BIOEN 3070/6070: Introduction to Statistics for Bioengineers

Advanced Topics: Survival Analysis

Introduction

Survival analysis is an area of statistics in which we are interested in studying the time until an event occurs. Survival analysis has applications in a diverse set of fields, including biostatistics, engineering, economics, sociology, and political science. This tutorial considers a common example of survival analysis - the study of the time until death for a set of patients who are receiving a particular treatment. We want to study the amount of time (generally referred to as the **survival time**) until the event of interest occurs.

Learning Objectives

In this tutorial you will learn how to:

1. Plot and interpret survival data using Kaplan-Meier curves.
2. Identify the probability distribution underlying the data at each time point and derive characteristic properties.
3. Develop a hypothesis test and derive an appropriate test statistic.
4. Draw conclusions about the underlying distributions of the survival data.

Additional Resources

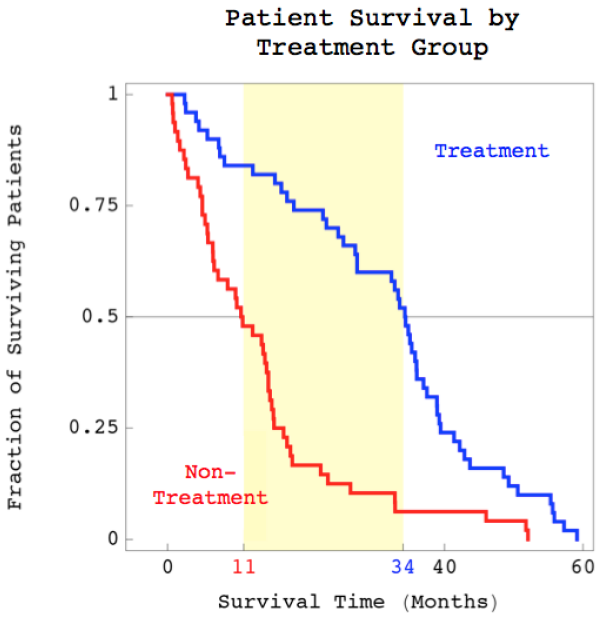
We highly recommend the following readings to help you with this tutorial. Online versions of both sources are available for free through the University of Utah J. Willard Marriott Library.

1. Liu, X. (2012). *Survival Analysis: Models and Applications* (pp. 46-51) . Wiley.
2. Kleinbaum, D. G. and Klein M. *Survival Analysis: A Self-Learning Text*, (pp. 45-82). Springer.

Kaplan-Meier Curves

Survival analysis is an area of statistics in which we are interested in studying the time until an event occurs. Consider the scenario of a set of patients who have been diagnosed with Condition Y. There is currently a clinical trial for Treatment X, a new drug therapy that is believed to treat Condition Y. We have a set of patients who are given the treatment and their survival time is recorded. These data can be represented as a vector comprising the survival time for each patient in the trial.

A useful way to visualize and interpret these data is with a **Kaplan Meier (KM) curve**. A KM curve is a plot of the fraction of surviving patients from the original cohort over time. The curve will be stepwise because it is generated from discrete data.



Suppose the clinical trial also includes a control group that is not given Treatment X. We can represent their survival times in a separate vector and use these values to create a second KM curve. We now have a way to visualize the survival curves for our two groups, as shown in the figure above.

1 Interpreting KM curves

One way to interpret the KM curves is to look at the separation between curves. If Treatment X is effective we would likely expect a large separation between the curves. However, if the treatment is no better than the current standard of care (assuming that is what is given to the non-treatment group) we would not expect to find a large separation between the groups.

a) What is the median survival time for the group of patients receiving Treatment X?

KM curves provide an intuitive way to visualize survival data and may give us an idea about whether or not the survival times for the treatment group and the non-treatment group are similar. However, we want to use our knowledge of statistics to make a quantitative comparison between the two groups.

Probability Distribution of Survival Data

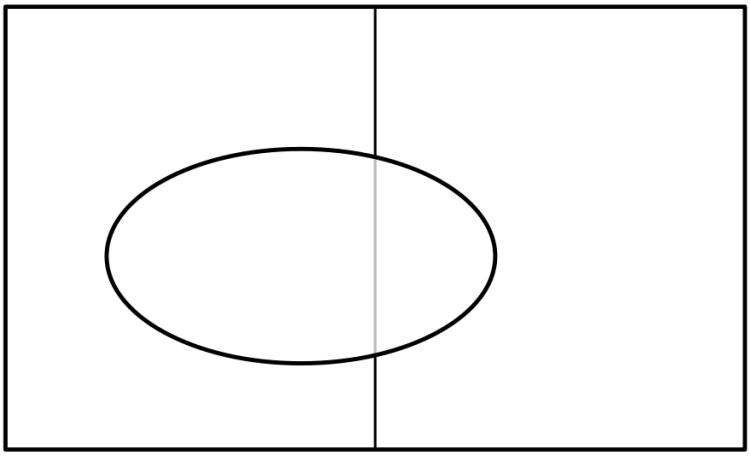
Consider our two survival groups at time t, where t is any observed survival time from either group. Suppose the total group of N patients that is comprised of patients from two distinct groups, G_1 and G_2. Of these patients, K die at time t. At time t there are n_1 and n_2 patients exposed to risk in G_1 and G_2, respectively. There are k_1 and k_2 patients who die at time t from groups G_1 and G_2, respectively.

2 Setting up the Distribution

a) Fill in the contingency table below using the variables introduced in the paragraph above.

Group	Die at $T = t$	Do not die at $T = t$	Total
G_1			
G_2			
Total			

b) Label the following variables in the Venn diagram below: N, K, n_1, n_2, k_1, and k_2.



c) For each time point, t_i, we can write a distribution for the number of patients from G_1 who die at t_i. Write down the name of the distribution.

d) Refer to your contingency table or Venn diagram and write the pmf for this distribution.

e) Why do we only need to consider G1 in our pmf? What does this say about the relationship between G1 and G2?

3 Characterizing the Distribution

In order to characterize our data, we need to calculate some properties of the distribution. In this problem, we will derive the expressions for the mean and variance of the hypergeometric distribution described above.

a) Show that the mean of the hypergeometric distribution described above is $\frac{Kn_1}{N}$ by using the following identity and the definition of mean as the expected value of x:

$$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$$

Hint: This proof relies on a very common trick in statistics we call summing without summing (or integrating without integrating for the continuous case). We know that the sum of a pmf over all possible values of a random variable is 1 (why?). Therefore we can complete this derivation by manipulating the equation such that everything left in the sum is a (properly normalized) pmf that we recognize, so it becomes 1. To sum without summing for this problem, first use the identity above to expand two of the terms in the pmf. Then perform a change of variable to $l = x-1$ and re-index the sum from $1 \leq x \leq n$ to $0 \leq l \leq n-1$. Name and explicitly label the parameters of the resulting distribution that sums to 1.

b) Show that the variance of the hypergeometric distribution described above is

$$\frac{Kn_1(K-1)(n_1-1)}{N(N-1)} + \frac{Kn_1}{N} - \left(\frac{Kn_1}{N}\right)^2.$$

Hint: Use the formula $\mathrm{Var}(x) = \mathrm{E}[x^2] - (\mathrm{E}[x])^2$ and follow steps similar to the derivation of the mean. In calculating $\mathrm{E}[x^2]$, we can apply the same process as we did in the mean twice. After applying the identity the first time, we will arrive at the same sum as we ended with in the mean, but now there is an x in it. If this x were $x-1$ instead, the expression would be for the mean of the distribution $P(x-1; N-1, k-1, n_1-1)$ (which we now know how to handle). We can make this happen with an old favorite trick - replace x with $x-1+1$. It is then possible to split the sum, leaving one term that is the mean of the distribution just mentioned, and one term that sums to 1.

4 The Standard Normal Distribution

One important theorem in statistics is the Central Limit Theorem (CLT) which states that a sum of many independent and identically distributed random variables will tend to follow a Gaussian distribution with mean and variance of the distribution that the random variable came from. In order to work with the Gaussian distribution, it is best to make it into a standard normal distribution (i.e. a Gaussian with mean 0 and variance 1). The pdf of a Gaussian is given by

$$P(x; \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean and σ is the standard deviation (or square root of the variance).

a) Show that the Gaussian given by $P(\frac{x-\mu}{\sigma}; \sigma, \mu)$ has mean 0 and variance 1, i.e. it is the standard normal distribution.

Hint: There are two ways to solve this one. The harder way is to actually do the integrals as we did in Problem 3. The easy way is to remember the following properties of mean and variance (where a, b are constants and X, Y are random variables):

$$\mathrm{E}[a] = a \tag{1}$$
$$\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y] \tag{2}$$

$$\mathrm{Var}[a] = 0 \tag{3}$$
$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}[X, Y] \tag{4}$$

Comparison of Survival Groups using the Log-Rank Test

In order to determine whether or not our two groups, G_1 and G_2, have different survival distributions we need to compare them quantitatively. We can do this by performing a **hypothesis test**. Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.

A hypothesis test considers two hypotheses - the **null hypothesis**, H_0, and the **alternate hypothesis**, H_a. We evaluate our hypothesis test with some predetermined significance level, α, which we will assume to be $\alpha = 0.05$. For a more complete review of hypothesis testing, see Chapter 9 in *Bulmer* (pp. 139-164).

Recall that we generally choose H_0 to be the hypothesis that the observed phenomenon is due to chance. Then, H_1 is the hypothesis that the experiment was designed to test. It is usually the hypothesis that the observed phenomenon is due to a real effect. Therefore, in our case the null hypothesis is that the two groups, G_1 and G_2, have identical survival distributions. The alternate hypothesis is that the two groups come from different underlying survival distributions.

Now that we have identified our null hypothesis and alternate hypothesis, we need a quantitative way to evaluate them. We do this with a **test statistic**, which is a mathematical formula that allows us to determine the likelihood of obtaining sample outcomes if H_0 were true.

5 Define the Hypothesis Test

a) Fill in the table below with the appropriate values and hypotheses:

Assuming the total number of time points is large, one appropriate test for this scenario is the chi-square (χ^2) test (see p. 154 in *Bulmer*). To perform this test, we need a statistic that

α	
H_0	
H_a	

has a χ^2 distribution. Conveniently, the χ^2 distribution with j degrees of freedom (denoted $\chi^2(j)$) is simply a sum of j random variables from the standard normal distribution.

Let $k_1(t_i)$ denote the number of patients from group G_1 who die at time t_i (this is the same as k_1 above, but with the time dependence written explicitly). Remember that the survival curve tells us the survival time of the patients. Since the groups are assumed to be identical under the null hypothesis, each patient is an independent observation from the survival distribution. Therefore the sum over time points of k_1, $\sum_i k_1(t_i)$, is Gaussian by the CLT.

6 Creating a Test Statistic

a) Recall that $k_1(t_i)$ has a hypergeometric distribution. Recall that $\sum_i k_1(t_i)$ is Gaussian by the CLT. What are the values of μ and σ for this Gaussian distribution?

b) How can we modify the statistic $\sum_i k_1(t_i)$ to have a standard normal distribution?

c) Use the statistic that you found in Problem 6 b) to generate a test statistic that is χ^2 distributed with 1 degree of freedom.

We can now use this formula to compute the test statistic for our data. However, we ultimately want a P-value that tells us the probability of obtaining the sample mean if H_0 is true. To make this conversion, we consult a χ^2 distribution table (p. 234 in *Bulmer*) or use a software package with basic statistical capabilities. We now have a P-value that we can compare to our predetermined value of α. If $p < \alpha$ we reject H_0. If $p \geq \alpha$ we do not reject H_0.

7 Interpreting a Hypothesis Test

It is important to note that the conclusions drawn from a hypothesis test are always given in terms of H_0. There are two possibilities: (a) we *reject* H_0 in favor of H_a, or (b) we *do not reject* H_0. We can never conclude that we accept (or prove) H_a. If our conclusion is that we do not reject H_0 this only means that there is not sufficient evidence to refute H_0 in favor of H_a for the given significance level. If our conclusion is that we reject H_0 we infer that H_a *may* be true.

What would you conclude from this hypothesis test for the following values of the test statistic where $Q \sim \chi^2(1)$? What can you infer about G_1 and G_2?

a) Q = 2.89

b) Q = 7.14

```
(* BIOEN 3070/6070: Introduction to Statistics for Bioengineers *)
```



```
(* Advanced Topics: Survival Analysis *)
```

```
(* General Commands *)

Clear["Global`*"]

(* Define Path *)

path = "http://www.alterlab.org/teaching/BIOEN3070/discussions/";

(* This block of code contains all the functions to create the Kaplan-
 Meier survival curves. Leave it as is and evaluate the cell. *)

monthConversion = (365.25 / 12);
Clear[framex, framey, labelx, labely, xlabel, ylabel];

annotationColumn[annotation_] :=
  Transpose[annotations][[Position[annotationnames, annotation][[1, 1]]]];
options[annotation_] := Intersection[annotationColumn[annotation]];

optionNumbers[annotation_] := Dimensions[options[annotation]][[1]];
optionCounts[annotation_] :=
  Table[Count[annotationColumn[annotation], options[annotation][[a]]],
   {a, 1, optionNumbers[annotation]}];

order[annotation_] := Sort[optionCounts[annotation], Greater];
first[annotation_] :=
  If[annotation == "Chemotherapy" ||
    annotation == "Probelet_2/Chemotherapy" ||
    annotation == "Arraylet_2/Chemotherapy",
   Position[optionCounts[annotation], order[annotation][[2]]][[1, 1]],
   Position[optionCounts[annotation], order[annotation][[1]]][[1, 1]]];
second[annotation_] :=
  If[annotation == "Chemotherapy",
   Position[optionCounts[annotation], order[annotation][[1]]][[1, 1]],
   If[displayNumber == 2,
    Position[optionCounts[annotation], order[annotation][[2]]][[1, 1]],
    Position[optionCounts[annotation], order[annotation][[3]]][[1, 1]]]];
third[annotation_] :=
  If[annotation == "Probelet_2/Chemotherapy" ||
    annotation == "Arraylet_2/Chemotherapy",
   Position[optionCounts[annotation], order[annotation][[1]]][[1, 1]],
   Position[optionCounts[annotation], order[annotation][[2]]][[1, 1]]];
fourth[annotation_] := Position[optionCounts[annotation],
    order[annotation][[4]]][[1, 1]];
groups[annotation_] := Sort[Transpose[
    Join[{annotationColumn[annotation]}, {times}, { status}]]];
optionCountsPosition[annotation_] := Join[{0}, Table[Sum[optionCounts[annotation][[a]],
     {a, 1, b}], {b, 1, Dimensions[optionCounts[annotation]][[1]]}]];
```

```
firstGroup[annotation_] := Take[groups[annotation],
    {optionCountsPosition[annotation][[first[annotation]]] + 1,
     optionCountsPosition[annotation][[first[annotation] + 1]]}];
secondGroup[annotation_] := Take[groups[annotation],
    {optionCountsPosition[annotation][[second[annotation]]] + 1,
     optionCountsPosition[annotation][[second[annotation] + 1]]}];
thirdGroup[annotation_] := If[displayNumber ≥ 3,
    Take[groups[annotation],
     {optionCountsPosition[annotation][[third[annotation]]] + 1,
      optionCountsPosition[annotation][[third[annotation] + 1]]}]];
fourthGroup[annotation_] := If[displayNumber ≥ 4,
    Take[groups[annotation],
     {optionCountsPosition[annotation][[fourth[annotation]]] + 1,
      optionCountsPosition[annotation][[fourth[annotation] + 1]]}]];

evaluateFirst[annotation_] := {
  nFirstGroup = optionCounts[annotation][[first[annotation]]];
  oFirstGroup = nFirstGroup;
  group = Transpose[firstGroup[annotation]][[2]];
  Do[If[firstGroup[annotation][[a, 3]] == 0, group = Drop[group, {a}]],
   {a, nFirstGroup, 1, -1}];
  firstMedian = 0;
  firstLine = {{0, 1}};
  firstLines = {RGBColor[0, 0, 1], Thickness[0.0075]};
  y = 1;
  Do[
   If[firstGroup[annotation][[a, 3]] == 0,
    {oFirstGroup = oFirstGroup - 1;
     firstLines = Join[firstLines, {Line[{{firstGroup[annotation][[a, 2]], y + 0.02},
          {firstGroup[annotation][[a, 2]], y - 0.02}}]}];
     If[a == optionCounts[annotation][[first[annotation]]],
      firstLine = Join[firstLine, {{firstGroup[annotation][[a, 2]], y}}]]},
    {firstLine = Join[firstLine, {{firstGroup[annotation][[a, 2]], y}}];
     y = y * (nFirstGroup - a) / (nFirstGroup - a + 1);
     If[firstMedian == 0,
      If[y ≤ 0.5, firstMedian = Round[firstGroup[annotation][[a, 2]]]]];
     firstLine = Join[firstLine, {{firstGroup[annotation][[a, 2]], y}}]}],
   {a, 1, optionCounts[annotation][[first[annotation]]]}];
  firstLine = Graphics[{RGBColor[0, 0, 1], Thickness[0.0075], Line[firstLine]}];
  firstLines = Graphics[firstLines];
  textFirstGroup = Graphics[Text[
     Style[ColumnForm[{
        StringReplace[options[annotation][[first[annotation]]], "_" → " "],
        StringJoin["N=", ToString[nFirstGroup]],
        StringJoin["O=", ToString[oFirstGroup]]},
       Center], {RGBColor[0, 0, 1], FontFamily → "Courier"}],
     If[annotation == "Chemotherapy" ||
       annotation == "Probelet_2/Chemotherapy" ||
       annotation == "Arraylet_2/Chemotherapy",
      {-0.065 * months, 0.1},
      {0.055 * months, 0.1}]]]}

evaluateSecond[annotation_] := {
  nSecondGroup = optionCounts[annotation][[second[annotation]]];
  oSecondGroup = nSecondGroup;
  group = Transpose[secondGroup[annotation]][[2]];
  Do[If[secondGroup[annotation][[a, 3]] == 0, group = Drop[group, {a}]],
   {a, nSecondGroup, 1, -1}];
  secondMedian = 0;
  secondLine = {{0, 1}};
```

```
  secondLines = {RGBColor[1, 0, 0], Thickness[0.0075]};
  y = 1;
  Do[
   If[secondGroup[annotation][[a, 3]] == 0,
    {oSecondGroup = oSecondGroup - 1;
     secondLines = Join[secondLines, {Line[{{secondGroup[annotation][[a, 2]], y + 0.02},
          {secondGroup[annotation][[a, 2]], y - 0.02}}]}];
     If[a == optionCounts[annotation][[second[annotation]]],
      secondLine = Join[secondLine, {{secondGroup[annotation][[a, 2]], y}}]]},
    {secondLine = Join[secondLine, {{secondGroup[annotation][[a, 2]], y}}];
     y = y * (nSecondGroup - a) / (nSecondGroup - a + 1);
     If[secondMedian == 0,
      If[y ≤ 0.5, secondMedian = Round[secondGroup[annotation][[a, 2]]]]];
     secondLine = Join[secondLine, {{secondGroup[annotation][[a, 2]], y}}]}],
   {a, 1, optionCounts[annotation][[second[annotation]]]}];
  secondLine = Graphics[{RGBColor[1, 0, 0], Thickness[0.0075], Line[secondLine]}];
  secondLines = Graphics[secondLines];
  textSecondGroup = Graphics[
    Text[Style[ColumnForm[{
        StringReplace[options[annotation][[second[annotation]]], "_" → " "],
        StringJoin["N=", ToString[nSecondGroup]],
        StringJoin["O=", ToString[oSecondGroup]]},
       Center], {RGBColor[1, 0, 0], FontFamily → "Courier"}],
     {If[annotation == "Arraylet_2/Chemotherapy" ||
         ylabel == "Patients from the Independent Set",
        0.7, If[secondMedian < 50, 0.8, 0.72]] * months, 0.9}]]}

evaluateThird[annotation_, displayNumber_] :=
 If[displayNumber ≥ 3, {
   nThirdGroup = optionCounts[annotation][[third[annotation]]];
   oThirdGroup = nThirdGroup;
   group = Transpose[thirdGroup[annotation]][[2]];
   Do[If[thirdGroup[annotation][[a, 3]] == 0, group = Drop[group, {a}]],
    {a, nThirdGroup, 1, -1}];
   thirdMedian = 0;
   thirdLine = {{0, 1}};
   thirdLines = {RGBColor[0, 0.5, 0], Thickness[0.0075]};
   y = 1;
   Do[
    If[thirdGroup[annotation][[a, 3]] == 0,
     {oThirdGroup = oThirdGroup - 1;
      thirdLines = Join[thirdLines, {Line[{{thirdGroup[annotation][[a, 2]], y + 0.02},
           {thirdGroup[annotation][[a, 2]], y - 0.02}}]}],
     {thirdLine = Join[thirdLine, {{thirdGroup[annotation][[a, 2]], y}}];
      y = y * (nThirdGroup - a) / (nThirdGroup - a + 1);
      If[thirdMedian == 0,
       If[y ≤ 0.5, thirdMedian = Round[thirdGroup[annotation][[a, 2]]]]];
      thirdLine = Join[thirdLine, {{thirdGroup[annotation][[a, 2]], y}}]}],
    {a, 1, optionCounts[annotation][[third[annotation]]]}];
   thirdLine = Graphics[{RGBColor[0, 0.5, 0], Thickness[0.0075], Line[thirdLine]}];
   thirdLines = Graphics[thirdLines];
   textThirdGroup = Graphics[
     Text[Style[ColumnForm[{
         StringReplace[options[annotation][[third[annotation]]], "_" → " "],
         StringJoin["N=", ToString[nThirdGroup]],
         StringJoin["O=", ToString[oThirdGroup]]},
        Center], {RGBColor[0, 0.5, 0], FontFamily → "Courier"}],
      {If[annotation == "Arraylet_2/Chemotherapy", 0.7, If[secondMedian < 50, 0.8, 0.72]] *
        months, If[secondMedian < 50, 0.65, 0.35]}]]},
  {thirdLine = Graphics[];
   thirdLines = Graphics[];
```

```
    textThirdGroup = Graphics[]}]

evaluateFourth[annotation_, displayNumber_] :=
 If[displayNumber ≥ 4, {
    nFourthGroup = optionCounts[annotation][[fourth[annotation]]];
    oFourthGroup = nFourthGroup;
    group = Transpose[fourthGroup[annotation]][[2]];
    Do[If[fourthGroup[annotation][[a, 3]] == 0, group = Drop[group, {a}]],
     {a, nFourthGroup, 1, -1}];
    fourthMedian = 0;
    fourthLine = {{0, 1}};
    fourthLines = {RGBColor[0.75, 0, 1], Thickness[0.0075]};
    y = 1;
    Do[
     If[fourthGroup[annotation][[a, 3]] == 0,
      {oFourthGroup = oFourthGroup - 1;
       fourthLines = Join[fourthLines, {Line[{{fourthGroup[annotation][[a, 2]], y + 0.02},
             {fourthGroup[annotation][[a, 2]], y - 0.02}}]}],
      {fourthLine = Join[fourthLine, {{fourthGroup[annotation][[a, 2]], y}}];
       y = y * (nFourthGroup - a) / (nFourthGroup - a + 1);
       If[fourthMedian == 0,
        If[y ≤ 0.5, fourthMedian = Round[fourthGroup[annotation][[a, 2]]]]];
       fourthLine = Join[fourthLine, {{fourthGroup[annotation][[a, 2]], y}}]}],
     {a, 1, optionCounts[annotation][[fourth[annotation]]]}];
    fourthLine = Graphics[{RGBColor[0.75, 0, 1], Thickness[0.0075], Line[fourthLine]}];
    fourthLines = Graphics[fourthLines];
    textFourthGroup = Graphics[Text[
       Style[ColumnForm[{
          StringReplace[options[annotation][[fourth[annotation]]], "_" → " "],
          StringJoin["N=", ToString[nFourthGroup]],
          StringJoin["O=", ToString[oFourthGroup]]},
         Center], {RGBColor[0.75, 0, 1], FontFamily → "Courier"}],
       If[annotation == "Probelet_2/Chemotherapy" ||
         annotation == "Arraylet_2/Chemotherapy",
        {-0.065 * months, 0.35},
        {0.055 * months, 0.35}]]]},
   {fourthLine = Graphics[];
    fourthLines = Graphics[];
    textFourthGroup = Graphics[]}]

evaluatePValue[annotation_] := {
    statistics = Sort[Transpose[Join[{times}, {annotationColumn[annotation]}, {status}]]];
    firstObservations = Table[If[statistics[[a, 2]] ==
        options[annotation][[first[annotation]]], statistics[[a, 3]], 0],
      {a, 1, patients}];
    firstEvents = Table[If[statistics[[a, 2]] == options[annotation][[first[annotation]]],
       ReplaceAll[statistics[[a, 3]], 0 → 1], 0],
      {a, 1, patients}];
    secondObservations = Table[If[statistics[[a, 2]] ==
        options[annotation][[second[annotation]]], statistics[[a, 3]], 0],
      {a, 1, patients}];
    secondEvents =
     Table[If[statistics[[a, 2]] == options[annotation][[second[annotation]]],
       ReplaceAll[statistics[[a, 3]], 0 → 1], 0],
      {a, 1, patients}];
```

```
    Do[If[statistics[[a, 1]] == statistics[[a - 1, 1]], {
       firstObservations[[a - 1]] = firstObservations[[a - 1]] + firstObservations[[a]];
       firstObservations = Drop[firstObservations, {a}];
       firstEvents[[a - 1]] = firstEvents[[a - 1]] + firstEvents[[a]];
       firstEvents = Drop[firstEvents, {a}];
       secondObservations[[a - 1]] = secondObservations[[a - 1]] + secondObservations[[a]];
       secondObservations = Drop[secondObservations, {a}];
       secondEvents[[a - 1]] = secondEvents[[a - 1]] + secondEvents[[a]];
       secondEvents = Drop[secondEvents, {a}]
      }], {a, patients, 2, -1}];
    timesNumbers = Dimensions[firstObservations][[1]];
    firstNumbers =
     Table[Total[firstEvents] - Total[Take[firstEvents, 1 ;; a - 1]], {a, 1, timesNumbers}];
    secondNumbers = Table[Total[secondEvents] - Total[Take[secondEvents, 1 ;; a - 1]],
      {a, 1, timesNumbers}];
    observations = firstObservations + secondObservations;
    numbers = firstNumbers + secondNumbers;
    firstExpectations = Table[If[numbers[[a]] == 0, 0,
       N[observations[[a]] * firstNumbers[[a]] / numbers[[a]]]],
      {a, 1, timesNumbers}];
    secondExpectations = Table[If[numbers[[a]] == 0, 0,
       N[observations[[a]] * secondNumbers[[a]] / numbers[[a]]]],
      {a, 1, timesNumbers}];
    variances = Table[If[numbers[[a]] ≤ 1, 0,
       N[firstNumbers[[a]] * secondNumbers[[a]] * observations[[a]] *
          (numbers[[a]] - observations[[a]]) / numbers[[a]]^2 / (numbers[[a]] - 1)]],
      {a, 1, timesNumbers}];
    z = (N[Total[firstObservations - firstExpectations]])^2 / N[Total[variances]];
    pValue = 1 - CDF[ChiSquareDistribution[1], {z}][[1]]};

w = 0.0075;
months = If[annotation == "Chemotherapy" ||
     annotation == "Probelet_2/Chemotherapy" ||
     annotation == "Arraylet_2/Chemotherapy", 54, 60];
framey[ylabel_] := ReplaceAll[ReplaceAll[
     Table[{a, Style[If[ylabel == "False", "     ", ToString[a]], FontFamily → "Courier"]},
      {a, 0, 1, 0.25}], "1." → "1"], "0." → "0"];
labelx[xlabel_] := If[xlabel == True, Style["Survival Time (Months)",
     FontFamily → "Courier"], ""];
plotlabelx[annotation_, xlabel_] :=
  Style[ColumnForm[
    If[hazardRatio == "", {StringJoin[xlabel, "  ", StringReplace[annotation, "_" → "  "]],
      StringJoin["     P-value = ", ToString[TraditionalForm[
         ScientificForm[pValue, 2, NumberPoint →
           If[Dimensions[Characters[ToString[NumberForm[pValue, 2, NumberFormat → (#1 &),
                  ExponentFunction → (# &)]]]][[1]] < 3, "", "."]]]]}, {
      StringJoin[xlabel, "  ", StringReplace[annotation, "_" → "  "]],
      StringJoin["     P-value = ", ToString[TraditionalForm[
         ScientificForm[pValue, 2, NumberPoint →
           If[Dimensions[Characters[ToString[NumberForm[pValue, 2, NumberFormat → (#1 &),
                  ExponentFunction → (# &)]]]][[1]] < 3, "", "."]]
        ]]], StringJoin["     Hazard Ratio", hazardRatio]}], Left],
   FontFamily → "Courier"];

labely[ylabel_] :=
 If[ylabel == "False", Style[ColumnForm[{" ", "  "}, Center], FontFamily → "Courier"],
  Style[ColumnForm[{title, ylabel}, Center], FontFamily → "Courier"]];
display[annotation_, displayNumber_,
  xplotlabel_, xlabel_, ylabel_, medianTicks_, highlight_] := {
  evaluateFirst[annotation];
```

```
  evaluateSecond[annotation];
  evaluateThird[annotation, displayNumber];
  evaluateFourth[annotation, displayNumber];
  framex = If[displayNumber == 2,
    {{0, Style["0", FontFamily → "Courier"]},
     {firstMedian, If[firstMedian == secondMedian, "",
       Style[StringJoin[medianTicks[[1]], ToString[firstMedian], medianTicks[[2]]],
        {RGBColor[0, 0, 1], FontFamily → "Courier"}]]},
     {secondMedian, Style[StringJoin[medianTicks[[3]], ToString[secondMedian],
        medianTicks[[4]]], {RGBColor[1, 0, 0], FontFamily → "Courier"}]},
     {40, Style["40", FontFamily → "Courier"]},
     If[annotation == "Chemotherapy",
      {50, Style["50", FontFamily → "Courier"]},
      {60, Style["60", FontFamily → "Courier"]}]},
    {{0, Style["0", FontFamily → "Courier"]},
     {fourthMedian, Style[StringJoin[medianTicks[[7]], If[fourthMedian == firstMedian, "",
         If[fourthMedian == thirdMedian, "", ToString[fourthMedian]]], medianTicks[[8]]],
       {RGBColor[0.75, 0, 1], FontFamily → "Courier"}]}, {thirdMedian,
      Style[StringJoin[medianTicks[[5]], ToString[thirdMedian], medianTicks[[6]]],
       {RGBColor[0, 0.5, 0], FontFamily → "Courier"}]},
     {firstMedian, Style[StringJoin[medianTicks[[1]], ToString[firstMedian],
        medianTicks[[2]]], {RGBColor[0, 0, 1], FontFamily → "Courier"}]},
     {secondMedian, Style[StringJoin[medianTicks[[3]], ToString[secondMedian],
        medianTicks[[4]]], {RGBColor[1, 0, 0], FontFamily → "Courier"}]},
     {40, Style["40", FontFamily → "Courier"]},
     If[annotation == "Arraylet_2/Chemotherapy",
      {40, Style["40", FontFamily → "Courier"]},
      If[annotation == "Probelet_2/Chemotherapy",
       {50, Style["50", FontFamily → "Courier"]},
       {60, Style["60", FontFamily → "Courier"]}]]}];
  If[displayNumber == 2, evaluatePValue[annotation]];
  Show[{Graphics[If[highlight == True, {RGBColor[1, 1, 0.8], Rectangle[{
          firstMedian, -0.025}, {secondMedian, 1.025}]}, {}]],
    fourthLine, fourthLines, textFourthGroup,
    thirdLine, thirdLines, textThirdGroup,
    firstLine, firstLines, textFirstGroup,
    secondLine, secondLines, textSecondGroup},
   GridLines → {None, {{0.5, Thickness[0.0025]}}},
   Frame → True,
   FrameTicks → {framex, framey[ylabel], None, None},
   FrameLabel → {labelx[xlabel], labely[ylabel], None, None},
   PlotLabel → plotlabelx[annotation, xplotlabel],
   AspectRatio → If[displayNumber == 2, 1, 1],
   PlotRange → If[annotation == "Chemotherapy" ||
      annotation == "Probelet_2/Chemotherapy" ||
      annotation == "Arraylet_2/Chemotherapy",
     {{-0.225 * months, 1.025 * months}, {-0.025, 1.025}},
     {{-0.1 * months, 1.025 * months}, {-0.025, 1.025}}],
   ImageSize → 250]}

title = "Fraction of Surviving Patients";

stream = path <> "Toy_Data.txt";
annotations = Import[stream, "Table"];
{patients, annotationsNumber} = Dimensions[annotations] - {1, 1};
annotationnames = Take[annotations, 1, {2, annotationsNumber + 1}][[1]];
annotations = Drop[Import[stream, "Table"], 1, 1];
Clear[stream];

times = Table[
   If[annotationColumn["Days Death"][[a]] == "Null",
```

```
     If[annotationColumn["Days_Followup"][[a]] == "Null",
      "Null",
      annotationColumn["Days_Followup"][[a]] / monthConversion],
     annotationColumn["Days_Death"][[a]] / monthConversion],
    {a, 1, patients}];
status = Table[If[annotationColumn["Days_Death"][[a]] == "Null",
     If[annotationColumn["Days_Followup"][[a]] == "Null", 2, 0], 1],
    {a, 1, patients}];
positionNull = Position[status, 2];
Do[{
   annotations = Drop[annotations, positionNull[[a]]],
   times = Drop[times, positionNull[[a]]],
   status = Drop[status, positionNull[[a]]]},
  {a, Dimensions[positionNull][[1]], 1, -1}]
patients = patients - Dimensions[positionNull][[1]];
```

```
(* Problem 1 (40% Extra Credit): Write a function (or a series of functions)
  that can be used to evaluate your test statistic on a sample data set. *)
```

```
? Union
```

Union[$list_1$, $list_2$, ...] gives a sorted list of all the distinct elements that appear in any of the $list_i$.
Union[*list*] gives a sorted version of a list, in which all duplicated elements have been dropped. ≫

```
? ChiSquareDistribution
```

ChiSquareDistribution[ν] represents a χ^2 distribution with ν degrees of freedom. ≫

```
? PDF
```

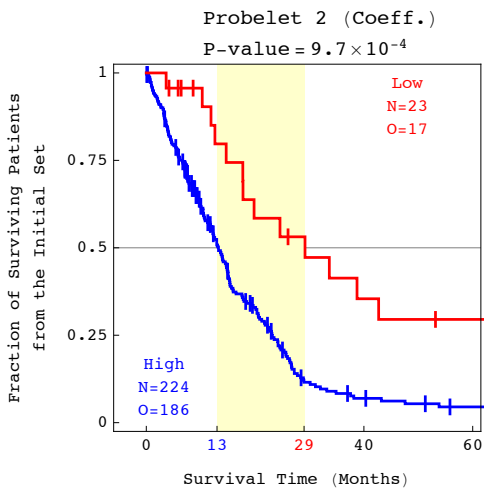
PDF[*dist*, x] gives the probability density function for the symbolic distribution *dist* evaluated at x.
PDF[*dist*, {x_1, x_2, ...}] gives the multivariate
    probability density function for a symbolic distribution *dist* evaluated at {x_1, x_2, ...}.
PDF[*dist*] gives the PDF as a pure function. ≫

```
? CDF
```

CDF[*dist*, x] gives the cumulative distribution function for the symbolic distribution *dist* evaluated at x.
CDF[*dist*, {x_1, x_2, ...}] gives the multivariate cumulative
    distribution function for the symbolic distribution *dist* evaluated at {x_1, x_2, ...}.
CDF[*dist*] gives the CDF as a pure function. ≫

```
(* Read the Toy Data *)

stream = path <> "Toy_Data.txt";
toyData = Import[stream, "Table"];

g1Data = toyData[[1 ;; 48]];
g1Times = g1Data[[All, 2]];
```

```
(* Problem 2 (15% Extra Credit):
 Calculate the p-value for the toy data set using your functions. *)

(* Read and pre-process the TCGA survival data from Lee,*
Alpert* et al. (PLoS One 2012). *)

title = "Fraction of Surviving Patients";
stream = path <> "251_Patients.txt";
annotations = Import[stream, "Table"];
{patients, annotationsNumber} = Dimensions[annotations] - {1, 1};
annotationnames = Take[annotations, 1, {2, annotationsNumber + 1}][[1]];
annotations = Drop[Import[stream, "Table"], 1, 1];
Clear[stream];

times = Table[
    If[annotationColumn["Days_Death"][[a]] == "Null",
     If[annotationColumn["Days_Followup"][[a]] == "Null",
      "Null",
      annotationColumn["Days_Followup"][[a]] / monthConversion],
     annotationColumn["Days_Death"][[a]] / monthConversion],
    {a, 1, patients}];
status = Table[If[annotationColumn["Days_Death"][[a]] == "Null",
     If[annotationColumn["Days_Followup"][[a]] == "Null", 2, 0], 1],
    {a, 1, patients}];
positionNull = Position[status, 2];
Do[{
   annotations = Drop[annotations, positionNull[[a]]],
   times = Drop[times, positionNull[[a]]],
   status = Drop[status, positionNull[[a]]]},
  {a, Dimensions[positionNull][[1]], 1, -1}]
patients = patients - Dimensions[positionNull][[1]];

(* Create Kaplan-Meier survival curves for the Probelet_2 Classification. *)

annotation = "Probelet_2_(Coeff.)";
displayNumber = 2;
highlight = True;
xplotlabel = "    ";
xlabel = True;
ylabel = "from the Initial Set";
medianTicks = Table["", {a, 1, 8}];
hazardRatio = "";
g1 = Show[display[annotation, displayNumber,
    xplotlabel, xlabel, ylabel, medianTicks, highlight], ImageSize → 250]
```



```
(* Problem 3 (15% Extra Credit):
 Use your functions to calculate the p-value for the TCGA data. *)

(* Problem 4 (30% Extra Credit):
 Create the Kaplan-Meier survival curves for an annotation of your
   choice. Change the string in the first line of the code below to one
   of the other annotations in 251_Patients.txt -- Gender, Age_(Years),
Chemotherapy, or IDH1_Mutation. Use your function to calculate the p-
 value for this annotation. Comment on your results. *)

annotation = "IDH1_Mutation";
displayNumber = 2;
highlight = True;
xplotlabel = "    ";
xlabel = True;
ylabel = "from the Initial Set";
medianTicks = Table["", {a, 1, 8}];
hazardRatio = "";
g1 = Show[display[annotation, displayNumber,
    xplotlabel, xlabel, ylabel, medianTicks, highlight], ImageSize → 250]
```