Test of Fit for Two Populations

jdt

1/2/2021

Contents

Theory

Goodness-of-Fit Tests

In this section the test of hypothesis concerns the CDF for the random variable X rather than its parameters; (location and scale). The test can be stated as

$$H_0 : F_X(x) = F_{X_0}(x)$$

when $F_{X_0}(x)$ is completely specified.[1]

SAS provides the following goodness of fit tests:

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Anderson-Darling test
- Cramer-von Mises test

The Kolmogorov-Smirnov test D statistic, the Anderson-Darling statistic, and the Cramer-von Mises statistic are based on the empirical distribution function (EDF).

[1] When you specify the NORMAL option in the PROC UNIVARIATE statement or you request a fitted parametric distribution in the HISTOGRAM statement, the procedure computes goodness-of-fit tests for the null hypothesis that the values of the analysis variable are a random sample from the specified theoretical distribution.

If you want to test the normality assumptions for analysis of variance methods, beware of using a statistical test for normality alone. A test's ability to reject the null hypothesis (known as the power of the test) increases with the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. Because small deviations from normality do not severely affect the validity of analysis of variance tests, it is important to examine other statistics and plots to make a final assessment of normality. The skewness and kurtosis measures and the plots that are provided by the PLOTS option, the HISTOGRAM statement, the PROBPLOT statement, and the QQPLOT statement can be very helpful. For small sample sizes, power is low for detecting larger departures from normality that may be important. To increase the test's ability to detect such deviations, you may want to declare significance at higher levels, such as 0.15 or 0.20, rather than the often-used 0.05 level. Again, consulting plots and additional statistics can help you assess the severity of the deviations from normality.

QQ and PP Plots

Probability plots can be used to assess normality of data, especially normality of the residuals in linear regression. Let $z_{(1)} \le z_{(2)} \le \ldots \le z_{(n)}$ represent the ordered values of n independent and identically distributed $N(0,1)$ random variables. It can be shown that the expected value of $z_{(i)}$ is

$$E(z_{(i)}) \approx \gamma_i = \Phi^{-1}[(i - 3/8)/(n + 1/4)]$$

where Φ is the cdf for the standard normal given by

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-1/2t^2} dt.$$

The QQ plot consists of plotting of the ordered data (standardized residuals), $z_{(i)}$ versus γ_i, i.e. $(z_{(i)}, \gamma_i)$. If the data are normal then the resulting scatterplot should fall on the diagonal degree line ($z_{(i)} = \gamma_i$). The PP plot is obtained when plotting of the ordered pairs $(\Phi(z_{(i)}), [i/n])$.

Shapiro-Wilk Statistic

The Shapiro-Wilk statistic, W (also denoted as W_n to emphasize its dependence on the sample size n) is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk, 1965). The statistic W is always greater than zero and less than or equal to one $(0 < W \le 1)$. When the data are normally distributed, one would expect the ratio of these two estimators for the variance to be close to one, in which case, small values of W lead to the rejection of the null hypothesis of normality. The distribution of W is highly skewed. Seemingly large values of W (such as 0.90) may be considered small and lead you to reject the null hypothesis. The method for computing the p-value (the probability of obtaining a W statistic less than or equal to the observed value) depends on n. For $n = 3$, the probability distribution of W is known and is used to determine the p-value. For $n > 4$, a normalizing transformation is computed:

$$Z_n = \begin{cases} (-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \le n \le 11 \\ (\log(1 - W_n) - \mu)/\sigma & \text{if } n \ge 12 \end{cases}$$

The values of σ, γ, and μ are functions of n obtained from simulation results. Large values of Z_n indicate departure from normality, and because the statistic Z_n has an approximately standard normal distribution, this distribution is used to determine the p-values for $n > 4$.[2]

EDF Goodness-of-Fit Tests

The EDF tests offer advantages over traditional chi-square goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints. For a thorough discussion, refer to D'Agostino and Stephens (1986).

[2] The Shapiro-Wilks procedure is based upon "moment-type" estimators and would not be as powerful as the procedures that are based upon the estimated CDF for X. It should probably not be used without considering the other procedures.

The **Empirical Distribution Function (EDF)** is defined for a set of n independent observations $X_1, \ldots, X_n$ with a common distribution function $F(x)$. Denote the observations ordered from smallest to largest as $X_{(1)}, \ldots, X_{(n)}$. The EDF, $F_n(x)$, is,

$$\begin{array}{lll} F_n(x) = 0, & x < X_{(1)} & \\ F_n(x) = \frac{i}{n}, & X_{(i)} \leq x < X_{(i+1)} & i = 1, \ldots, n-1 \ . \\ F_n(x) = 1, & X_{(n)} \leq x & \end{array}$$

Note: $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. This function estimates the distribution function $F(x)$. At any value x, $F_n(x)$ is the proportion of observations less than or equal to x, while $F(x)$ is the probability of an observation less than or equal to x. EDF statistics measure the discrepancy between $F_n(x)$ and $F(x)$.[3]

Kolmogorov D Statistic

The Kolmogorov-Smirnov statistic (D) is defined as

$$D = \sup_x |F_n(x) - F(x)|$$

The Kolmogorov-Smirnov statistic is computed as the maximum of D^+ and D^-, where D^+ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function, and D^- is the largest vertical distance when the EDF is less than the distribution function[4].

$$\begin{array}{lll} D^+ & = & \max_i \left(\frac{i}{n} - U_{(i)} \right) \\ D^- & = & \max_i \left(U_{(i)} - \frac{i-1}{n} \right) \\ D & = & \max \left(D^+, D^- \right) \end{array}$$

Quadratic EDF Statistics

The Anderson-Darling statistic and the Cramer-von Mises statistic are special cases of the general quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F_n(x) - F(x))^2$. The general form of the quadratic class of EDF statistics is

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \psi(x) dF(x) \tag{1}$$

where $\psi(x)$ is a weight function defined on the squared difference $(F_n(x) - F(x))^2$.

Anderson-Darling Statistic

The Anderson-Darling statistic considers $\psi(x) = [F(x)(1 - F(x))]^{-1}$ in which case equation (1) is

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

where $U_{(i)} = F_X(X_{(i)})$.

Cramer-von Mises Statistic

The Cramer-von Mises statistic considers $\psi(x) = 1$ in which case equation (1) is

$$W^2 = \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}.$$

[3]The computational formulas for the EDF statistics are based upon the probability integral transformation $U = F(X)$. That is, if $X \sim F(X)$ then $U \sim U(0,1)$. In the test of fit problems, $F(X)$ is the null (or specified) distribution function.

[4]PROC UNIVARIATE uses a modified Kolmogorov D statistic to test the data against a normal distribution with mean and variance equal to the sample mean and variance.

Tests Based on the Empirical Distribution Function (EDF)

This section describes three nonparametric tests that are based on the empirical distribution function[5]. The procedures are; the Kolmogorov-Smirnov and Cramer-von Mises tests, and also the Kuiper test for two-sample data.[6] The null hypothesis is $H_0 : F_X(\cdot) = F_Y(\cdot) = F(\cdot)$. The (EDF) of a sample $\{x_j\}$, $j = 1, 2, \ldots, n$ is defined as

$$\hat{F}(x) = \frac{1}{n}(\text{number of } x_j \leq x) = \frac{1}{n}\sum_{j=1}^{n} I(x_j \leq x)$$

where $I(\cdot)$ is an indicator function. Let $\hat{F}_i$ denote the sample EDF for the i^{th} group. The EDF for the overall sample, pooled over groups, can also be expressed as

$$\hat{F}(x) = \frac{1}{n}\sum_{i}\left(n_i \hat{F}_i(x)\right)$$

where n_i is the number of observations in the i^{th} group, and n is the total number of observations.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic measures the maximum deviation of the EDF within the groups from the pooled EDF. The Kolmogorov-Smirnov statistic is computed as,

$$KS = \max_j \sqrt{\frac{1}{n}\sum_i n_i\left(\hat{F}_i(x_j) - \hat{F}(x_j)\right)^2} \quad \text{for} \quad j = 1, 2, \ldots, n$$

The asymptotic Kolmogorov-Smirnov statistic is computed as, $KS_a = KS \times \sqrt{n}$. If there are only two class levels, the two-sample Kolmogorov-Smirnov test statistic D is

$$D = \max_j \left|\hat{F}_1(x_j) - \hat{F}_2(x_j)\right| \quad \text{for} \quad j = 1, 2, \ldots, n$$

The p-value for this test is the probability that D is greater than the observed value d under the null hypothesis of no difference between class levels (samples). The asymptotic p-value for D is approximated as,

$$\Pr(D > d) = 2\sum_{i=1}^{\infty}(-1)^{(i-1)}\, e^{(-2i^2z^2)}$$

where

$$z = d\sqrt{n_1 n_2/n}$$

See Hodges (1957) for information about this approximation.

Cramer-von Mises Test

The Cramer-von Mises statistic is

$$CM = \frac{1}{n^2}\sum_i\left(n_i\sum_{j=1}^{p} t_j\left(\hat{F}_i(x_j) - \hat{F}(x_j)\right)^2\right)$$

where t_j is the number of ties at the j^{th} distinct value and p is the number of distinct values. The asymptotic value is computed as

$$CM_a = CM \times n.$$

[5] [PROC NPAR1WAY - EDF option].

[6] For further information about the formulas and the interpretation of EDF statistics, see Hollander and Wolfe (1999) and Gibbons and Chakraborti (2010). For details about the k-sample analogs of the Kolmogorov-Smirnov and Cramer-von Mises statistics, see Kiefer (1959).

Kuiper Test

For data with two class levels, the Kuiper statistic is

$$K = \max_j \left(\hat{F}_1(x_j) - \hat{F}_2(x_j)\right) - \min_j \left(\hat{F}_1(x_j) - \hat{F}_2(x_j)\right) \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic value is

$$K_a = K\sqrt{n_1 n_2 / n}$$

The p-value for the Kuiper test is the probability of observing a larger value of K_a under the null hypothesis of no difference between the two classes Owen (1962, p 441).

Simple Linear Rank Tests for Two-Sample Data

Statistics of the form

$$S = \sum_{j=1}^{n} a(R_j)$$

are called *simple linear rank statistics*, where R_j is the rank of observation j, $a(R_j)$ is the score based on the rank of observation j, and n is the total number of observations[7].

To compute an asymptotic test for a linear rank sum statistic, use the standardized test statistic z, which has an asymptotic standard normal distribution under the null hypothesis as

$$z = \frac{(S - E_0(S))}{\sqrt{Var_0(S)}}$$

where $E_0(S)$ is the expected value of S under the null hypothesis, and $Var_0(S)$ is the variance under the null hypothesis. As shown in Randles and Wolfe (1979),

$$E_0(S) = \frac{n_1}{n} \sum_{j=1}^{n} a(R_j)$$

where n_1 is the number of observations in the first class level (sample), n_2 is the number of observations in the other class level, and

$$Var_0(S) = \frac{n_1 n_2}{n(n-1)} \sum_{j=1}^{n} \left(a(R_j) - \bar{a}\right)^2$$

where $\bar{a}$ is the average score,

$$\bar{a} = \frac{1}{n} \sum_{j=1}^{n} a(R_j)$$

Wilcoxon and Mann-Whitney Test

The Mann-Whitney and Wilcoxon test assumes that

- The data consists of a random sample of n_1 values, denoted $X_1, X_2, \ldots, X_{n_1}$ from $F_X(x)$ with $median(X) = \theta_X$, and a random sample of n_2 values, denoted $Y_1, Y_2, \ldots, Y_{n_2}$ from $F_Y(y)$ with $median(Y) = \theta_Y$.
- The random samples are at least ordinal.
- F_X and F_Y differ only with respect to the their median. That is, $\theta_X = \theta_Y + \delta$.

The null hypothesis is $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ (the usual one-sided tests are possible). The procedure is

1. Combine the data for the two random samples and rank the combined data where $r_j = rank(Y_j)$.

[7] For two-sample data (where the observations are classified into two levels), PROC NPAR1WAY calculates simple linear rank statistics for the scores that you specify.

2. Compute the Wilcoxon statistics as, $W = \sum_{j=1}^{n_2} r_j$.
3. Reject H_0 if W is either too small ($\theta_Y < \theta_X$) or too large ($\theta_Y > \theta_X$).
4. The large sample distribution of $W^* = \frac{W-E(W)}{\sqrt{Var(W)}} \sim N(0,1)$, where

$$E(W) = \frac{n_2(N+1)}{2}$$

and

$$Var(W) = \frac{n_1 n_2(N+1)}{12}$$

for $N = n_1 + n_2$.

The Mann-Whitney test statistics is

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

where $\phi(x,y) = I_{x<y}$ the indicator function. The two statistics are similar in that

$$W = U + \frac{n_2(n_2+1)}{2}.$$

Note: $\Pr[X < Y] = E(I_{X<Y})$ in which case the statistic U can be used as a nonparametric estimate of $\Pr[X < Y]$.[8]

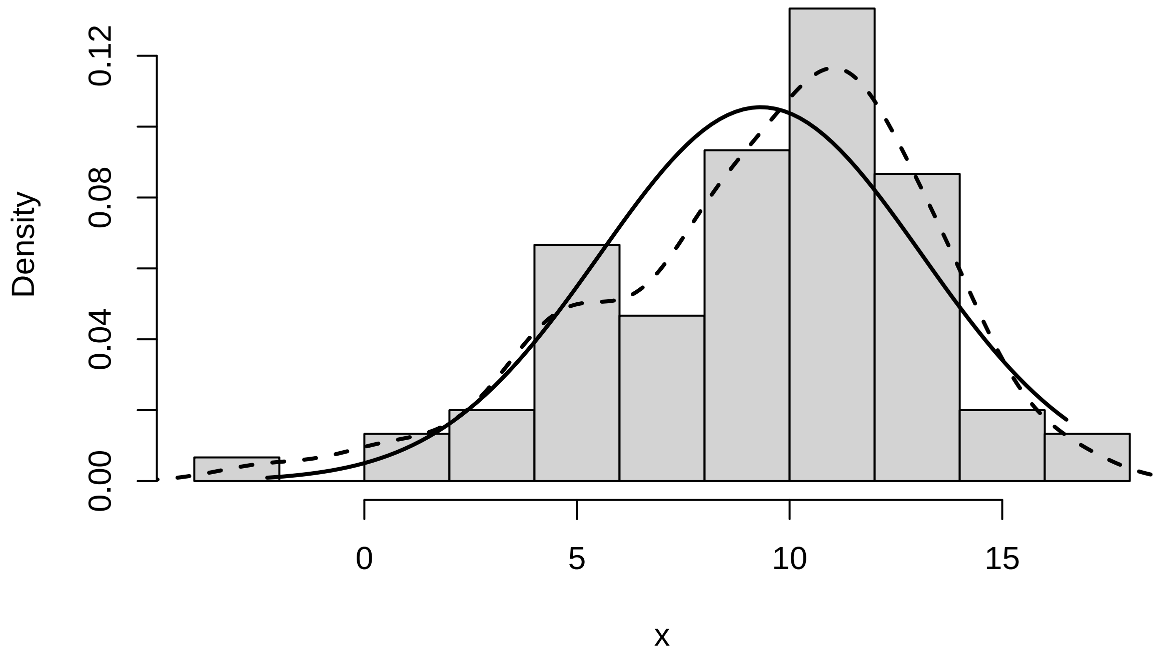
Simulated Data

R

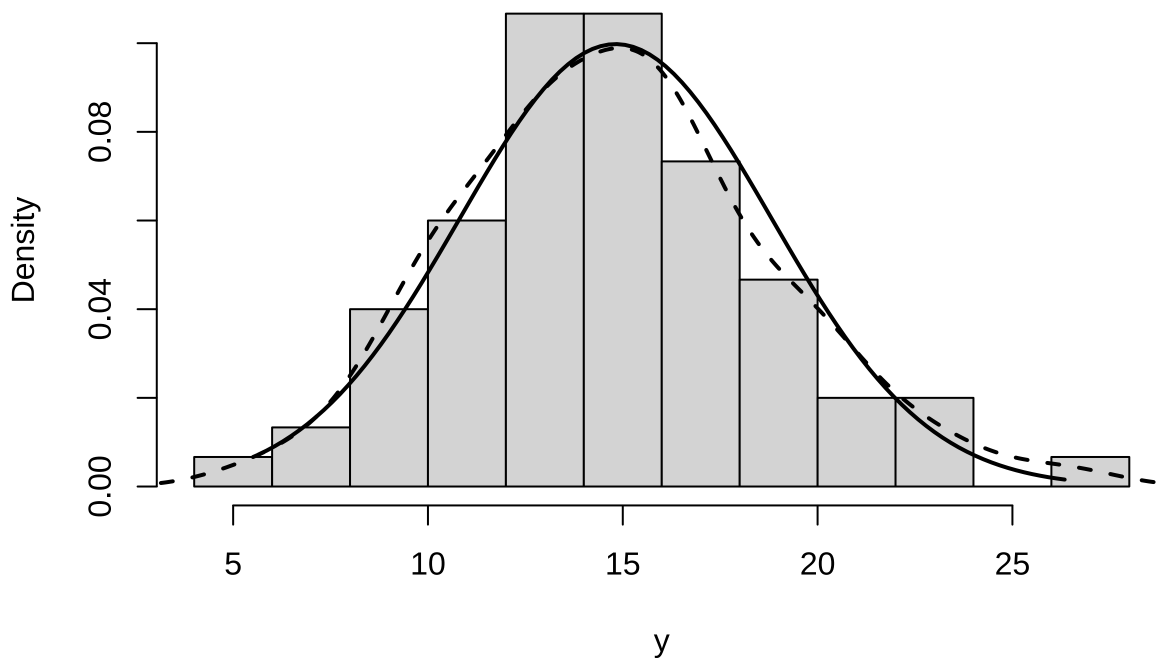
Normal simulated data

```
x = rnorm(75, mean=10, sd = 4)
y = rnorm(75, mean=14, sd = 4)
cert = data.frame(x,y)
#hist(x)
#hist(y)

with(cert, hist(x, main="", freq=FALSE))
with(cert, lines(density(x), main="X", lty=2, lwd=2))
xvals = with(cert, seq(from=min(x), to=max(x), length=100))
with(cert, lines(xvals, dnorm(xvals, mean(x), sd(x)), lwd=2))
```

[8] Since $\Pr[X < Y]$ = AUC, the Mann-Whitney statistics is a nonparametric estimate for the area under the ROC curve (AUC).




```r
with(cert, hist(y, main="", freq=FALSE))
with(cert, lines(density(y), main="", lty=2, lwd=2))
xvals = with(cert, seq(from=min(y), to=max(y), length=100))
with(cert, lines(xvals, dnorm(xvals, mean(y), sd(y)), lwd=2))
```



Descriptive statistics

```r
library("mosaic")
favstats(x)
```

```
##         min      Q1  median       Q3     max     mean      sd  n missing
##   -2.285134 7.41193 10.1152 11.66413 16.5094 9.321553 3.78183 75       0
```

```r
mean(x, trim=.05)
```

```
## [1] 9.470389
```

```r
quantile(x, seq(from=.025, to= .975, by=.1))
```

```
##       2.5%      12.5%      22.5%      32.5%      42.5%      52.5%      62.5%
##  0.6014589  4.5654297  7.0005118  8.3247608  9.1960820 10.4706844 11.0338863
##      72.5%      82.5%      92.5%
## 11.5762074 12.7045181 13.9266712
```

```r
t.test(x, mu=12, conf.level=.9)    #test for mu=12 and 90 percent ci
```

```
##
##  One Sample t-test
##
## data:  x
## t = -6.1335, df = 74, p-value = 3.859e-08
## alternative hypothesis: true mean is not equal to 12
## 90 percent confidence interval:
##   8.594159 10.048948
## sample estimates:
## mean of x
##  9.321553
```

```r
favstats(y)
```

```
##       min       Q1  median       Q3      max     mean       sd  n missing
##  5.654374 12.33702 14.7939 17.21898 26.34099 14.81781 3.997653 75       0
```

```r
mean(y, trim=.05)
```

```
## [1] 14.74651
```

```r
quantile(y, seq(from=.025, to= .975, by=.1))
```

```
##      2.5%     12.5%     22.5%     32.5%     42.5%     52.5%     62.5%     72.5%
##  7.831725 10.122533 11.686866 12.969433 13.756949 14.977590 15.763748 16.976849
##     82.5%     92.5%
## 18.659638 20.553840
```

```r
t.test(y, mu=12, conf.level=.9)    #test for mu=12 and 90 percent ci
```

```
##
##  One Sample t-test
##
## data:  y
## t = 6.1043, df = 74, p-value = 4.359e-08
## alternative hypothesis: true mean is not equal to 12
## 90 percent confidence interval:
##  14.04890 15.58671
## sample estimates:
## mean of x
##  14.81781
```

```r
library("coin")
wilcox.test(x,y)
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  x and y
## W = 881, p-value = 3.926e-13
## alternative hypothesis: true location shift is not equal to 0
```

```r
t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -8.6496, df = 147.55, p-value = 8.062e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.751984 -4.240520
## sample estimates:
## mean of x mean of y
##  9.321553 14.817806
```

```r
ks.test(x,y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.54667, p-value = 1.229e-10
## alternative hypothesis: two-sided
```

Test of fit

```r
library("nortest")

ad.test(x)
```

```
##
##  Anderson-Darling normality test
##
## data:  x
## A = 0.8728, p-value = 0.02408
```

```r
cvm.test(x)
```

```
##
##  Cramer-von Mises normality test
##
## data:  x
## W = 0.15273, p-value = 0.02127
```

```r
lillie.test(x)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.10459, p-value = 0.04115
```

```r
pearson.test(x)
```

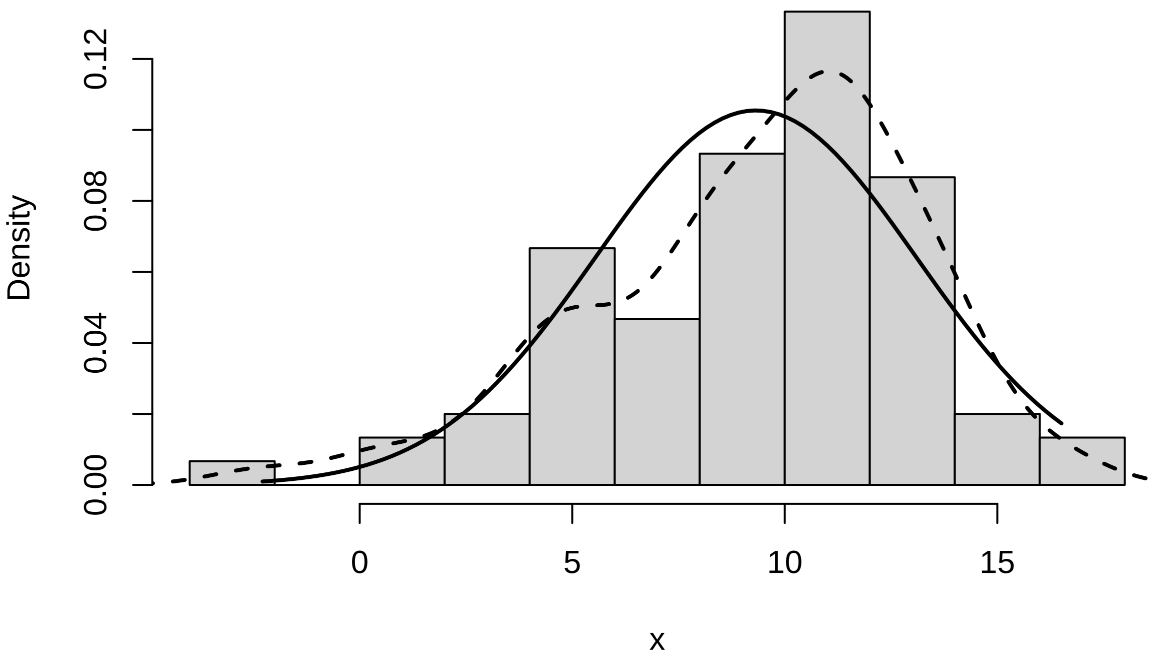
```
##
```

```
##  Pearson chi-square normality test
##
## data:  x
## P = 14.76, p-value = 0.09774
```

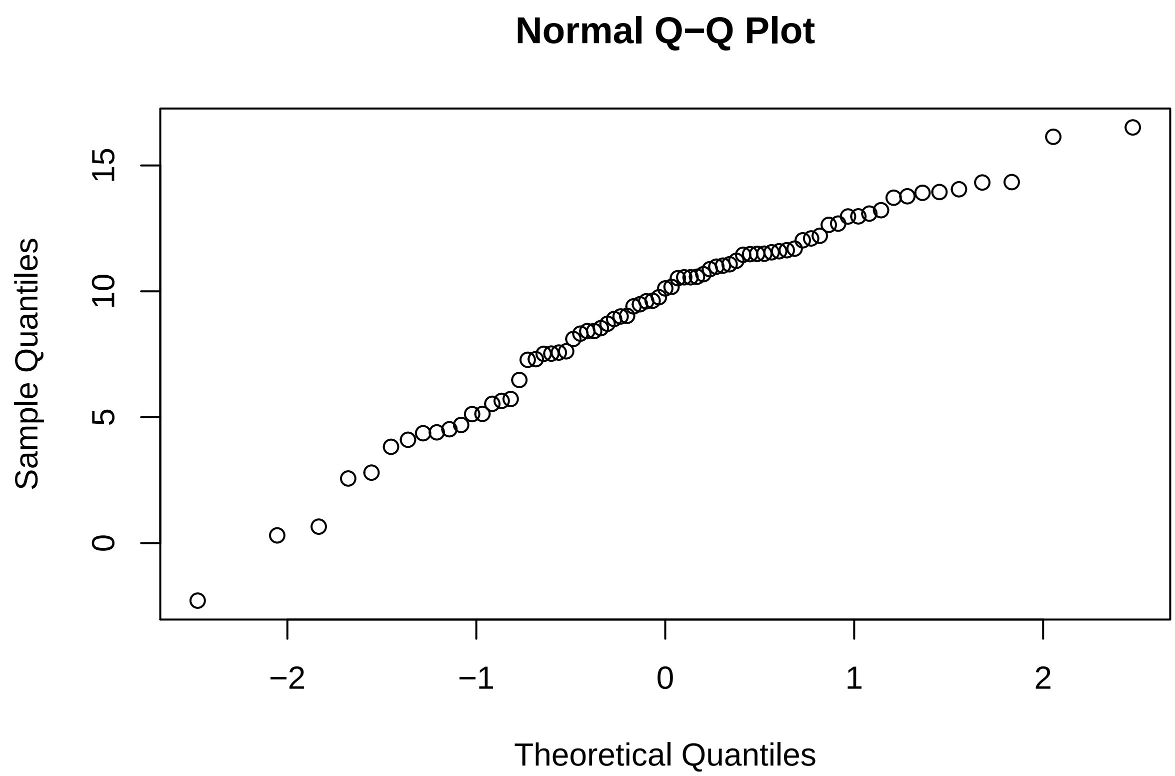
```r
sf.test(x)
```

```
##
##  Shapiro-Francia normality test
##
## data:  x
## W = 0.96309, p-value = 0.02904
```

```r
with(cert, hist(x, main="", freq=FALSE))
with(cert, lines(density(x), main="X", lty=2, lwd=2))
xvals = with(cert, seq(from=min(x), to=max(x), length=100))
with(cert, lines(xvals, dnorm(xvals, mean(x), sd(x)), lwd=2))
```




```r
qqnorm(x)
```



Lognormal simulated data

```
x = 2*rlnorm(75, mean=0, sd = 1) + 10
y = 2*rlnorm(75, mean=0, sd = 1) + 12
cert = data.frame(x,y)
#hist(x)
#hist(y)
```

Descriptive statistics

```
favstats(x)
```

```
##       min       Q1   median       Q3     max     mean       sd  n missing
##  10.06103 11.23886 12.32273 14.67958 40.1339 13.68235 4.502776 75       0
```

```
mean(x, trim=.05)
```

```
## [1] 13.07115
```

```
quantile(x, seq(from=.025, to= .975, by=.1))
```

```
##     2.5%    12.5%    22.5%    32.5%    42.5%    52.5%    62.5%    72.5%
## 10.38123 10.93878 11.17899 11.41408 11.88395 12.44388 13.48679 14.38446
##    82.5%    92.5%
## 15.02332 18.62166
```

```
t.test(x, mu=12, conf.level=.9)    #test for mu=12 and 90 percent ci
```

```
##
##  One Sample t-test
##
## data:  x
```

```
## t = 3.2357, df = 74, p-value = 0.001816
## alternative hypothesis: true mean is not equal to 12
## 90 percent confidence interval:
##  12.81629 14.54841
## sample estimates:
## mean of x
##  13.68235
```

```
favstats(y)
```

```
##       min       Q1   median       Q3     max     mean       sd  n missing
##  12.07774 13.23737 14.49034 17.35879 27.7801 15.68706 3.447846 75       0
```

```
mean(y, trim=.05)
```

```
## [1] 15.37771
```

```
quantile(y, seq(from=.025, to= .975, by=.1))
```

```
##     2.5%    12.5%    22.5%    32.5%    42.5%    52.5%    62.5%    72.5%
## 12.41773 12.82579 13.17634 13.37436 13.95301 14.72350 15.44446 16.99944
##    82.5%    92.5%
## 18.09518 20.89182
```

```
t.test(y, mu=12, conf.level=.9)    #test for mu=12 and 90 percent ci
```

```
##
##  One Sample t-test
##
## data:  y
## t = 9.2611, df = 74, p-value = 5.353e-14
## alternative hypothesis: true mean is not equal to 12
## 90 percent confidence interval:
##  15.02390 16.35022
## sample estimates:
## mean of x
##  15.68706
```

```
library("coin")
wilcox.test(x,y)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 1413, p-value = 1.453e-07
## alternative hypothesis: true location shift is not equal to 0
```

```
t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -3.0613, df = 138.58, p-value = 0.002647
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.2995102 -0.7099112
## sample estimates:
```

```
## mean of x mean of y
##  13.68235  15.68706
```

```r
ks.test(x,y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.49333, p-value = 1.192e-08
## alternative hypothesis: two-sided
```

Test of fit

```r
library("nortest")
ad.test(x)
```

```
##
##  Anderson-Darling normality test
##
## data:  x
## A = 7.7079, p-value < 2.2e-16
```

```r
cvm.test(x)
```

```
## Warning in cvm.test(x): p-value is smaller than 7.37e-10, cannot be computed
## more accurately
```

```
##
##  Cramer-von Mises normality test
##
## data:  x
## W = 1.3818, p-value = 7.37e-10
```

```r
lillie.test(x)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.22496, p-value = 3.972e-10
```

```r
pearson.test(x)
```

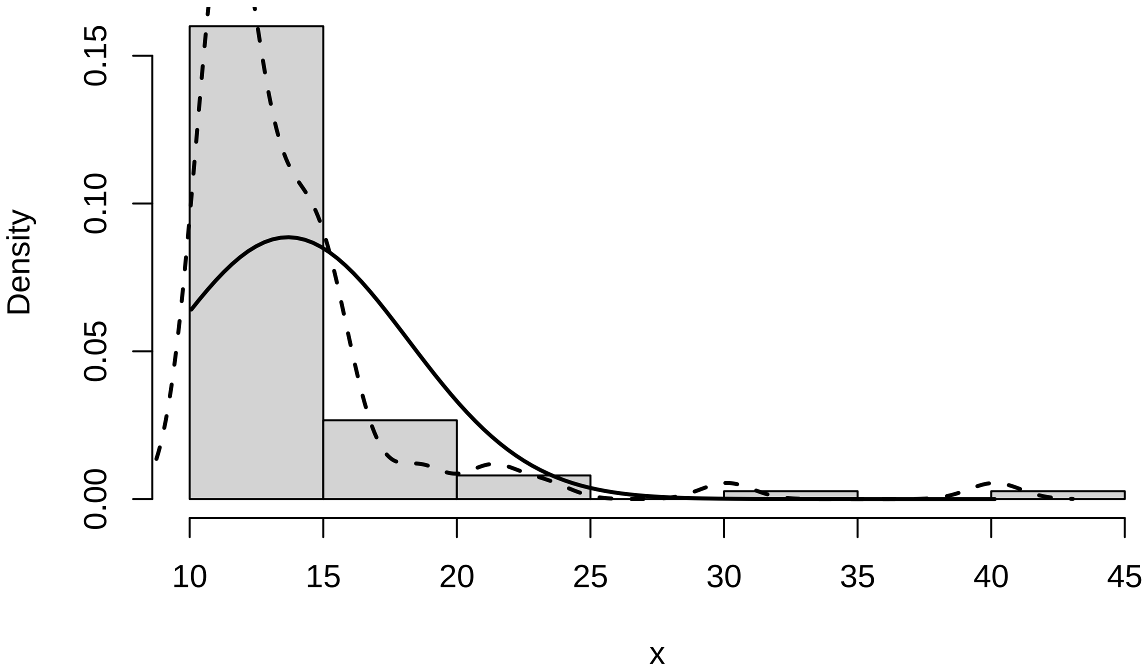
```
##
##  Pearson chi-square normality test
##
## data:  x
## P = 105, p-value < 2.2e-16
```

```r
sf.test(x)
```

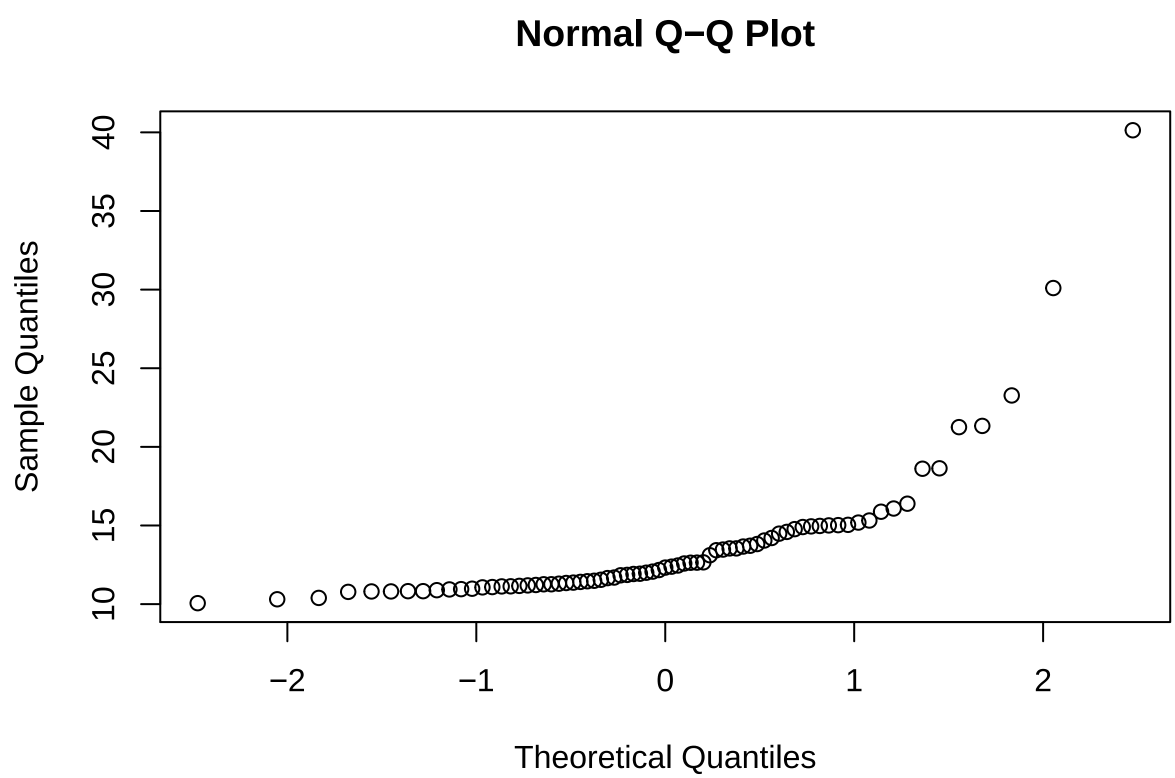
```
##
##  Shapiro-Francia normality test
##
## data:  x
## W = 0.598, p-value = 3.341e-11
```

```r
with(cert, hist(x, main="", freq=FALSE))
with(cert, lines(density(x), main="X", lty=2, lwd=2))
```

```r
xvals = with(cert, seq(from=min(x), to=max(x), length=100))
with(cert, lines(xvals, dnorm(xvals, mean(x), sd(x)), lwd=2))
```




```r
qqnorm(x)
```




```r
ad.test(y)
```

```
##
##  Anderson-Darling normality test
```

```
##
## data:  y
## A = 3.9696, p-value = 5.589e-10
```

```
cvm.test(y)
```

```
##
##  Cramer-von Mises normality test
##
## data:  y
## W = 0.68043, p-value = 8.146e-08
```

```
lillie.test(y)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  y
## D = 0.16505, p-value = 2.797e-05
```

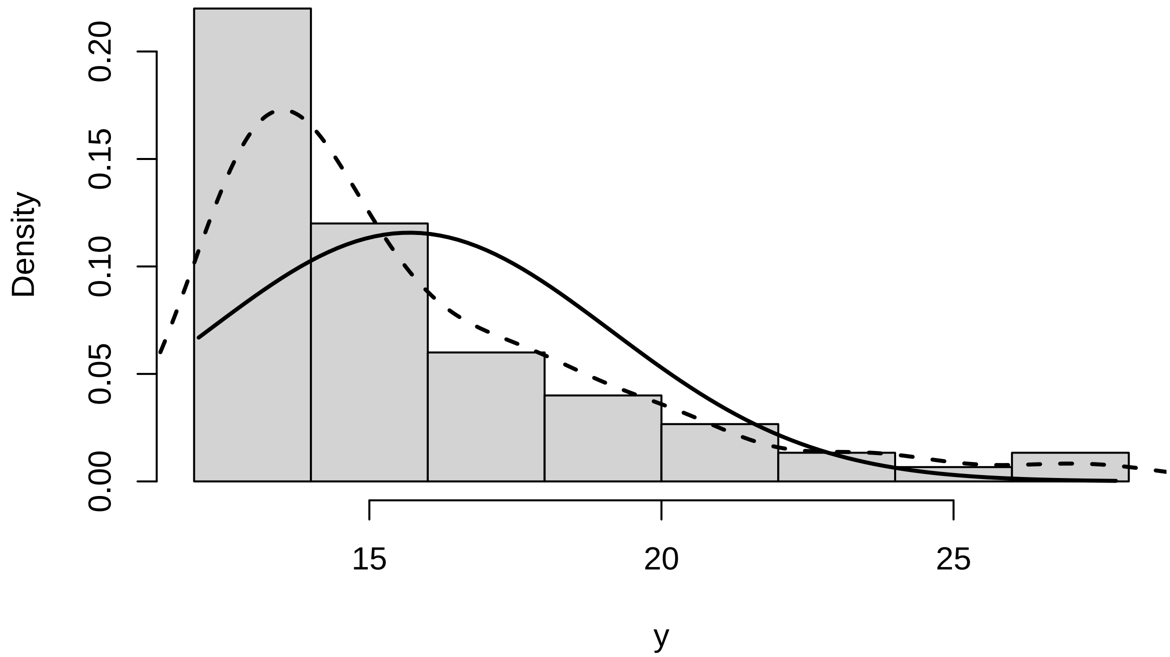
```
pearson.test(y)
```

```
##
##  Pearson chi-square normality test
##
## data:  y
## P = 61.16, p-value = 8.008e-10
```

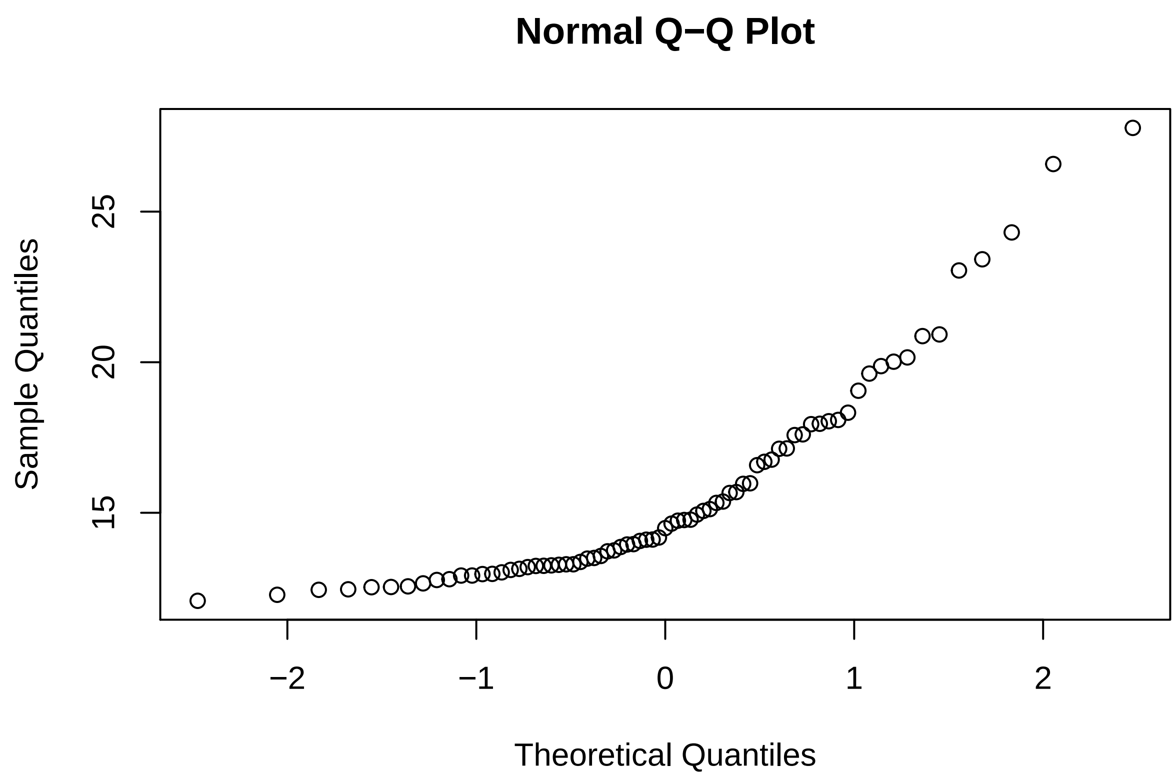
```
sf.test(y)
```

```
##
##  Shapiro-Francia normality test
##
## data:  y
## W = 0.82961, p-value = 5.798e-07
```

```
with(cert, hist(y, main="", freq=FALSE))
with(cert, lines(density(y), main="Y", lty=2, lwd=2))
xvals = with(cert, seq(from=min(y), to=max(y), length=100))
with(cert, lines(xvals, dnorm(xvals, mean(y), sd(y)), lwd=2))
```




```r
qqnorm(y)
```



SAS

In this section, I have added the SAS code with similar output as to the R output using simulated data. I have used a real data set called the the certification data set for the SAS example. Originally, I have did not include any results for simulated data using SAS. I have added that the end. The cases considered in SAS differ from those considered in R.

Code

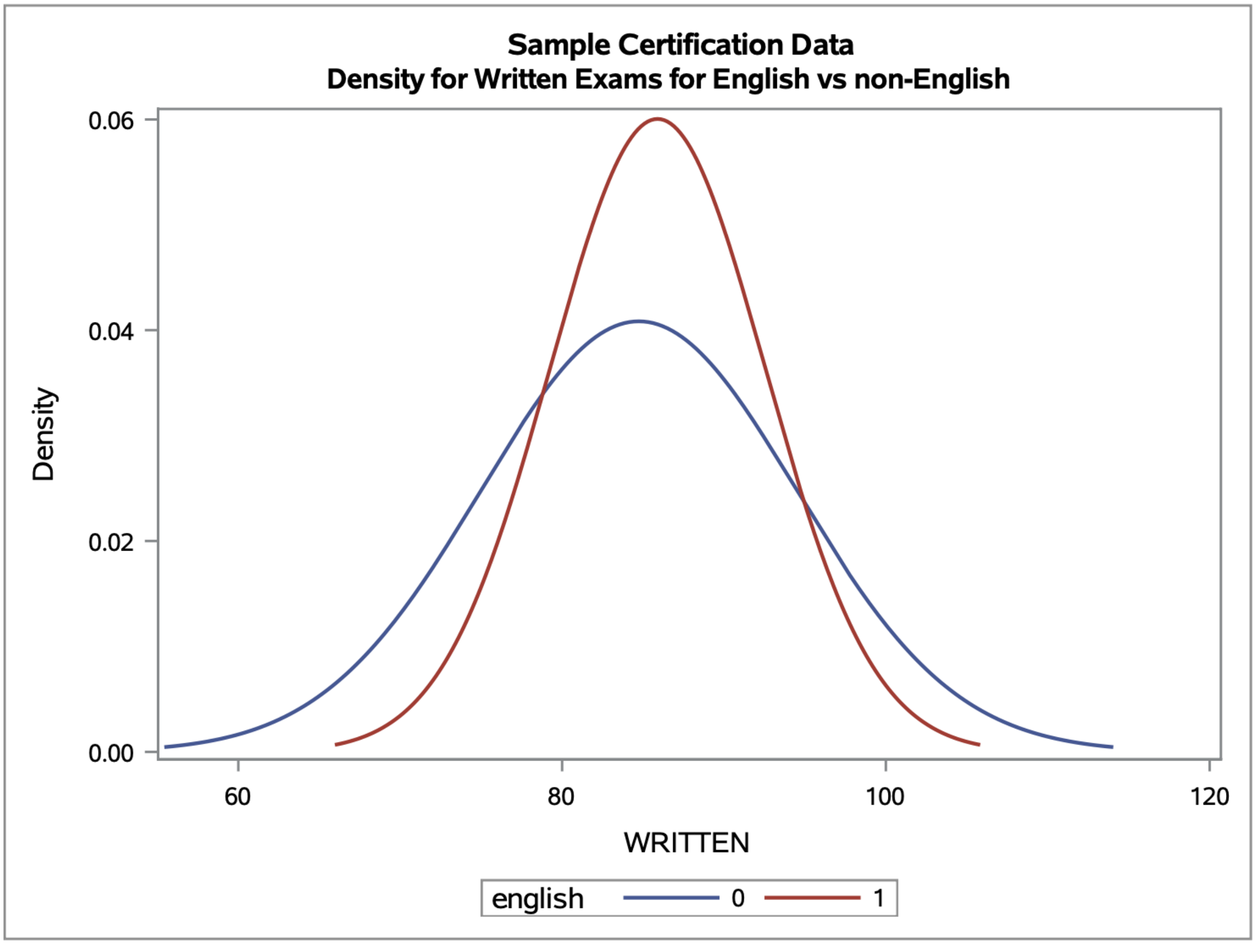
```
options center nodate pagesize=100 ls=70;

title 'Sample Certification Data';
data cert; set LDATA.certification;
english = (language = 'English');
run;
proc sort data=cert; by year; run;

ods graphics on;
title2 'Density for Written Exams for English vs non-English';
proc sgplot data=cert; where year = 2008;
density written /group=english;
run;

title2 'Descriptive Statistics';
proc univariate data=cert normal trim=.05 winsor=.05 mu0=75; where year = 2008;
var written;
run;

title2 'Nonparametric Test of Hypothesis';
proc npar1way data=cert wilcoxon edf ; where year = 2008;
var written;
run;

title2 'T Test';
proc ttest data=cert ; where year = 2008;
class english;
var written;
run;

title2 'Nonparametric Tests for 2 populations';
proc npar1way data=cert wilcoxon edf; where year = 2008;
class english;
var written;
run;

title2 'Fit for 2 populations';
proc univariate data=cert normal mu0=75; where year = 2008;
class english;
var written;
run;
quit;
```

The SAS code for the analysis of the simulated data would be as follows (no output is included).

```
title 'Simulated Normal Data';
%let n=50;

data cert;
  do group = 1 to 2;
     do i = 1 to &n;
        x = rand('normal', 0, 1);
      output;
      end;
  end;
run;
data cert; set cert;
   if group = 1 then x = 2*x + 10;
   else x = 2*x + 14;
run;

title2 'Density for Group 1 vs Group 2';
proc sgplot data=cert;
density x /group=group;
run;

title2 'Descriptive Statistics';
proc univariate data=cert normal trim=.05 winsor=.05 mu0=75;
var x;
run;

title2 'Nonparametric Test of Hypothesis';
proc npar1way data=cert wilcoxon edf ;
class group;
var x;
run;

title2 'T Test';
proc ttest data=cert ;
class group;
var x;
run;

title2 'Nonparametric Tests for 2 populations';
proc npar1way data=cert wilcoxon edf;
class group;
var x;
run;

title2 'Fit for 2 populations';
proc univariate data=cert normal mu0=75;
class group;
var x;
run;

title 'Simulated Exponential Data';
%let n=50;

data cert;
  do group = 1 to 2;
     do i = 1 to &n;
        x = rand('EXPONENTIAL', 1);
```

```
      output;
      end;
  end;
run;
data cert; set cert;
   if group = 1 then x = 2*x + 5;
   else x = 5*x + 5;
run;

title2 'Density for Group 1 vs Group 2';
proc sgplot data=cert;
density x /group=group;
run;

title2 'Descriptive Statistics';
proc univariate data=cert normal trim=.05 winsor=.05 mu0=75;
var x;
run;

title2 'Nonparametric Test of Hypothesis';
proc npar1way data=cert wilcoxon edf ;
class group;
var x;
run;

title2 'T Test';
proc ttest data=cert ;
class group;
var x;
run;

title2 'Nonparametric Tests for 2 populations';
proc npar1way data=cert wilcoxon edf;
class group;
var x;
run;

title2 'Fit for 2 populations';
proc univariate data=cert normal mu0=75;
class group;
var x;
run;
```

Output for Certification Data



Sample Certification Data
Descriptive Statistics
The UNIVARIATE Procedure
Variable: WRITTEN (WRITTEN)

Moments			
N	392	Sum Weights	392
Mean	85.2653061	Sum Observations	33424
Std Deviation	8.49558456	Variance	72.1749569
Skewness	−1.5546607	Kurtosis	3.36775798
Uncorrected SS	2878128	Corrected SS	28220.4082
Coeff Variation	9.96370616	Std Error Mean	0.42909182

Basic Statistical Measures			
Location		Variability	
Mean	85.26531	Std Deviation	8.49558
Median	87.00000	Variance	72.17496
Mode	90.00000	Range	53.00000
		Interquartile Range	9.00000

Tests for Location: Mu0=75				
Test	Statistic		p Value	
Student's t	t	23.92333	Pr > |t|	<.0001
Sign	M	155.5	Pr >= |M|	<.0001
Signed Rank	S	32271	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.884435	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.140605	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.842764	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	11.00673	Pr > A-Sq	<0.0050

Trimmed Means								
% in Tail	# in Tail	Trimmed Mean	Std Error Trimmed Mean	95% CI		DF	$H_0 : \mu_0 = 75.$	Pr > |t|
5.10	20	86.02273	0.392290	85.25119	86.79426	351	28.09845	<.0001

Winsorized Means								
% in Tail	# in Tail	Winsorized Mean	Std Error Winsorized Mean	95% CI		DF	$H_0 : \mu_0 = 75.$	Pr > |t|
5.10	20	85.66327	0.392347	84.89162	86.43491	351	27.17818	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	98
99%	98
95%	95
90%	94
75% Q3	91
50% Median	87
25% Q1	82
10%	75
5%	68
1%	55
0% Min	45

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
45	866	98	752
48	943	98	784
51	753	98	828
55	817	98	831
55	792	98	873

Sample Certification Data
T Test
The TTEST Procedure

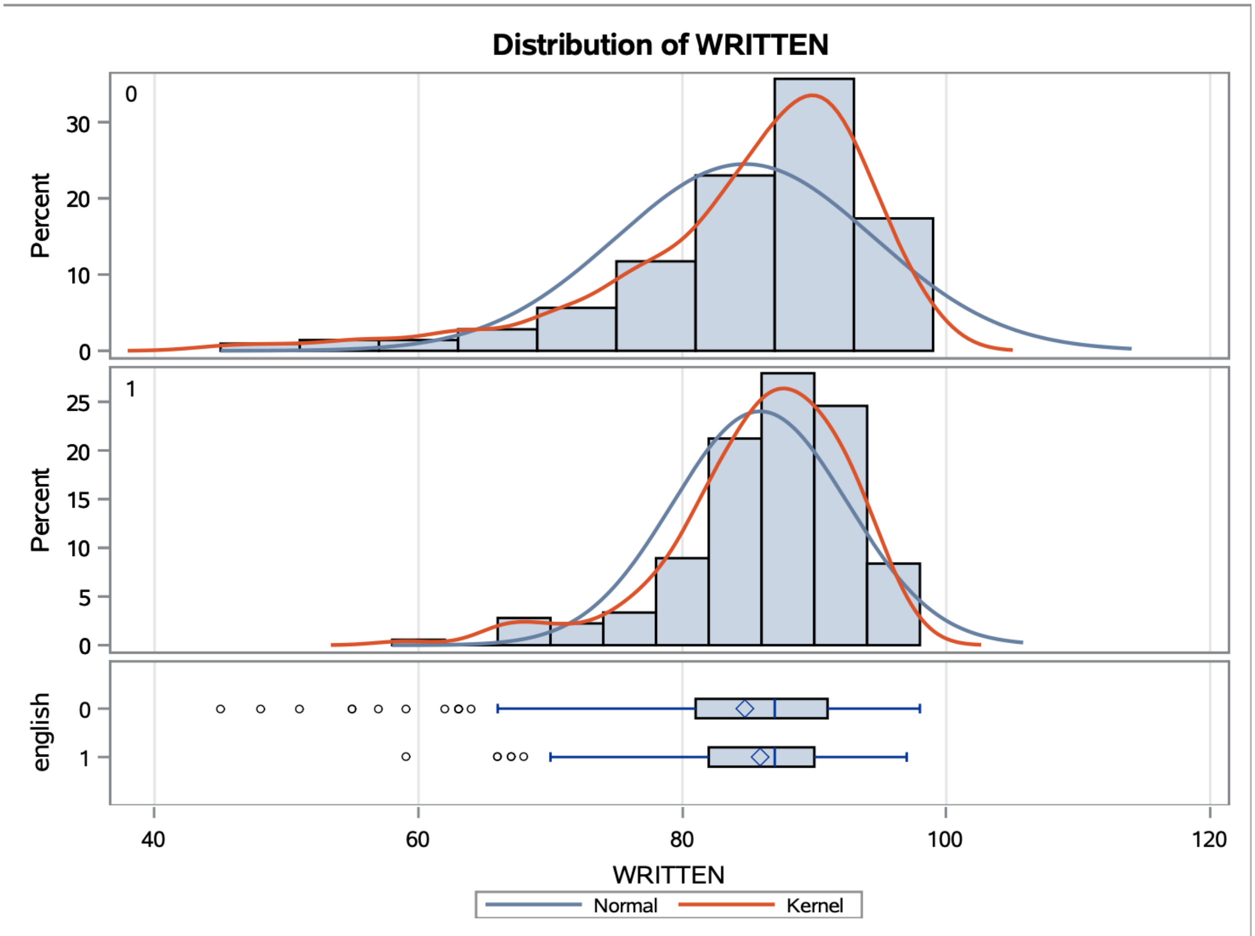
Variable: WRITTEN (WRITTEN)

english	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		213	84.7371	9.7693	0.6694	45.0000	98.0000
1		179	85.8939	6.6438	0.4966	59.0000	97.0000
Diff (1-2)	Pooled		−1.1568	8.4868	0.8605		
Diff (1-2)	Satterthwaite		−1.1568		0.8335		

english	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		84.7371	83.4176	86.0566	9.7693	8.9213	10.7969
1		85.8939	84.9139	86.8738	6.6438	6.0195	7.4137
Diff (1-2)	Pooled	−1.1568	−2.8486	0.5351	8.4868	7.9308	9.1274
Diff (1-2)	Satterthwaite	−1.1568	−2.7956	0.4821			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	390	−1.34	0.1797
Satterthwaite	Unequal	374.47	−1.39	0.1660

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	212	178	2.16	<.0001



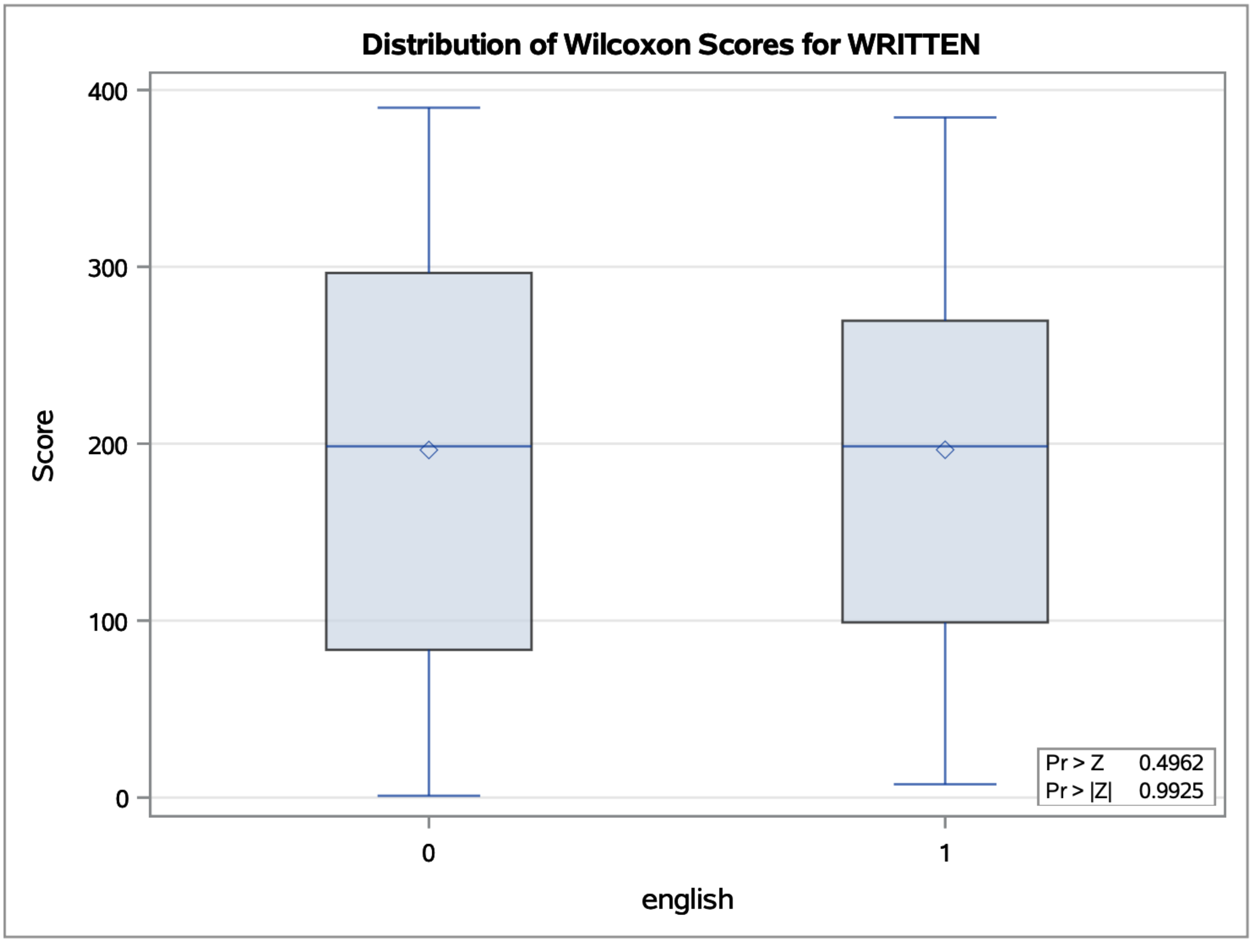
Distribution of WRITTEN
0
1
Percent
english
WRITTEN
40
60
80
100
120
Normal
Kernel

Sample Certification Data
Nonparamtric Tests for 2 populations
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable WRITTEN Classified by Variable english					
english	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	213	41843.50	41854.50	1115.69634	196.448357
1	179	35184.50	35173.50	1115.69634	196.561453
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
Statistic	Z	Pr > Z	Pr > |Z|	t Approximation	
				Pr > Z	Pr > |Z|
35184.50	0.0094	0.4962	0.9925	0.4962	0.9925
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
0.0001	1	0.9921


Distribution of Wilcoxon Scores for WRITTEN
Score
400
300
200
100
0
0
1
english
Pr > Z 0.4962
Pr > |Z| 0.9925

Sample Certification Data
Fit for 2 populations
The UNIVARIATE Procedure
Variable: WRITTEN (WRITTEN)
english = 0

Moments			
N	213	Sum Weights	213
Mean	84.7370892	Sum Observations	18049
Std Deviation	9.76933907	Variance	95.4399858
Skewness	−1.4951427	Kurtosis	2.62137594
Uncorrected SS	1549653	Corrected SS	20233.277
Coeff Variation	11.5290001	Std Error Mean	0.66938408

Basic Statistical Measures			
Location		Variability	
Mean	84.73709	Std Deviation	9.76934
Median	87.00000	Variance	95.43999
Mode	90.00000	Range	53.00000
		Interquartile Range	10.00000

Tests for Location: Mu0=75				
Test	Statistic		p Value	
Student's t	t	14.54634	Pr > |t|	<.0001
Sign	M	78	Pr >= |M|	<.0001
Signed Rank	S	8743.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.878002	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.149233	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.194293	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7.011716	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	98
99%	98
95%	96
90%	94
75% Q3	91

Quantiles (Definition 5)	
Level	Quantile
50% Median	87
25% Q1	81
10%	71
5%	63
1%	51
0% Min	45

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
45	866	98	752
48	943	98	784
51	753	98	828
55	817	98	831
55	792	98	873

Sample Certification Data
Fit for 2 populations
The UNIVARIATE Procedure
Variable: WRITTEN (WRITTEN)
english = 1

Moments			
N	179	Sum Weights	179
Mean	85.8938547	Sum Observations	15375
Std Deviation	6.64382083	Variance	44.1403553
Skewness	−1.178462	Kurtosis	1.92250441
Uncorrected SS	1328475	Corrected SS	7856.98324
Coeff Variation	7.73491987	Std Error Mean	0.49658248

Basic Statistical Measures			
Location		Variability	
Mean	85.89385	Std Deviation	6.64382
Median	87.00000	Variance	44.14036
Mode	86.00000	Range	38.00000
		Interquartile Range	8.00000

Tests for Location: Mu0=75				
Test	Statistic		p Value	
Student's t	t	21.93765	Pr > |t|	<.0001
Sign	M	77.5	Pr >= |M|	<.0001
Signed Rank	S	7458.5	Pr >= |S|	<.0001

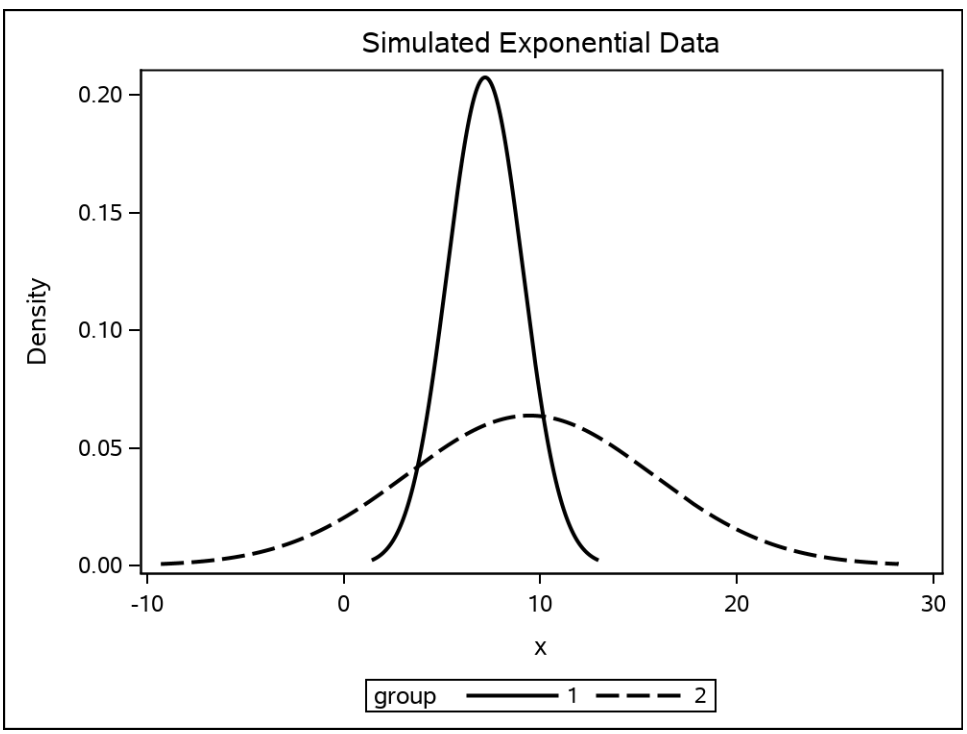
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.923474	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.116879	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.466086	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.124944	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	97
99%	96
95%	94
90%	93
75% Q3	90

Quantiles (Definition 5)	
Level	Quantile
50% Median	87
25% Q1	82
10%	78
5%	71
1%	66
0% Min	59

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
59	988	96	992
66	1091	96	1032
66	1048	96	1058
67	1056	96	1066
67	1025	97	1053

Output for SAS Simulated data



Simulated Normal Data
Descriptive Statistics
The UNIVARIATE Procedure
Variable: x

Moments			
N	100	**Sum Weights**	100
Mean	11.4915596	**Sum Observations**	1149.15596
Std Deviation	2.81544939	**Variance**	7.92675526
Skewness	−0.1346856	**Kurtosis**	−0.5457219
Uncorrected SS	13990.3429	**Corrected SS**	784.74877
Coeff Variation	24.5001505	**Std Error Mean**	0.28154494

Basic Statistical Measures			
Location		**Variability**	
Mean	11.49156	**Std Deviation**	2.81545
Median	11.81443	**Variance**	7.92676
Mode	.	**Range**	13.41163
		Interquartile Range	4.45710

Tests for Location: Mu0=75				
Test	**Statistic**		**p Value**	
Student's t	**t**	−225.571	**Pr > |t|**	<.0001
Sign	**M**	−50	**Pr >= |M|**	<.0001
Signed Rank	**S**	−2525	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.987944	**Pr < W**	0.5041
Kolmogorov-Smirnov	**D**	0.05815	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.075131	**Pr > W-Sq**	0.2411
Anderson-Darling	**A-Sq**	0.453718	**Pr > A-Sq**	>0.2500

Trimmed Means								
% in Tail	**# in Tail**	**Trimmed Mean**	**SE Trimmed Mean**	**95% CI**		**DF**	**H0:Mu0=75.00**	**Pr > |t|**
5.00	5	11.50140	0.286118	10.93289	12.06991	89	−221.932	<.0001

Winsorized Means								
% in Tail	# in Tail	Winsor Mean	SE	95% CI		DF	H0: Mu0=75.00	Pr > |t|
5.00	5	11.48208	0.286278	10.91325	12.05091	89	−221.875	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	18.18694
99%	17.74974
95%	15.59268
90%	14.88071
75% Q3	13.65445
50% Median	11.81443
25% Q1	9.19735
10%	7.62784
5%	7.01972
1%	5.21826
0% Min	4.77530

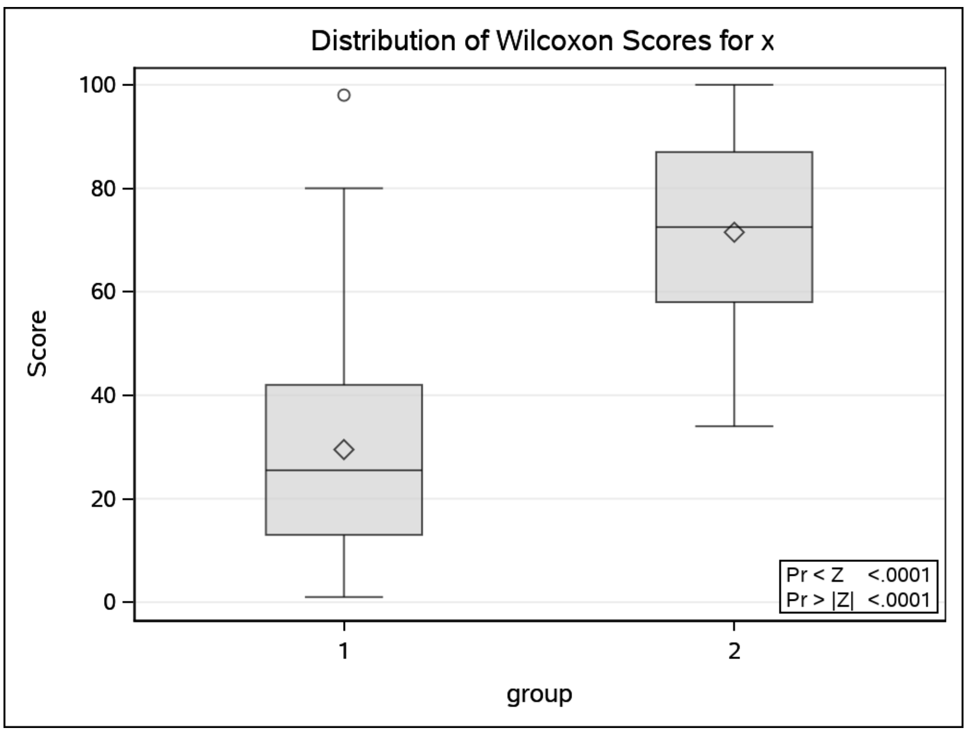
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
4.77530	13	15.7223	84
5.66122	3	16.3695	62
5.77145	33	16.5955	4
6.74925	41	17.3125	53
6.88607	43	18.1869	91

Simulated Normal Data
Nonparametric Test of Hypothesis
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable x			Classified by Variable group		
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	50	1475.0	2525.0	145.057460	29.50
2	50	3575.0	2525.0	145.057460	71.50

Wilcoxon Two-Sample Test					
Statistic	Z	Pr < Z	Pr > |Z|	t Approximation	
				Pr < Z	Pr > |Z|
1475.000	−7.2351	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

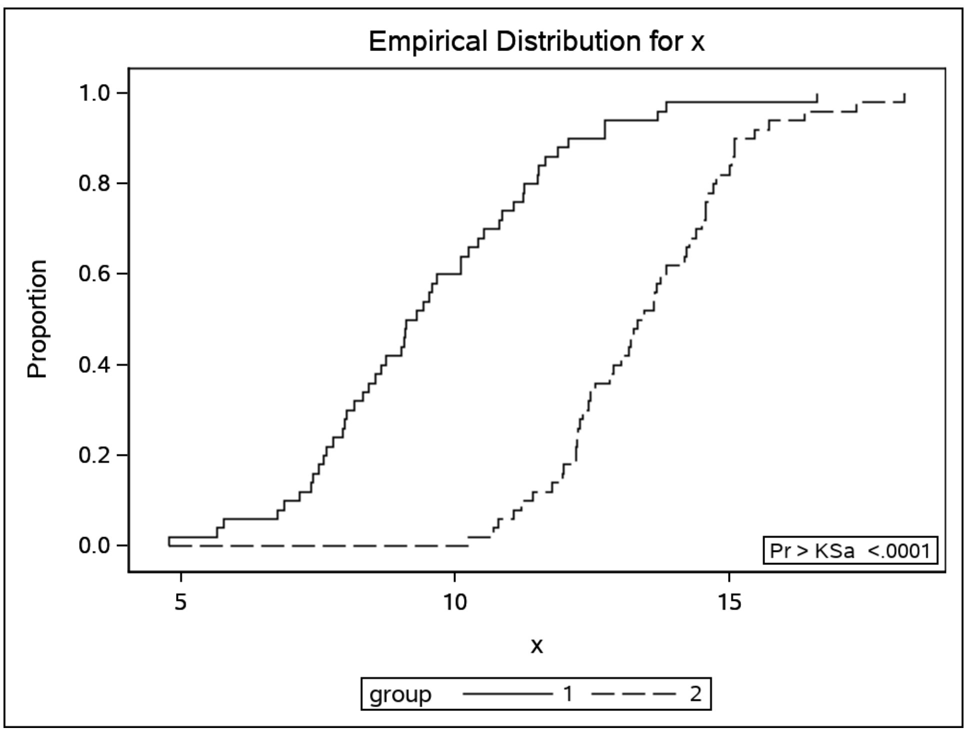
Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
52.3960	1	<.0001



Simulated Normal Data
Nonparametric Test of Hypothesis
The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable x Classified by Variable group			
group	**N**	**EDF at Maximum**	**Deviation from Mean at Maximum**
1	50	0.860	2.616295
2	50	0.120	−2.616295
Total	100	0.490	
Maximum Deviation Occurred at Observation 8			
Value of x at Maximum = 11.631965			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.370000	**D**	0.740000
KSa	3.700000	**Pr > KSa**	<.0001



Cramer-von Mises Test for Variable x Classified by Variable group		
group	**N**	**Summed Deviation from Mean**
1	50	2.83750
2	50	2.83750

Cramer-von Mises Statistics (Asymptotic)			
CM	0.056750	**CMa**	5.675000

Kuiper Test for Variable x Classified by Variable group		
group	**N**	**Deviation from Mean**
1	50	0.740
2	50	0.000

Kuiper Two-Sample Test (Asymptotic)					
K	0.740000	**Ka**	3.700000	**Pr > Ka**	<.0001

Simulated Normal Data
T Test
The TTEST Procedure

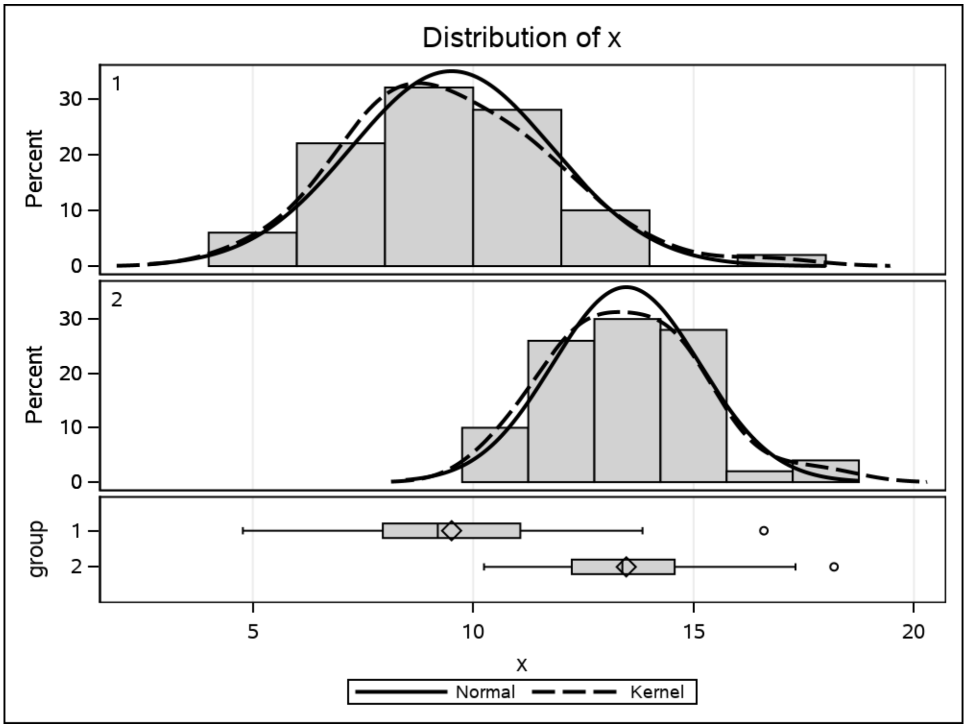
Variable: x

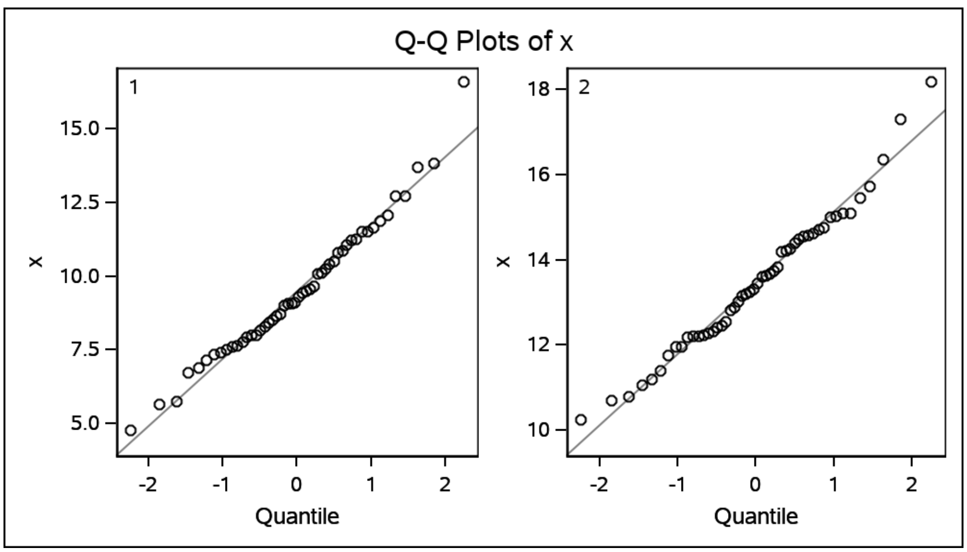
group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		50	9.5104	2.2847	0.3231	4.7753	16.5955
2		50	13.4728	1.6688	0.2360	10.2461	18.1869
Diff (1-2)	Pooled		−3.9624	2.0006	0.4001		
Diff (1-2)	Satterthwaite		−3.9624		0.4001		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		9.5104	8.8610	10.1597	2.2847	1.9085	2.8471
2		13.4728	12.9985	13.9470	1.6688	1.3940	2.0795
Diff (1-2)	Pooled	−3.9624	−4.7564	−3.1684	2.0006	1.7554	2.3259
Diff (1-2)	Satterthwaite	−3.9624	−4.7573	−3.1675			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	98	−9.90	<.0001
Satterthwaite	Unequal	89.698	−9.90	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	49	49	1.87	0.0300


Distribution of x
1
2
Percent
group
x
Normal
Kernel


Q-Q Plots of x
1
2
x
Quantile

Simulated Normal Data

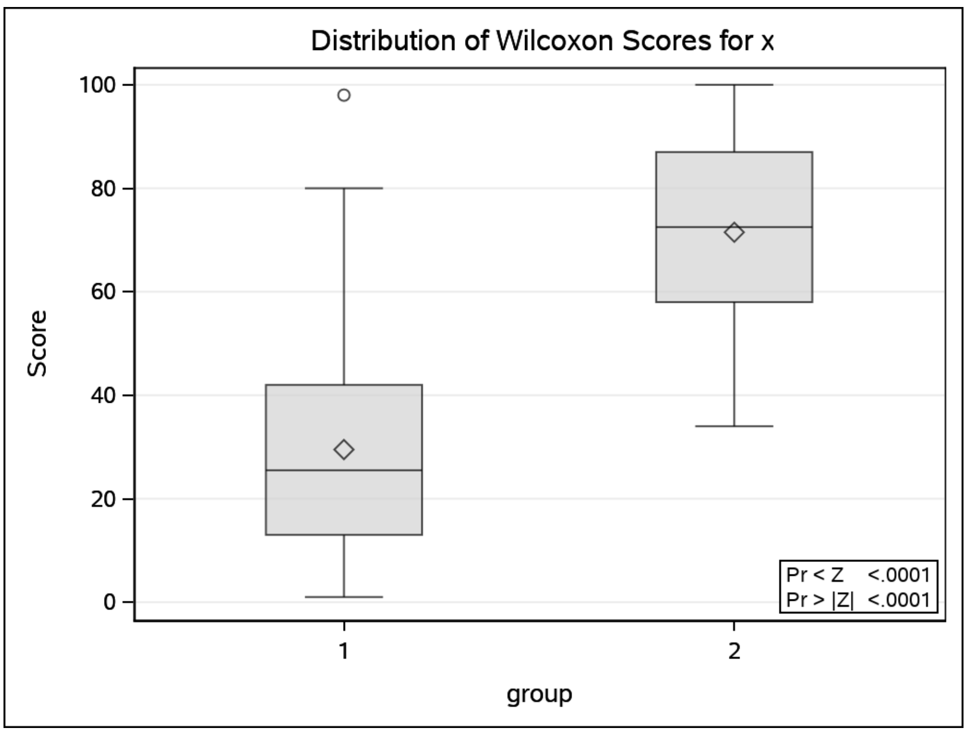
Nonparametric Tests for 2 populations

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable x			Classified by Variable group		
group	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
1	50	1475.0	2525.0	145.057460	29.50
2	50	3575.0	2525.0	145.057460	71.50

Wilcoxon Two-Sample Test					
Statistic	**Z**	**Pr < Z**	**Pr > |Z|**	**t Approximation**	
				Pr < Z	**Pr > |Z|**
1475.000	−7.2351	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	**DF**	**Pr > ChiSq**
52.3960	1	<.0001



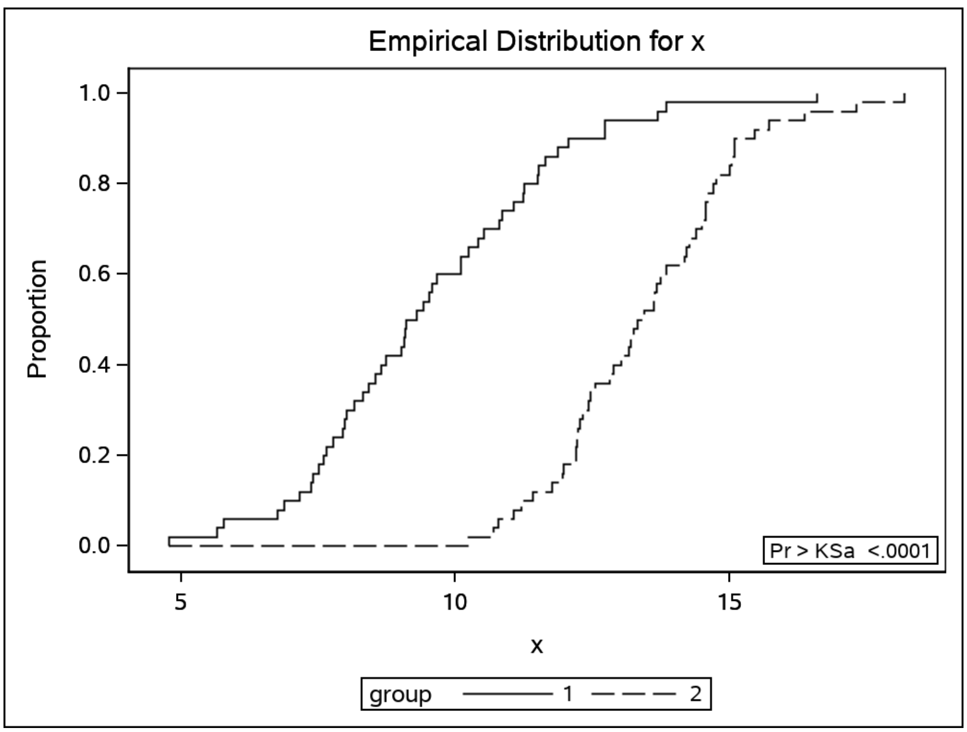
Simulated Normal Data

Nonparametric Tests for 2 populations

The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable x Classified by Variable group			
group	**N**	**EDF at Maximum**	**Deviation from Mean at Maximum**
1	50	0.860	2.616295
2	50	0.120	−2.616295
Total	100	0.490	
Maximum Deviation Occurred at Observation 8			
Value of x at Maximum = 11.631965			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.370000	**D**	0.740000
KSa	3.700000	**Pr > KSa**	<.0001



Cramer-von Mises Test for Variable x Classified by Variable group		
group	**N**	**Summed Deviation from Mean**
1	50	2.83750
2	50	2.83750

Cramer-von Mises Statistics (Asymptotic)			
CM	0.056750	CMa	5.675000

Kuiper Test for Variable x Classified by Variable group		
group	N	Deviation from Mean
1	50	0.740
2	50	0.000

Kuiper Two-Sample Test (Asymptotic)					
K	0.740000	Ka	3.700000	Pr > Ka	<.0001

Simulated Normal Data
Fit for 2 populations
The UNIVARIATE Procedure
Variable: x
group = 1

Moments			
N	50	**Sum Weights**	50
Mean	9.51035836	**Sum Observations**	475.517918
Std Deviation	2.28473136	**Variance**	5.21999739
Skewness	0.57236853	**Kurtosis**	0.80353939
Uncorrected SS	4778.12568	**Corrected SS**	255.779872
Coeff Variation	24.0236096	**Std Error Mean**	0.32310981

Basic Statistical Measures			
Location		**Variability**	
Mean	9.510358	**Std Deviation**	2.28473
Median	9.197347	**Variance**	5.22000
Mode	.	**Range**	11.82022
		Interquartile Range	3.11885

Tests for Location: Mu0=75				
Test	**Statistic**		**p Value**	
Student's t	**t**	−202.685	**Pr > |t|**	<.0001
Sign	**M**	−25	**Pr >= |M|**	<.0001
Signed Rank	**S**	−637.5	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.978149	**Pr < W**	0.4768
Kolmogorov-Smirnov	**D**	0.074654	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.045459	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.297529	**Pr > A-Sq**	>0.2500

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	16.59552
99%	16.59552
95%	13.69239

Quantiles (Definition 5)	
Level	**Quantile**
90%	12.39527
75% Q3	11.06938
50% Median	9.19735
25% Q1	7.95052
10%	7.01972
5%	5.77145
1%	4.77530
0% Min	4.77530

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
4.77530	13	12.7270	49
5.66122	3	12.7318	25
5.77145	33	13.6924	5
6.74925	41	13.8417	47
6.88607	43	16.5955	4

Simulated Normal Data
Fit for 2 populations
The UNIVARIATE Procedure
Variable: x
group = 2

Moments			
N	50	Sum Weights	50
Mean	13.4727608	Sum Observations	673.63804
Std Deviation	1.66875899	Variance	2.78475657
Skewness	0.44144605	Kurtosis	0.36256628
Uncorrected SS	9212.21725	Corrected SS	136.453072
Coeff Variation	12.3861695	Std Error Mean	0.23599816

Basic Statistical Measures			
Location		Variability	
Mean	13.47276	Std Deviation	1.66876
Median	13.38415	Variance	2.78476
Mode	.	Range	7.94083
		Interquartile Range	2.33377

Tests for Location: Mu0=75				
Test	Statistic		p Value	
Student's t	t	−260.711	Pr > |t|	<.0001
Sign	M	−25	Pr >= |M|	<.0001
Signed Rank	S	−637.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.979908	Pr < W	0.5481
Kolmogorov-Smirnov	D	0.069628	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.033405	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.257146	Pr > A-Sq	>0.2500

Quantiles (Definition 5)	
Level	Quantile
100% Max	18.1869
99%	18.1869
95%	16.3695

Quantiles (Definition 5)	
Level	Quantile
90%	15.2786
75% Q3	14.5709
50% Median	13.3842
25% Q1	12.2372
10%	11.3039
5%	10.7851
1%	10.2461
0% Min	10.2461

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10.2461	80	15.4630	83
10.6926	65	15.7223	84
10.7851	96	16.3695	62
11.0663	77	17.3125	53
11.2030	56	18.1869	91