat <- apply(orig.mat, 1, bc, theta)

## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
## Warning:  the condition has length > 1 and only the first element will be used
```

VI. (20 Points) A random sample of 500 data values are selected from four separate processes having cdf's, F_1, F_2, F_3, F_4. The plot of the sample quantile versus a standard normal quantile for each of the four samples is given below. For each of these plots, **SELECT ONE** of the following distributions to describe the pdf which generated the data. Hint: make sure to take into consideration the size of values associated with each distribution.

(A) Cauchy($\theta_1 = .35, \theta_2 = 1$)

(B) t with df=35

(C) Logistic($\theta_1 = .35, \theta_2 = 1$)

(D) Beta($\alpha = 2, \beta = 6$)

(E) Uniform(0,.7)

(F) Normal($\mu = .35, \sigma = 1$)

(G) Exponential($\beta = 80$)

(H) Weibull($\gamma = 0.7, \beta = 20$)

(I) Gamma($\alpha = 1.2, \beta = 25$)

(J) Mixture of 90% Normal$(10, 1)$ & 10% Normal$(30, (3)^2)$

(K) Mixture of 10% Normal$(10, 1)$ & 90% Normal$(30, (1)^2)$

(L) Normal($\mu = 20, \sigma = 1$)

Plot 1 _____

Plot 2 _____

Plot 3 _____

Plot 4 _____

VII. (18 points) Multiple Choice Questions Select the letter of the **BEST** answer. Justify your answer in 20 words or less.

1. The Anderson-Darling (AD) GOF statistic is prefered to the Kolmogorov-Smirnov (KS) GOF statistic for testing the goodness-of-fit of a continuous pdf because

   A. AD is a more modern procedure.

   B. AD has a more accurate p-value than does KS.

   C. AD is less likely to falsely declare that a distribution does not fit the collected data.

   D. AD is more likely to declare that a distribution function does not fit the edf, especially in the tails of the distribution.

   E. All of the above are true.

2. An entomologist is recording Y, the saturation deficit (a function of temperature and relative humidity) of Lone Star ticks. In order to be very precise in his study of these ticks, the entomologist wants to determine if the cdf of Y, $F(y)$ has a particular form $F_o(y)$. The entomologist measures the value of Y for each of 213 randomly selected Lone Star ticks. Which of the following procedures is best for determining whether $F = F_o$?

   A. Anderson-Darling statistic.

   B. Kolmogorov-Smirnov statistic.

   C. Chi-square Goodness-of-Fit statistic.

   D. Shapiro-Wilks statistic.

   E. depends on the shape of F_o.

3. The Anderson-Darling GOF statistic is referred to as a distribution-free statistic when the continuous cdf F_o is completely specified because

   A. the distribution of the statistic depends only on location-scale parameters.

   B. the distribution of the statistic has a $N(0,1)$ distribution for large n.

   C. the distribution of the statistic does not depend on location-scale parameters.

   D. the distribution of the statistic is the same for choices of F_o, provided it is continuous.

   E. none of the above are valid reasons

4. The Anderson-Darling statistic is preferred to the Chi-square statistic in testing $H_o : F = F_o$ based on a random sample $X_1, \cdots, X_n$ from a continuous cdf F because

   A. the Anderson-Darling is more likely to maintain its level of Type I error.

   B. the number of degrees of freedom for the Anderson-Darling are larger than those for the Chi-squared test.

   C. the probability of Type I error is higher for the Chi-squared test than for the Anderson-Darling.

   D. the Anderson-Darling test is more likely to correctly detect that F is not equal to F_o.

   E. none of the above

5.) An plant physiologist is studying the infestation rate of potato bud insects on genetically altered potato plants. The researcher measures the number, $Y_1, \cdots, Y_{300}$, of potato bud insects on 300 randomly selected genetically altered plants. In order to model the factors affecting the infestation rate, the researcher wants to determine which of five possible cdfs, $F_1, \cdots, F_5$ best fits the cdf of Y, $F(y)$. Which of the following tests would be the most appropriate procedure for determining which of the five cdfs is the best fit?

   A. Anderson-Darling statistic

   B. Kolmogorov-Smirnov statistic

   C. Chi-square Goodness-of-Fit statistic

   D. Shapiro-Wilks statistic

   E. none of the above would be appropriate, they all deal with a single cdf

6. An estimator $\hat{\theta}_{max}$ of the maximum stress load of a new alloy has been developed. A statistician conjectures that the sampling distribution of $\hat{\theta}_{max}$ is Weibull with parameters $(\beta = 10, \gamma = .2)$. A large scale simulation study was conducted to evaluate this claim. A simulation was run with 5000 replications of $n = 10$ data values resulting in 5000 values of $\hat{\theta}_{max}$. A Weibull reference distribution plot was constructed yielding the following plot:

   The Anderson-Darling GOF of the Weibull $(\beta = 10, \gamma = .2)$ model yielded a p-value of 0.001. Why does the reference distribution plot and the Anderson-Darling statistic appear to be contradictory in their assessment of the sampling distribution of $\hat{\theta}_{max}$?

   A. There is no contradiction.

   B. The vertical scale on the reference distribution plot is deceptive.

   C. The distribution must be incorrect.

   D. With 5000 values of $\hat{\theta}_{max}$, the Anderson-Darling test is extremely sensitive to slight deviations from the Weibull model.

   E. The Anderson-Darling test is not a very powerful test for the Weibull distribution.