Statistics 3: Logistic Regression, Survival Analysis, Cluster Analysis

Logistic Regression

In Epi Info, either the TABLES command or logistic regression (LOGISTIC command) can be used when the outcome variable is dichotomous (for example, disease/no disease). Analysis with the TABLES command in Epi Info is adequate if there is only one "risk factor." Logistic regression is needed when the number of explanatory variables ("risk factors") is more than one. The method is often called "multivariate logistic regression." Logistic regression shows the relationship between an outcome variable with two values and explanatory variables that can be categorical or continuous.

Logistic regression can be applied to case-control, follow-up, and cross-sectional data. All types of variables (categorical and continuous) can be included in a logistic regression model. The outcome variable must be dichotomous. Categorical data must be coded in "0/1", "0" being the class with the least exposure (Epi Info takes care of this for YES/NO variables).

The purpose of logistic regression is to relate the probability of an outcome to particular values of risk factor variables. A model might predict the probability of occurrence of a myocardial infarction (MI) over a 5-year period, given a patient's age, sex, race, blood pressure, cholesterol level, and smoking status.

How Can I Use Categorical Variables with More Than 2 Values?

Categorical variables with more than 2 values must be recoded with dummy variables. If not, these variables will be considered to be quantitative variables, which may make it difficult to interpret the results of the logistic regression model. A variable with "k" values must be coded in "k – 1" dichotomous dummy variables that have two values, with "0" value indicating no or least exposure.

1. Read in the COURSE.PRJ project. Open LEADPRAC
2. Define TOXIC; (If LEAD0 > 10 then TOXIC = 1 else Toxic = 0). This codes outcome as dichotomous. Make sure you indicate that TOXIC is a number under the Variable Type settings[1].
3. Because Location is coded in the following way

Nassarawa	N
Gangare	G
Dogon Agogo	D

4. Dichotomous dummy variables (Gang and Dogo) for Location will need to replace the original coding:

	Design Variables	
Location	Gang	Dogo
N	0	0
G	1	0
D	0	1

[1] In Epi Info 7, you need to explicitly declare the variable type while defining them or you might get unexpected results.

Two dichotomous dummy variables are enough to locate the three initial values of the "Location" variable. In the logistic regression model, only the variables Gang (Location = G) and Dogo (Location = D) will appear, and the risk computed for each will be in reference to Nassarawa.

```
DEFINE GANG NUMERIC
DEFINE DOGO NUMERIC
ASSIGN GANG = 0
ASSIGN DOGO = 0
IF LOCATION = "G" THEN
   ASSIGN GANG = 1
END
IF LOCATION = "D" THEN
   ASSIGN DOGO = 1
END
LIST LOCATION DOGO GANG GRIDTABLE
```

Procedure

Under Advanced Statistics choose Logistic Regression. Select the dichotomous outcome variable Toxic, then select all the other variables you will want to evaluate as explanatory variables. Include Dogo, Gang, EyePencil, Sex, Religion, Age, EduM, HAZ, WHZ. Make Sex, Religion, EyePencil as Dummy Variables. You can also use this button to create dummy variables like we did from location above, but it will not necessarily choose the reference category you want. Click OK when done.

Results of the multiple logistic regression analysis appear. Variable coefficients, standard errors of the coefficients, P values, odds ratios with upper and lower confidence limits, log likelihood ratios, and any warning messages. Note the variable with the least significant P value, which will be removed next in the backward elimination. Note the result of the Final –2*Log Likelihood and the Likelihood Ratio. If the Final –2*Log Likelihood does not change much upon removing a variable, that variable adds little to the model. Generally, if the difference in 2*Log Likelihood between 2 models that differ by one variable is less than 3.84 (the χ^2 distribution value corresponding to P=0.05 for one degree of freedom), then the difference in the 2 models is not significant.

Remove the variable with the least significant P value from the command line in the Program Editor. Highlight the line and click Run Commands. Continue to remove variables until all variables in the model are significant (backwards elimination).

How Can I Use Explanatory Continuous Variables?

If there are continuous variables in the dataset (age, or duration of a treatment, for example), it is possible to include them as continuous variables in the logistic regression model. The results are expressed as the risk for each unit of the continuous variable (each year of age or each day of a treatment, for example) and the risk for the number of units desired will have to be computed.

Continuous variables can also be recoded into categories to make interpretation easier, using the median as a cut point to recode them in two values, or quartiles as a cut point to recode them in four values, or other meaningful categories (e.g., age = adult / child or 10 year age groups, treatment = none, partial, or full).

NOTE: The interpretation of the results for a continuous variable is more difficult than for a categorical variable, but in recoding a continuous variable in a categorical variable, some information is lost from the data. It is never a simple choice.

What is Interaction and Why Should It Be Addressed Early in the Model?

Interaction means that the odds ratio (OR) for a variable varies with the value of another variable. For example, if the outcome is mesothelioma, a form of lung cancer caused by asbestos exposure, the effect (OR) of asbestos exposure differs greatly between smokers and nonsmokers.

Renal failure = race + hypertension

does not tell the whole story, and another term, called an "interaction term," is needed:

Renal failure = race + hypertension +race*hypertension

Interaction must be addressed early in forming a model, because the model must contain all single variables that are found in significant interaction terms to be "hierarchically well structured." All pertinent interaction terms are evaluated before eliminating any individual variables.

How Do I Construct a Good Logistic Regression Model?

One modeling strategy to find risk factors for an outcome involves two stages:

A. Variable and interaction specification: Choose the variables to include in the multivariate logistic regression model.

If you have only a few variables, start with all of them. Otherwise, include variables that may be risk factors or control variables, based on literature review, and then add all variables for which the p-value of the chi square, Fisher exact, or t-test is less than 0.25 (not the usual threshold of 0.05) in the Epi Info TABLES or MEANS commands. (Note: If you have several exposure and control variables, you have to build a model with all of them and their pertinent interactions.) Now that you have an outcome variable and a list of confounder/risk variables, choose one variable as your primary effect variable for analysis.

In the LEADPRAC example, the outcome variable is Toxic (yes/no). To examine the role of eye pencil, choose it as the effect variable. Construct the model as follows:

Toxic = age + eyepencil + sex + age*eyepencil + eyepencil*sex + sex*age

NOTE: If there are a lot of exposure variables, build a first model with all of the variables, choose those variables with the smallest p-value (around 7 variables) and build another model with these variables and their pertinent interactions (biologically meaningful subsets).

Use backward elimination to find the best model. This means eliminating variables or groups of variables one at a time, keeping only those that are "meaningful" in the model. "Meaningful" in this case stands for a p-value < 0.05 and the likelihood ratio test < 0.05 for the model containing this "chunk" versus the one from which it has been removed. Frequently there is "interaction" among variables so that the OR for one depends on the value of another (e.g., the effect of hypertension or drugs might vary in different races).

If you have two or more categorical explanatory variables in the best model, you must systematically test the interaction term in the logistic regression model. The interaction term in the logistic regression model corresponds to the mathematical product of the two variables (variable 1*variable 2). In this case, interaction variables are included in the model, as hypertension*race and drug*race. If you have more than two categorical variables included in the model, you must test the interaction term for biologically meaningful subsets (which may be all of them, depending on the literature). If there is an interaction term, it is important information for the study: the relationship between the exposure variable and the outcome variable is not the

same for all levels of the explanatory variable (the risk factor varies).

All variables in interaction terms must be included elsewhere in the model as single terms, or Epi Info will not allow the interaction term. This results in what is known as a "hierarchically well formulated" model. Once an interaction has been found to be significant, all of its smaller subcomponents must be left in subsequent models.

B. Confounding assessment.

Confounding is assessed by determining whether the estimated OR changes meaningfully when comparing the OR of the best model versus the model without one or more possible confounders. If there is no interaction, the assessment of confounding is carried out by monitoring changes in the OR of the explanatory variable. However, if there is interaction, the assessment of confounding is more subjective, because it requires comparison of the OR of the exposure variable with the significant interaction terms. The goal is to find a model that gives OR estimates for the exposure variable and interaction terms similar to those given by the "Gold Standard" model. If confounders can be removed without changing these ORs, and the precision (width of the confidence interval) of the OR for the exposure improves, this should be done. Care should be taken not to remove any confounders used in interaction terms.

Cluster Sample Analysis

The Frequency, Tables, and Means commands in the Analysis program perform statistical calculations that assume the data come from simple random (or unbiased systematic) samples. In many survey applications, more complicated sampling strategies are used. These may involve sampling features like stratification, cluster sampling, and the use of unequal sampling fractions. The Complex Sample functions compute proportions or means with standard errors and confidence limits for studies in which the data did not come from a simple random sample. The actual proportions and mean values do not change, but the confidence intervals become wider, depending on the degree of clustering as measured by the design effect. Data from complex sample designs should be analyzed with methods that account for the sampling design.

1. In Classic Analysis, Read the **Smoke** dataset in the **Sample.prj** project.
2. Under Statistics, select Frequencies and obtain frequencies for the variable **SMOKE**.
   - This analysis is designed for a simple random sample.
3. Under Advanced Statistics select Complex Sample Frequencies, and enter the following:

   Frequency of: *SMOKE* Stratify by: *STRATA*

   PSU (Primary Sampling Unit): *PSUID* Weight: *SAMPW*

   These variables describe a study with several different strata (independently sampled areas), cluster sampling, and weighting of individual records to adjust for survey design and/or response rates. This analysis will give weighted estimates of the proportion of people who smoke (SMOKE=1) and do not smoke (SMOKE=2).
4. Now run the program and note that the results do not appear to differ from those for simple random sampling. Because the design effect is close to a value of 1, the confidence intervals do not differ much from those for simple random sampling. To determine if differences in strata are significant (P<0.05), check if the confidence intervals overlap.
5. Choose the command Complex Sample Tables, and enter the following:

   Outcome variable: *SMOKE* Stratify by: *STRATA*

   Exposure variable: *RACE*

   PSU (Primary Sampling Unit): *PSUID* Weight: *SAMPW*

   Output to Table: type CTableSMOKE

   Note that the tabulation now produces a 2 by 2 or 2 by n table. SMOKE is a categorical (binary) variable (1=yes, 2=no). The tabulation shows the upper and lower confidence levels (UCL and LCL) and the design effect for RACE=1 (White) and RACE=2 (Black). Note that the design effects differ for each category of RACE. To determine if differences in strata are significant (P<0.05), check if the confidence intervals overlap. Conclusion: There is very little difference in smoking rate between the two races.
6. To use the number of cigarettes (NUMCIGAR), a continuous variable, as the Outcome variable choose Complex Sample Means, and enter the following:

   Means of: *NUMCIGAR* Stratify by: *STRATA*

   Cross-tabulate by value of: *RACE*

   PSU (Primary Sampling Unit): *PSUID* Weight: *SAMPW*

Notice that the means and confidence intervals are returned. The mean number of cigarettes per day is higher in RACE 1 smokers than in RACE 2 smokers. The confidence intervals do not overlap, so the difference of 6.4 cigarettes per day is significant.

7. Some variables such as AGE may be analyzed either way, although AGE would usually be grouped before using the TABLES approach. Repeat the analysis to calculate the mean age of smokers and nonsmokers:

   Means of: *AGE* Stratify by: *STRATA*

   Cross-tabulate by value of: *SMOKE*

   PSU (Primary Sampling Unit): *PSUID* Weight: *SAMPW*

Sampling Concepts and Terminology

Calculations in Complex Sample functions assume that individual records are members of a sample in which random (or complete or unbiased systematic) sampling has been used in some part of the design. There are three other basic features of sample design that might be used and that must be accounted for in analysis: *cluster sampling, stratification,* and *unequal sampling rates*. When none of these features is used, methods for simple random samples are appropriate.

A sample is chosen to represent a larger universe called the sample's *target population*. This population consists of individuals called *members* who are the object of study. A sample of population members is chosen from one or more lists called the *sampling frame* whose individual entries are called *sampling units.* In some samples (e.g., a sample chosen for a physician survey from a list of physicians), the population members and sampling units are the same, but in other samples this is not necessarily the case (a list of households for a survey of individuals). Simple random sampling is a form of list sampling in which selections are made at random and with equal probability from a complete population frame. Simple random selection can be done *with replacement,* meaning that a member can be chosen again once it is selected the first time, or it can be done *without replacement,* in which case a selected member is not allowed to be chosen again. The calculations in Complex Sample functions assume that random sampling has been done *with replacement.* If the sample is a small fraction of the population, the assumption of sampling with replacement may be made for practical purposes, as described below, even if this was not the original sampling method.

Stratification

Stratification is a common feature in sampling designs. In sampling terminology, stratification means that the frame is subdivided into mutually exclusive and complete (exhaustive) groups called strata and that samples are chosen separately from each stratum. This use of the word “stratum” should be clearly distinguished from "stratification" during data analysis — the process that occurs when you add a third dimension to a TABLES command in ANALYSIS and a separate table are made for each stratum. In sampling the strata are determined prior to data collection; in epidemiology, data are generally stratified after data collection. Sample stratification is commonly used in list sampling, with simple random sampling being the selection method within each stratum. The result is what is called a stratified simple random sample. Stratification is also often used in conjunction with cluster sampling. In general stratification tends to reduce the variance (narrow the confidence limits), at least partially offsetting the opposite effect of cluster sampling. While, in principle, stratification can be used in each stage of a multi-stage cluster sample, it is most commonly used for the first stage of selection. Stratification for choosing Primary Sampling Units (PSUs) is called primary stratification, and is described below.

Cluster Sampling

A sampling design is said to involve *cluster sampling* if at some point in the selection process the sampling units consist of one or more mutually exclusive groups, called clusters. The clusters used for survey sampling are typically spatial (e.g., a sample of residential households that is obtained by selecting local governments or villages), organizational (e.g., a sample of students that is identified by sampling schools), or temporal (e.g., a sample of patients visiting a health clinic chosen by sampling days the clinic is open). The Expanded Program in Immunization (EPI) coverage surveys are cluster surveys. All villages and cities (i.e., clusters) are listed, and then a sample of villages and cities is selected for the survey. The clusters in real populations rarely have the same number of members and may vary greatly in size. Cluster sampling is often done in more than one step or stage of selection. This type of design produces what is called a multi-stage cluster sample. To do so requires the existence of a hierarchical configuration of clusters, so that the clusters at any given stage consist of members or clusters in the subsequent stage(s). For example, a three-stage sample of households might be chosen by designating local governments (LGAs) to be the sampling units in the first stage (called primary sampling units, or PSU's), by assigning wards to be the second stage sampling units for sampling within each PSU, and by designating households to be chosen separately within selected wards as the third stage. Cluster sampling usually increases the variance (widens confidence intervals) of survey estimates. This happens because members of the same cluster tend to be more alike than the population as a whole. Members of a sample from the same cluster therefore tend to provide less information about the population than do members from different clusters. This reduction in information from a clustered sample translates into estimates that are likely to be less precise than estimates obtained using simple random sampling. Primary sampling units (PSU's) are the units chosen from a list in the first (upper) stage of sampling that involves choosing clusters.

For example: In a national survey, the country is stratified by Regions, 2 States randomly selected within each region, 3 LGAs randomly selected within each state, 4 Wards randomly selected within each LGA, 5 Streets randomly selected within each Ward, 6 Households randomly selected from each Street, 4 persons randomly chosen within each Household. Stage two contains the Primary Sampling Units--the "clusters" at the first stage where clusters are randomly chosen (the States in our example). Each record must contain an identifier for the state from which it came. The file may contain identifiers for each of the other levels as well, but those below stage two will not be used by Complex Sample functions. They may be necessary for the calculation of weights prior to running Complex Sample functions, however.

Unequal Selection Probabilities

The third feature that complicates a sampling design is having unequal selection probabilities for population members. This happens when the ratio of sample to population size differs for different parts of the sample. In stratified sampling this can happen if the sample sizes are not proportional to the population of the strata. Unequal selection probabilities may be found in cluster samples in a variety of ways (e.g., sampling clusters with unequal probabilities in a one-stage sample, using simple random samples of the same size in both stages of a two-stage design where the clusters vary in size, etc.). They also occur during the course of a survey through differing response rates in different areas, and other factors that may be ascertained through the survey itself, such as response rates or household size. Unequal selection probabilities are accounted for in the analysis of data by computing sample weights for each member of the sample (i.e., record of the data set). Producing these weights is usually done just prior to analysis. Determining selection probabilities requires that good records be kept for each selection step of the sampling process so that the selection sequence can be explicitly recreated. For multi-stage

cluster samples this means knowing the first stage selection probability for the PSU of which the member is a part, the selection probability for the second stage cluster of which the member is a part, and so on up to the selection probability of the member in the final stage cluster. For designs with stratified sampling this implies having separate information for sampling in each stratum (e.g., sampling rates for stratified simple random sampling). A sample weight for a sample member in its most basic form is simply 1 divided by the member's selection probability. More intuitively, it is the number of population members whom this member represents. The sample weights found in survey data sets are rarely seen in this form, however. They are often adjusted (i.e., multiplied by some appropriate adjustment factor) to compensate for such things as imbalance in the sample due to failure of the frame to fully cover the population (i.e., undercoverage), failure to secure participation from all sample members (i.e., nonresponse), and departures from the demographic composition of the population due to randomized sampling. For example, if each person represents 100 people but participation rates differed among LGAs, additional weighting would be done to compensate for this—an LGA with 90% participation would have a weight of 100/90 in comparison with a LGA with 70% participation having a weight of 100/70. The weights may also be normalized so that they sum to a designated value (e.g., the total sample size), although this is not required by Complex Sample functions.

Complex Sample calculations

The Complex Sample functions enable you to account for the design, regardless of how complex it is, by telling the program three things about each sample member (i.e., data record): 1) The *Primary Sampling Unit (PSU)* from which it came, 2) The primary *stratum* from which its PSU was chosen, and 3) Its sample *weigh*t. To do so requires that either of the following hold: 1) PSUs were selected with replacement, or 2) First stage sampling rates for without-replacement PSU selection were small (less than 5 percent). If sampling rates are greater than 5 percent for without-replacement PSU sampling, the variance estimates will tend to overstate the actual variance. The amount of the overstatement will be directly proportional to the size of the sampling rate. Note that overstated variances will generally produce conservative statistical results (e.g., confidence intervals that indicate less than actual precision for survey estimates; tests of hypothesis that return larger than actual significance levels). This is not necessarily disadvantageous but it should be kept in mind. The bottom line is that sampling *without* replacement, while it will not produce misleading results, may not have quite as much statistical power as sampling *with* replacement, if the sample is a sizable fraction of the population.

Design Effect

A useful design-related measure for surveys is the so-called “design effect”, the ratio of the variance of the estimate under the actual design used to produce the estimate to the variance of the estimate assuming the same data to have come from a simple random sample. The design effect reflects the estimated variance of the survey data relative to that of a simple random sample. Design effect for multi-stage cluster samples will usually exceed 1, sometimes substantially, while for stratified simple random samples and other list samples design effect will be near or slightly less than 1. Generally stratification tends to reduce the design effect and cluster sampling to increase it. Widely variable sample weights tend to increase design effect.

Preparing the Data

Complex Sample functions have several important data requirements. The data must be organized with records in the Epi Info 7 file representing population members. When the sample design features primary stratification, the *stratum identifier* must be included in each record. If cluster sampling is used, a *PSU identifier* must be included. Finally, if unequal selection probability

(samples not proportional to population) has occurred, *sample weights* should be included in each record, which you must compute ahead of time. None of these variables is strictly required. If, for example, you have used multi-stage cluster sampling, where no primary stratification was used, only the PSU identifier and the weight variable are needed. Moreover, if the weight variable was not computed since selection probabilities are the same, the program will still properly account for the design but will assume equal probabilities of selection.

Example: A World Health Organization Cluster Survey of Vaccination Status

Surveys that include some form of complex sampling include the coverage surveys of the WHO Expanded Program on Immunization (EPI) (Lemeshow and Robinson, 1985).

1. From **Sample.prj** Read the form **Epi1**. There are 210 children, with 7 in each of 30 different clusters. Each subject has two measured variables: PRENATAL and VAC. The former assesses whether the mother received prenatal care (Y=1, N=2), while the later indicates if the child has received a complete set of vaccinations (Y=1, N=2). Accompanying the two variables is a third variable CLUSTER, which identifies the cluster where the child resided. The clusters are numbered from 1 to 30.
2. The data will be analyzed in two ways:
   - With Frequencies (which assumes random sampling);
   - With Complex Sample Frequencies, specifying the CLUSTER identifier as the PSUID, taking into account the cluster design.
3. Using the Frequencies command, the analysis indicates 73.8% of the children are vaccinated (95% CI 67.3 – 79.6%).
4. Using Complex Sample Frequencies to analyze the file, the items in the input dialog box are:

   Frequency of: *VAC* Stratify by: *none*

   PSU (Primary Sampling Unit): *CLUSTER* Weight: *none*

The analysis indicates 73.8% of the children are vaccinated, the same result as the simple Frequencies command. Notice that the bounds of the LCL 64.4% and UCL 83.2% are wider than those assuming simple random sampling. This is because the design effect is 2.3 (greater than 1).

Whether or not the mother received prenatal care was also recorded in this survey. The authors were interested in the hypothesis that receiving prenatal care is associated with a higher rate of vaccination among children.

Using the Tables command and (incorrectly) treating the sample as a simple random sample, enter PRENATAL as the exposure variable and VAC as the outcome variable. The results would be: Odds ratio of 5.18 with 95% confidence limits 2.37 – 11.3, and a risk ratio of 1.43 with 95% confidence limits 1.23 – 1.67. This could lead to the conclusion that children whose mothers had prenatal care are 1.43 times more likely to be vaccinated than children whose mothers did not have prenatal care, with (incorrect) confidence limits of 1.23 and 1.67 for the ratio.

To perform the necessary cross tabulation in Complex Sample Tables, using the cluster-sample design in the calculations, the following choices are made:

Outcome variable: VAC Stratify by: none

Exposure variable: PRENATAL

PSU (Primary Sampling Unit): CLUSTER Weight: none

The results report the same odds ratio of 5.18, but now the 95% confidence limits are 2.21 – 12.1, wider than assuming a simple random sample. Similarly, the risk ratio is 1.43 with 95% confidence limits 1.14 – 1.80, also wider. The 95% confidence limits do not include 1.0, indicating that the association is statistically significant.

Example: Combining Several WHO Cluster Surveys

1. Read the **Epi10** form from the **Sample.prj** project.
2. Use the List command to view the records. Here there are 10 different EPI surveys. The subjects all have the same three measured variables included in Epi1: PRENATAL, VAC and CLUSTER. The record for each person contains a number that identifies the survey (LOCATION) and the population in the area where the survey took place (POPW). The immunization levels varied in the different surveys, from 31.1% to 84.8%.

As mentioned previously, the two-stage cluster surveys used by WHO/EPI are self-weighted samples. As such, each person or household (i.e., the sampling unit) in the population is selected with equal probability. We can say, therefore, that each sampling unit represents some given number of units in the population. For example, if a survey was done of 225 children in a population of 9,870 children, each child in the sample would represent 9,870/225 or 43.87 children in the sampled population. In this case, 43.87 would be the population weight for each child in the survey. The 225 children in the sample would represent (225 x 43.87) or 9,870 children in the population.

In the dataset Epi10, there are two variables necessary for combining surveys: LOCATION (the survey identification number) and POPW (the population weight). To combine the results from the ten locations, LOCATION is used as the Strata variable.

3. In Complex Sample Frequencies, enter the following:

   Frequency of: *VAC* Stratify by: *LOCATION*

   PSU (Primary Sampling Unit): *CLUSTER* Weight: *POPW*

   The vaccination level for the 10 areas is 55.3% (95% CI 50.0 – 60.5) and the design effect is 5.79. If the data are analyzed with Frequencies, the vaccination rate is incorrectly reported as 57.7% (95% CI 55.6 – 59.8).

4. Enter the following in Complex Sample Tables to examine the effect of PRENATAL:

   Outcome variable: VAC Stratify by: LOCATION

   Exposure variable: PRENATAL

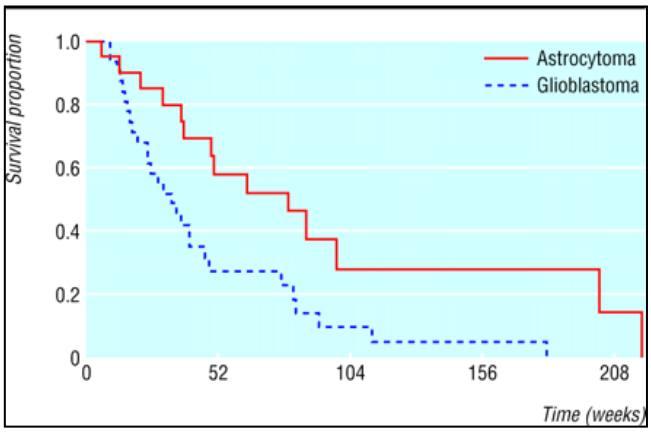
   PSU (Primary Sampling Unit): CLUSTER Weight: POPW

   The results are: Odds Ratio (OR) 2.088 (95% CI 1.50 – 2.90), Risk Ratio (RR) 1.427 95% (CI 1.23 – 1.66), Risk Difference (RD) 18.174% (95% CI 10.26 – 26.09).

Survival analysis

Kaplan-Meier method

In clinical epidemiology, particularly in the study of usually fatal chronic diseases, the measurement of patient survival has become an important criterion in evaluating the effectiveness of therapeutic modalities. For example, we could compare the survival of patients with two types of brain tumors.



The objective of the Kaplan-Meier (KM) methodology is to estimate the probability of survival of a defined group at a designated time interval (conditional probability). KM uses a non-parametric survival function for a group of patients (in other words their survival probability after the time *t*) and therefore does not make assumptions about the survival distribution.

Each time a survival study is done, the KM methodology could be used to estimate the probability of survival over a given time period. "Survival" means that the event of interest has not occurred. The event can be death, a complication of a treatment, or other defined adverse event. KM therefore provides an estimate of being free of the event at time *t*. Conversely, "1 minus the probability of being free of the event at time *t*" is the probability of having the event at time *t*.

Examples:

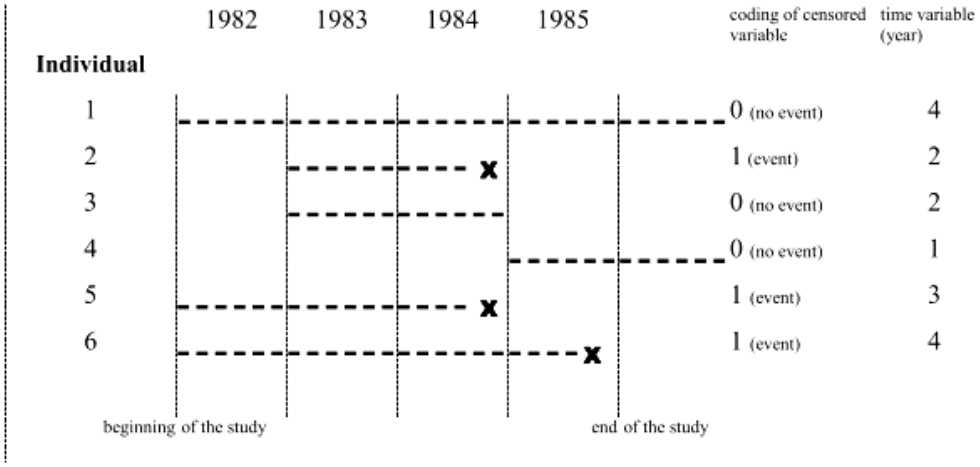
- Comparing survival time in 2 diseases (e.g. cervical vs. uterine cancer)
- Comparing time to an event in 2 diseases (development of diabetes with exercise vs. no exercise, hospitalization for CHF on digoxin vs. no digoxin, time to fever clearance with artemisinin vs. SP, time to pregnancy after HSG vs. no HSG)

The variable to be studied is the time delay until the occurrence of an event (death, disease, treatment outcome, etc.). This time delay corresponds to survival duration (the difference between the beginning study date and the event date).

What distinguishes survival analysis from most other statistical methods is that the event of interest will not have occurred in all patients at the end of follow up. For these individuals the survival time is "censored," indicating that the observation period was cut off before the event occurred. For example, in a study of survival following two different treatment regimens, analysis of the trial typically occurs well before all the patients have died. For those still alive at the time of analysis, the true survival time is known only to be greater than the time observed to date. Such an observation is said to be "censored." There are two other sorts of incomplete observation: the "lost to follow-up" (patient missing during the study duration) or the appearance of an event other than the event being studied (e.g. death from an accident). These observations are also considered censored. Patients with censored data contribute valuable information and they should not be omitted from the analysis. In most survival studies patients are recruited over a period and followed up to a fixed date beyond the end of recruitment. Thus the last patients recruited will be observed for a shorter period than those recruited first and will be less likely to experience the event.

For survival analysis, the censored variable, the time variable, the units of time (day, month, year), and the group of patients (if studying the effect of a treatment) must be specified. The time variable is numeric. The censored variable is coded: "1" if the patient experiences the event (uncensored data), "0" if the event is not known to occur (censored data). Survival data is often presented using a "+" for the censored observation, so that a set of times might be 8, 11+, 14, 2, 36+, etc.

Table 1: Coding of censored variable for 6 patients with bladder cancer.



The censored variable has the value of "0" for individuals 1, 3, and 4, and "1" for individuals 2, 5, and 6.

The KM survival function is a decreasing series of straight line steps, constant between two consecutive death times, with a step for each time of observed death. This function is not defined after the last observation if this observation is censored.

In the first step Epi Info sorts the records by time, t_i, then for each time interval t_i up to, but not including, t_{i+1} it counts:

- the number of deceased patients d_i at time t_i,
- the number of censored patients c_i at time t_i,
- the number of risk patient n_i (number of patients living just before t_i)

$$n_i = n_{i-1} - c_{i-1} - d_{i-1}$$

For each time interval, we estimate the probability that those who survived to the beginning of the interval will survive to the end:

$$\left(1 - \frac{d_i}{n_i}\right)$$

Survival to any time point is calculated as the product of the conditional probabilities of surviving each time interval. This is called the KM estimator:

$$S = \Pi\left(1 - \frac{d_i}{n_i}\right)$$

("Π" means "product of")

The probability of surviving two months is the probability of surviving the first month times the

probability of surviving the second month given that the first month was survived. So if one started with 38 women, and 32 survived the first month, the conditional probability of surviving the first month would be 32/38=0.842. Of 32 women at the start of the second month, 27 were still alive at the end of the second month. The conditional probability of surviving the second month would be 27/32=0.844. The overall probability of survival after two months is 0.842×0.844=0.711. We continue this way until we reach the last event.

The example provided uses the Leukemia study (Freireich, 1963), a study about the remission time delay for patients with leukemia who are given different treatments. This example is based on a randomized clinical trial to evaluate if patients assigned to treatment with 6-mercaptopurine would fare better than untreated (Placebo) patients.

1. Read in the **LEUKEM2** data in the **Sample,prj** project file
2. Use the list command to view the data. The variables are:

   SURVTIME: time delay (in weeks)

   CENSORED: censored variable ("1" for censored individual, "0" for uncensored individual)

   TREATMENT: group of patients (6-MP or Placebo)
3. Under Advanced Statistics, choose Kaplan-Meier Survival, and enter the following:

   Censored Variable: *CENSORED* Value for Uncensored: 0

   Time Variable: *SURVTIME* Time Unit: *weeks*

   Test Group Variable: *TREATMENT* Graph Type: *Survival Probability*

   Note the graphical output of the survival plot and the statistical tests reported:

Test	Statistic	D.F.	P-Value
Log-Rank	16.7929	1	0.0000
Wilcoxon	13.4579	1	0.0002

How do we determine if the difference between two survival curves is greater than would be expected by chance alone? We could simply compare the proportion of patients surviving in each group at a particular time point using a chi-square statistic, but this does not account for the total survival experience of both groups.

Epi Info uses the log-rank test, a large-sample chi square test that uses as its test criterion an overall comparison of the KM curves being compared. The log-rank statistic, like many other statistics used in other kinds of chi square tests, makes use of observed events versus expected events over categories of outcomes. The log-rank statistic takes into account the entire follow-up period, and it requires no assumptions about the shape of the survival curves. The log-rank test is used to test the null hypothesis that there is no difference between groups in the probability of an event at any time point.

Epi Info also uses the generalized Wilcoxon test, another non-parametric test. This test gives greater weighting to the earlier observations (which have more subjects and precision).

A p-value <0.05 suggests a difference in survival between the two groups. Both tests show that 6-MP results in significantly longer survival than placebo.

Cox Proportional Hazards

The Cox Proportional Hazards command in Analysis is a form of survival analysis that relates covariates to an event (e.g. death) through hazard ratios. A covariate with a hazard ratio less than one improves survival. At any given time, some of the subjects may be "censored", that is, not have information available on their status. Cox Proportional Hazards is especially constructed to deal with this situation.

The hazard function for a subject is of the form: $h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2}$; β_1 and β_2 are unknown regression coefficients that are estimated from the data, similar to logistic regression.

1. Read the LEUKEM2 data in the Sample.prj project.
2. Under Advanced Statistics, choose Cox Proportional Hazards, and enter the following:
3. Censored Variable: CENSORED Value for Uncensored: 0
4. Time Variable: SURVTIME Time Unit: weeks
5. Group Variable: TREATMENT Graph Type: Survival Probability
6. Weight: (leave blank) Output to Table: (leave blank)
7. Confidence Limits: 95%

Note the graphical output of the survival plot. Close the graph window.

Note the statistical tests reported:

Term	Hazard Ratio	95% C.I.		Coefficient	S. E.	Z-Statistic	P-Value
TREATMENT (Placebo/6-MP)	4.5231	2.0269	10.0932	1.5092	0.4095	3.6851	0.0002

Convergence:	**Converged**
Iterations:	4
-2 * Log-Likelihood:	172.7592

Test	Statistic	D.F.	P-Value
Score	15.9305	1	0.0001
Likelihood Ratio	15.2109	1	0.0001

The hazard ratio 4.5 (95% CI 2.0 – 10.1) indicates that placebo relative to 6-MP is associated with a greater hazard of death. In other words, a hazard ratio is a relative risk measure (cohort study). This test is most useful when examining more than one covariate (e.g. treatment group, sex, age).