Nonparametric Statistical Methods

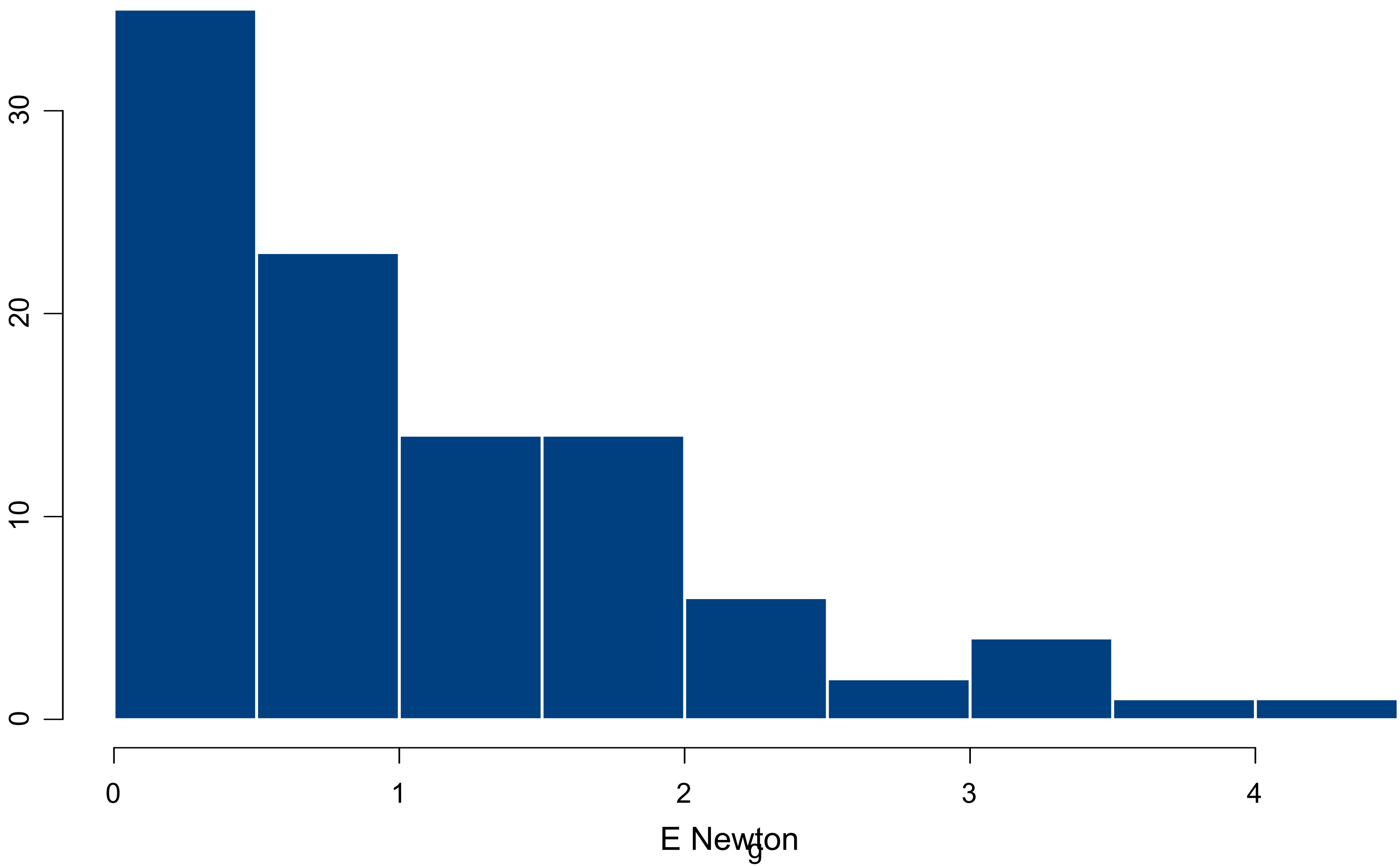
Corresponds to Chapter 14 of Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT)

Nonparametric Methods

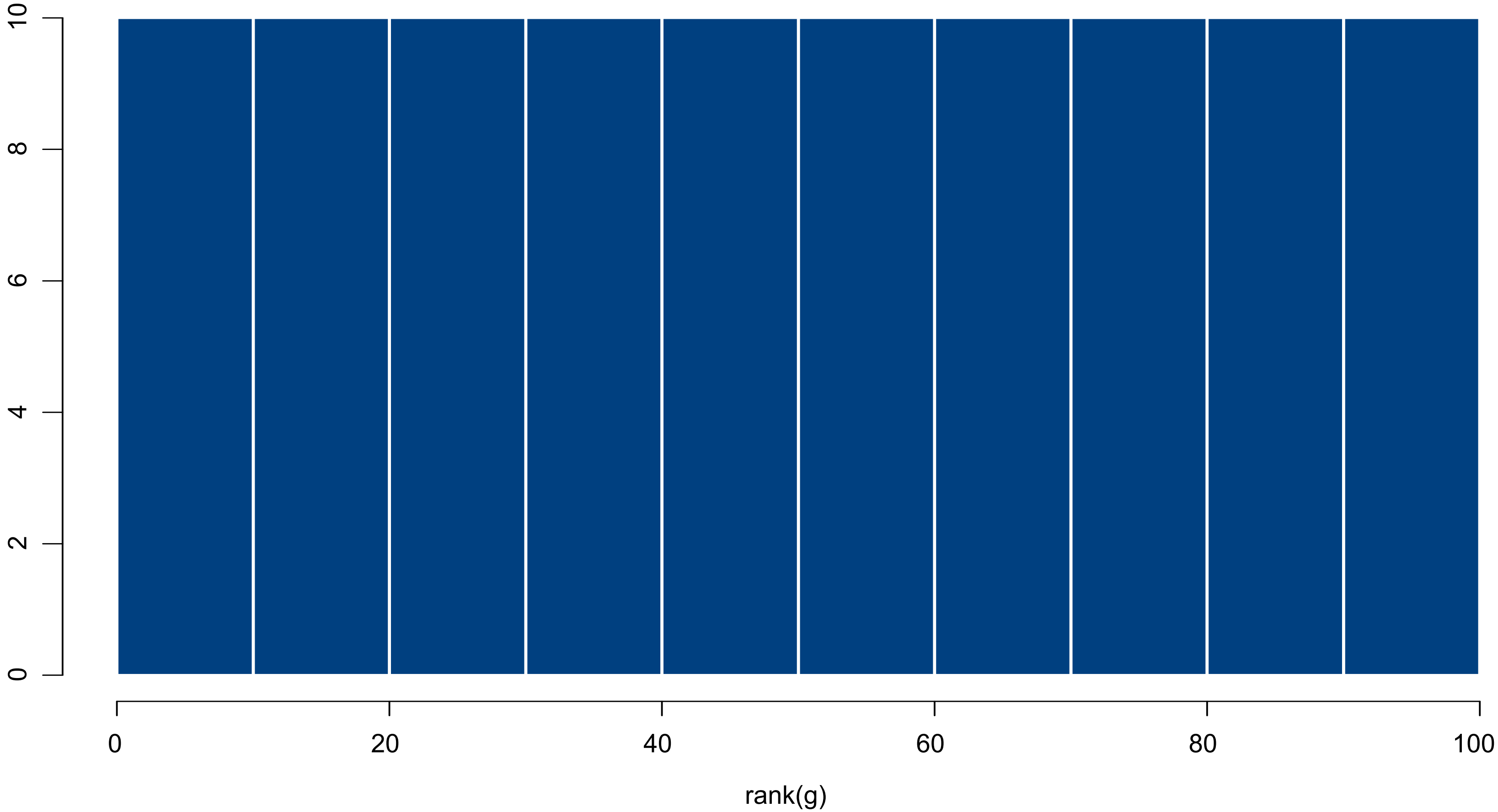
- Most NP methods are based on ranks instead of original data
- Reference: Hollander & Wolfe, Nonparametric Statistical Methods

Histogram of 100 gamma(1,1) r.v.'s



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Histogram of ranks of 100 r.v.'s



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Parametric and Nonparametric Tests

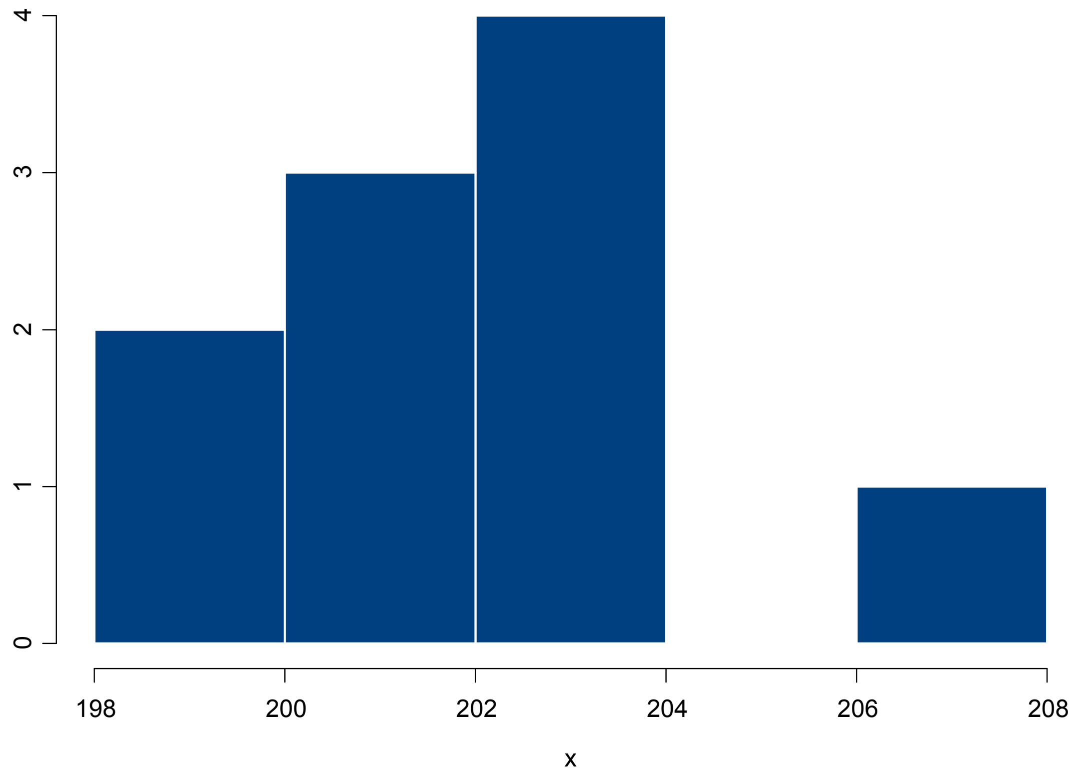
Type of test	**Parametric**	**Nonparametric**
Single Sample	z and t tests	Sign test
		Wilcoxon Signed Rank Test
Two independent samples	z and t tests	Wilcoxon Rank Sum Test
		Mann Whitney U Test

Type of test	Parametric	Nonparametric
Several Independent Samples	ANOVA CRD	Kruskal-Wallace Test
Several Matched Samples	ANOVA RBD	Friedman Test
Correlation	Pearson	Spearman Rank Correlation
		Kendall's Rank Correlation

Sign Test

- Inference on median (u) for a single sample, size n
- H_0: $u=u_0$ vs. H_1 $u \neq u_0$
- Count the number of x_i's that are greater than u_0 and denote this s+
- The number of x_i's less than u are s- = n - s+
- Reject H_0 if s+ is large or if s- is small.
- Under H_0, s+ (and s-) has binomial(n,1/2) distribution
- Large sample z test

Histogram of thermostat data



E Newton

Sign Test in S-Plus

```
> thermostat
 [1] 202.2 203.4 200.5 202.5 206.3 198.0 203.7 200.8
   201.3 199.0

> thermostat<200
 [1] F F F F F T F F F T

> sum(thermostat<200)
[1] 2

> 2*pbinom(sum(thermostat<200),10,0.5)
[1] 0.109375
```

Wilcoxon Signed Rank Test

- Inference on median (u), single sample, size n
- Assumes population distribution is symmetric
- H_0: $u=u_0$ vs. H_1 $u \neq u_0$
- $d_i = x_i - u_0$
- Rank order $|d_i|$
- W+ = sum of ranks of positive differences
- W- = sum of ranks of negative differences
- W_{max} = maximum (W+, W-)
- Reject H_0 if W_{max} is large.
- Null Distribution – see text
- Large sample z test

S-Plus wilcox.test for thermostat data

```
> thermostat
 [1] 202.2 203.4 200.5 202.5 206.3 198.0 203.7 200.8
   201.3 199.0

> sum(rank(abs(thermostat-200))[-c(6,10)])
[1] 47

> wilcox.test(thermostat,mu=200)

   Exact Wilcoxon signed-rank test

data:  thermostat
signed-rank statistic V = 47, n = 10, p-value =
   0.0488
alternative hypothesis: true mu is not equal to 200
```

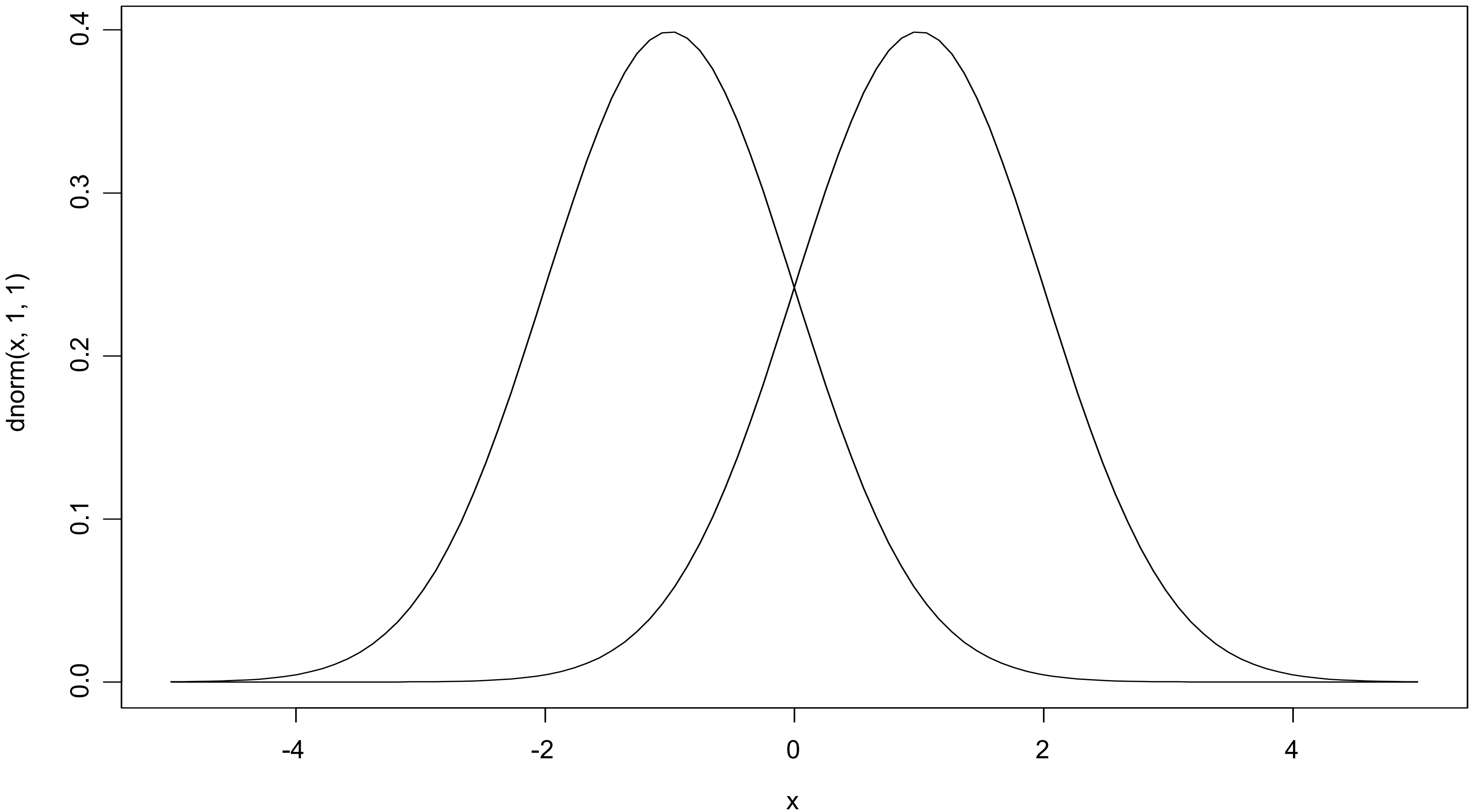
S-Plus parametric t-test for thermostat data

```
> t.test(thermostat, mu=200)

    One-sample t-Test

data:  thermostat
t = 2.3223, df = 9, p-value = 0.0453
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
 200.0459 203.4941
sample estimates:
 mean of x
    201.77
```

Location-Scale Families

- See course textbook, page 575.

2 normal pdf's with location parameters = -1 and 1, scale parameter =1



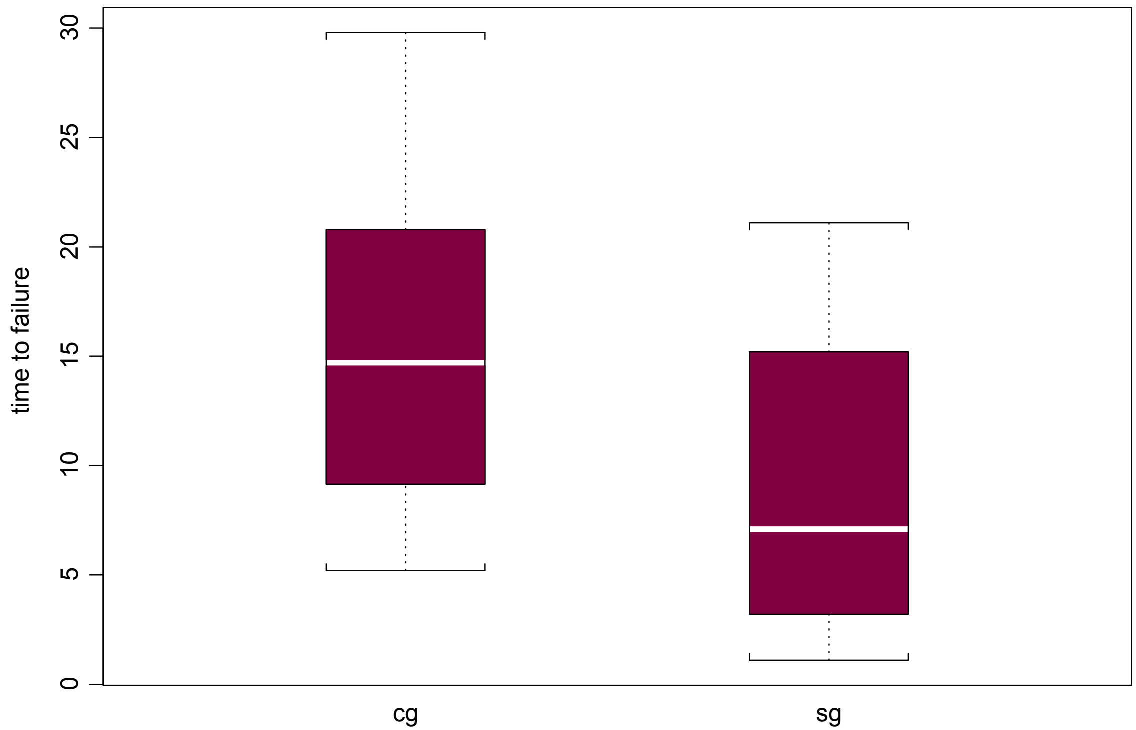
Wilcoxon Rank Sum Test

- Inference on location of distribution of 2 independent random samples X and Y (e.g. from control and treatment population).
- Assume $X \sim Y + \Delta$
- H_0: $\Delta = 0$ vs. H_1: $\Delta \neq 0$
- Rank all N = n1 + n2 observations
- W=sum of ranks assigned to the Y's (or X's, whichever has smaller sample size)
- Reject H_0 if W is extreme

Mann-Whitney U test

- Equivalent to Wilcoxon rank sum test
- Compare each x_i with each y_i.
- There are n_x*n_y such comparisons
- U= number of pairs in which $x_i<y_i$.
- Icbst $W = U + (n*(n+1))/2$ (when no ties)
- Reject H_0 if U is extreme.

Boxplots of times to failure for control and stressed capacitors



S-Plus wilcox.test

```
> wilcox.test(cg, sg)

   Exact Wilcoxon rank-sum test

data:  cg and sg
rank-sum statistic W = 95, n = 8, m = 10, p-value =
   0.1011
alternative hypothesis: true mu is not equal to 0
```

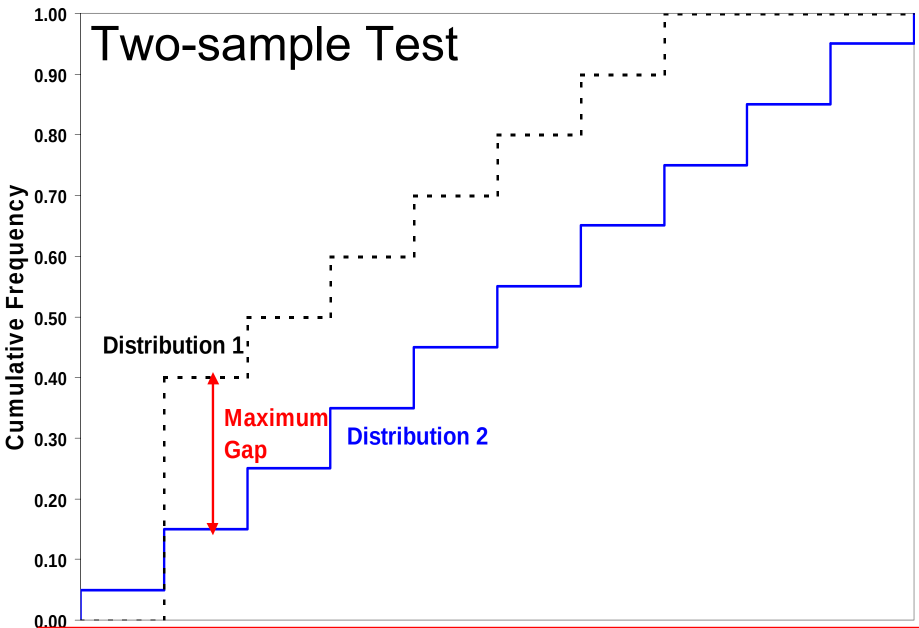
S-Plus parametric t-test

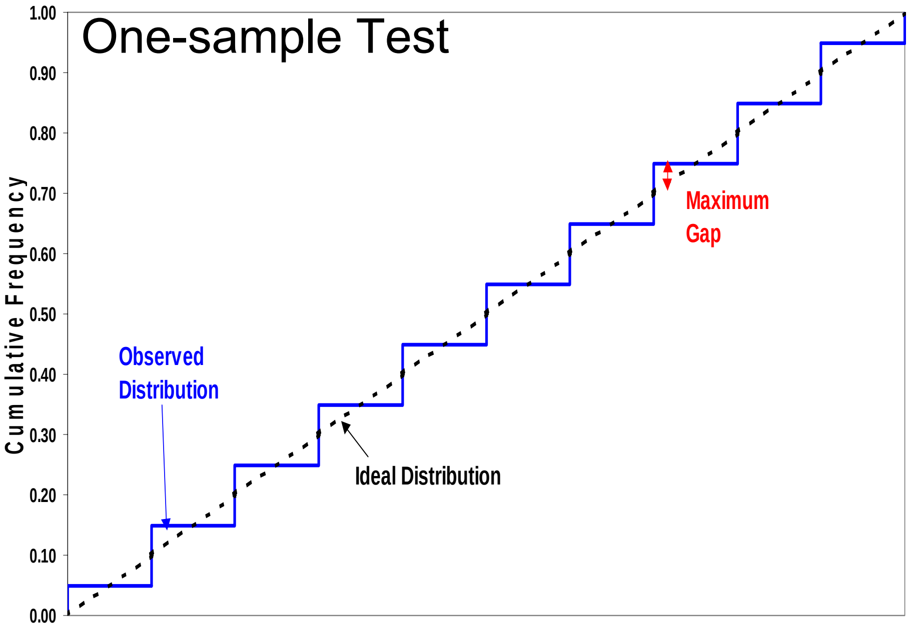
```
> t.test(cg,sg)

    Standard Two-Sample t-Test

data:  cg and sg
t = 1.8105, df = 16, p-value = 0.089
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.103506  14.018506
sample estimates:
 mean of x mean of y
   15.5375      9.08
```

Kolmogorov-Smirnov Tests

The Kolmogorov-Smirnov test detects differences in location, scale, skewness, or whatever (any differences between two distributions), uses two empirical cumulative distribution functions (step functions).

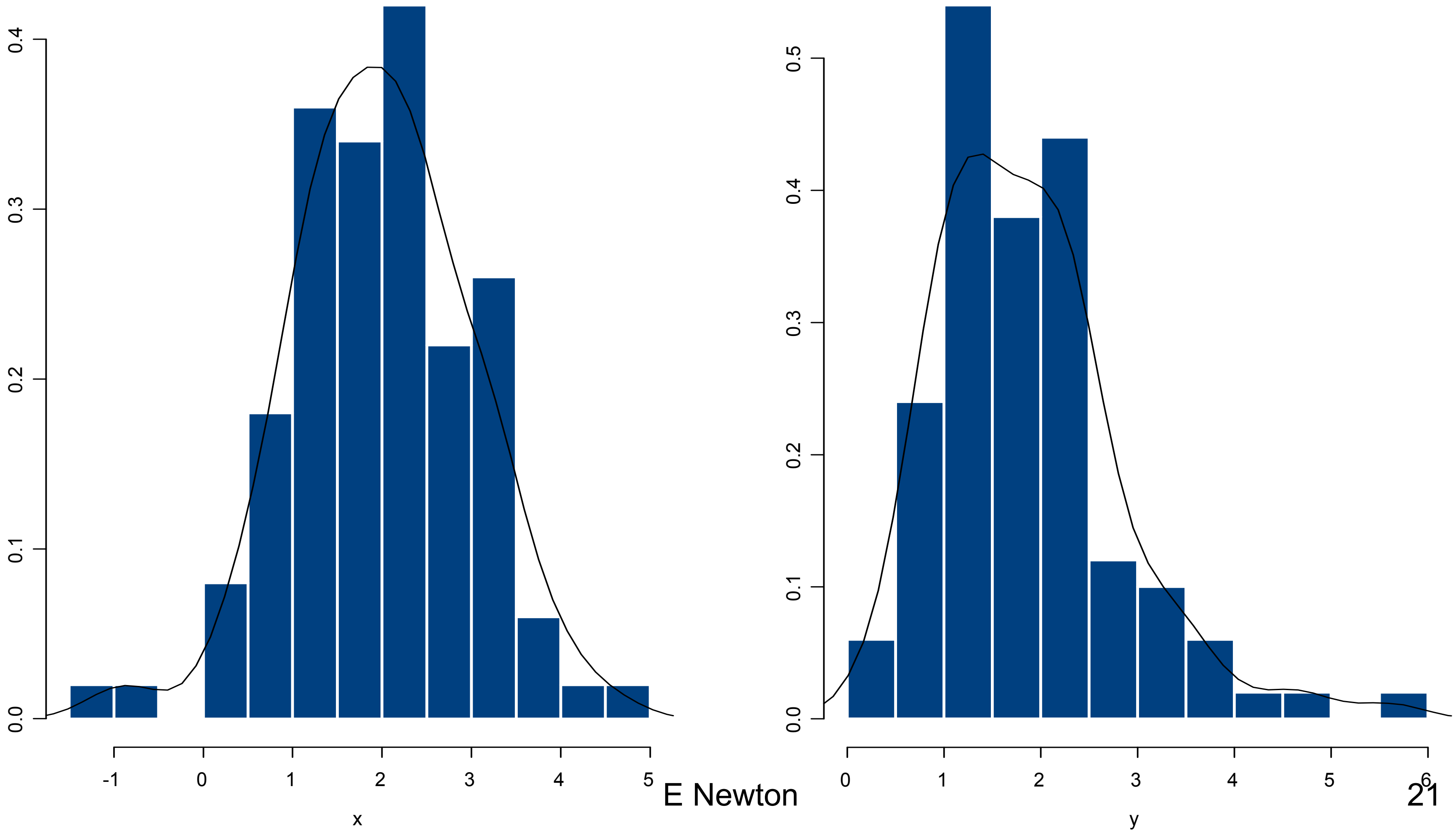
There is also a one-sample version for testing the distance between some observed data and a specified (ideal) distribution.





Tests the maximum gap between the observed distribution and the hypothesized distribution as a function of sample size (tables or p-values).

Histograms of 100 random normal (2,1) deviates and 100 random gamma(4,2) deviates



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Kolmogorov-Smirnov Tests

```
> ks.gof(x,y)

   Two-Sample Kolmogorov-Smirnov Test

data:  x and y
ks = 0.15, p-value = 0.2112
alternative hypothesis: cdf of x does not equal the
              cdf of y for at least one sample point.


> ks.gof(y)

   One sample Kolmogorov-Smirnov Test of Composite Normality

data:  y
ks = 0.0969, p-value = 0.0216
alternative hypothesis: True cdf is not the normal distn. with
   estimated parameters
sample estimates:
 mean of x standard deviation of x
  1.865857               0.9421928
```

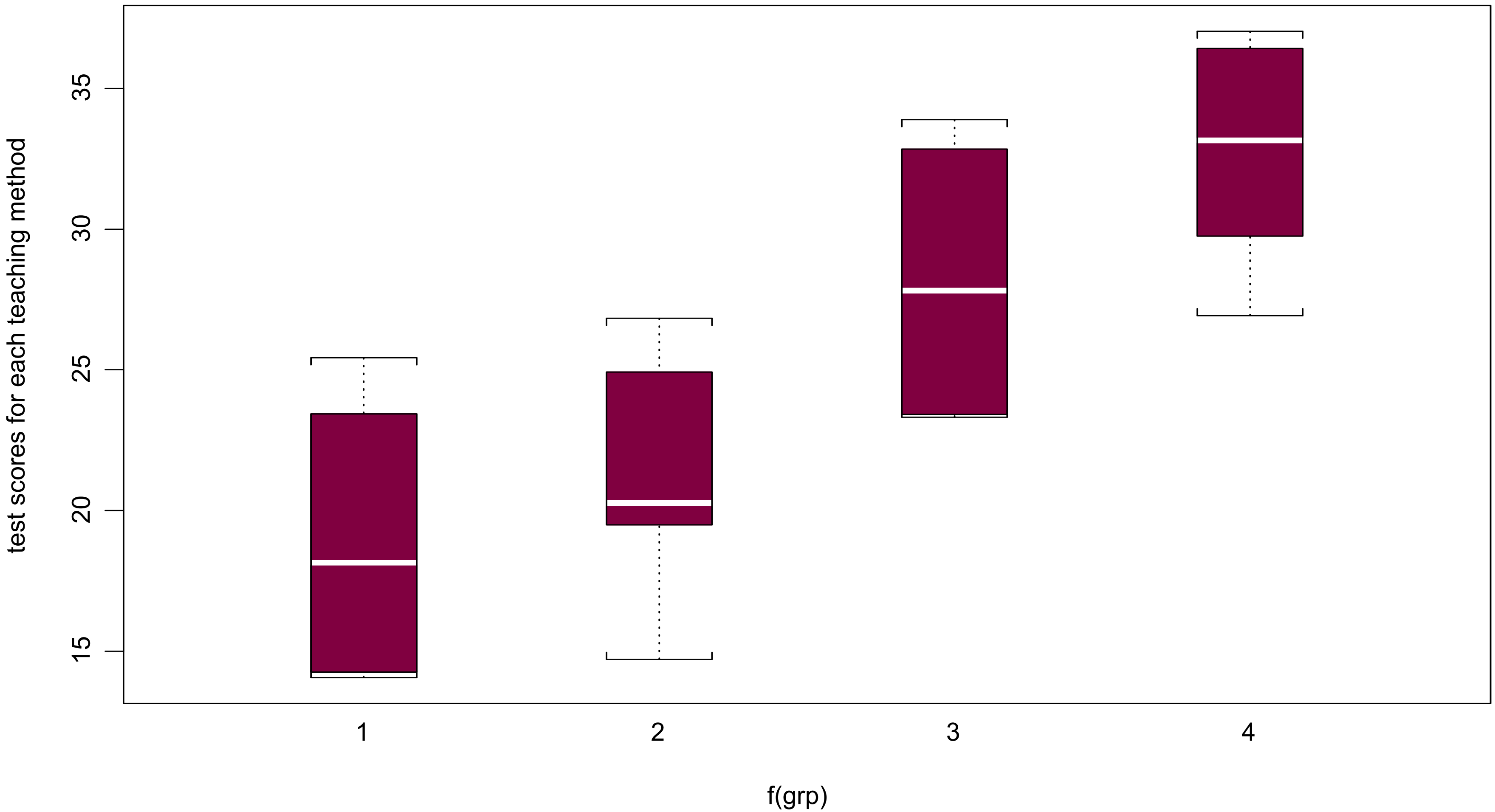
Kruskal-Wallis Test

- Inference for several independent samples
- Assume distributions of each of the samples differ only possibly in location.
- $X_{ij} = \theta + \tau_j + e_{ij}$.
- H_0: $\tau_1=\tau_2=..= \tau_k$, vs. H_1: $\tau_i \neq \tau_j$ for some $i \neq j$
- Rank all $N=n_1+n_2..+n_a$ observations.
- Calculate rank sums and averages in each group
- Calculate KW test statistic=kw (see text)
- Reject H_0 for large values of kw
- For large n_i's, null dist'n of kw χ^2_{a-1}

Test scores for four different teaching methods (page 582)

```
scm<-matrix(score,7,4)
> scm
      [,1]  [,2]  [,3]  [,4]
[1,] 14.06 14.71 23.32 26.93
[2,] 14.26 19.49 23.42 29.76
[3,] 14.59 20.20 24.92 30.43
[4,] 18.15 20.27 27.82 33.16
[5,] 20.82 22.34 28.68 33.88
[6,] 23.44 24.92 32.85 36.43
[7,] 25.43 26.84 33.90 37.04
```

Plot.factor(f(grp),score)



Ranks of Test Scores

```
> scmr<-matrix(rank(score),7,4)
> scmr
     [,1] [,2] [,3] [,4]
[1,]    1  4.0 11.0   18
[2,]    2  6.0 12.0   21
[3,]    3  7.0 14.5   22
[4,]    5  8.0 19.0   24
[5,]    9 10.0 20.0   25
[6,]   13 14.5 23.0   27
[7,]   16 17.0 26.0   28

> tmp<-apply(scmr,2,sum)
> tmp
[1]  49.0  66.5 125.5 165.0

> (12/(28*29))*sum((tmp^2)/7)-3*29
[1] 18.13406
```

Kruskal-Wallis test in S-Plus

```
> kruskal.test(scm, col(scm))

  Kruskal-Wallis rank sum test

data:  scm and col(scm)
Kruskal-Wallis chi-square = 18.139, df = 3,
  p-value = 0.0004
alternative hypothesis: two.sided
```

ANOVA for test scores

```
summary(aov(score~f(grp)))
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
   f(grp)  3  830.1914 276.7305 15.93607 6.509182e-006
Residuals 24  416.7609  17.3650
```

Friedman Test

- Inference for several matched samples
- a treatments, b blocks
- H_0: $\tau_1=\tau_2=..=\tau_k$, vs. H_1: $\tau_i \neq \tau_j$ for some $i \neq j$
- Rank observations separately within each block
- Calculate rank sums
- Calculate the Friedman statistic, fr (see text)
- Reject H_0 for large values of fr
- For b large, fr $\sim \chi^2_{a-1}$

Ranks within Blocks (rows)

```
> scmrb<-t(apply(scm,1,rank))
> scmrb
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    1    2    3    4
[3,]    1    2    3    4
[4,]    1    2    3    4
[5,]    1    2    3    4
[6,]    1    2    3    4
[7,]    1    2    3    4

> tmp<-apply(scmrb,2,sum)
[1]  7 14 21 28

> (12/(4*7*5))*sum(tmp^2)-3*7*5
[1] 21
```

Friedman test in S-Plus

- > friedman.test(scm, col(scm), row(scm))
- Friedman rank sum test
- data: scm and col(scm) and row(scm)
- Friedman chi-square = 21, df = 3, p-value = 0.0001
- alternative hypothesis: two.sided

ANOVA test score data with blocks

```
> summary(aov(score~f(grp)+f(blk)))
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
   f(grp)  3  830.1914 276.7305 260.4768 5.220000e-015
   f(blk)  6  397.6377  66.2729  62.3804 4.558276e-011
Residuals 18   19.1232   1.0624
```

Correlation Methods

- Pearson Correlation: measures only linear association.
- Spearman Correlation: correlation of the ranks
- Kendall's Tau: based on number of concordant and discordant pairs.

Kendall's Tau

- Assume: the n bivariate observations $(X_1,Y_1),\ldots,(X_n,Y_n)$ are a random sample from a continuous bivariate population.
- H_0: X_i, Y_i are independent
- H_0: $F(x,y) = F(x)F(y)$
- Measure dependence by finding the number of concordant and discordant pairs.
- Population correlation coefficient:

  $\tau = 2*P\{X_2-X_1)(Y_2-Y_1)>0\}-1$

Kendall's Tau

$\textit{For } 1 \le \mathrm{i} < \mathrm{j} \le \mathrm{n}:$

$$Q((X_i, Y_i),(X_j, Y_j)) = \begin{cases} 1, \text{if } (X_i - X_j)(Y_i - Y_j) > 0 \\ 0, \text{if } (X_i - X_j)(Y_i - Y_j) = 0 \\ -1, \text{if } (X_i - X_j)(Y_i - Y_j) < 0 \end{cases}$$

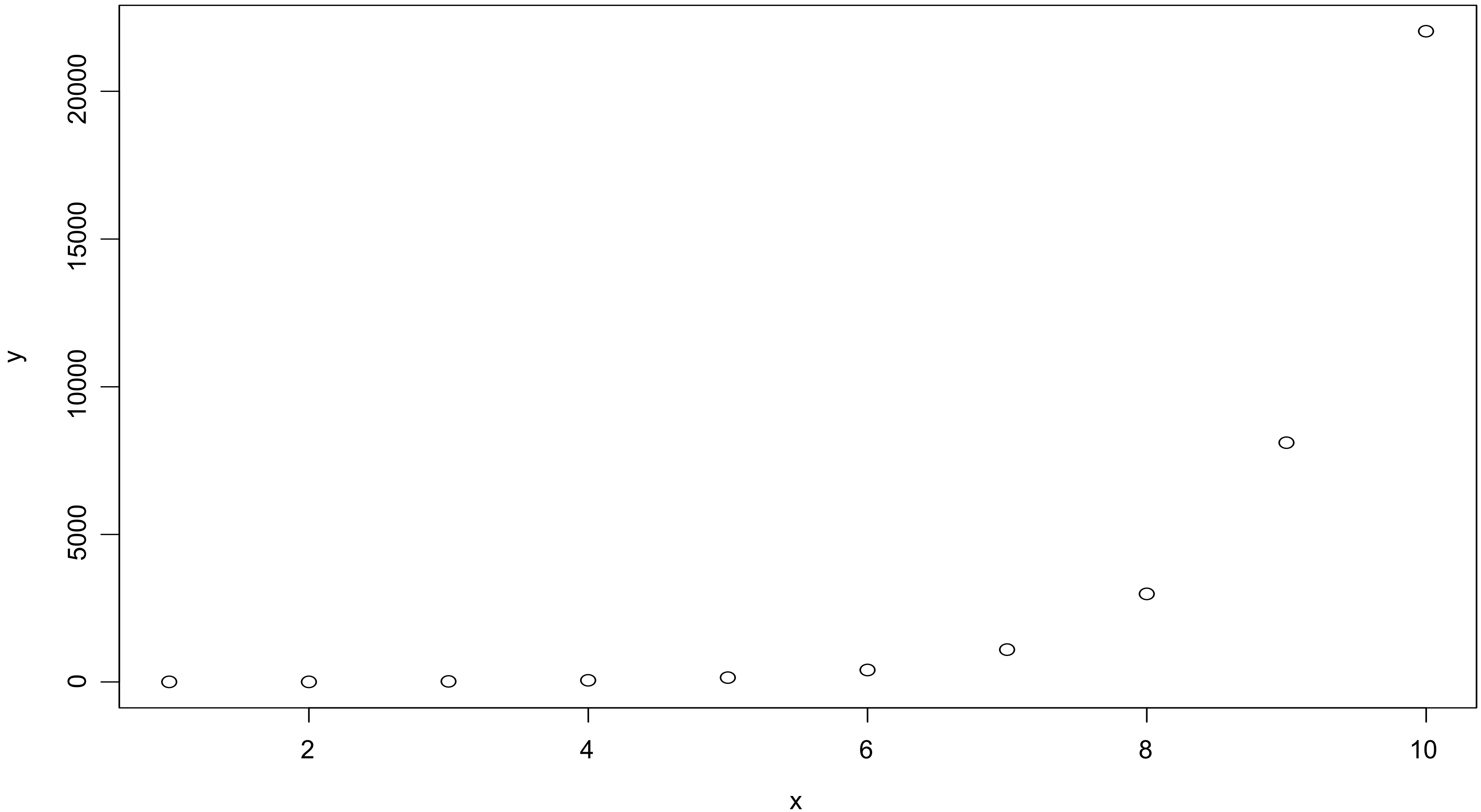
$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((X_i, Y_i),(X_j, Y_j))$$

$$\hat{\tau} = \frac{2K}{n(n-1)}$$

Kendall's Tau example

```
> m
   1  3  2  4
1 NA  1  1  1
2 NA NA -1  1
3 NA NA NA  1
4 NA NA NA NA

> 2*sum(m,na.rm=T)/12
[1] 0.6666667

> cor.test(c(1,2,3,4),c(1,3,2,4),method="k")

   Kendall's rank correlation tau

data:  c(1, 2, 3, 4) and c(1, 3, 2, 4)
normal-z = 1.3587, p-value = 0.1742
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
 0.6666667
```

x=1:10
y=exp(x)



Pearson Correlation

```
> cor.test(x,y,method="p")

  Pearson's product-moment correlation

data:  x and y
t = 2.9082, df = 8, p-value = 0.0196
alternative hypothesis: true coef is not
  equal to 0
sample estimates:
       cor
 0.7168704
```

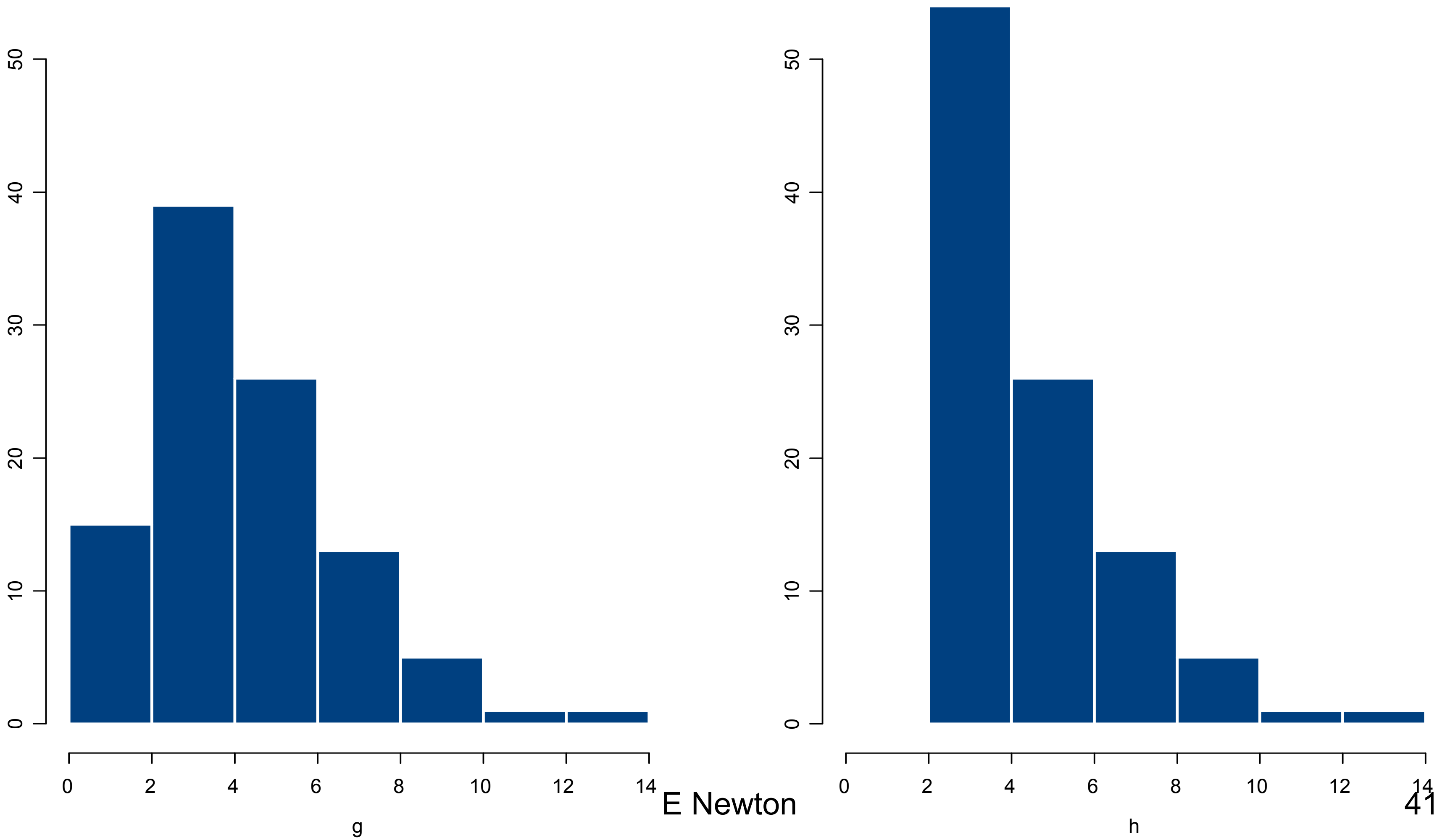
Spearman Correlation

```
> cor.test(x,y,method="s")

  Spearman's rank correlation

data:  x and y
normal-z = 2.9818, p-value = 0.0029
alternative hypothesis: true rho is not
  equal to 0
sample estimates:
 rho
   1
```

Kendall Correlaton

```
> cor.test(x,y,method="k")

  Kendall's rank correlation tau

data:  x and y
normal-z = 4.0249, p-value = 0.0001
alternative hypothesis: true tau is not
  equal to 0
sample estimates:
 tau
   1
```

Example - Environmental Data – Censored below LOD



E Newton

Resampling Methods

- Parametric methods – Inference based on assumed population distribution
- Resampling methods – No assumption about functional form of population distribution.
- Permutation Tests – 2 sample problem
- Jackknife – Delete one observation at a time
- Bootstrap – resample with replacement

Permulation Tests

- Goal: estimate difference in means (2 sample problem)
- $(x_1, x_2 \ldots x_{n1})$ and $(y_1, y_2 .. y_{n2})$ are independent samples drawn from F_1 and F_2.
- H_0: $F_1=F_2$ => all assignments of labels x and y equally likely.
- Choose SRS of size n1 from n1+n2 observations and label as x, label rest as y.
- Calculate value of test statistic (e.g. difference in means) for each assignment -> permutation distribution.
- There are (n1+n2) choose (n1) possible distinct assignments (capacitor data set Ex14.7, n1=8, n2=10, number of assignments=43,758)

Jackknife

- Goal: estimate distribution and standard error of statistic (e.g. median or mean)
- Draw n samples of size n-1 from original sample, by deleting one observation at a time.
- Calculate m_j^*=mean (median) from each sample

$$JSE(m) = \sqrt{\frac{n-1}{n} \sum_{j=1}^{n} (m_j^* - \overline{m}^*)^2}$$

- JSE is exact for mean, not necessarily very good for median

Bootstrap

- Goal: estimate distribution, standard error, confidence interval of statistic (e.g. mean, median, correlation)
- Draw B samples of size n, with replacement, from original sample
- Calculate test statistics from each sample

$$BSE(m) = \sqrt{\frac{\sum_{j=1}^{B}(m_j^* - \overline{m}^*)^2}{B-1}}$$

Swiss Data Set in S-Plus

Fertility Data for Switzerland in 1888

SUMMARY:

The swiss.fertility and swiss.x data sets contain fertility data for Switzerland in 1888.

ARGUMENTS:

swiss.fertility

standardized fertility measure I[g] for each of 47 French-speaking provinces of Switzerland in approximately 1888.

swiss.x

matrix with 5 columns that contain socioeconomic indicators for the provinces: 1) percent of population involved in agriculture as an occupation; 2) percent of "draftees" receiving highest mark on army examination; 3) percent of population whose education is beyond primary school; 4) percent of population who are Catholic; and, 5) percent of live births who live less than 1 year (infant mortality).

SOURCE:

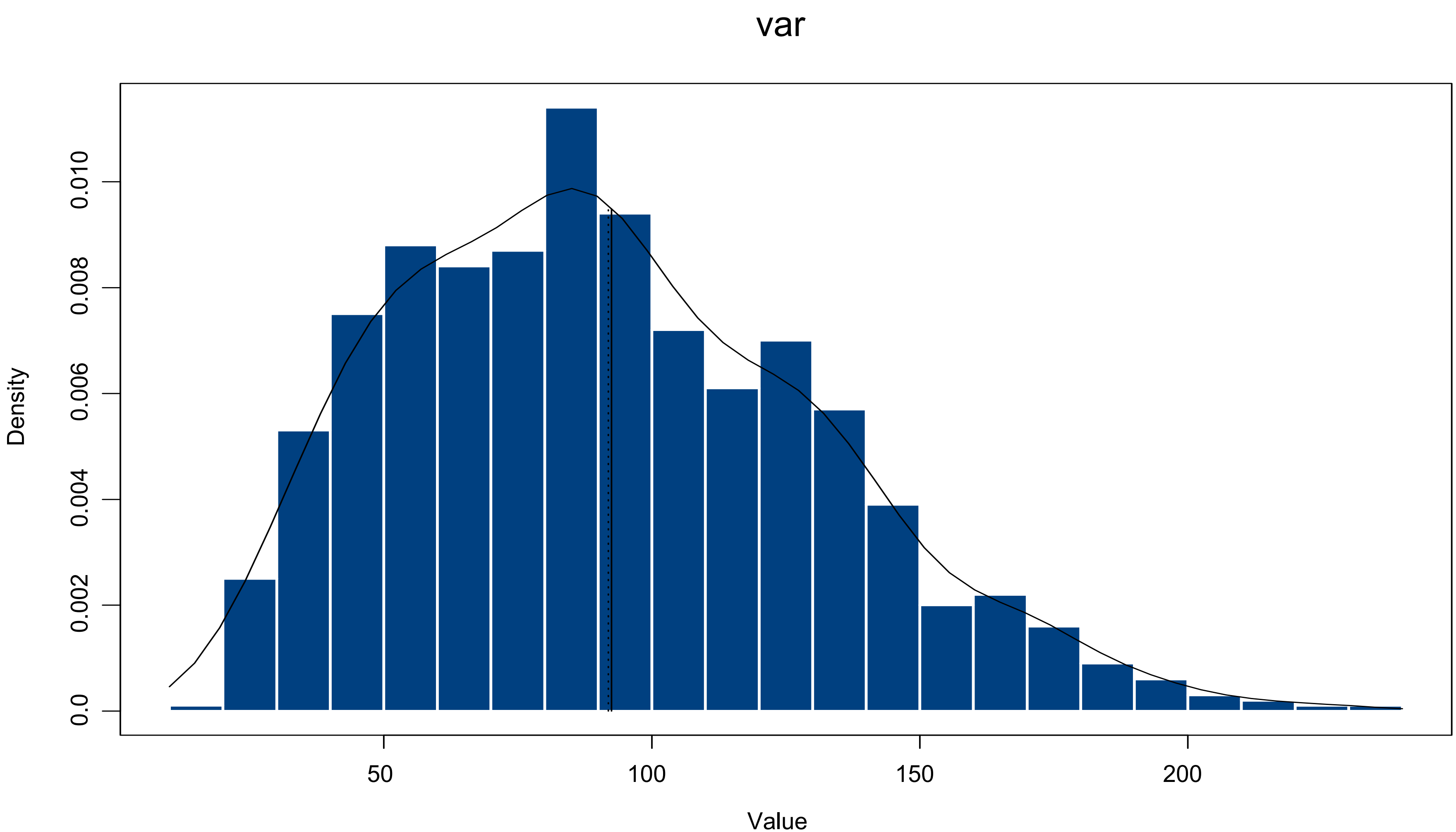
Mosteller and Tukey (1977). *Data Analysis and Regression.* Addison-Wesley.

Unpublished data used by permission of Francine van de Walle. Population Study Center, University of Pennsylvania, Philadelphia, PA.

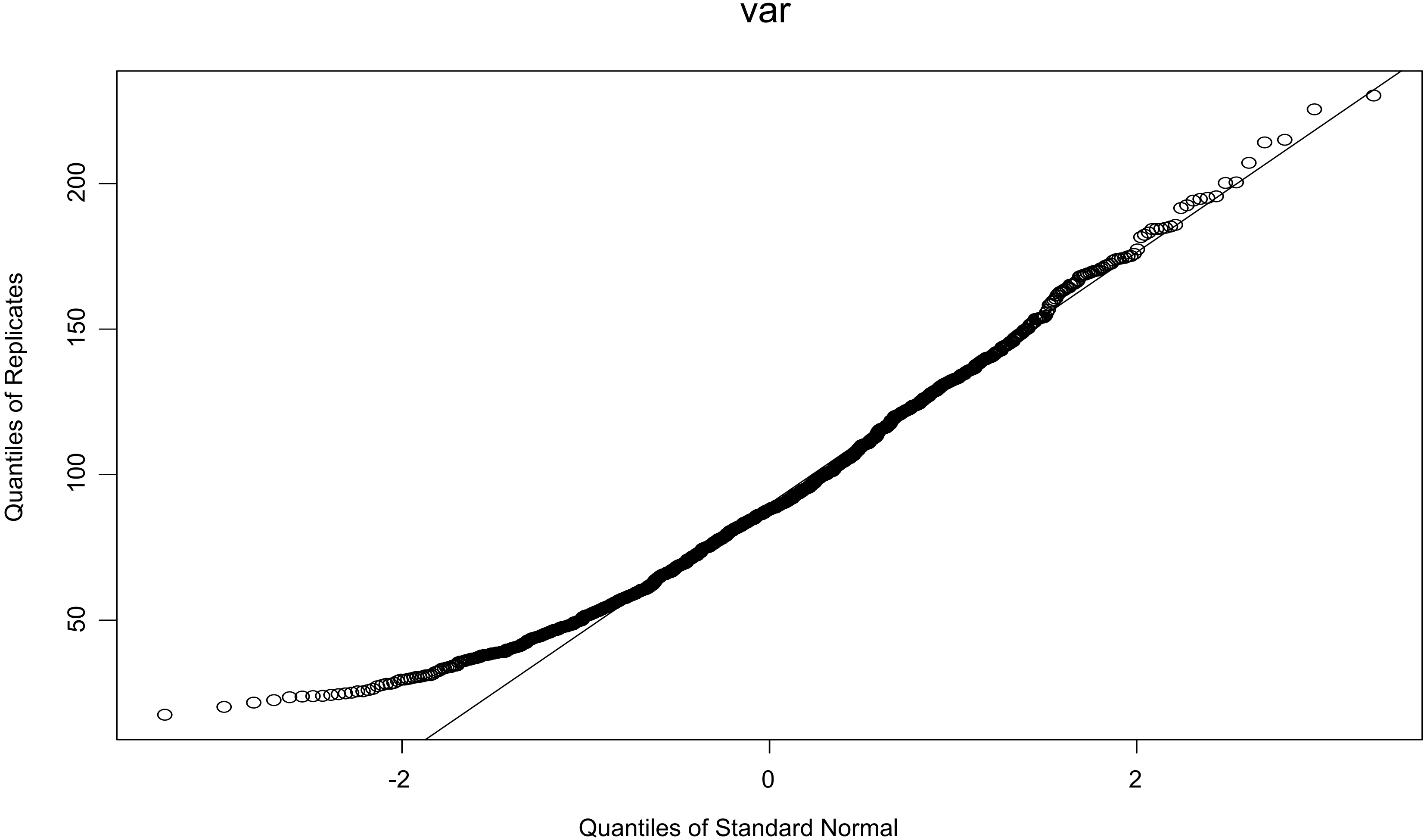
Bootstrap estimates and CI for variance of education

```
> educ<-swiss.x[,3]
> var(educ)
[1] 92.45606

> educ.boot<-bootstrap(educ,var,trace=F)
> summary(educ.boot)
Call:
bootstrap(data = educ, statistic = var, trace = F)

Number of Replications: 1000

Summary Statistics:
    Observed    Bias  Mean    SE
var    92.46 -0.5972 91.86 39.14

Empirical Percentiles:
     2.5%    5%   95% 97.5%
var 29.98 36.26 165.3   175
```

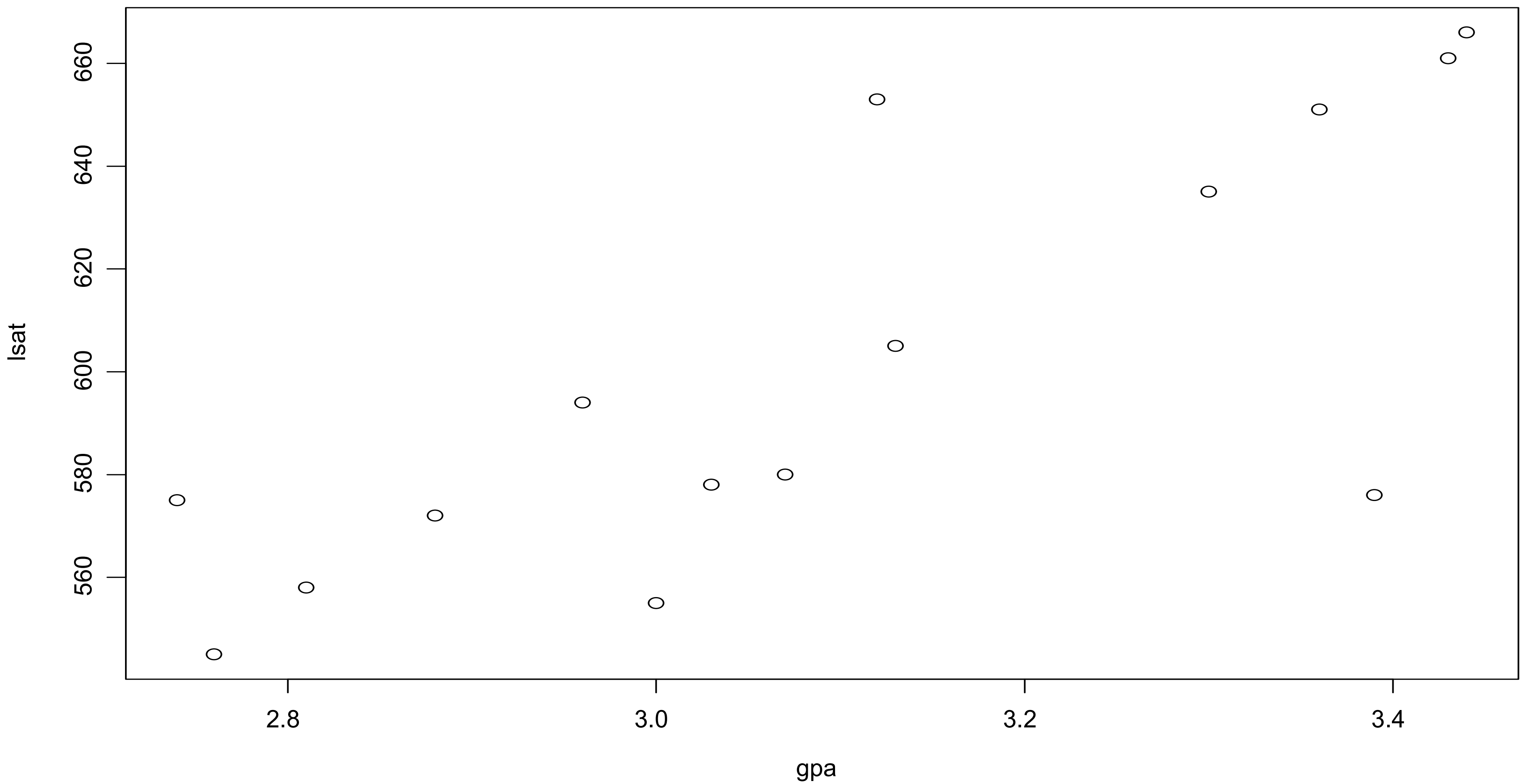
Histogram of variance estimates obtained from 1000 bootstrap samples



QQ plot of variance estimates



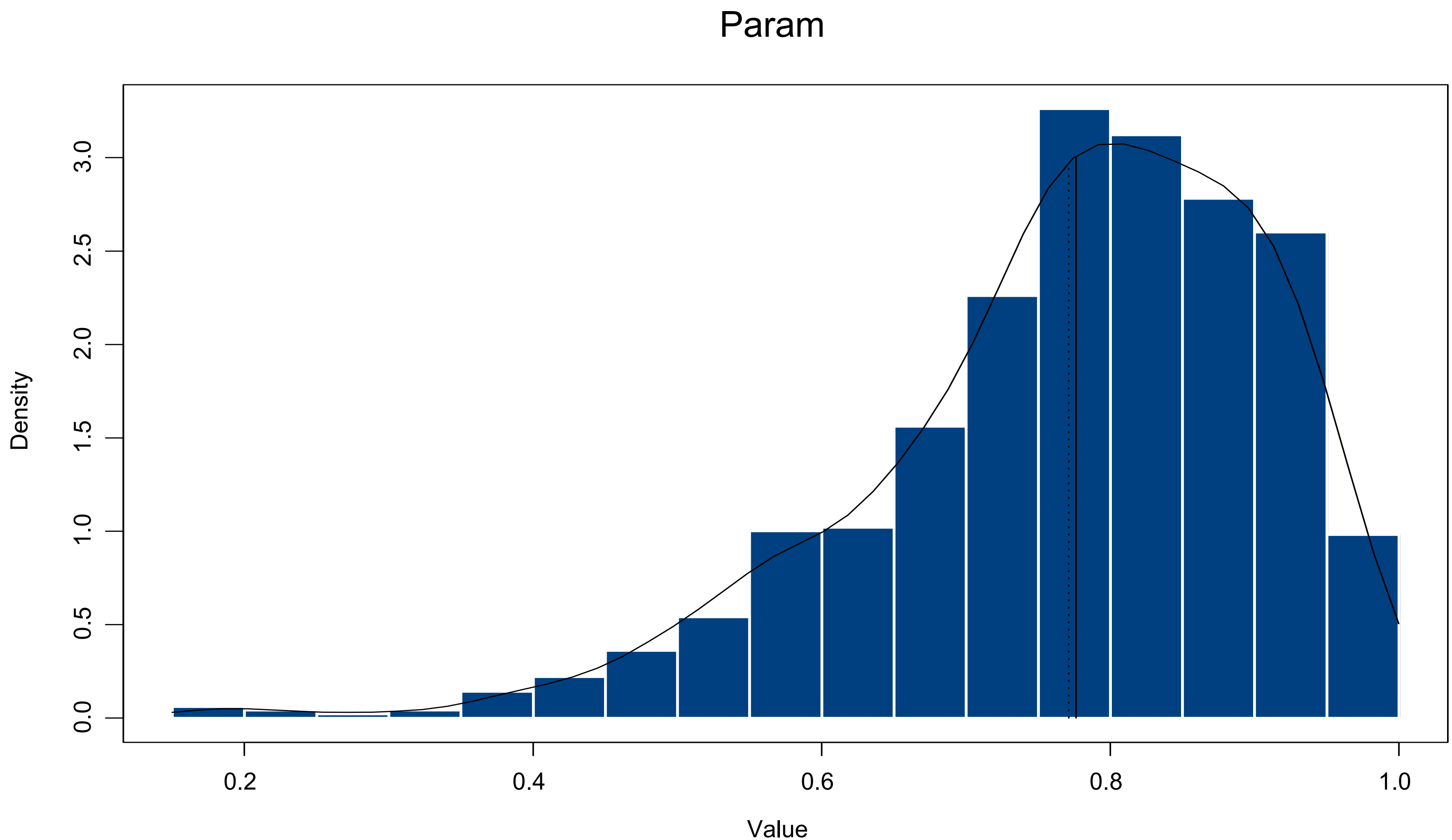
Plot of LSAT scores by GPA for a sample of 15 schools



Bootstrap estimates and CI for correlation between LSAT and GPA

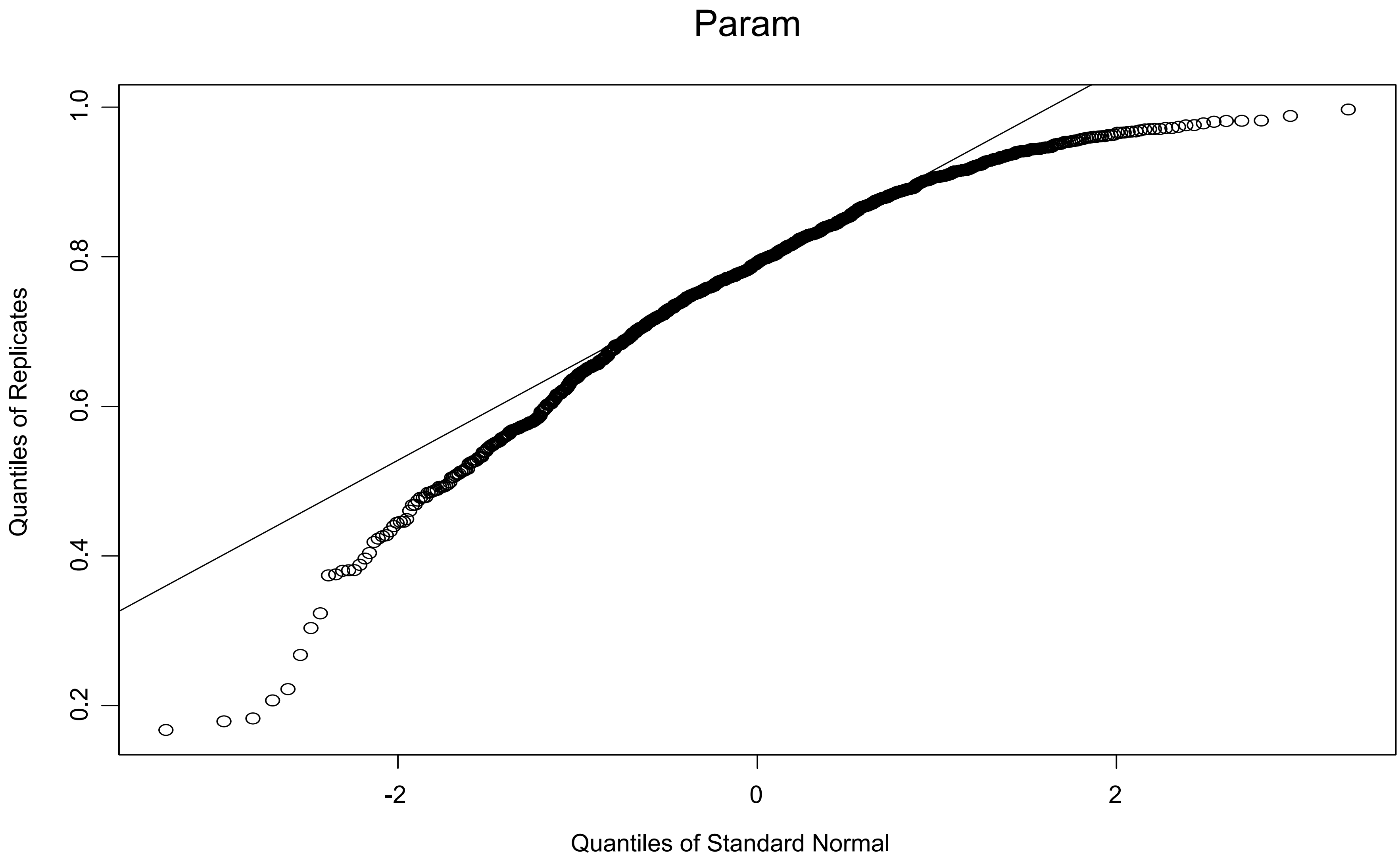
```
> law.boot<-bootstrap(law.data, cor(lsat,gpa), trace=F)
> summary(law.boot)
Call:
bootstrap(data = law.data, statistic = cor(lsat, gpa), trace = F)

Number of Replications: 1000

Summary Statistics:
      Observed     Bias   Mean     SE
Param   0.7764 -0.00506 0.7713 0.1368

Empirical Percentiles:
        2.5%     5%   95%  97.5%
Param 0.449 0.5133 0.947 0.9623

BCa Confidence Limits:
         2.5%     5%    95%  97.5%
Param 0.2623 0.4138 0.9232 0.9413
```

Histogram of correlation estimates obtained from 1000 bootstrap samples



E Newton

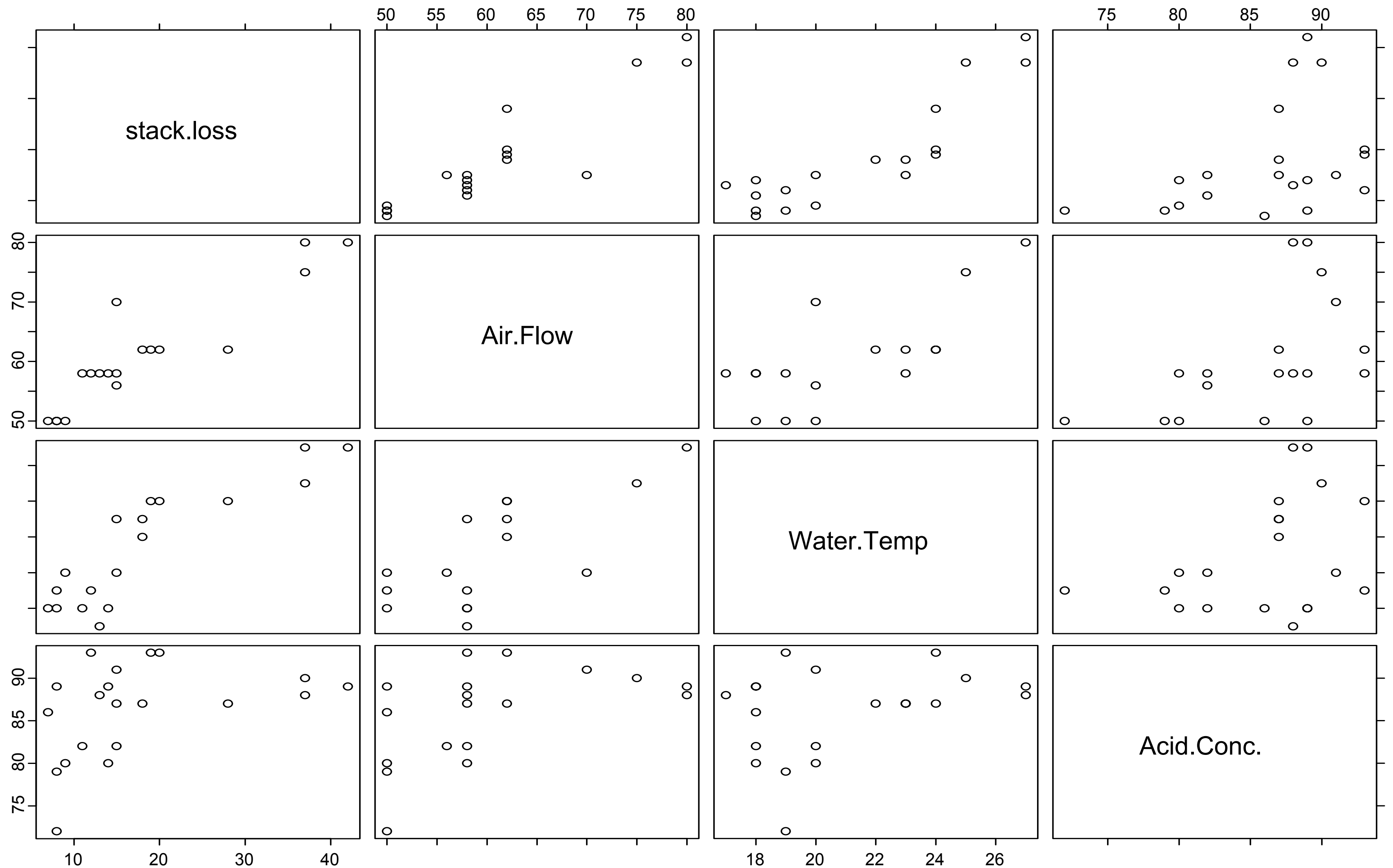
QQ Plot of correlation estimates



S-Plus Stack-loss data set

- **Stack-loss Data**
- **SUMMARY:**
- The stack.loss and stack.x data sets are from the operation of a plant for the oxidation of ammonia to nitric acid, measured on 21 consecutive days.
- **ARGUMENTS:**
- **stack.loss**
  - percent of ammonia lost (times 10).
- **stack.x**
  - matrix with 21 rows and 3 columns representing air flow to the plant, cooling water inlet temperature, and acid concentration as a percentage (coded by subtracting 50 and then multiplying by 10).
- **SOURCE:**
- Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering.* New York: John Wiley & Sons, Inc.
- Draper and Smith (1966). *Applied Regression Analysis.* New York: John Wiley & Sons, Inc.
- Daniel and Wood (1971). *Fitting Equations to Data.* New York: John Wiley & Sons, Inc.

S-Plus stack loss data set



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Summary of stack loss regression

```
> summary(tmp)

Call: lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc., data =
   stack)
Residuals:
    Min     1Q  Median    3Q   Max
 -7.238 -1.712 -0.4551 2.361 5.698

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -39.9197  11.8960    -3.3557   0.0038
   Air.Flow   0.7156   0.1349     5.3066   0.0001
 Water.Temp   1.2953   0.3680     3.5196   0.0026
 Acid.Conc.  -0.1521   0.1563    -0.9733   0.3440

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-Squared: 0.9136
F-statistic: 59.9 on 3 and 17 degrees of freedom, the p-value is 3.016e-009

Correlation of Coefficients:
           (Intercept) Air.Flow Water.Temp
  Air.Flow  0.1793
Water.Temp -0.1489     -0.7356
Acid.Conc. -0.9016     -0.3389   0.0002
```

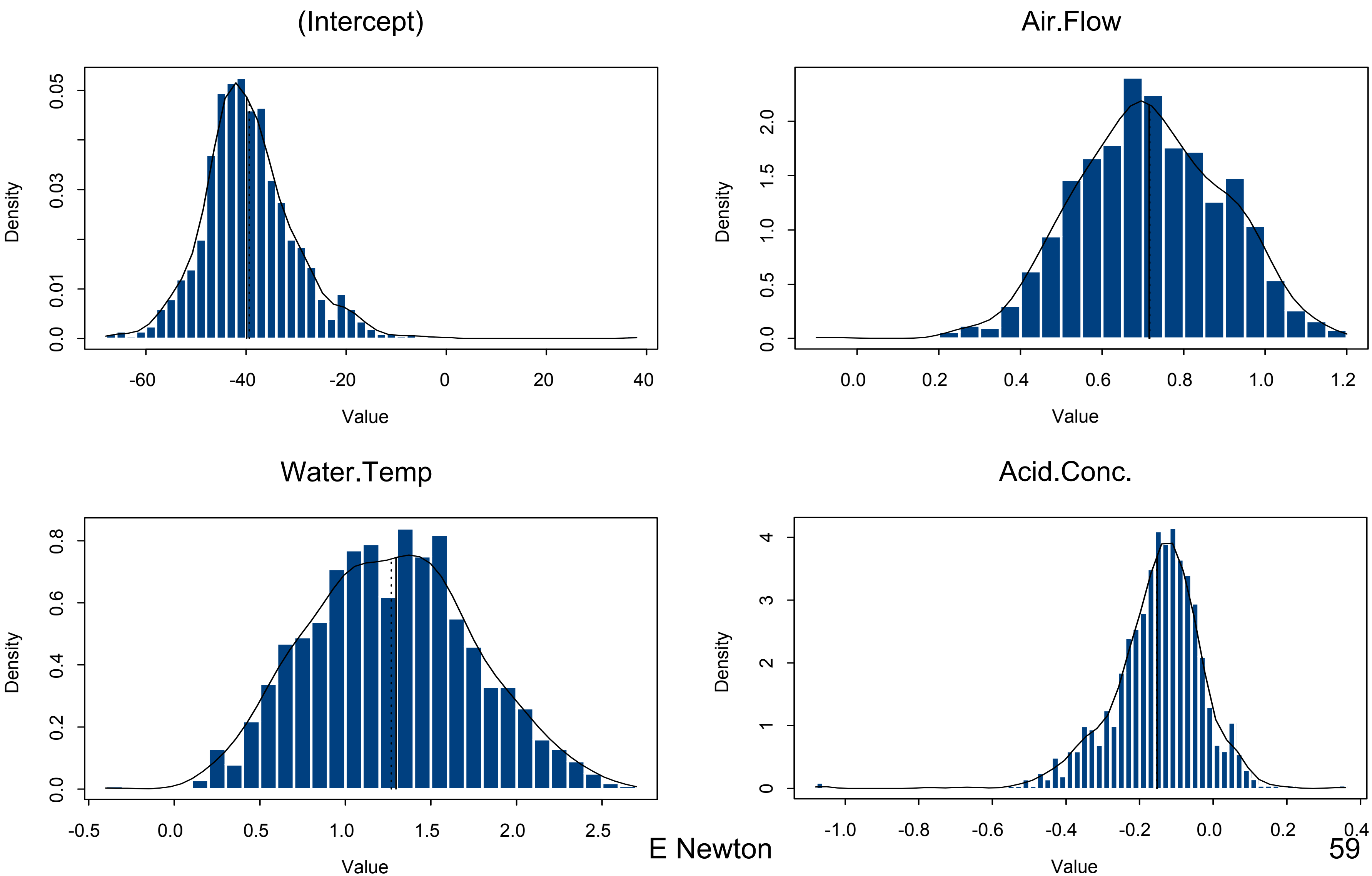
Summary of stack loss bootstrap output

```
summary(stack.boot)
Call:
bootstrap(data = stack, statistic = coef(lm(stack.loss ~ Air.Flow
   + Water.Temp + Acid.Conc., stack)), trace = F)

Number of Replications: 1000

Summary Statistics:
            Observed       Bias     Mean     SE
(Intercept) -39.9197  0.5691396 -39.3505 9.3731
   Air.Flow   0.7156  0.0016734   0.7173 0.1777
 Water.Temp   1.2953 -0.0264873   1.2688 0.4798
 Acid.Conc.  -0.1521 -0.0006978  -0.1528 0.1261

Empirical Percentiles:
                2.5%        5%       95%     97.5%
(Intercept) -56.0109 -53.4216 -21.92994 -18.75262
   Air.Flow   0.3903   0.4366   1.00261   1.04605
 Water.Temp   0.4004   0.5131   2.07381   2.23633
 Acid.Conc.  -0.4285  -0.3740   0.03282   0.05912
```

Summary of stack loss bootstrap output

```
summary(stack.boot)

BCa Confidence Limits:
                2.5%       5%        95%     97.5%
(Intercept) -55.6465 -52.6606 -21.451125 -18.55810
   Air.Flow   0.3266   0.4120   0.992007   1.01855
 Water.Temp   0.5244   0.6193   2.264165   2.40956
 Acid.Conc.  -0.4629  -0.4101  -0.007724   0.04459

Correlation of Replicates:
            (Intercept) Air.Flow Water.Temp Acid.Conc.
(Intercept)     1.00000 -0.17636    0.09902   -0.80236
   Air.Flow    -0.17636  1.00000   -0.78822   -0.07635
 Water.Temp     0.09902 -0.78822    1.00000   -0.24463
 Acid.Conc.    -0.80236 -0.07635   -0.24463    1.00000
```

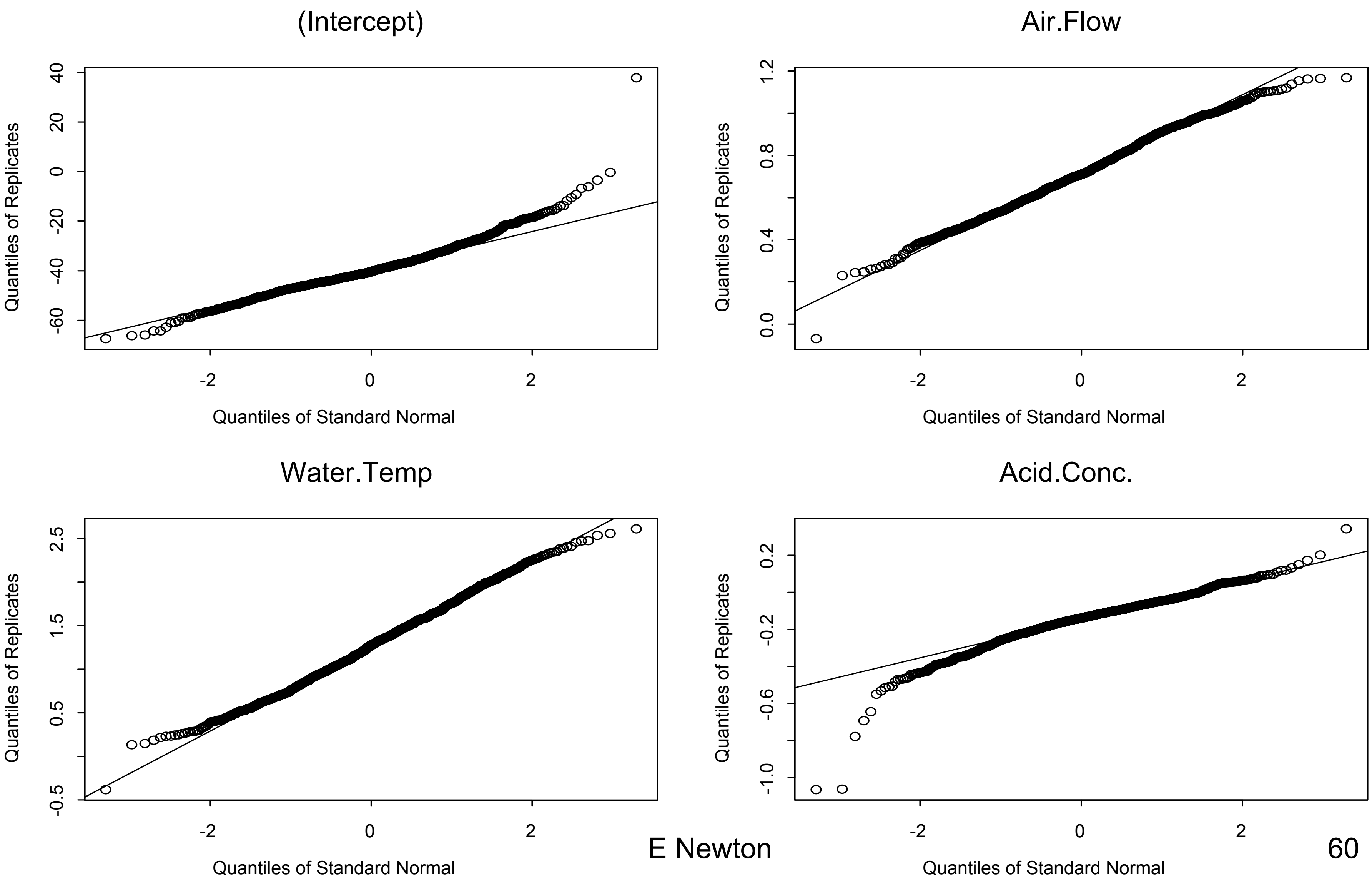
Histograms of regression coefficients



E Newton

This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

QQ Plots of regression coefficients



E Newton