Goodness-of-Fit Tests with Censored Data

Edsel A. Pena
Statistics Department
University of South Carolina
Columbia, SC
[E-Mail: pena@stat.sc.edu]

Talk at Cornell University, 3/13/02

Research support from NIH and NSF.

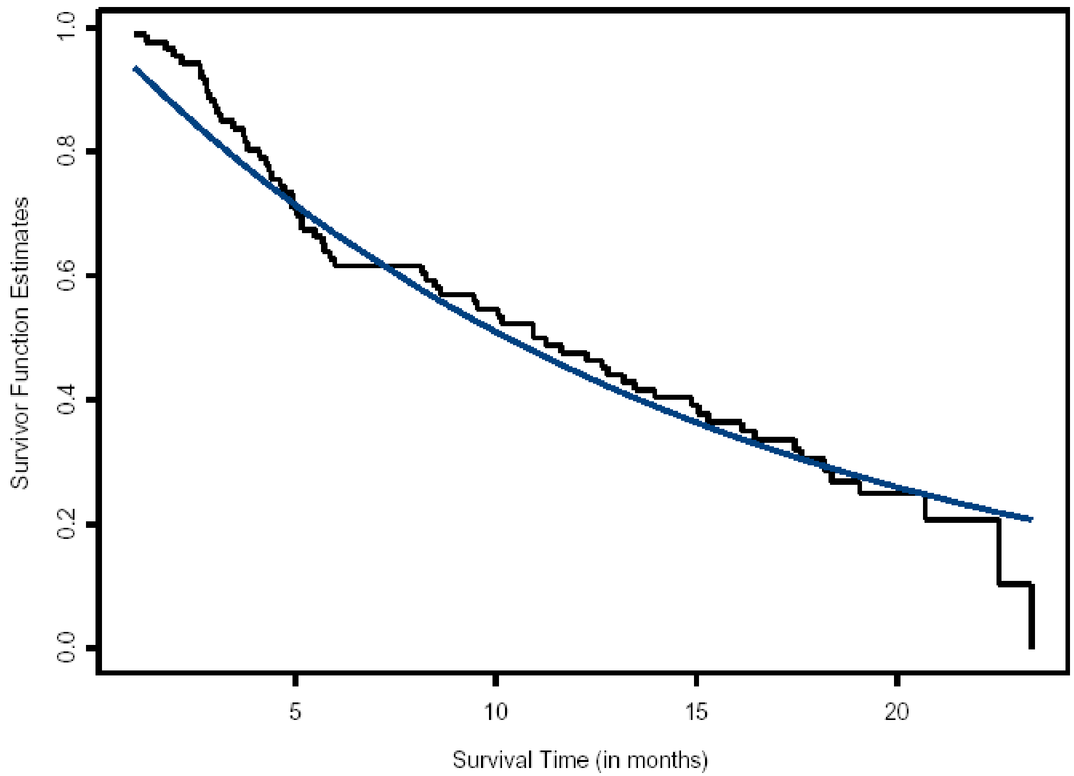
Practical Problem

- Right-censored survival data for lung cancer patients from Gatsonis, Hsieh and Korwar (1985).

- Survival times (in months): $n = 86$ with 23 right-censored.

  0.99, 1.28, 1.77, 1.97, 2.17, 2.63, 2.66, 2.76, 2.79, 2.86, 2.99, 3.06, 3.15, 3.45, 3.71, 3.75, 3.81, 4.11, 4.27, 4.34, 4.40, 4.63, 4.73, 4.93, 4.93, 5.03, 5.16, 5.17, 5.49, 5.68, 5.72, 5.85, 5.98, 8.15, 8.26, 8.48, 8.61, 9.46, 9.53, 10.05, 10.15, 10.94, 10.94, 11.04+, 11.24, 11.63, 12.26, 12.65, 12.78, 13.18, 13.47, 13.53+, 13.96, 14.23+, 14.65+, 14.88, 14.91+, 15.05, 15.31, 15.47+, 16.13, 16.46, 16.49+, 17.05+, 17.28+, 17.45, 17.61, 17.88+, 17.97+, 18.20, 18.37 18.83+ 19.06 19.55+ 19.58+ 19.75+ 19.78+ 19.95+ 20.04+, 20.24+, 20.70, 20.73+, 21.55+, 21.98+, 22.54, 23.36

Product-Limit Estimator and Best-Fitting Exponential Survivor Function



Question: Is the underlying survivor function modeled by a family of exponential distributions? a Weibull distribution?

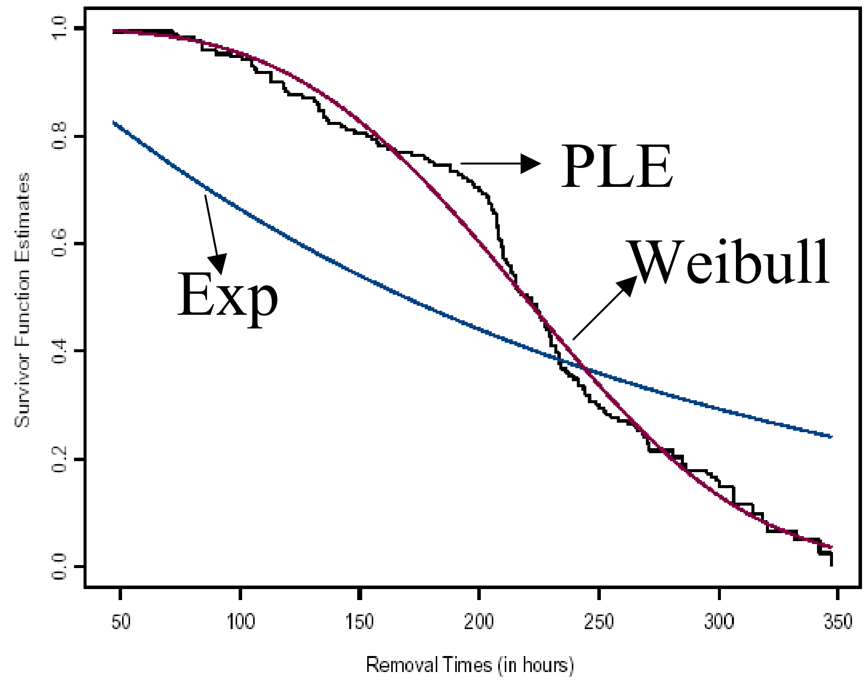
A Car Tire Data Set

- Times to withdrawal (in hours) of 171 car tires, with withdrawal either due to failure or right-censoring.

- Reference: Davis and Lawrance, in *Scand. J. Statist.*, 1989.

- Pneumatic tires subjected to laboratory testing by rotating each tire against a steel drum until either failure (several modes) or removal (right-censoring).

Product-Limit Estimator, Best-Fitting Exponential and Weibull Survivor Functions



Question: Is the Weibull family a good model for this data?

Goodness-of-Fit Problem

- $T_1, T_2, \ldots, T_n$ are IID from an unknown distribution function F
  - Case 1

Statement of the GOF Problem

On the basis of the data $(Z_1, \delta_1), (Z_2, \delta_2), \ldots, (Z_n, \delta_n)$:

<u>Simple GOF Problem</u>: For a pre-specified F_0, to test the null hypothesis that

$$H_0: F = F_0 \quad \text{versus} \quad H_1: F \neq F_0.$$

<u>Composite GOF Problem</u>: For a pre-specified family of dfs $\mathscr{F} = \{F_0(.;\eta): \eta \in \Gamma\}$, to test the hypotheses that

$$H_0: F \in \mathscr{F} \quad \text{versus} \quad H_1: F \notin \mathscr{F}.$$

Generalizing Pearson

With complete data, the famous Pearson test statistics are:

Simple Case: $$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

Composite Case: $$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

where O_i is the # of observations in the i^{th} interval; E_i is the expected number of observations in the i^{th} interval; and

$$\hat{E}_i = nF_0(I_i;\hat{\eta})$$

is the estimated expected number of observations in the i^{th} interval under the null model.

Obstacles with Censored Data

- With right-censored data, determining the exact values of the O_j's is not possible.
- Need to estimate them using the product-limit estimator (Hollander and Pena, ‘92; Li and Doss, ‘93), Nelson-Aalen estimator (Akritas, ‘88; Hjort, ‘90), or by self-consistency arguments.
- Hard to examine the power or optimality properties of the resulting Pearson generalizations because of the *ad hoc* nature of their derivations.

In Hazards View: Continuous Case

For T an abs cont +rv, the hazard rate function λ(t) is:

$$\lambda(t)dt \approx P\{t \leq T < t + dt \mid T \geq t\} = \frac{f(t)dt}{\overline{F}(t)}$$

Cumulative hazard function Λ(t) is:

$$\Lambda(t) = \int_0^t \lambda(w)\,dw$$

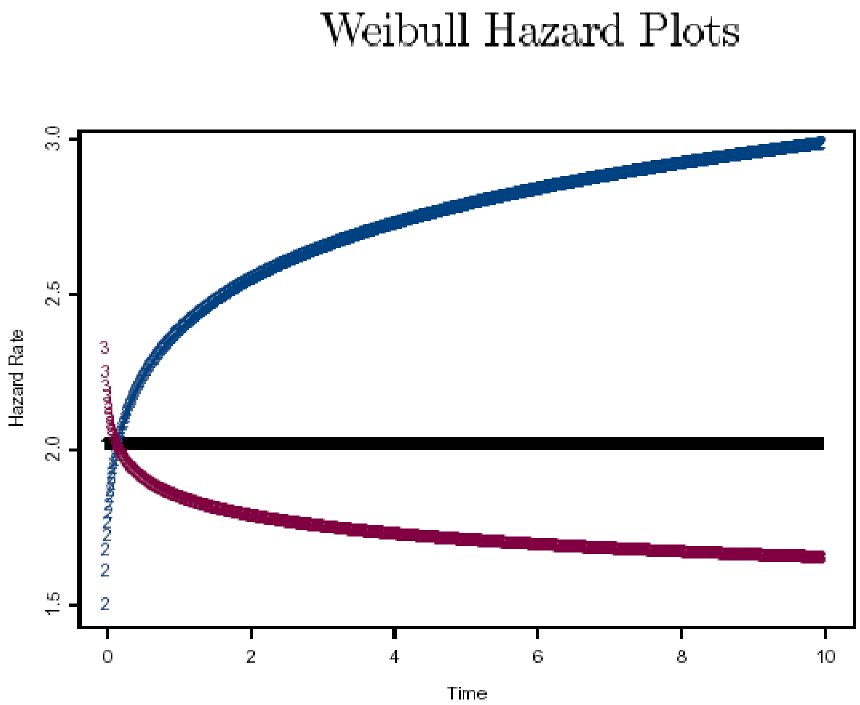
Survivor function in terms of Λ:

$$\overline{F}(t) = P\{T > t\} = \exp\{-\Lambda(t)\}$$

Two Common Examples

Exponential: $\lambda(t;\eta) = \eta$

Two-parameter Weibull: $\lambda(t;\alpha,\eta) = (\alpha\eta)(\eta t)^{\alpha-1}$



Counting Processes and Martingales

$$N(t) = \sum_{i=1}^{n} I\{Z_i \le t, \delta_i = 1\}$$

$$Y(t) = \sum_{i=1}^{n} I\{Z_i \ge t\}$$

$$M(t) = N(t) - \int_0^t Y(w)\lambda(w)dw$$

$\{M(t): 0 < t < \tau\}$ is a square-integrable zero-mean martingale with predictable quadratic variation (PQV) process

$$\langle M, M\rangle(t) = \int_0^t Y(w)\lambda(w)dw$$

Idea in Continuous Case

- For testing H_0: $\lambda(.) \in \mathcal{C} = \{\lambda_0(.;\eta): \eta \in \Gamma\}$, if H_0 holds, then there is some η_0 such that the true hazard $\lambda_0(.)$ is such

$$\lambda_0(.) = \lambda_0(.;\eta_0)$$

- Let $\kappa(t;\eta) = \log\left\{\dfrac{\lambda_0(t)}{\lambda_0(t;\eta)}\right\} \in$ some space $\mathcal{K}$

- Basis Set for $\mathcal{K}$: $\{\psi_1(.;\eta), \psi_2(.;\eta), \ldots\}$

- Expansion: $\kappa(t;\eta) = \sum_{k=1}^{\infty} \theta_i \psi_i(t;\eta)$

- Truncation: $\kappa(t;\eta) \approx \sum_{k=1}^{p} \theta_k \psi_k(t;\eta)$, p is smoothing order

Hazard Embedding and Approach

- From this truncation, we obtain the approximation

$$\lambda_0(t) \approx \lambda_0(t;\eta)\exp\left\{\sum_{k=1}^{p}\theta_k\psi_k(t;\eta)\right\} = \lambda_0(t;\eta)\exp\{\theta^t\Psi(t;\eta)\}$$

- Embedding Class

$$\mathcal{C}_{\mathrm{p}} = \left\{\lambda_p(t;\theta,\eta) = \lambda_0(t;\eta)\exp\left\{\sum_{k=1}^{p}\theta_k\psi_k(t;\eta)\right\} : \ \theta \in \Re^p\right\}$$

- Note: $H_0 \subseteq \mathcal{C}_{\mathrm{p}}$ obtains by taking $\theta = 0$.
- GOF Tests: Score tests for H_0: $\theta = 0$ versus H_1: $\theta \neq 0$.
- Note that η is a nuisance parameter in this testing problem.

Class of Statistics

- Estimating equation for the nuisance η:

$$\rho(t;\eta) = \nabla_\eta \log \lambda_0(t;\eta)$$

$$\int_0^\tau \rho(w;\hat{\eta})\left\{dN(w) - Y(w)\lambda_0(w;\hat{\eta})dw\right\} = 0$$

- Quadratic Statistic:

$$\hat{Q} = \int_0^\tau \Psi_p(w;\hat{\eta})\left\{dN(w) - Y(w)\lambda_0(w;\hat{\eta})dw\right\}$$

$$S_p = \frac{1}{n}\hat{Q}^t\left\{\hat{\Xi}^-\right\}\hat{Q}$$

- $\hat{\Xi}$ is an estimator of the limiting covariance of $\frac{1}{\sqrt{n}}\hat{Q}$

Asymptotics and Test

- Under regularity conditions,

$$\frac{1}{\sqrt{n}}\hat{Q} \xrightarrow{d} N_p(0,\ \Xi \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

- Estimator of Ξ obtained from the matrix:

$$\hat{\Sigma} = \frac{1}{2n}\int_0^\tau \begin{bmatrix} \Psi_p(w;\hat{\eta}) \\ \rho(w;\hat{\eta}) \end{bmatrix}^{\otimes 2} \left\{dN(w) + Y(w)\lambda_0(w;\hat{\eta})dw\right\}$$

- Test: Reject H_0 if $\quad S_p > \chi^2_{p^*;\alpha}$.

A Choice of Ψ Generalizing Pearson

- Partition $[0,\tau]$ into $0 = a_1 < a_2 < \ldots < a_p = \tau$, and let

$$\Psi_p(t) = (I_{[0,a_1]}(t), I_{(a_1,a_2]}(t), \ldots, I_{(a_{p-1},a_p]}(t))^t$$

- Then

$$\hat{Q} = \left(O_1 - \hat{E}_1, O_2 - \hat{E}_2, \ldots, O_p - \hat{E}_p\right)^{\mathrm{t}}$$

$$O_j = \int_{a_{j-1}}^{a_j} dN(w) = N(a_j) - N(a_{j-1})$$

$$\hat{E}_j = \int_{a_{j-1}}^{a_j} Y(w)\lambda_0(w;\hat{\eta})dw$$

- $\hat{E}_j's$ are *dynamic* expected frequencies

Special Case: Testing Exponentiality

- Exponential Hazards: $\mathcal{C} = \{\lambda_0(t;\eta)=\eta\}$

$$O_j = \int_{a_{j-1}}^{a_j} dN(w) \quad \text{and} \quad \hat{E}_j = \hat{\eta} \int_{a_{j-1}}^{a_j} Y(w)dw \qquad \text{with} \quad \hat{\eta} = \frac{\int_0^\tau dN(w)}{\int_0^\tau Y(w)dw}$$

- Test Statistic ("generalized Pearson"):

$$S_p = \hat{E}_\bullet (p - \hat{\pi})^t \{Dg(\hat{\pi}) - \hat{\pi}\hat{\pi}^t\}^- (p - \hat{\pi})$$

where

$$\hat{E}_\bullet = \sum_{k=1}^p \hat{E}_k; \quad p = \frac{1}{\hat{E}_\bullet}(O_1, O_2, \ldots, O_p)^t; \quad \hat{\pi} = \frac{1}{\hat{E}_\bullet}(\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_p)^t$$

A Polynomial-Type Choice of Ψ

$$\Psi_p(t;\eta) = \left(1, \Lambda_0(t;\eta), \ldots, [\Lambda_0(t;\eta)]^{p-1}\right)^{\mathrm{t}}$$

• Components of $\hat{Q}$

$$\hat{Q}_k = \int_0^{\tau^*} w^{k-1}\left\{dN^R(w) - Y^R(w)dw\right\}, \; k = 1,2,\ldots,p;$$

where

$$R_i = \Lambda_0(Z_i;\hat{\eta}); \quad N^R(t) = \sum_{i=1}^{n} I\{R_i \le t; \delta_i = 1\}; \text{ and } Y^R(t) = \sum_{i=1}^{n} I\{R_i \ge t\}.$$

• Resulting test based on the 'generalized' residuals. The framework allows correcting for the estimation of nuisance η.

Simulated Levels

(Polynomial Specification, K = p)

Null Dist.	Exponential(η)	
Parameters	$\eta = 2$	$\eta = 5$

% Uncensored	75%	50%	% Uncensored		75%	5(
5%	5%	5%	Level		5%	
			n	K		
.65	5.00	5.60		2	4.65	6
.50	4.35	4.95	50	3	5.45	5
.50	5.10	5.85		4	6.55	5
.00	5.20	4.20		5	6.40	5
.75	4.45	4.35		2	4.90	4
.35	4.65	4.25	100	3	4.55	4
.30	5.10	4.80		4	5.70	5
.90	5.30	4.75		5	5.75	4

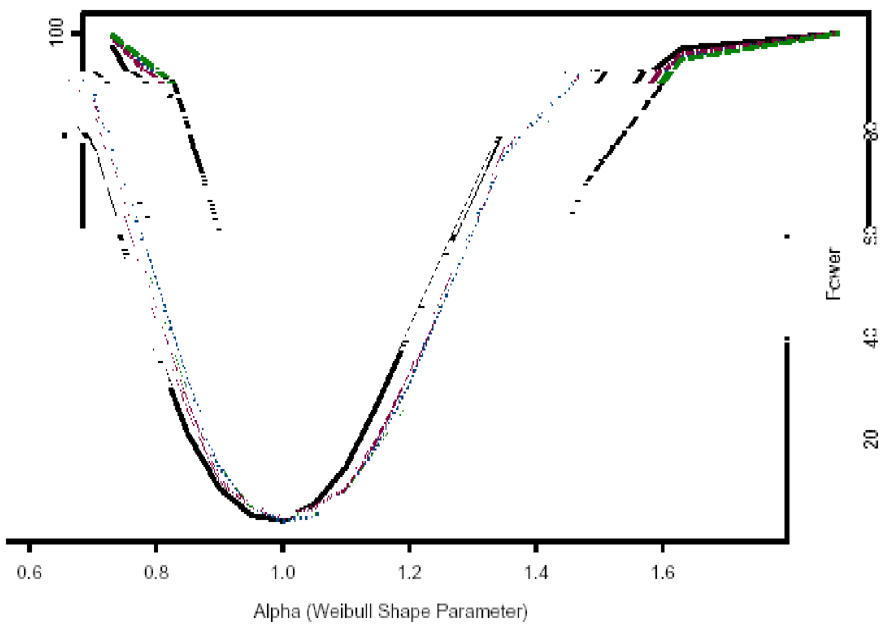
Null Dist.		Weibull(α, η)			
Parameters		$(\alpha, \eta) = (2, 1)$		$(\alpha, \eta) = (3, 2)$	
% Uncensored		75%	50%	75%	50%
Level		5%	5%	5%	5%
n	K				
50	2	4.30	4.80	4.60	6.20
	3	5.40	5.15	6.30	5.70
	4	4.80	4.80	6.10	5.30
	5	5.25	3.45	6.70	4.60
100	2	3.90	4.95	4.20	4.80
	3	5.75	4.65	5.15	5.25
	4	5.55	4.15	5.00	4.30
	5	6.00	4.65	5.80	5.30

Simulated Powers

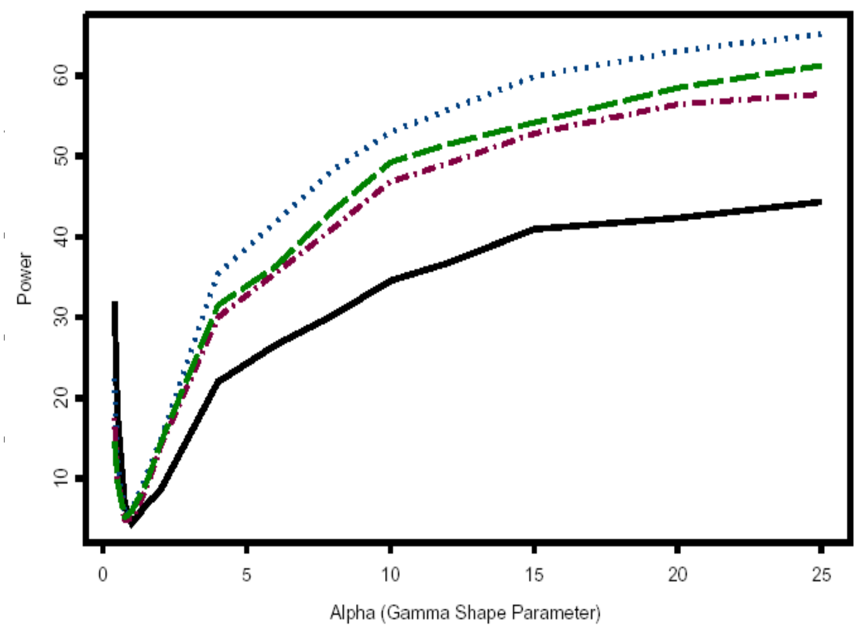
Simulation Parameters: $n = 100$; 25% censoring.

Legend: Solid: p=2; Dots: p=3; Short Dashes: p = 4; Long Dashes: p=5

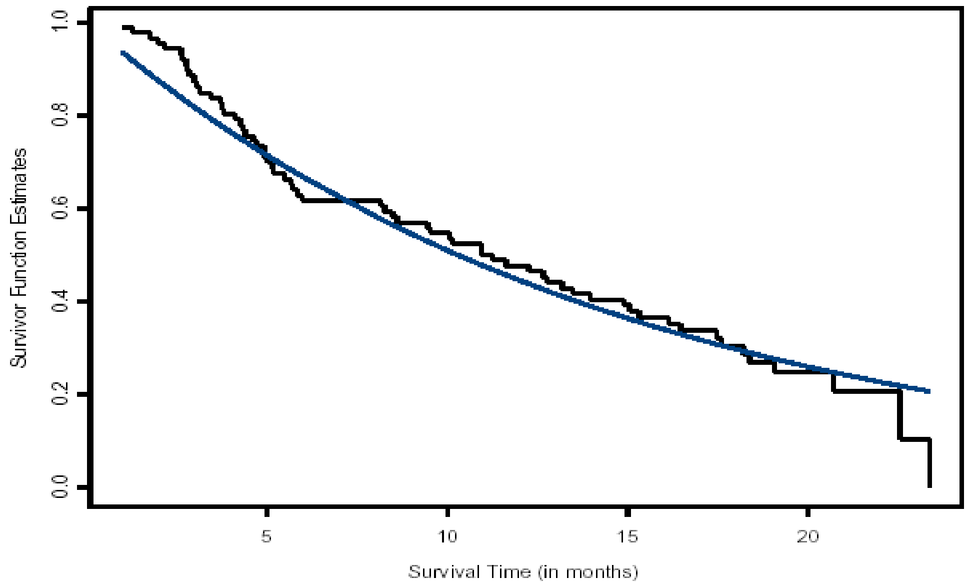
Null: **Exponential** vs. Alt: **2-Weibull**



Null: **2-Weibull** vs. Alt: **2-Gamma**



Back to Lung Cancer Data



Test for Exponentiality

$S_2 = 1.92 (p = .1661);\quad S_3 = 1.94 (p = .3788);$

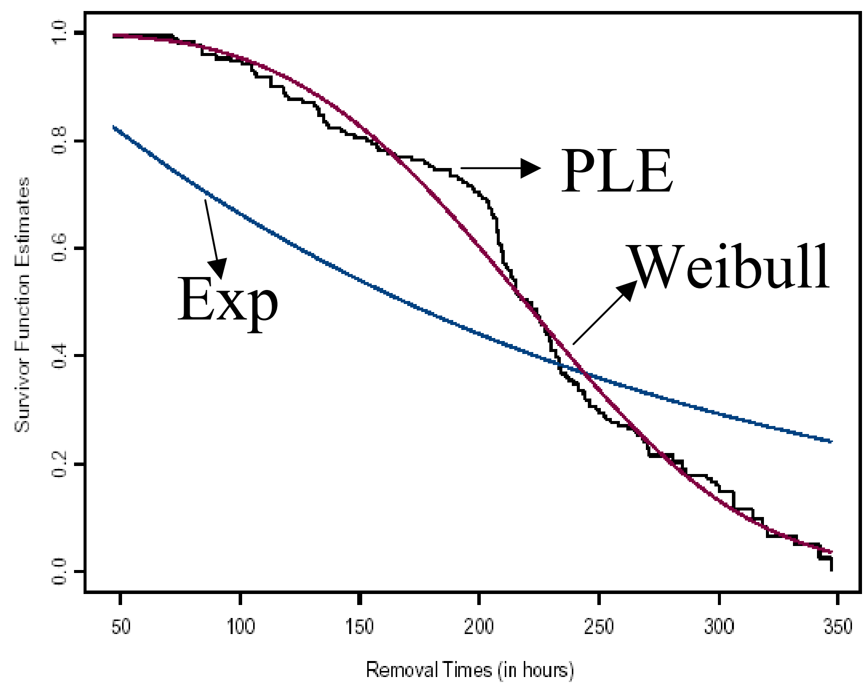
$S_4 = 7.56 (p = .0561);\quad S_5 = 12.85 (p = .0121).$

Test for Weibull

$S_3 = 8.35 \quad (p = .0153)$

S_4 and S_5 also both indicate rejection of Weibull family.

Back to Davis & Lawrance Car Tire Data



Test of Exponentiality

```
Results for Testing Exponentiality

Summary of Results of Smooth Goodness-of-Fit Test
 Testing for Exponential Distribution

 Input FileName:  C:\Talks\TalkATUG\davislawrancecartiredata.txt
 Output FileName: C:\Talks\TalkATUG\cartire.out

 Sample Size =          171
 Sum of Z_i =    36457.0000000000
 Sum of Delta_i = [illegible]
 of Eta = 4.3174[illegible]                                  Estimate
 of Mean =            243                                    Estimate

_k              DF        P-Value                            k        S
000             0         1.00000                            1      0.0
710             1         0.00000                            2    172.3
297             2         0.00000                            3    174.9
617             3         0.00000                            4    175.7
[illegible]
```

Conclusion: Exponentiality does not hold as in graph!

Test of Weibull Family

```
Results for Testing the Weibull Class

Summary of Results of Smooth Goodness-of-Fit Test
 Testing for the Two-Parameter Weibull Distribution
 Using the Polynomial Specification for Psi

 Input FileName:  C:\Talks\TalkATUG\davislawrancecartiredata.txt
 Output FileName: C:\Talks\TalkATUG\cartireWeibull.txt

 Sample Size =           171
 # of Uncensored Values =    150.000000000000
 Estimate of alpha =    3.41809555692639
 Estimate of eta =   4.08516600613130 5E-003

 k      S_k             DF        P-Value
 1     0.0000           0         1.00000
 2     0.5013           1         0.47891
 3     0.5550           2         0.75769
 4     6.4203           3         0.09286
 5     6.5183           4         0.16364
```

Conclusion: Cannot reject Weibull family of distributions.

Simple GOF Problem: Discrete Data

- T_i's are discrete +rvs with jump points $\{a_1, a_2, a_3, \ldots\}$.
- Hazards: $\lambda_i = \mathbf{P}\{T = a_i | T > a_i\}$
- Let $\{\lambda_1^0, \lambda_2^0, \ldots, \lambda_j^0, \ldots\}$ be such that $\lambda_j^0 \in [0, 1]$
- Problem: To test the hypotheses

$$H_0 : \lambda_j = \lambda_j^0, \quad j = 1, 2, \ldots, \quad \text{versus} \quad H_1 : \lambda_j \neq \lambda_j^0 \text{ for some } j \in \{1, 2, \ldots\}$$

based on the right-censored data $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$.

- True and hypothesized hazard odds:

$$\rho_j = \frac{\lambda_j}{(1-\lambda_j)} \quad \text{and} \quad \rho_j^0 = \frac{\lambda_j^0}{(1-\lambda_j^0)}$$

- For p a pre-specified order, let

$$\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_J) = \begin{bmatrix} \psi_1 \\ \psi_{\mathbf{2}} \\ \vdots \\ \psi_{\mathbf{p}} \end{bmatrix}$$

be a p x J (possibly random) matrix, with its p rows linearly independent, and with $[0, a_J]$ being the maximum observation period for all n units.

Embedding Idea

- To embed the hypothesized hazard odds $(\rho_1^0, \rho_2^0, \ldots, \rho_J^0)$ into

$$\mathcal{C}_p = \left\{ (\rho_1(\theta), \rho_2(\theta), \ldots, \rho_J(\theta)) : \theta = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{t}} \in \Re^p \right\}$$

$$\rho_j(\theta) = \rho_j^0 \exp\{\theta^{\mathrm{t}} \boldsymbol{\Psi}_j\}, \quad j = 1, 2, \ldots, J.$$

- Equivalent to assuming that the log hazard odds ratios satisfy

$$\log \left\{ \frac{\rho_j(\theta)}{\rho_j^0} \right\} = \theta^{\mathrm{t}} \boldsymbol{\Psi}_j = \sum_{k=1}^{p} \theta_k \Psi_{kj}, \quad j = 1, 2, \ldots, J.$$

- Class of tests are the score tests of H_0: $\boldsymbol{\theta} = \mathbf{0}$ vs. H_1: $\boldsymbol{\theta} \neq \mathbf{0}$ as p and Ψ are varied.

Class of Test Statistics

$$: \sum_{i=1}^{n} I\{Z_i = a_j, \delta_i = 1\} \quad \text{and} \quad R_j = \sum_{i=1}^{n} I\{Z_i \geq a_j\} \quad O_j =$$

$$\mathbf{O} = (O_1, O_2, \ldots, O_J)^t \qquad \mathbf{E}_0 = (E_1^0, E_2^0, \ldots, E_J^0)^t$$

$$E_j^0 = R_j \lambda_j^0 \qquad V_{jj}^0 = R_j \lambda_j^0 (1 - \lambda_j^0)$$

$$V_0 = Dg(V_{11}^0, V_{22}^0, \ldots, V_{JJ}^0)$$

- Quadratic Score Statistic:

$$S_p^2(\Psi) = (\mathbf{O} - \mathbf{E}_0)^t \Psi^t \left(\Psi \mathbf{V}_0 \Psi^t\right)^- \Psi(\mathbf{O} - \mathbf{E}_0)$$

- Under H_0, this converges in distribution to a chi-square rv.

A Pearson-Type Choice of Ψ

Partition {1,2,…,J}: $A_1, A_2, \ldots, A_p$ with $A_i \neq \emptyset, i = 1, 2, \ldots, p$

$\mathbf{1}_A$ the $J \times 1$ vector whose jth element is $I\{j \in A\}$

$$\Psi_{\mathbf{Indi}} = [\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \ldots, \mathbf{1}_{A_p}]^{\mathrm{t}}$$

$$O_\bullet(A) = \mathbf{1}_A^{\mathrm{t}}\mathbf{O} = \textstyle\sum_{j \in A} O_j,\ E_\bullet^0(A) = \mathbf{1}_A^{\mathrm{t}}\mathbf{E}_0 = \sum_{j \in A} E_j^0$$

$$V_\bullet^0(A) = \mathbf{1}_A^{\mathrm{t}}\mathbf{V}_0\mathbf{1}_A = \textstyle\sum_{j \in A} V_{jj}^0 = \sum_{j \in A} E_j^0(1 - \lambda_j^0)$$

$$S^2(\Psi_{\mathrm{Indi}}) = \sum_{i=1}^{p} \frac{[O_\bullet(A_i) - E_\bullet^0(A_i)]^2}{V_\bullet^0(A_i)}$$

A Polynomial-Type Choice

$$\Psi^p_{\text{Poly}} = \left[\left(\frac{\mathbf{R}}{n}\right)^0, \left(\frac{\mathbf{R}}{n}\right)^1, \ldots, \left(\frac{\mathbf{R}}{n}\right)^{p-1} \right]^{\mathrm{t}}$$

$$\begin{aligned}
U_i^*(\Psi^p_{\text{Poly}}) &= \left[\left(\frac{\mathbf{R}}{n}\right)^{i-1} \right]^{\mathrm{t}} (\mathbf{O} - \mathbf{E}_0) = \sum_{j=1}^{J} \left(\frac{R_j}{n}\right)^{i-1} (O_j - E_j^0); \\
I_{i_1 i_2}^*(\Psi^p_{\text{Poly}}) &= \left[\left(\frac{\mathbf{R}}{n}\right)^{i_1-1} \right]^{\mathrm{t}} \mathbf{V}_0 \left[\left(\frac{\mathbf{R}}{n}\right)^{i_2-1} \right] = \sum_{j=1}^{J} \left(\frac{R_j}{n}\right)^{i_1+i_2-2} V_{jj}^0.
\end{aligned}$$

$S^2(\Psi^p_{\text{Poly}}) =$ quadratic form from the above matrices.

Hyde's Test: A Special Case

When p = 1 with polynomial specification, we obtain:

$$S^2(\psi_1) = \frac{[\sum_{j=1}^{J}(O_j - E_j^0)]^2}{\sum_{j=1}^{J} R_j \lambda_j^0 (1-\lambda_j^0)} = \left[\frac{O_\bullet - E_\bullet^0}{\sqrt{V_\bullet^0}}\right]^2$$

$$K_i^* = \max\{j \in \{1, 2, \ldots, J\} : a_j \leq Z_i\}$$

$$O_\bullet - E_\bullet^0 = \sum_{i=1}^{n}\left(\delta_i - \sum_{j=1}^{K_i^*}\lambda_j^0\right)$$

Resulting test coincides with Hyde's ('77, Btka) test.

Adaptive Choice of Smoothing Order

$L_p^*(\theta_p)$ = partial likelihood of $\theta_p = (\theta_1, \ldots, \theta_p)$

$\mathbf{I}_p^*(\theta_p)$ = associated observed information matrix

$\hat{\theta}_p$ = partial MLE of $\theta_p = (\theta_1, \ldots, \theta_p)$

Adjusted Schwarz ('78, Ann. Stat.) Bayesian Information Criterion

$$p_{BICAdj} = \text{argmax}_{1 \le p \le P_m} \left\{ \log L_p(\hat{\theta}_p) - \frac{p}{2} \left[\log(n) + \hat{\zeta}_1 \right] \right\}$$

$\hat{\zeta}_1$ is the largest eigenvalue of $\mathbf{I}_p^*(\hat{\theta}_p)$

Simulation Results for Simple Discrete Case

Table 1: Empirical levels and powers (in percents) of the 5% asymptotic level fixed-order and adaptive tests for testing the geometric distribution. The second column contains the achieved levels, while the other columns contain the achieved powers for different hazard alternatives.

Test Statistic	Geometric (Null)	Geometric (Different Mean)	Negative Binomial	'Polynomial' Hazards	'Trigonometric' Hazards
S_1^2	4.6	52.5	2.0	11.7	8.8
S_2^2	5.1	45.8	92.8	58.2	33.9
S_3^2	5.9	41.5	90.6	53.7	83.6
S_4^2	7.6	40.3	87.9	54.3	92.1
$S_{p^*}^2$ (Adaptive)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Note: Based on polynomial-type specification. Performances of Pearson type tests were not as good as for the polynomial type.

Concluding Remarks

- Framework is general enough so as to cover both continuous and discrete cases.
- Mixed case dealt with via hazard decomposition.
- Since tests are score tests, they possess local optimality properties.
- Enables automatic adjustment of effects due to estimation of nuisance parameters.
- Basic approach extends Neyman's 1937 idea by embedding hazards instead of densities.
- More studies needed for adaptive procedures.