Censored and truncated data

Eric Feigelson (Penn State) edf@astro.psu.edu

2nd East Asian Workshops in Astrostatistics Summer 2018

Adapted from R scripts in Appendix B, *Modern Statistical Methods for Astronomy With R Applications*, Eric D. Feigelson & G. Jogesh Babu 2012 http://astrostatistics.psu.edu/MSMA (http://astrostatistics.psu.edu/MSMA)

We start with a problem involving left-censoring that astronomers call `upper limits`. We consider a sample of optically selected quasars from the Sloan Digital Sky Survey that are observed, but not all detected, in the X-ray ROSAT All-Sky Survey (RASS). We then seek to estimate the X-ray luminosity function. Note that, in this case, the RASS did not report exact upper limits for every location in the sky, so we use an approximate survey sensitivity.

In [ ]:

```
setwd('/Users/ericfeigelson/Desktop/Rdir')

# Construct sample of X-ray luminosities of low-redshift (z<0.3) Sloan
quasars
qso <- read.table("SDSS_QSO.dat", head=T, fill=T)

dim(qso) ; names(qso) ; summary(qso)
qso.lowz <- qso[qso[,2]<0.3,]
dim(qso.lowz) ; names(qso.lowz) ; summary(qso.lowz)
attach(qso.lowz)
```

In [ ]:

```
# X-ray luminosities for detections and (approximately) for upper limi
ts

ROSAT[ROSAT < (-8.900)] <- NaN  # X-ray nondetections
Dist <- (3*10^{10})*z*(3.09*10^{24}) / (73*10^5)
XLum <- 8.3*10^(-12)*10^(ROSAT) * (4*pi*Dist^2)
XLum[is.na(ROSAT)] <- 3.*10^(-13)* (4*pi*Dist[is.na(ROSAT)]^2)

par(mfrow=c(1,2))
plot(log10(XLum[!is.na(ROSAT)]), ylim=c(42.5,45), xlab='',
   ylab='log(Lx) erg/s', pch=20)
plot(log10(XLum[is.na(ROSAT)]), ylim=c(42.5,45), xlab='',
   ylab='log(Lx) erg/s', pch=25)
par(mfrow=c(1,1))
```

In [ ]:

```
# Construct quasar Kaplan-Meier luminosity function and 95% confidence
band

library(survival)
Xstatus <- seq(1,1,length.out=400)  # 1 = detected
Xstatus[is.na(ROSAT)] <- 0  # 0 = right-censored
survobj <- Surv(-XLum, Xstatus)
KM.XLF<- survfit(survobj~1, conf.int=0.95, conf.type='plain',
   conf.lower='modified')
KM.output <- summary(survfit(survobj~1))

# Plot Kaplan-Meier estimator with confidence bands

plot(log10(-KM.XLF$time), KM.XLF$surv,  ylim=c(0,1), pch=20, cex=0.5,
main='',
   xlab=expression(log*L[x]), ylab='Kaplan-Meier estimator')
lines(log10(-KM.XLF$time), KM.XLF$upper, lty=2)
lines(log10(-KM.XLF$time), KM.XLF$lower, lty=2)
```

We now consider a problem that combines univariate 2-sample tests and bivariate correlation tests in the presence of censoring. It concerns beryllium and lithium abundances in the atmosphere of stars with and without exoplanets. (The goal is to test whether metal enhancements have occurred by stars engulfing planets in the past.) The dataset is from Santos et al. (2002).

In [ ]:

```
# Read dataset on beryllium and lithium abundances in stars
abun <- read.table('censor_Be.dat', header=T)
dim(abun) ; names(abun) ; attach(abun)

# Boxplot of abundances for stars with and without planets

install.packages('NADA', repos='https://cloud.r-project.org') ; library(NADA)
cen_Be <- seq(FALSE, FALSE, length=68) ; cen_Be[Ind_Be==0] <- TRUE
cen_Li <- seq(FALSE, FALSE, length=68) ; cen_Li[Ind_Li==0] <- TRUE
Type_factor <- as.factor(Type)
par(mfrow=c(1,2))
cenboxplot(logN_Be, cen_Be, Type_factor, log=FALSE, ylab='log N(Be)',
   names=c('Planets','No planets'), boxwex=0.5, notch=TRUE, varwidth=TRUE,
   cex.axis=1.5, cex.lab=1.5, lwd=2)
cenboxplot(logN_Li, cen_Li, Type_factor, log=FALSE, ylab='log N(Li)',
   names=c('Planets','No planets'), boxwex=0.5, notch=TRUE, varwidth=TRUE,
   cex.axis=1.5, cex.lab=1.5, lwd=2)
par(mfrow=c(1,1))

# Test significance of possible lithium abundance effect

logN_Li1 <- logN_Li[-c(1,23)] ; cen_Li1 <- cen_Li[-c(1,23)]
Type_factor1 <- Type_factor[-c(1,23)] # remove NaN values
cendiff(logN_Li1, cen_Li1, Type_factor1, rho=0)
cendiff(logN_Li1, cen_Li1, Type_factor1,rho=1)
```

Just for fun, let's reproduce the figure in Santos et al. comparing stellar beryllium vs. lithium. This requires lots of manual adjustments to the plot. Then calculate the probability of correlation and show the Akritas-Thiel-Sen semi-parametric regressio line from Helsel's book and `NADA` CRAN package.

In [ ]:

```
# Reproduce Santos et al. (2002) plot of stellar beryllium vs. lithium abundance

ind_det1 <- which(Ind_Li==1 & Ind_Be==1 & Type==1) # filled circles
ind_det2 <- which(Ind_Li==1 & Ind_Be==1 & Type==2) # open circles
ind_left1 <- which(Ind_Li==0 & Ind_Be==1 & Type==1)
ind_left2 <- which(Ind_Li==0 & Ind_Be==1 & Type==2)
ind_down1 <- which(Ind_Li==1 & Ind_Be==0 & Type==1)
ind_down2 <- which(Ind_Li==1 & Ind_Be==0 & Type==2)
ind_both1 <- which(Ind_Li==0 & Ind_Be==0 & Type==1)
ind_both2 <- which(Ind_Li==0 & Ind_Be==0 & Type==2)

plot(logN_Li[ind_det1], logN_Be[ind_det1], xlim=c(-0.6,3.5), ylim=c(-0.2,1.5),
   main="", xlab="log N(Li)",
   ylab="log N(Be)", pch=16, lwd=2) # plot detections
points(logN_Li[ind_det2], logN_Be[ind_det2], pch=1)
arrows(logN_Li[ind_left1], logN_Be[ind_left1], logN_Li[ind_left1]-0.2,
   logN_Be[ind_left1],length=0.1) # plot left arrows
arrows(logN_Li[ind_left2], logN_Be[ind_left2], logN_Li[ind_left2]-0.2,
   logN_Be[ind_left2],length=0.1)
points(logN_Li[ind_left1], logN_Be[ind_left1], pch=16)
points(logN_Li[ind_left2], logN_Be[ind_left2], pch=1)
arrows(logN_Li[ind_down1], logN_Be[ind_down1], logN_Li[ind_down1],
   logN_Be[ind_down1]-0.1, length=0.1)
arrows(logN_Li[ind_down2], logN_Be[ind_down2], logN_Li[ind_down2],
   logN_Be[ind_down2]-0.1, length=0.1)
points(logN_Li[ind_down1], logN_Be[ind_down1], pch=16)
points(logN_Li[ind_down2], logN_Be[ind_down2], pch=1)

arrows(logN_Li[ind_both1], logN_Be[ind_both1],
   logN_Li[ind_both1]-0.2, logN_Be[ind_both1], length=0.1) # plot double arrows
arrows(logN_Li[ind_both1], logN_Be[ind_both1], logN_Li[ind_both1],
   logN_Be[ind_both1]-0.1,length=0.1)
arrows(logN_Li[ind_both2], logN_Be[ind_both2], logN_Li[ind_both2]-0.2,
   logN_Be[ind_both2],length=0.1)
arrows(logN_Li[ind_both2], logN_Be[ind_both2], logN_Li[ind_both2],
   logN_Be[ind_both2]-0.1,length=0.1)
points(logN_Li[ind_both1], logN_Be[ind_both1], pch=16)
points(logN_Li[ind_both2], logN_Be[ind_both2], pch=1)

# Bivariate correlation and regression using Akritas-Thiel-Sen procedure

logN_Li1 <- logN_Li[-c(1,23)] # remove two points with NaN entries
Ind_Li1 <- Ind_Li[-c(1,23)] ;
logN_Be1 <- logN_Be[-c(1,23)]
Ind_Be1 <- Ind_Be[-c(1,23)]
Li_cen <- seq(FALSE, FALSE, length=66) # construct censoring indicator variables
Li_cen[which(Ind_Be1==0)]  <- TRUE
Be_cen=seq(FALSE, FALSE, length=66)
Be_cen[which(Ind_Li1==0)] <- TRUE
cenken_out <- cenken(logN_Be1, Be_cen, logN_Li1, Li_cen)
abline(a=cenken_out$intercept, b=cenken_out$slope, lwd=2)
cenken_out
```

Next we estimate the luminosity of normal stars from a truncated, flux-limited survey from the ESA Hipparcos satellite. Luminosities can be derived from the satellite photometry and parallax measurements. We compare the nonparametric unbinned Lynden-Bell-Woodroofe (LBW) estimator with the (commonly used but less accurate) binned Schmidt (1968) estimator, and compare both to the well-established Wielen (1983) stellar luminosity function from the (nearly) complete Gliese sample of nearby stars. The LBW estimator performs better than the Schmidt estimator using this truncated dataset. We do not show the bootstrap confidence intervals around the LBW curve as it is computationally expensive; it is shown on the cover of the MSMA book.

In [ ]:

```
# Construct a sample of bright nearby Hipparcos stars

hip <- read.table('HIP1.tsv',header=T, fill=T)
attach(hip) ; dim(hip); summary(hip)
hip <- na.omit(hip)
summary(hip)
```

In [ ]:

```
# Plot luminosity distribution of stars and their truncation limits

AbsMag <- na.omit(Vmag + 5*log10(Plx/1000) + 5)
Lum <- 2.512^(4.84 - AbsMag)
plot(density(log10(Lum)),ylim=c(0,1.7), main='',
   xlab='log L (solar, V band)', lwd=2, cex.lab=1.2)
AbsMaglim <- na.omit(10.5 + 5*log10(Plx/1000) + 5)
Lumlim <- 2.512^(4.84 - AbsMaglim)
lines(density(log10(Lumlim)), lty=2, lwd=2, col=2)
text(0.7, 0.5, 'Hipparcos sample', pos=4, font=2)
text(-0.5, 1.2, 'Truncation limits', pos=4, font=2)
```

In [ ]:

```
# Compute Lynden-Bell-Woodroofe estimator
# See cover of MSMA textbook for bootstrap generated 90% confidence bands

install.packages('DTDA', repos='https://cloud.r-project.org') ; library(DTDA)
LBW.hip <- efron.petrosian(log10(Lum), log10(Lumlim), boot=F, B=100, alpha=0.1)
summary(LBW.hip)
plot(LBW.hip$time, LBW.hip$survival, pch=20, cex=0.6, main='',
   xlab='log L (solar, V band)', ylab='Density', cex.lab=1.2)
upper <- LBW.hip$upper.Sob[-(1000:1013)]
lower <- LBW.hip$lower.Sob[-(1000:1013)]
lines(LBW.hip$time,upper, lty=2) ; lines(LBW.hip$time, lower, lty=2)

# Compare with observed Wielen 1983 local star LF

Wielen.MV <- seq(0, 12, 1)
Wielen.LF <- c(35,126,209,380,676,955,1050,891,1120,1410,2140,2510,4470)
points(log10(2.512^(4.84-Wielen.MV)), Wielen.LF/2500)

# Compare with binned 1/V.max luminosity function (Schmidt 1968)

Vol.max <- (4*pi/3)*(1000/Plx)^3
bin.sum <- function(vol) sum(1/vol)
Lum.bins <- cut(log10(Lum), breaks=seq(-2.5,3.0,0.5), ord=T)
Schmidt.LF <- by(Lum, Lum.bins, bin.sum)
lines(seq(-2.25, 2.75, 0.5), Schmidt.LF/4500, pch=3, lwd=2, col=3)
```