Overview of Nonparametric Statistics

February 7, 2016

Statisticool.com

Motivation for This Talk

- We already use nonparametric methods in sampling, exploratory data analysis, outlier detection, imputation, variance estimation, simulation, goodness of fit tests, …

- However, nonparametric statistics itself is rarely discussed

What to Get Out of This Talk

- An overview of nonparametric statistics
- Learn advantages and disadvantages
- Learn a variety of SAS procedures
- Uses of nonparametric statistics in survey work

Outline

- Definition of nonparametric statistics
- History of nonparametric statistics
- Preliminaries: Note on Survey Data, Order Statistics, and Ranks
- Wilcoxon rank sum test

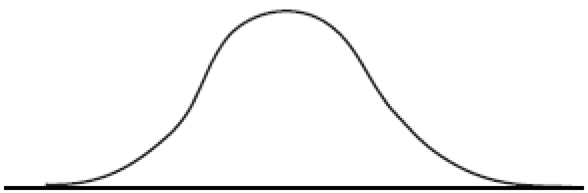
Outline (cont.)

- Various nonparametric techniques used in survey work
- Summary
- References

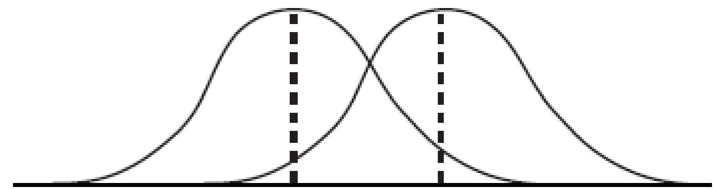
Hypothesis Testing

- Choose α
- Formulate hypotheses

H_0: $\mu_1 = \mu_2$



H_a: $\mu_1 \neq \mu_2$



- Calculate test statistic *under H_0*:

$$t_{m+n-2} = \frac{\bar{y} - \bar{x}}{\sqrt{S_p\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Hypothesis Testing (cont.)

- Errors

	H_0 is true	H_0 is false
Reject H_0	Type 1 error (α)	Correct decision
Fail to reject H_0	Correct decision	Type 2 error (β)

- <u>If</u> model is misspecified, error rates and inferences can be wrong

Definition of nonparametric statistics

- "Distribution free" – random variable has a sampling distribution that does not depend on the distribution function of the population
- "Nonparametric test" – hypothesis test which does not concern a parameter
  - e.g. tests of randomness, goodness of fit tests, tests for independence

Definition of nonparametric statistics (cont.)

- "Flexible models" – relaxed model structure

Parametric	Nonparametric
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_{i,}$ $\varepsilon_i \sim$ i.i.d. $N(0,\sigma)$	1. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_{i,}$ i.i.d. with median(ε_i) = 0 2. $Y_i = \mu(X_i) + \varepsilon_i$

History of nonparametric statistics

- Karl Pearson’s x^2 for goodness of fit (1900)

- Rank correlation coefficients
  - Spearman’s r (1904)
  - Kendall’s t (1938)

- Beginning of modern subject in mid 1930’s, says Savage (1953, 1962)

History of nonparametric statistics (cont.)

- “nonparametric” term first used by Wolfowitz (1942)
- Two-sample rank sum test by Wilcoxon (1945)
- Mann and Whitney extended Wilcoxon’s test for unequal sample sizes (1947)

History of nonparametric statistics (cont.)

- Pitman efficiency (1948)

- Jackknife by Quenouille (1949) for bias reduction and Tukey (1958, 1962) for variance estimation

- Hodges and Lehmann derived estimators from rank tests (1963)

History of nonparametric statistics (cont.)

- Bootstrap by Efron (1979)
- Locally weighted regression by Cleveland (1979) and Cleveland and Devlin (1988)
- *And much more !*

Note on Survey Data

- Note, “i.i.d.” is generally not valid for sample survey data

- Adjustments exist that take survey design into account for x^2 tests and other procedures:

  - **PROC SURVEYFREQ**
  - **PROC SURVEYREG**
  - **PROC SURVEYLOGISTIC**

Note on Survey Data (cont.)

- The main point is that tests need to be modified for use with survey data

- A simple option is to generate w_i observations per unit, where w_i is the final weight

Order Statistics

- Let $X_1, X_2, \ldots, X_n$ be your data

- Ordering these data from smallest to largest gives: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$

- $X_{(1)}$ is the minimum
- $X_{(n)}$ is the maximum
- If n is odd, $X_{((n+1)/2)}$ is the median
- If n is even, $(X_{(n/2)} + X_{((n/2)+1)})/2$ is the median

Ranks

- The rank of the i^{th} observation X_i, in a sample of n observations, is equal to the number of observations that are less than or equal to X_i

$$\text{rank}\left(\text{X}_{\text{i}}\right) = \sum_{j=1}^{n} I\left(X_j \le X_i\right)$$

- **PROC RANK data=mydata ties=mean**

Ranks (cont.)

- $X = \{5,6,7\}$, $\text{rank}(X) = \{1,2,3\}$

- In practice, ties in ranks occur

- $X = \{5, 6, 6, 7\}$
- Midrank method: $\text{rank}(X) = \{1, 2.5, 2.5, 4\}$

- Need to adjust variance because of ties

Ranks (cont.)

- Often use T(rank(X)) instead of T(X)
- We might be concerned about loss of efficiency – the “throwing away data” issue
- What is Corr(X, rank(X)) ?

Ranks (cont.)

- Stuart (1954, 1955) showed

$$\lim_{N->\infty} \rho\left[X, rank(X)\right] = \frac{2\sqrt{3}}{\sigma_X}\left\{E\left[XF_X(X)\right] - \frac{1}{2}E(X)\right\}$$

- I simulated various symmetrical and skewed F

Ranks (cont.)

Type	Distribution	Corr(X, rank(X))
Symmetric	Binomial(n=100, p=.5)	.978
Symmetric	Normal(μ=0, σ=1)	.977
Symmetric	T(df=10)	.961
Symmetric	Uniform(a=0, b=1)	.999
Skewed Right	Beta(α=2, β=5)	.975
Skewed Right	Binomial(n=100, p=.1)	.977
Skewed Right	Chi-square(df=2)	.865
Skewed Right	Exponential(λ=1)	.867
Skewed Right	F(df_{num}=10, df_{den}=10)	.811
Skewed Right	Gamma(θ=2)	.918
Skewed Right	Lognormal(0, 1)	.689
Skewed Right	Poisson(λ=4)	.973

Wilcoxon rank sum test

- Is there a difference between the means of two groups?

- Typically, we'd use a 2-sample t-test:

$$t_{m+n-2} = \frac{\bar{y} - \bar{x}}{\sqrt{S_p\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

X Sales ($)	Y Sales ($)
9,000	8,500
9,500	6,000
9,200	4,900
	6,900

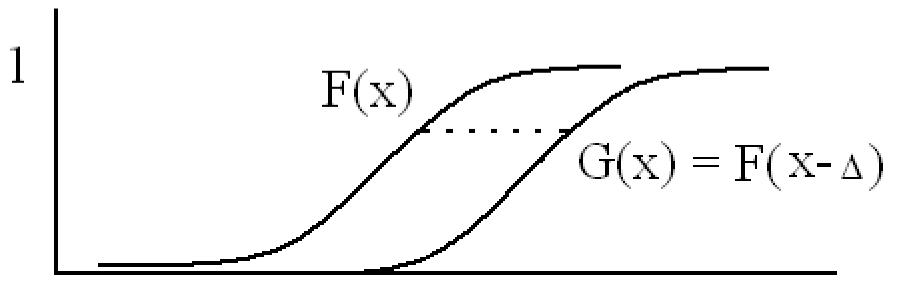
Wilcoxon rank sum test (cont.)

- Assumptions
  - $X_1, X_2, \ldots, X_m$ random i.i.d. sample from G
  - $Y_1, Y_2, \ldots, Y_n$ random i.i.d. sample from F
  - F and G are continuous
  - F and G differ only in location, i.e. $G(X) = F(X - \Delta)$

- $H_0: \Delta = 0$ vs. $H_1: \Delta > 0$

Wilcoxon rank sum test (cont.)

- Distributions differ only in location



- the data from one distribution is systematically larger than the data from the other

Wilcoxon rank sum test (cont.)

- Combine $N = m + n$ X-values and Y-values and calculate their ranks.

- W is the sum of ranks assigned to the X-values

$$W = \sum_{j=1}^{m} rank\left(X_j \right)$$

Wilcoxon rank sum test (cont.)

- Calculating the ranks

X Sales ($)	rank(X)	Y Sales ($)	rank(Y)
9,000	5	8,500	4
9,500	7	6,000	2
9,200	6	4,900	1
		6,900	3

Wilcoxon rank sum test (cont.)

$$W = \sum_{j=1}^{m} rank\left(X_j\right) = 5 + 7 + 6 = 18$$

- Under H_0,

$$E(W) = \frac{m(m+n+1)}{2} \qquad V(W) = \frac{mn(m+n+1)}{12}$$

Wilcoxon rank sum test (cont.)

- Exact null probability distribution of W can be obtained by systematic enumeration

- m = 3 and n = 4
  - configurations for the rank of the X's = $\dfrac{7!}{3!(7-3)!} = 35$

  - W will range between 6 and 18, symmetric about E(W) = 12

Wilcoxon rank sum test (cont.)

W	Possible rank of X's	Frequency
18	5,6,7	1
17	4,6,7	1
16	3,6,7 ; 4,5,7	2
15	2,6,7 ; 3,5,7 ; 4,5,6	3
14	1,6,7 ; 2,5,7 ; 3,4,7 ; 3,5,6	4
13	1,5,7 ; 2,4,7 ; 2,5,6 ; 3,4,6	4
12	1,4,7 ; 2,3,7 ; 1,5,6 ; 2,4,6 ; 3,4,5	5

- $P(W \geq 18) = 1/35 = .0286$

Wilcoxon rank sum test (cont.)

- P-value$_{exact}$ = .0286, so reject H_0, and conclude that the distribution of X is shifted to the right of Y at the 10% level

- **PROC NPAR1WAY data=mydata WILCOXON hl;**

  Var variable;

  Exact;

Wilcoxon rank sum test (cont.)

- Estimate of Δ (Hodges and Lehmann)

$$\hat{\Delta} = median\left(X_j - Y_i\right) = \$2{,}800$$

- 90% confidence interval for Δ
  - U is ordered list of the mn X – Y differences
  - $CI = \left(U_{(C_\alpha)}, U_{(mn+1-C_\alpha)}\right) = (\$500, \$4{,}600)$
  - Where $C_\alpha \approx \frac{mn}{2} - z_{\frac{\alpha}{2}}\left(\frac{mn(m+n+1)}{12}\right)^{\frac{1}{2}}$ for large m and n

Wilcoxon rank sum test (cont.)

- How does this Wilcoxon rank sum test compare relative to a two-sample t-test?

- And how do we carry out such a comparison?

Asymptotic Relative Efficiency

- Pitman (1948)
  - asymptotic relative efficiency (“A.R.E”)
    - limit of the ratio of sample sizes required for the two tests to achieve the same power under the same level of significance as the sample sizes tend to infinity

$$E_{W,t}(F) = 12\sigma_F^2\left(\int f^{\,2}\right)^2$$

Asymptotic Relative Efficiency (cont.)

F	Normal	Uniform	Logistic	Double Exponential	Exponential
E(W, t)	.955	1	1.097	1.5	3

- For <u>all</u> populations (i.e. for any F), $E_{W,t}(F) \geq .864$
  - Hodges and Lehmann, 1956

Nonparametric Methods Used in Survey Work

- A sampling of methods from
  - Correlation
  - Outlier detection
  - Variance estimation
  - Simulation
  - Goodness of fit
  - Regression

Correlation

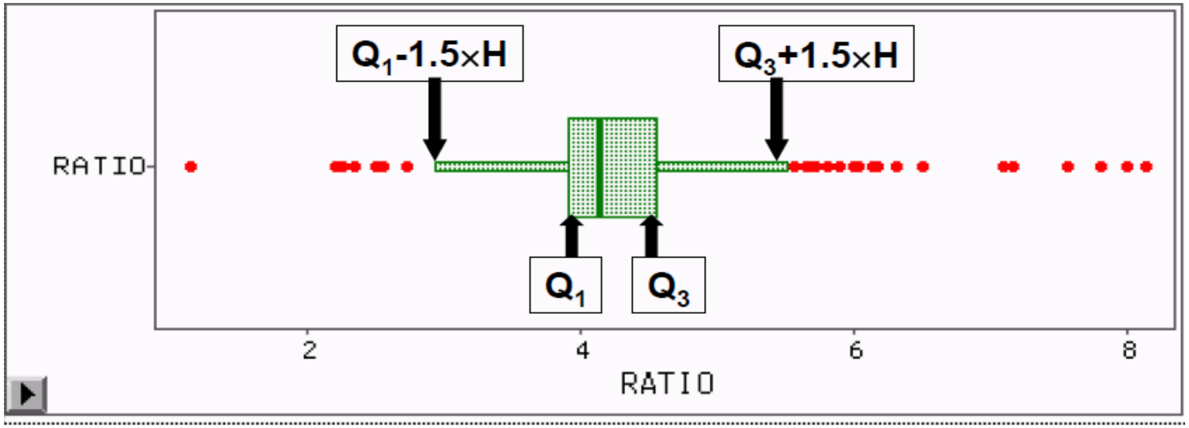
- Spearman correlation coefficient
  - Spearman(X,Y) = Pearson(rank(X), rank(Y))
  - let s = rank(X), and t = rank(Y), then

$$r = \frac{\sum_i (s_i - \bar{s})(t_i - \bar{t})}{\sqrt{\sum_i (s_i - \bar{s})^2 \sum_i (t_i - \bar{t})^2}} = 1 - \frac{6\sum_i (s_i - t_i)^2}{n(n^2 - 1)}$$

  - **PROC CORR data=mydata SPEARMAN;**

Resistant Fences

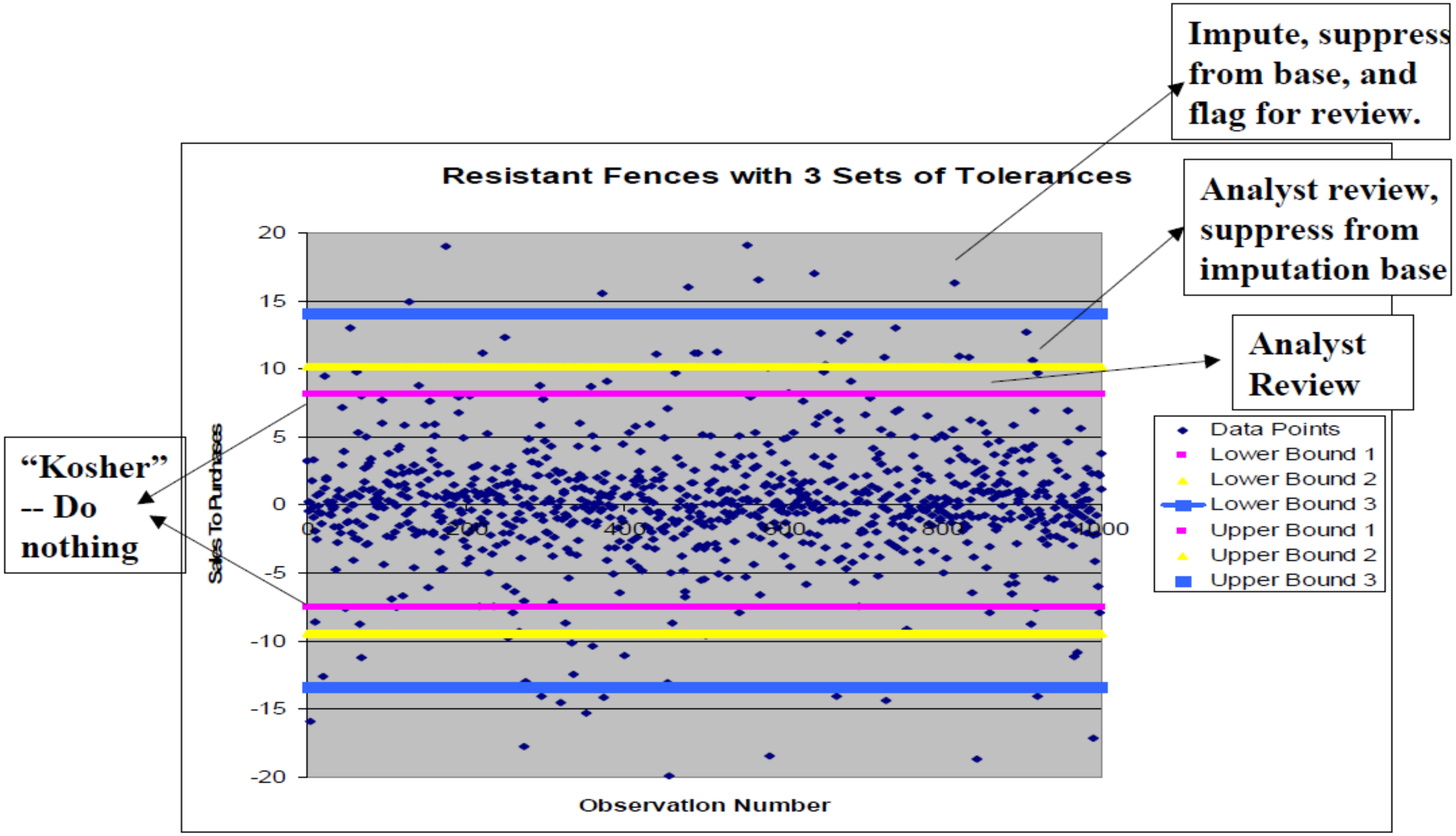
- symmetric →
- asymmetric
- flexible
- item or ratio



Q_1= the 1st quartile of a distribution of cell ratios
Q_3= the 3rd quartile of a distribution of cell ratios
$H = (Q_3 - Q_1)$ = the interquartile range

- take actions depending on region the point lies in

Resistant Fences (cont.)



Hidiroglou-Berthelot ("HB") Edit

- Generates tolerances that identify ratios as outlying or not

- Two positively correlated items
  - Q2 Sales / Q1 Sales

- Three-step process
  - Centering transformation
  - Magnitude transformation
  - Quartile test

HB Edit (cont.)

- Centering transformation

$$R_i = \frac{x_i}{y_i} \qquad S_i = \begin{cases} 1 - \dfrac{R_m}{R_i}, & 0 < R_i < R_m \\ \dfrac{R_i}{R_m} - 1, & R_i \geq R_m \end{cases}$$

  - R_m = median of R_i

HB Edit (cont.)

- Magnitude transformation

$$E_i = S_i \cdot \{\max(x_i, y_i)\}^u$$

- u is size parameter (0 q u q 1)
  - u = 1 gives full importance to unit's size
  - u = 0 gives no importance to unit's size
  - default is u = .5

HB Edit (cont.)

- Quartile Test

  – Calculate

$$D_{Q1} = \max\left(E_m - E_{Q1}, \left|A \cdot E_m\right|\right) \quad D_{Q3} = \max\left(E_{Q3} - E_m, \left|A \cdot E_m\right|\right)$$

  – A is a multiplier, say .05

- Flag E_i as outlier if
  – less than E_m-cD_{Q1}, or greater than E_m+cD_{Q3}

Variance Estimation

- Replication methods
  - divide parent sample into subsamples (R replicates)
  - calculate replicate weights (to represent full sample)
  - repeat estimation process on each subsample
  - estimate variance as $\hat{V}\left(\hat{Y}\right) = c\sum_{r=1}^{R}\left(\hat{Y}_r - \hat{Y}\right)^2$
- Ongoing work to implement stratified jackknife

Variance Estimation (cont.)

- Bootstrap
  - randomly resample B samples with replacement from the original sample
  - bootstrap samples : original sample :: original sample : population
  - each resample is same size as original sample
  - compute point estimate, confidence intervals, etc.

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* \qquad \hat{v}_{boot}\left(\hat{\theta}\right) = \frac{\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2}{B-1}$$

Simulation

- evaluating statistical properties of parameter or variance estimators over repeated samples

- generalized population simulation programs

- nearest neighbors technique to simulate a multivariate population with an unknown distribution

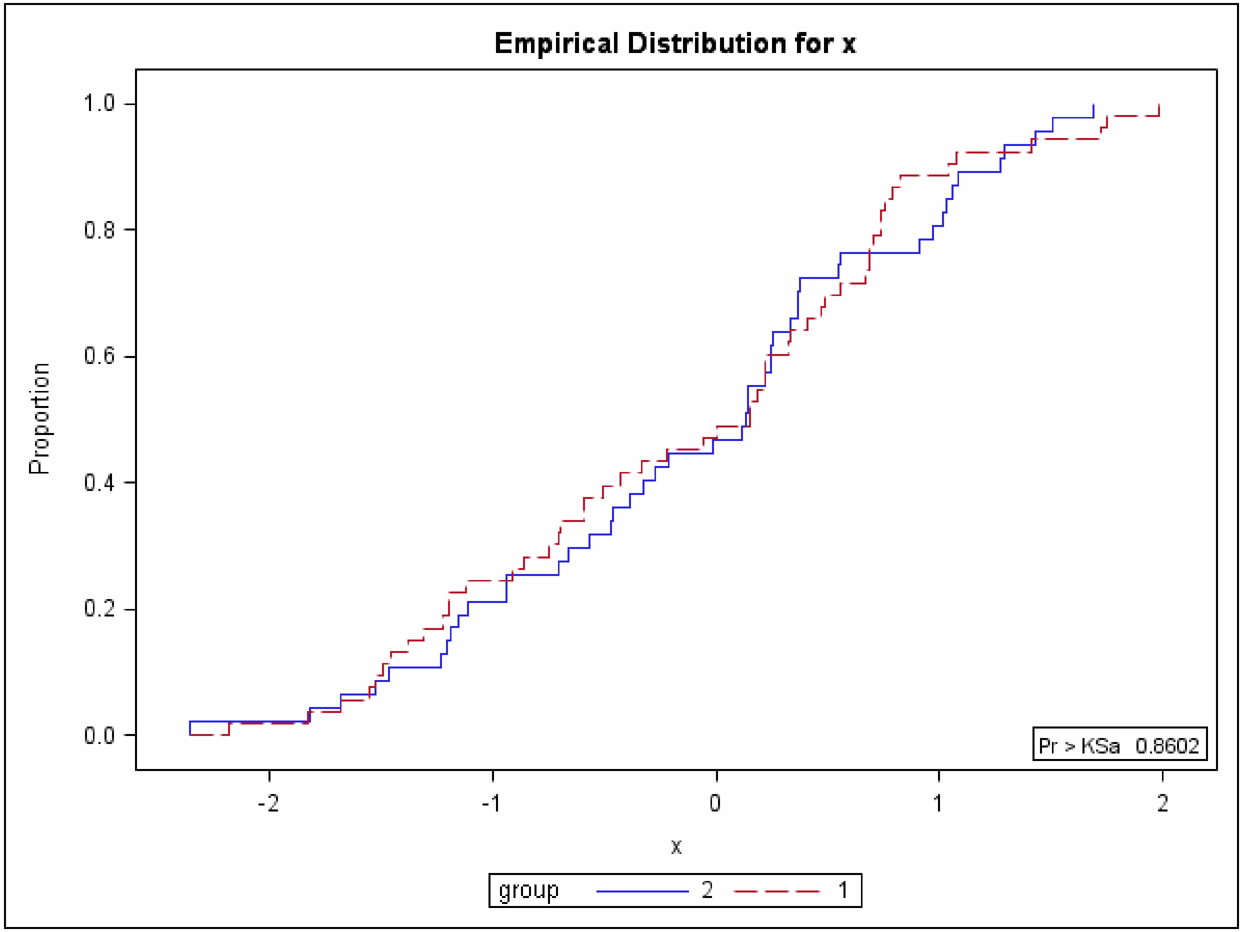
Kolmogorov-Smirnov Goodness of Fit Test

- largest vertical distance between empirical CDFs:

$$D_{n,m} = \max\left|F_n(x) - G_m(x)\right| \qquad \hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X \le X_i\right)$$

- **PROC NPAR1WAY data=mydata edf plots=edfplot**

Kolmogorov-Smirnov Test (cont.)

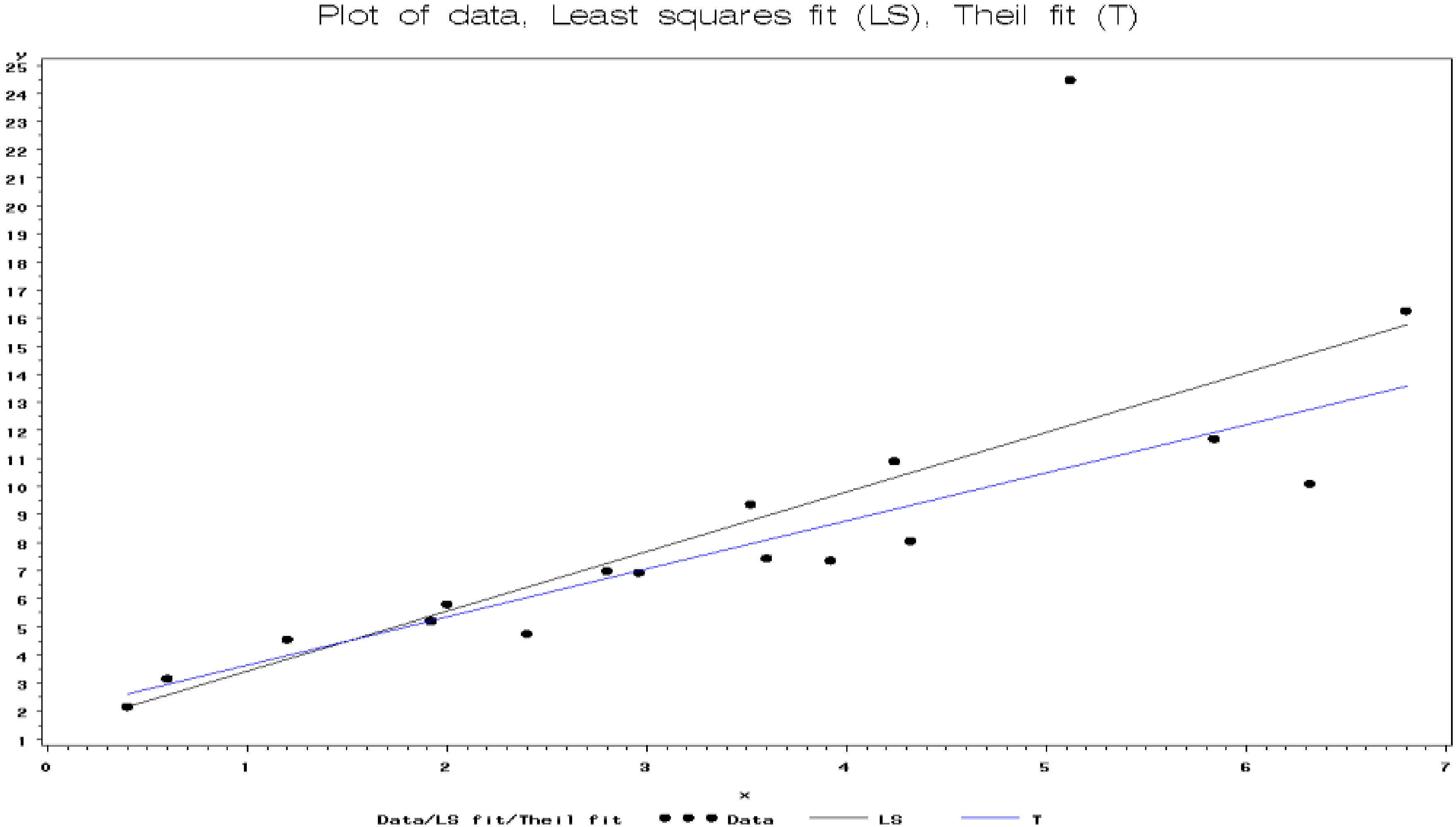


Theil estimator

- Calculate all possible slopes $S_{ij} = \dfrac{y_j - y_i}{x_j - x_i}, x_j \neq x_i$

- Then calculate
  - Slope: $\tilde{\beta}_1 = median\left(S_{ij}\right)$
  - Intercept:
    i) $\tilde{\beta}_0 = \tilde{y} - \tilde{\beta}_1 \tilde{x}$
    ii) $\tilde{\beta}_0 = median\left(y_i - \tilde{\beta}_1 x_i\right)$

Theil estimator (cont.)



Local regression

- Cleveland (1979)
  - Locally weighted least squares fit
  - Specify degree, smoothing parameter, and weighting function
  - **PROC LOESS**
- Process
  - k = floor(smoothing parameter * n)
    - 5 = .05 * 100
  - For each x_0 find the k closest points

Local regression (cont.)

– Calculate the max width of the neighborhood:

$$\Delta\left(x_0\right) = \max\left|x_0 - x_i\right|$$

– Assign a weight to each of the k points in the neighborhood:

$$w_i\left(x_0\right) = W\left(\frac{\left|x_0 - x_i\right|}{\Delta\left(x_0\right)}\right)$$

– Note, $W(u) = \left(1 - u^3\right)^3, 0 \le u < 1$ is the Tri-cube function

Local regression (cont.)

- W(x) > 0 for |x| < 1

(negative weights don't make sense)

- W(-x) = W(x)

(no reason to treat points on the left of x_i differently than those on the right)

- W(x) is non-increasing function for x ≥ 0

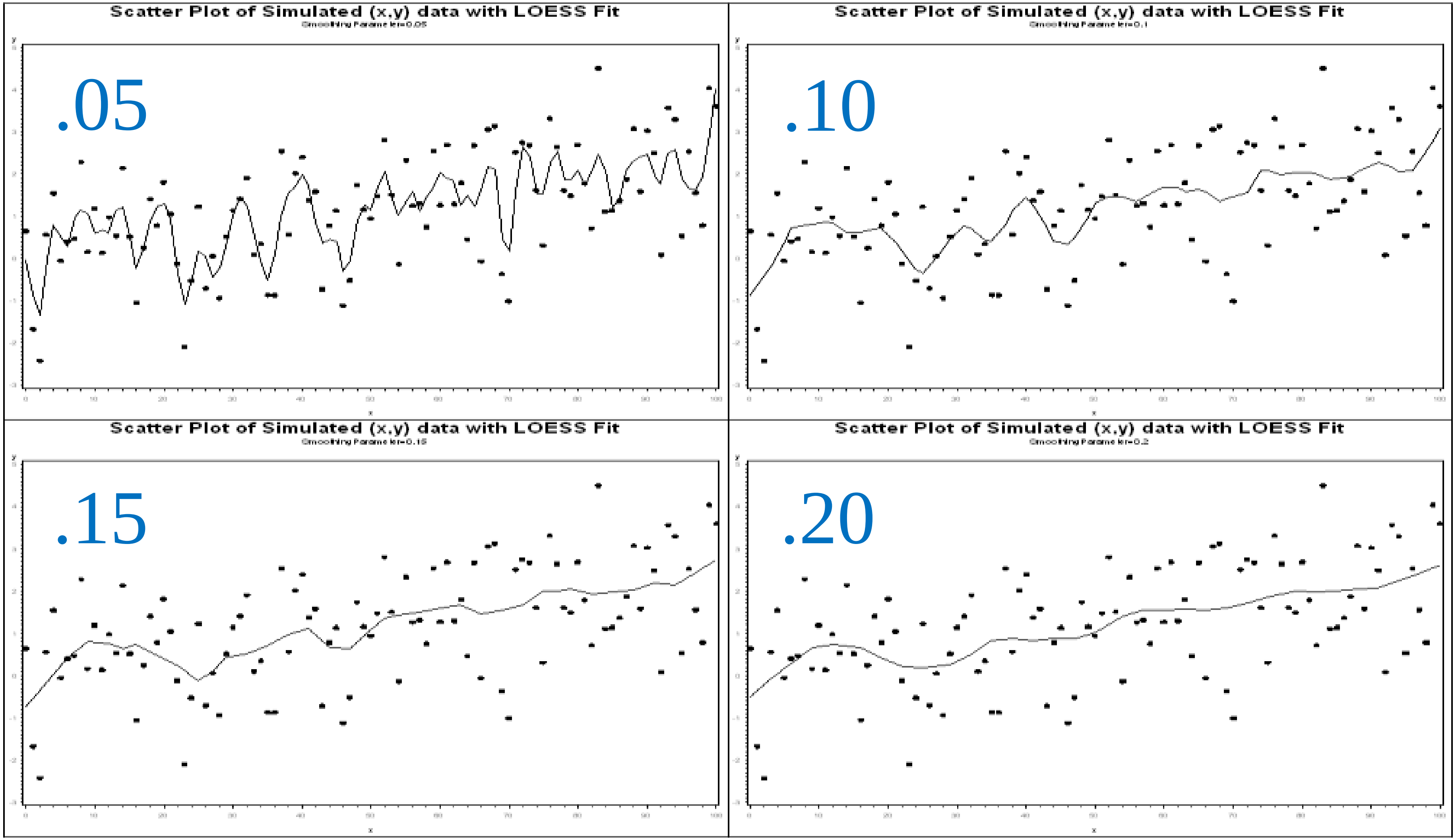
(unreasonable to allow a close point to have less weight than one that is further from x_i)

- W(x) = 0 for |x| ≥ 1

(for computational reasons)

- Minimize $\sum_{i=1}^{k} W_i(x_0)\left(y_i - \sum_{j=0}^{p} \beta_j x^j\right)^2 \mid x_0$

Local regression (cont.)



Summary

- Nonparametric statistics often requires few assumptions about the underlying population from which the data are obtained

- Can often obtain exact p-values
  - However, this may take a long time with large sample sizes

- Need to adjust procedures for survey data

Summary (cont.)

- Procedures that use ranks and medians are relatively insensitive to outlying observations

- The jackknife and bootstrap can be used in complicated situations where the distribution theory needed to support parametric methods is intractable

Summary (cont.)

- Some tests are slightly less efficient than their parametric counterparts even on the parametric “home turf”, but can be much more efficient

- Can lose power if underlying distribution is actually normal (for example)

Summary (cont.)

- Nonparametric methods have produced good results in survey processing

- I expect continued use of nonparametric methods in exploratory data analysis, hypothesis testing, imputation, and variance estimation in survey work

References

- Stuart, A. (1954). The correlation between variate-values and ranks in samples from a continuous distribution. *British Journal of Statistical Psychology, 7, 37-44.*

- Rao, J. N. K., Scott, A. J., (1987). On Simple Adjustments to Chi-Square Tests with Sample Survey Data. *The Annals of Statistics*, Vol. 15, No. 1, 385-397.

- <u>The Jackknife and Bootstrap</u>, Shao, Tu, 1995

- <u>Practical Nonparametric Statistics</u>, Conover, 1999

- <u>Nonparametric Statistical Methods</u>, Hollander, Wolfe, 1999

- <u>Sampling: Design and Analysis</u>, Lohr, 1999

- <u>Nonparametric Simple Regression</u>, Fox, 2000

- <u>Nonparametrics: Statistical Methods Based on Ranks</u>, Lehmann, 2006

- <u>Nonparametric Statistical Inference</u>, Gibbons, Chakraborti, 2011

Contact Information

Statisticool.com

justin@statisticool.com