MWSUG 2016 - Paper RF08

The Power of Cumulative Distribution Function (CDF) Plot in Assessing Clinical Outcomes

Min Chen and Patricia Kultgen, Cook Research Inc., West Lafayette, IN 47906

ABSTRACT

Clinical endpoints (e.g., antibody concentrations, quality of life scores) are generally collected quantitatively at baseline and follow-up windows in clinical trials. One popular way of assessing efficacy is to use a prospectively determined criterion to define success. One caveat of this method is that the pre-defined criterion is generally subjective and might not be the best for specific studies. This is especially true for trials at exploratory phases. In addition, the success/failure outcome does not illustrate the whole story. The purpose of this paper is to encourage the use of the Cumulative Distribution Function (CDF) plot to display the empirical data, and evaluate outcomes at multiple standards. All the graphs and statistics are generated using SAS® 9.2 in window system, and only basic SAS skills are required to repeat these efforts.

INTRODUCTION

In clinical trials, treatment outcomes are often presented as the percentage of patients who met a specific criterion. Examining the outcome data in this way is simple, but limited, especially in exploratory phase studies. An additional method that may be used is the cumulative distribution function (CDF), which is a powerful tool to graphically illustrate data range and distribution since the plots present the probabilities of all possible X values.

CDF plots have been effectively used to assess a variety of clinical topics[3, 4]. Despite their strengths, CDF plots are generally under-utilized. To encourage the use of CDF, this paper provides step-by-step instructions for generating CDF plots using SAS. For this example, one random dataset, DATASET, was simulated. Variable ARM is a two-level trial arm indicator, and ENDPOINT represents the primary endpoint of interest which could simply be the raw measurements or changes in clinical outcomes at time of interest.

BASIC CDF PLOT

Basic CDF plots, which is the probability of the value on x-axis to be less or equal to x, that is $P(X \leq x)$, can be produced with Proc UNIVARIATE or GPLOT procedure. Compared to UNIVARIATE, GPLOT requires more programming, but provides more flexibility. Hereafter, all the examples are demonstrated using GPLOT.

Cumulative percentage of the clinical outcome ENDPOINT is obtained from the output dataset with FREQ procedure by ARM variable. The CDF is then plotted against ENDPOINT values on the x-axis with GPLOT procedure (Figure 1A):

```
proc freq data = dataset;
      by arm;
      tables Endpoint / outcum out=cdf;
run;

data cdf;
      set cdf;
      by arm Endpoint;
      cdf = Cum_Pct;
      output;
```

```
run;

proc sort data=cdf;
      by arm Endpoint cdf;
data cdf2;
      set cdf;
      cdf2 = 100 - cdf;
      if cdf <= 50 then cdf3 = cdf;
            else if cdf > 50 then cdf3 = cdf2;
      keep arm Endpoint cdf cdf2 cdf3;
run;

axis1 value = (h=1.5) label = (a = 90 r = 0 h = 1.5 'Cumulative percentage
(%)') order=0 to 100 by 10;
axis2 value = (h=1.5) label = (h = 1.5 "Endpoints");
axis3 value = (h=1.5) label=(a=90 h = 1.5 'Difference') order=0 to 100 by 10;

symbol1 h = 1 f = marker c = black w = 2.0 l = 1 i = steplj;
symbol2 h = 1 f = marker c = black w = 2.0 l = 2 i = steplj;
symbol3 h = 0.5 f = marker c = grey w = 1.2 l = 33 i = steplj;

legend1 order=("A" "B") value=(h=1.5 "A" "B") label=none ;
legend2 order=("Difference") value=(h=1.5 "Difference") label=none ;

goptions border;
proc gplot data=cdf2;
      plot cdf*Endpoint = arm /vaxis=axis1 haxis=axis2 lhref=4 lvref=4 grid
legend = legend1 ; *basic CDF;
      plot cdf2*Endpoint = arm /vaxis=axis1 haxis=axis2 lhref=4 lvref=4  grid
legend = legend1 ; *CCDF;
      plot cdf2*Endpoint = arm /vaxis=axis1 haxis=axis2 lhref=4 lvref=4  grid
legend = legend1 hreverse ; *CCDF with reverse x-axis;
      plot cdf3*Endpoint = arm /vaxis=axis1 haxis=axis2 lhref=4 lvref=4  grid
legend = legend1 ; *FCDF;
run;
quit;
```

COMPLEMENTARY CDF PLOT

Frequently, the interest is the opposite of basic CDF, that is $P(X \geq x)$. For example, in clinical studies examining quality of life scores, the primary interest is the percentage of participants with at least certain improvement post-intervention.[4] In this case, a complementary (or reverse) CDF plot is easier to comprehend. Complementary CDF (CCDF), denoted as CDF2 in the code, is simply 100-CDF in the percentage scale (Figure 1B), and the plot is a mirror image of Figure 1A. Additionally, a HREVERSE option could be added to the PLOT statement, which would reverse the order of x-axis values, as shown in Figure 1C.

FOLDED CDF PLOT

A Folded CDF (FCDF) plot, also called mountain plot, combines basic (lower half) and complementary (upper half) CDFs. Instead of an S-shaped plot, the upper half of the basic CDF is "folded over" to produce a mountain-shape curve (Figure 1D). This plot emphasizes the median, which is represented by the peak of the curves. In addition, the area under an FCDF curve is the mean absolute deviation (MAD) from the median[5], which is a measurement for dispersion of the distribution.

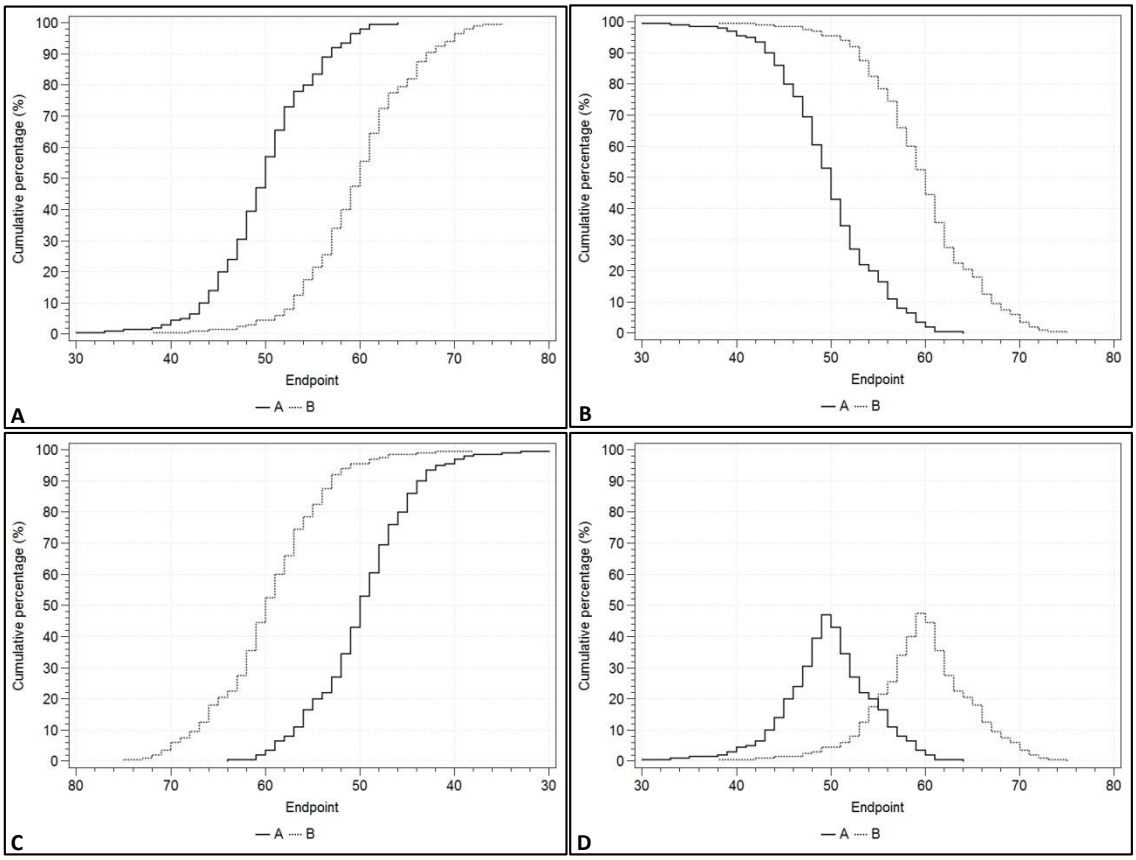

Figure 1. A: basic CDF; B: complementary CDF; C: Complementary CDF with reversed x-axis; D: Folded CDF

ADD REFERENCE LINES OR RIGHT Y-AXIS

In clinical trials, treatment effect is based on the difference in outcomes between groups. In addition to looking at a pre-defined cutoff, CDF plots could be used to explore the treatment effects across all x-axis values. Treatment effects across the whole X spectrum could be plotted against the right y-axis, which may be in the same or different scales with the left y-axis (Figure 2). This is extremely useful in early phase trials where the cut-offs for success and failure are among the parameters explored.

The output dataset CDF from the previous analyses was transposed from long into wide format so that the CDF difference at each ENDPOINT value, DIFFERENCE variable, could be easily calculated. A RETAIN statement is used to carry the last observed CDF over a missing value within each arm. Under GPLOT procedure, the PLOT statement is specified to obtain the basic CDF, which is slightly different from the earlier example as the data format changes. A second plot statement, PLOT2, is needed to plot the DIFFERENCE on the right y-axis against the same x-axis. The corresponding ENDPOINT value with maximum DIFFERENCE is identified using Proc SQL procedure, and is added as one of the two reference lines, with the predefined one assumed to be 50. Here are the codes to produce Figure 2:

```
proc sort data=cdf;
      by Endpoint arm;
proc transpose data=cdf out=cdf_tr;
      by Endpoint;
      id arm;
      var cdf;
run;
```

```
data cdf tr;
      set cdf tr;
      by  name ;
      retain  A  B;
      if first. name_ then do;
            A = A;
           _B = 0;
      end;
      else do;
      if missing(A) then A=_A;
            else  A = A;
      if missing(B) then B=_B;
            else _B = B;
      end;
      Difference = _A - _B;
      drop _A _B;
run;

proc sql;
      select Endpoint into :maxref
      from cdf tr
      having difference = max(difference);
quit;

proc gplot data=cdf tr;
      plot (A B)*Endpoint /overlay vaxis=axis1 haxis=axis2 href=50 &maxref.
            lhref=3 lvref=3 grid legend = legend1 ;
      plot2 Difference*Endpoint /overlay vaxis=axis3 legend = legend2 ;
run;
```

Keep in mind, however, that the clinical interpretation of the cutoff is equally, or in some sense more, important than statistical significance. A final decision should be based on both clinical and statistical significance.

KOLMOGOROV-SMIRNOV (K-S) TEST

Kolmogorov-Smirnov (K-S) test may be used to formally differentiate the location of two CDF curves. Notably, the K-S test is not based on any data distribution assumptions, and does not rely on a predefined cutoff to compare rates. In SAS, the NPAR1WAY procedure computes the empirical distribution function (EDF), and tests whether the distribution of the specified variables varies across groups:

```
proc npar1way data=dataset;
      class arm;
      var Endpoint;
run;
```

In addition to a p-value from the K-S test, the output provides the corresponding ENDPOINT value where the maximum DIFFERENCE is achieved (output not shown), which is same as the one requested from the SQL procedure above.

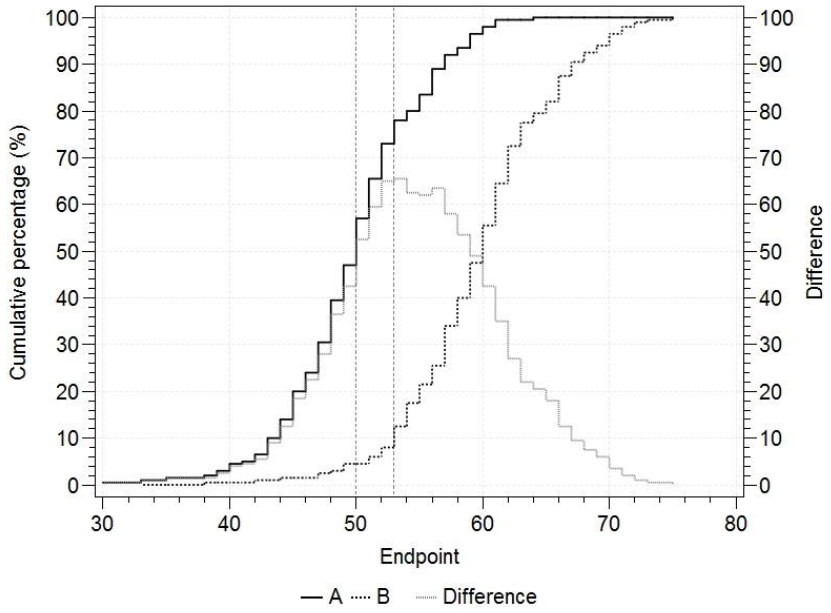

Figure 2. A CDF plot with an additional right y-axis and reference lines.

CONCLUSION

CDF plots are powerful in visually presenting a complete data set, and providing insights on assessing clinical outcomes across groups.

REFERENCES

1. Liu X and Millar S. Producing Cumulative Distribution Frequency Figures with Minimal Important Difference Reference Lines to Assess Quality of Life Treatment Effects. http://www.wuss.org/proceedings11/Papers_Liu_X_75059.pdf

2. Yang AM and Chen HL. The Power of PROC GPLOT in Statistical Analysis of Clinical Studies. http://www.lexjansen.com/pharmasug/2008/po/PO18.pdf

3. Zimmerman RK, Lin CJ, Raviotta JM, Nowalk MP. Do vitamin D levels affect antibody titers produced in response to HPV vaccine? Hum Vaccin Immunother. 2015;11(10):2345-9.

4. Cella D, Ivanescu C, Holmstrom S, Bui CN, Spalding J, Fizazi K. Impact of enzalutamide on quality of life in men with metastatic castration-resistant prostate cancer after chemotherapy: additional analyses from the AFFIRM randomized clinical trial. Ann Oncol. 2015 Jan;26(1): 179-85

5. Xue, JH; Titterington, DM; (2011) The p-folded cumulative distribution function and the mean absolute deviation from the p-quantile. STAT PROBABIL LETT, 81 (8) 1179 - 1182.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Min Chen
Cook Research Inc.
1 Geddes Way
West Lafayette, IN 47906
Min.Chen@CookMedical.com

Patricia Kultgen
Cook Research Inc.
1 Geddes Way
West Lafayette, IN 47906
Patricia.Kultgen@CookMedical.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.