$P(X > Y)$ as an effect size measure for censored data: From ROC curves to Kaplan-Meier

Llorenç Badiella[1]

[1]Servei d'Estadística Aplicada, UAB, Cerdanyola (Barcelona), Spain.
E-mail: llorenc.badiella@uab.cat

Jornades de Consultoria Estadística II
October 25th, 2013

Contents

Effect size measures

In order to quantify the strenght of a relationship or the magnitude of the difference between different groups, we need effect size measures.

Table : Different effect size measures

Quantitative vs binary	$\delta = \frac{\mu_1 - \mu_2}{\sigma_c}$ (T-Test)
	$P(X > Y)$ (Mann-Whitney-Wilcoxon)
Categorical vs binary	Cramer's V
Binary vs binary	Odds Ratio
	Relative risk
Quantitative vs quantitative	Pearson correlation
	Spearman correlation
	β coefficients or R_2 in Linear Regression
Binary vs Quantitative	Odds Ratio in Logistic Regression
Survival time vs Binary	Hazard ratio - Cox PH models

Effect sizes measures are fundamental in sample size determination: they quantify magnitudes in addition to provide the basis for statistical testing.

Effect size measures

Survival Analysis and $P(X > Y)$

- Hazard Ratio is opaque.
- Hazard Ratio assumes proportional hazards
- Log-Rank test is a weighted version of Gehan's test
- Gehan's test is identical to Mann-Whitney-Wilcoxon U test (for complete cases without censored data).
- The MWW test's statistic is the non-parametric estimate of $P(X > Y)$.
- $P(X > Y)$ is the reference effect size measure for diagnostic accuracy, $P(X > Y) = AUC$
- Could we use $P(X > Y)$ as en effect size measure in survival analysis?
- How can we estimate this parameter?
- How de we proceed with inference?

Non-parametric estimation

A continuous diagnostic test measured on m healthy subjects and n diseased individuals. Let X_i and Y_j denote the observations for healthy subjects $(i = 1, ..., m)$ and diseased individuals $(j = 1, ..., n)$. Let F and G be their survival functions, and f and g their respective density functions.

$$P(X > Y) = \int_{\infty}^{-\infty} F(s)dG(s)$$

Non-parametric estimation

The empirical nonparametric estimation of $P(X > Y)$ is given by:

$$\widehat{P}_W(X > Y) = \int_{\infty}^{-\infty} \widehat{F}(s) d\widehat{G}(s) = \sum_{j=1}^{n} \widehat{F}(y_j)\widehat{g}(y_j)$$

where

$\widehat{F}(y_j) = \widehat{P}_W(X > y_j) = \frac{1}{m} \sum_{i=1}^{m} I(x_i > y_j)$ and $\widehat{g}(y_j) = \frac{1}{n}$

Non-parametric estimation

Moreover, in the presence of tied data we define:
$P(X \geq Y) = P(X > Y) + \frac{1}{2}P(X = Y)$.

$$\widehat{P}_W(X \geq Y) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Psi_W(X_i, Y_j) = \frac{W}{mn}$$

where

$$\Psi_W(X_i, Y_j) = \begin{cases} 1 & X_i > Y_j, \\ 0.5 & X_i = Y_j, \\ 0 & X_i < Y_j. \end{cases}$$

and W is the Mann-Whitney-Wilcoxon statistic for the two-sample problem.

Properties of $P_W(X > Y)$

- W/mn represents the natural non-parametric estimate of $P(X > Y)$.
- The statistic W can be seen as a two-sample U-Statistic having kernel $h(x, y) = I(x > y)$ (Hetmansperger, 1984). This approach leads to obtaining an estimate of its asymptotic variance (DeLong et al, 1988).
- $P(X > Y)$ is an important effect size measure:
  - X, Y normally distributed with common σ, $\delta = (\mu_X - \mu_y)/\sigma$ then $P(X > Y) = \Phi(\delta/\sqrt{2})$
  - X, Y exponentially distributed with parameters θ_1 and θ_2 then $P(X > Y) = \frac{\theta_1}{\theta_1 + \theta_2}$
  - X, Y binary variables then $P(X > Y) + 0.5P(X = Y) = Se/2 + Sp/2$

Notation

Now consider, $U_1, \ldots, U_m$ and $V_1, ..., V_n$ with $K(s)$ and $J(s)$ as their survival functions, also continuous.
Instead of directly observing X_i and Y_j as before, we only observe:

$$X_i^c = \min(X_i, U_i) \quad \xi_i = I(X_i < U_i)$$
$$Y_j^c = \min(Y_j, V_j) \quad \nu_j = I(Y_j < V_j)$$

Assume that $X_1^c < \cdots < X_m^c$ and $Y_1^c < \cdots < Y_n^c$. Denote their survival distributions as $F^c(s) = F(s)K(s)$ and $G^c(s) = G(s)J(s)$ respectively.

Non-parametric tests to compare survival distributions

Different nonparametric test statistics can be constructed to contrast the null hypothesis: $H_0 : F = G$ (Prentice & Marek, 1979; Leton, 2007).

- Gehan test (assuming $K = J$)
  - consistent against alternatives where G is stochastically greater than F: $F(s) < G(s), \forall s$
- Log-Rank test (assuming $K = J$ and proportional Hazards)
  - It is a weighted version of Gehan's test. These weights represent the expected values of exponential order statistics.
  - Log Rank test can be shown to be the most efficient when hazard or survival functions are proportional to each other.
- Peto-Peto (efficient under proportional odds)

However, they do not provide any reliable effect size measure.

Naive estimator: Gehan's test and Harrell's c

Gehan's test can be seen as a generalization of Mann-Whitney's test in the following sense:

$$W_G = \sum_{i=1}^{m} \sum_{j=1}^{n} \Psi_G(X_i^c, Y_j^c),$$

where

$$\Psi_G(X^c, Y^c) = \begin{cases} 1 & X_i^c \geq Y_j^c \text{ and } \nu_j = 1 \\ 0.5 & X_i^c = Y_j^c \text{ and } \xi_i = \nu_j \\ 0.5 & X_i^c > Y_j^c \text{ and } \nu_j = 0, \\ 0.5 & X_i^c < Y_j^c \text{ and } \xi_i = 0 \\ 0 & X_i^c \leq Y_j^c \text{ and } \xi_i = 1. \end{cases}$$

Naive estimator: Gehan's test and Harrell's c

In order to estimate $P(X > Y)$ under random censorship, Harrell (1982) proposed to exclude from computations uninformative pairs. Harrell's C estimator can be obtained as:

$$\widehat{P}_G(X > Y) = \frac{\sum_{i,j \in Ip} \Psi_G(X_i^c, Y_j^c)}{\# Ip}$$

where Ip denotes de set (i, j) of informative pairs.

Gehan's test and Harrell's c

Harrell's c is commonly used to obtain an estimate of $P(X > Y)$ in the context of reliability, however

- This estimate is biased except when $P(X > Y) = 1/2$ (Koziol and Jia, 2009) and depends on the censorship distributions K and J.
- Thus, Gehan's test is not formally a non-parametric test (asymptotically it is), since under H_0 its distribution will depend on the relation between K and J.

Efron's test

Efron (1967) obtained the maximum likelihood estimator of the parameter $P(X > Y)$ in the presence of random right-censorship and proposed an estimator of its variance.
Efron's estimation method consists in substituing from

$$P(X > Y) = \int_{\infty}^{-\infty} F(s)dG(s)$$

the distributions F and G by their Kaplan-Meier's estimates.

Efron's test

Let $\widehat{F}$ and $\widehat{G}$ be the Kaplan-Meier estimates of both survival functions, treating the large observation for each group as if it were uncensored, i.e. $\xi_m = 1$ and $\nu_n = 1$. Then,

$$\widehat{P}_E(X > Y) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \Psi_E(X_i^c, Y_j^c; \xi_i, \nu_j)$$

where the values of the function Ψ_E are:

Efron's test

Table : Values of $\Psi_E(X_i^c, Y_j^c; \xi_i, \nu_j)$

(ξ_i, ν_j)	$X_i^c > Y_j^c$	$X_i^c < Y_j^c$
$(1,1)$	1	0
$(0,1)$	1	$\frac{\widehat{F}(Y_j^c)}{\widehat{F}(X_i^c)}$
$(1,0)$	$1 - \frac{\widehat{G}(X_i^c)}{\widehat{G}(Y_j^c)}$	0
$(0,0)$	$1 - \frac{\widehat{G}(X_i^c)}{\widehat{G}(Y_j^c)} - \int_{X_i^c}^{\infty} \frac{F(s)dG(s)}{F(X_i^c)G(Y_j^c)}$	$-\int_{Y_j^c}^{\infty} \frac{F(s)dG(s)}{F(X_i^c)G(Y_j^c)}$

Efron's test

- Efron's estimate is an asymptotically unbiased estimate of $P(X > Y)$ under censored data.
- It does not depend on the relationship between K and J.
- Efron claimed that the test based on $\widehat{P}_E(X > Y)$ should be more powerful than Gehan's test, however, being sensitive to heavy censoring.
- Computational requirements inhibited its use in practice.
- Latta (1977) showed the equivalency of Peto and Peto's test to a modification of Efron's test under the null hypothesis $P(X > Y) = 1/2$.

Efron's test

More recently,

- Bose & Sen, 2002 showed that this statistic can be viewed as a Kaplan-Meier U-Statistic for censored data (with the kernel $h(x, y) = I(x > y)$) and derived its asymptotic normality,
- Stute, 1993 showed its consistency.
- Datta, 2010 expressed Kaplan-Meier U-Statistics using inverse probability of censoring weights (IPWC), i.e. weights based on the inverse of the censoring survival distribution.

Efron's test

- The K-M estimators can be seen as: the mass for censored data is redistributed equally among all greater values.
- This is equivalent to reweight each event with the inverse of the probability of censorsghip.

Thus,

$$\widehat{P}_E(X > Y) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{h(X_i^c, Y_j^c)\xi_i \nu_j}{\widehat{K}(X_i^c)\widehat{J}(Y_j^c)}}{mn}$$

with $h(x, y) = I(x > y)$.

Efron's test

- Under this approach, only event times are taken into account.
- Both survival distributions for the censoring times K and J can be estimated using Kaplan-Meier approach reversing the status indicator.
- Thus, asymptotic normality and an expression for the variance of $\widehat{P}_E(X > Y)$ under random censorship can be derived from theory developed for IPWC U-statistics.

Efron's test

Using Datta's notation, define:

$$V_i^{01} = \frac{\widehat{h}_{01}(X_i)\xi_i}{\widehat{K}(X_i)} + \widehat{w}_{01}(X_i)(1-\xi_i) - \sum_{k=1}^{m} \frac{\widehat{w}_{01}(X_k)(1-\xi_k)I(X_i \geq X_k)}{\sum_{l=1}^{m} I(X_l \geq X_k)}$$

$$V_j^{10} = \frac{\widehat{h}_{10}(Y_j)\nu_j}{\widehat{J}(Y_j)} + \widehat{w}_{10}(Y_j)(1-\nu_j) - \sum_{k=1}^{n} \frac{\widehat{w}_{10}(Y_k)(1-\nu_k)I(Y_j \geq Y_k)}{\sum_{l=1}^{n} I(Y_l \geq Y_k)}$$

Efron's test

where

$$\widehat{h}_{01}(X) = \frac{\sum_{j=1}^{n} h(X, Y_j) \frac{\nu_j}{\widehat{J}(Y_j)}}{n}$$

$$\widehat{h}_{10}(Y) = \frac{\sum_{i=1}^{m} h(X_i, Y) \frac{\xi_i}{\widehat{K}(X_i)}}{m}$$

$$\widehat{w}_{01}(X) = \frac{1}{\sum_{l=1}^{m} I(X_l > X)} \sum_{k=1}^{m} \frac{\widehat{h}_{01}(X_k)\xi_k}{\widehat{K}(X_k)} I(X_k > X)$$

$$\widehat{w}_{10}(Y) = \frac{1}{\sum_{l=1}^{n} I(Y_l > Y)} \sum_{k=1}^{n} \frac{\widehat{h}_{10}(Y_k)\nu_k}{\widehat{J}(Y_k)} I(Y_k > Y)$$

Efron's test

Then,

$$\widehat{\mathrm{Var}}[\widehat{P}_E(X > Y)] = \frac{\mathrm{Var}[V_i^{01}]}{4m} + \frac{\mathrm{Var}[V_j^{10}]}{4n} \tag{1}$$

Finally, since $\widehat{P}_E(X > Y)$ is asymptotically normal, the null hypothesis $H_0 : P(X > Y) = 1/2$ (equivalently $H_0 : F = G$) can be tested through:

$$\frac{\widehat{P}_E(X > Y)}{\sqrt{\widehat{\mathrm{Var}}[\widehat{P}_E(X > Y)]}} > \Phi^{-1}(1 - \alpha/2) \tag{2}$$

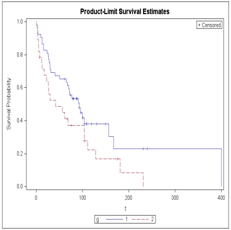
where α is the desired significance level and Φ is the standard Normal cumulative distribution function.

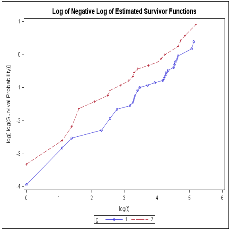
Efron's test

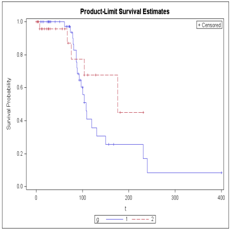
- We have provided an effect size measure for comparing two survival distributions.
- We have obtained confidence intervals.
- Testing $H_0 : F = G$ using Efron's test does not assume any relationship between K and J, and is consistent against alternatives where $P(X > Y) \neq 1/2$.

Example 1

Data on 80 males diagnosed with cancer of the tongue Tumor, time to death is compared between two different DNA profiles (1=Aneuploid Tumor or 2=Diploid Tumor). Reference: Sickle-Santanello et al. Cytometry 9 (1988): 594-599. Extracted from Klein & Moeschberger (2003).







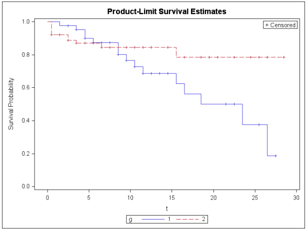
Example 1

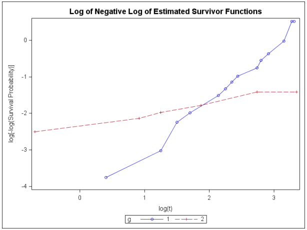
Table : Example 1 results

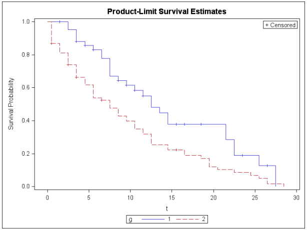
Test	P-Value	$P(X > Y)$	CI (95%)
Gehan	0.0666	0.6364	
Log-Rank	0.0601		
Peto-Peto	0.0626		
Efron	0.1023	0.6339	(0.4733 - 0.7945)

Example 2

Data on 119 kidney dialysis patients. Time to infection is compared between two different Catheter placements (1=surgically, 2=percutaneously).
Reference: Nahman el at. J. Am Soc. Nephrology 3 (1992): 103-107.
Extracted from Klein & Moeschberger (2003).







Example 2

Table : Example 2 results

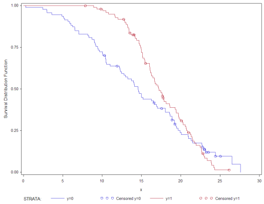
Test	P-Value	$P(X > Y)$	CI (95%)
Gehan	0.9636	0.5054	
Log-Rank	0.1117		
Peto-Peto	0.2369		
Efron	0.1147	0.1898	(0 - 0.5752)

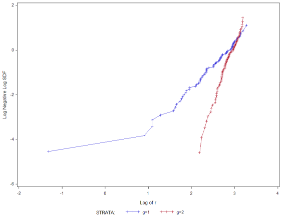
Example 2

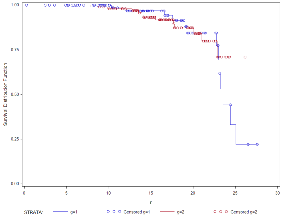
Synthetic Example:

$X \sim N(17, 4)$ $Y \sim N(13, 8)$ and common censorship

$U \sim N(5, 25)$, $m = m = 100$







Example 3

Table : Example 3 results

Test	P-Value	$P(X > Y)$	CI (95%)
Gehan	0.0004	0.6524	
Log-Rank	0.3389		
Peto-Peto	0.0008		
Efron	0.0026	0.6306	(0.5443 - 0.7168)

Simulation study

In this section, we compare the performance of several tests for two sample comparisons in presence random censorship: Gehan, Log-Rank, Peto-Peto and Efron's test using the proposed variance estimator. Note that under the null hypothesis, Gehan and Peto tests are equivalent. In absence of censored data, Gehan, Peto-Peto and Efron's test are similar. The power of these tests is obtained under different scenarios:

- Scenario 1: $X \sim Exp(\rho)$, $Y \sim Exp(1)$, without censorship.
- Scenario 2: $X \sim Exp(\rho\phi)$, $Y \sim Exp(\phi)$, $U \sim Exp(1)$, $V \sim Exp(1)$

where $P(X > Y) = \rho/(1+\rho)$, $\phi/(1+\phi)$ represents the percentage of censored data in the second population. Each scenario is based on 2000 replications for each combination of of $P(X > Y) = 0.5$ (null hypothesis), 0.6 (small differences) and 0.7 (high differences), with (m, n)=(50,50), (30,100), (100,30).

Simulation study

Results of the simulation study are presented in Tables 6- **??**.

Table : Power (%) of several test for $H_0 : F = G$. Scenario 1

Sample	Y Censorship	$P(X > Y)$	Efron	Log-Rank	Gehan	Peto-Peto
$m = n = 50$	0%	0.5	0.050	0.055	0.046	0.046
		0.6	0.412	0.505	0.401	0.401
		0.7	0.943	0.983	0.940	0.940
$m = 100, n = 30$	0%	0.5	0.056	0.062	0.053	0.053
		0.6	0.378	0.473	0.359	0.359
		0.7	0.902	0.972	0.910	0.910
$m = 100, n = 30$	0%	0.5	0.055	0.057	0.048	0.048
		0.6	0.416	0.507	0.413	0.413
		0.7	0.959	0.983	0.947	0.947
$m = n = 50$	30%	0.5	0.050	0.057	0.047	0.048
		0.6	0.323	0.372	0.28	0.313
		0.7	0.851	0.898	0.792	0.843
$m = 100, n = 30$	30%	0.5	0.056	0.061	0.059	0.055
		0.6	0.257	0.308	0.234	0.260
		0.7	0.762	0.851	0.710	0.779
$m = 100, n = 30$	30%	0.5	0.052	0.060	0.054	0.055
		0.6	0.367	0.382	0.306	0.325
		0.7	0.893	0.914	0.811	0.852

Results

Simulation Results

- When no censorship is present all tests perform equivalently.
- Considering a moderate censorship rate, Log-Rank performs better than the other tests.

Discussion

A new approach to constructing nonparametric confidence intervals for the $P(X > Y)$ index in presence of censored data has been presented.

$P(X > Y)$ Index

- The estimation of $P(X > Y)$ and its variance are based on IPCW U-Statistics.
- Confidence intervals can be easily computed, allowing non-inferiority evaluations.
- This method is not affected by the nature of the censorship distributions.
- In can be an interesting effect size measure to report in survival analysis.

Discussion

Comparison to other tests

- Gehan's test is optimal against location shift alternatives: $F(x - \theta)$. However it is highly sensitive to differences in censorship patterns.
- Log Rank test is optimal against scale shift alternatives: $F(x/(1 + \theta))$. Unequal censorship patterns, may give inflated power.
- Efron's test is efficient in both situations.

References

- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387-415.
- DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837-845.
- Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. New York: Springer.
- Newcombe, R.G. (2006a). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1, general issues and tail-area-based methods. *Statistics in Medicine*, **25**, 543-557.
- Newcombe, R.G. (2006b). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Statistics in Medicine*, **25**, 559-573.
- Pepe, M.S. (2003).*The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

$P(X > Y)$ as an effect size measure for censored data: From ROC curves to Kaplan-Meier

Llorenç Badiella[1]

[1]Servei d'Estadística Aplicada, UAB, Cerdanyola (Barcelona), Spain.
E-mail: llorenc.badiella@uab.cat

Jornades de Consultoria Estadística II
October 25th, 2013