Estimating abundance and trend

Anna Tucker

March 13, 2018

### Unmarked individuals

#### Canonical population estimator

There are many different methods for estimating population abundance from counts of unmarked individuals. At their core is the idea that if we count some number of individuals and know the detection probability, we can estimate the total abundance: \* C = number of individuals counted  
\* p = detection probability  
\* N = true abundance

This is known as the canonical population estimator. Note the “hat” over and , which indicates that we estimate these parameters. Occasionally we might have estimates of detection probability from other studies, for example a mark-recapture study, in which case that estimate could be used to calculate based on our counts. In most cases though, either estimates of detection probability are not available or not appropriate to use. For example, if the detection probability was estimated for a site with different habitat structure, or if the level of sampling effort was different between the mark-recapture and count study, it may not be appropriate. The many different methods for estimating abundance (which we won’t cover here) differ mainly in the assumptions made about the how the count data are collected and the method used for estimating detection probability from the count data.

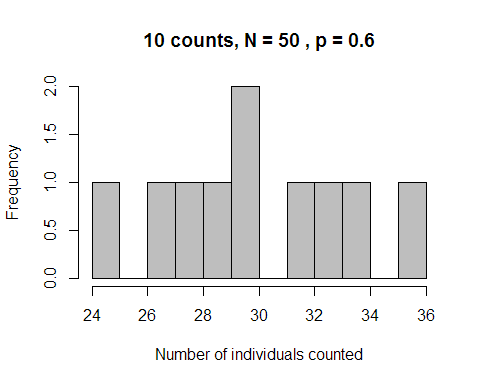
We can consider the count data as arising from a series of coin flips in which each individual has some probability of being detected. This type of data can be described using the Binomial distribution.

If you flip a coin times, and that coin has a probability of landing on heads, is the number of times the coin will land heads. If you repeated this and flipped the coin another times, you would likely get a different value for . Statistical methods have been developed that use many samples of replicated across space and time to estimate the most likely values for and .

N = 50  
p = 0.6  
n.counts = 10  
  
C = rbinom(n.counts, N, p) # simulate counts  
C # look at counts

## [1] 30 29 32 33 36 24 30 34 27 28

hist(C, breaks = 10, col = "gray", main = paste(n.counts, "counts, N =", N, ", p =", p),  
 xlab = "Number of individuals counted")



If we know (e.g. from another study), we can estimate based on our observed count .

C = 10  
p = 0.7  
Nhat = C/p  
Nhat

## [1] 14.28571

**Question 1: You conduct an auditory survey for northern bobwhite in a small field and count 7 individuals calling. Previous studies using auditory surveys for bobwhite in similar habitats have estimated a detection probability of 0.56. What is the estimated abundance of bobwhite in this field?**

#### Population indices

It is usually very difficult to estimate abundance from population indices, but they can be useful for comparing the relative abundance between different sites or years.

We often log-transform counts to estimate the trend in population indices over time, which helps the data meet the assumptions of a linear model. The effects of site or year (or anything else of interest) can then be easily estimated in R, Excel, or any other statistical software.

Recall the exponential population growth model. If I is a good index of N, and the relationship between the index and true population size doesn’t change over time, we can model the index I using the exponential growth model. Taking the log of both sides transforms it into a linear model.

$$ I = \beta N \\
N\_t = N\_0e^{rt} \rightarrow log(N\_t) = log(N\_0) + rt $$

One important distinction between a linear regression model and the model that we use for counts is the assumption about the distribution of the model residuals. In linear models, we assume that the residuals follow a Normal distribution, but here we assume the residuals follow a Poisson distribution. This distribution is commonly used to model count data because it only allows integer values greater than 0.

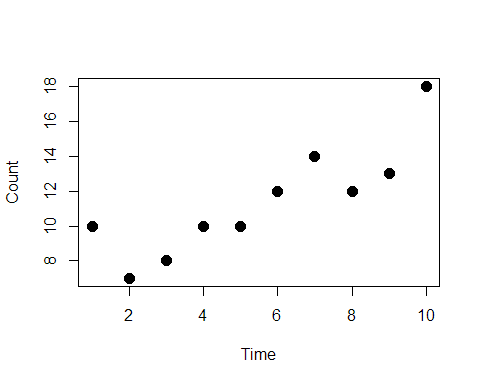
We can use the generalized linear model function in R to calculate over the time series. Under this method we are assuming that remains the same from year to year. We could also back-transform our estimate of if we wanted to report .

As with all statistical estimation methods, it is important to report not just the point estimate but also the precision of that estimate, usually provided in the R output as standard error (SE). The results will also include a statistical test for whether the estimate is significantly different from 0 and an associated p-value (we typically consider an effect to be statistically significant if p < 0.05).

Run the R code to enter population index data, plot the counts, and fit the model. The summary output in your console shows the model results. The intercept is equal to lna(I\_0 ) and the coefficient for “time” is the slope, which is equal to r. The p-value associated with the slope (under the column “Pr(>|t|)”) indicates whether the slope is significantly different from 0.

Run the R code to generate predicted values from the model and plot the model predictions and data on both the log scale and natural scale. Plotting model predictions is a good first step in assessing how well the model fits the data.

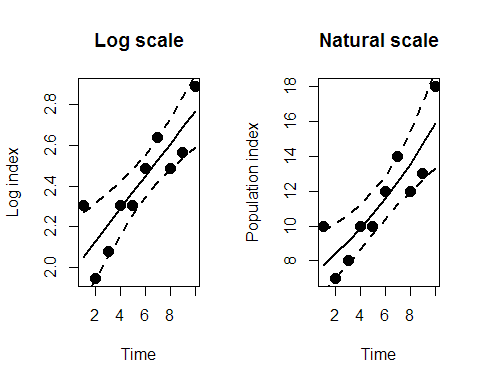
# manually enter counts of active wood duck nests  
year = c(1:10)  
index = c(10, 7, 8, 10, 10, 12, 14, 12, 13, 18)  
  
# plot counts over time  
plot(index ~ year, pch = 16, cex = 1.5, xlab = "Time", ylab = "Count")



# fit the generalized linear regression model  
mod = glm(index ~ year, family = "poisson")  
summary(mod)

##   
## Call:  
## glm(formula = index ~ year, family = "poisson")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.50068 -0.41587 -0.08205 0.34111 0.76692   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.97034 0.22421 8.788 <2e-16 \*\*\*  
## year 0.07952 0.03313 2.400 0.0164 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 7.7163 on 9 degrees of freedom  
## Residual deviance: 1.8631 on 8 degrees of freedom  
## AIC: 48.396  
##   
## Number of Fisher Scoring iterations: 4

# get predicted values and confidence limits  
pred = predict.lm(mod, interval = "confidence", newdata = data.frame(year = year))  
  
# plot population size over time on the log scale and add predicted line  
# dashed lines are the 95% confidence intervals  
par(mfrow = c(1,2))  
plot(log(index) ~ year, pch = 16, cex = 1.5, xlab = "Time", ylab = "Log index", main = "Log scale")  
lines(pred[,1], lwd = 2)   
lines(pred[,2], lwd = 2, lty = 2)  
lines(pred[,3], lwd = 2, lty = 2)  
  
# we can also back-transform the estimates to look at them on the natural scale  
pred\_nat = exp(pred)  
  
plot(index ~ year, pch = 16, cex = 1.5, xlab = "Time", ylab = "Population index",  
 main = "Natural scale")  
lines(pred\_nat[,1], lwd = 2)  
lines(pred\_nat[,2], lwd = 2, lty = 2)  
lines(pred\_nat[,3], lwd = 2, lty = 2)



# convert r to lambda  
r = -0.01 # change value of r here  
lambda = exp(r)  
lambda

## [1] 0.9900498

**Question 2: What are the model estimates for intercept and slope? What do those numbers correspond to in the exponential population growth model?**

**Question 3: Is this population increasing, decreasing, or stable? Report your estimate of population growth rate (r) and the standard error. Use the output from the linear model to support your conclusion. (Is the estimated growth rate significantly different from 0?) Calculate λ based on your estimate of r.**

Question 4 (answer independently): Below are counts of coyote scat in a WMA conducted following consistent methodology over 12 years. Year Count 2006 26 2007 27 2008 25 2009 26 2010 24 2011 27 2012 26 2013 26 2014 27 2015 28 2016 29 2017 27

1. Is the population of coyotes in this WMA increasing, decreasing, or stable? Report the appropriate estimate(s) with precision and results of statistical tests to support your conclusion.
2. After speaking with the WMA biologists, you realize that the survey methodology changed in 2013. All WMA roads were surveyed, but in 2013 a few roads were washed out and never restored, so the total length of road surveyed was shorter for all subsequent years. Can these data still be used to estimate the population growth rate? Why or why not?

### Marked individuals

#### Two-sample mark-recapture (Lincoln Petersen estimator)

We can estimate the total population size with these three pieces of information, collected over two sampling periods and assuming the population was closed between those periods (no births, deaths, emigration, or immigration): \* M = the number of individuals captured, marked, and released in the first period (Marked)  
\* C = the number of individuals captured in the second period (Captured)  
\* R = the number of previously-marked individuals captured in the second period (Recaptured)

$$\hat N = \frac{(M+1)(C+1)}{(R+1)}-1 \\
\\
var(\hat N) = \frac{(M+1)(C+1)(M-R)(C-R)}{(R+1)^2(R+2)}\\
\\
SE= \sqrt {var} \\ CV = \frac{SE}{\hat N}\*100 $$

The coefficient of variation (CV) is a relative measure of uncertainty in our estimate; a larger CV indicates more uncertainty. Often people will interchange whether they talk about precision or uncertainty. Precision is the opposite of uncertainty; the more precise an estimate, the less uncertainty we have in its value.

M = 10 # number captured and marked in first sampling period  
C = 14 # total number captured in second sampling period  
R = 3 # number of recaptures (previously-marked) in second samplng period  
  
# 2-sample Lincoln-Petersen population size estimate  
Nhat = (((M+1)\*(C+1))/(R+1))-1 # Lincoln-Petersen estimator adjusted for small sample size  
  
# estimate variance in estimate of population size and compute confidence intervals  
var = ((M+1)\*(C+1)\*(M-R)\*(C-R))/(((R+1)^2)\*(R+2))  
se = sqrt(var)  
conf.low = Nhat - 1.96\*se  
conf.high = Nhat + 1.96\*se  
  
# coefficient of variation   
cv = (se/Nhat)\*100  
  
# print results  
print(c("Nhat" = Nhat, "conf.low" = conf.low, "conf.high" = conf.high, "cv" = cv))

## Nhat conf.low conf.high cv   
## 40.25000 15.54992 64.95008 31.30952

We can also estimate detection probabilities (capture probabilities) for each sampling period.

$$ \hat{p\_1} = \frac{R}{C} \\ \hat{p\_2} = \frac{R}{M}$$

p1 = R/C  
p2 = R/M  
  
p1

## [1] 0.2142857

p2

## [1] 0.3

We often use information gained from previous studies to help plan future ones. For planning a study to estimate abundance with the Lincoln-Petersen estimator, we can estimate the expected CV using guestimates of the true population size N and the expected detection probability, assuming detection probability at each sampling period are equal.

p = 0.15 # detection probability  
N = 100 # expected population size  
  
exp\_cv = (1-p)/(p\*sqrt(N))\*100  
exp\_cv

## [1] 56.66667

**Question 5: A study was conducted to estimate the size of a population of common snapping turtles on a 3-km stretch of a large river. In the first sampling period, 10 snapping turtles were individually marked with PIT tags and shell notching and released. In the second sampling period a few days later, 14 snapping turtles were captured, 3 of which were recaptures from the first period.Use the R script provided to you to calculate the estimated population size, variance, and CV of that estimate for snapping turtles in this river and report those numbers. How would you describe the uncertainty/precision of this estimate? Use R to calculate detection probability for each sampling period and report those values. Was detection probability the same between the two periods?**

Question 6 (independent): A similar study was conducted on a different, larger stretch of river with more intensive sampling effort. In the first period, 41 turtles were captured and marked. In the second sampling period, 38 total individuals were captured, and 14 of them were recaptures from the first period. a. Using the Lincoln-Petersen estimator, what is the estimate of population size for this river? b. What were the detection probabilities for each sampling period?

Question 7: Compare precision (CV) and detection probabilities between the two studies. a. Why are they different? b. How are precision and detection probability related?

Question 8: A student is planning a two-period closed population mark-recapture survey with live traps to estimate abundance of marsh rice rats at a study site in central Alabama. The student will establish a grid of live traps placed at some fixed spacing. The student is charged with picking a grid size and spacing that maximizes his expected precision (i.e. minimizes the expected CV of the abundance estimate). The student has 100 traps to use. The bigger the grid size selected, the wider the spacing needs to be among traps to cover the grid. Moreover, as the distance between traps on the grid increases, the detection (capture) probability of rats on the grid is expected to decrease. All grid sizes being considered are much bigger than the short-term home range size of a rice rat. As trap spacing increases, each individual rat on the grid is exposed to fewer traps and is therefore has a lower probability of being caught on each trap-night. However, the student knows that, all things being equal, it is much easier to get precise abundance estimates for a larger population than a small one, and the bigger the grid, the bigger the population being sampled. Thus, there is a trade-off between grid size and trap spacing – and therefore between the size of the population being sampled and the expected nightly capture probability.

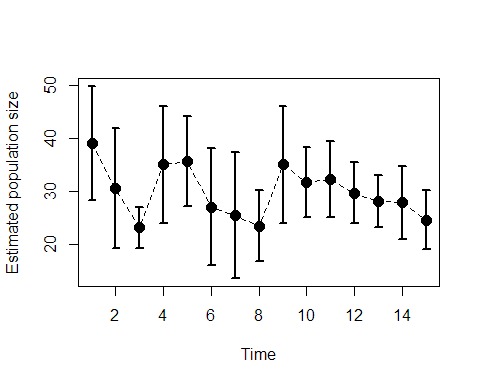
The student is considering three grid sizes (options A, B, C). Based on previous work in similar habitats, the student is given the following information about how true population size and expected nightly capture probability are expected to change among the three options for grid size and trap spacing.

Grid layout options Expected population size Distance between each trap Expected detection probability A: 0.8 ha 15 10 m 0.45 B: 1.8 ha 40 15 m 0.40 C: 4 ha 100 22 m 0.15

Calculate and report the expected CV of the abundance estimates for each of the above three sampling options. Which sampling option is expected to produce the most precise estimates (lowest CV)? Why?

##### Estimating abundance over several years

# calculate N and lambda over several years   
dat = read.csv("marked\_dat.csv")  
  
# calculate Nhat and variance using same equations as above  
dat$Nhat = (((dat$M+1)\*(dat$C+1))/(dat$R+1))-1  
dat$var = ((dat$M+1)\*(dat$C+1)\*(dat$M-dat$R)\*(dat$C-dat$R))/(((dat$R+1)^2)\*(dat$R+2))  
dat$conf.low = dat$Nhat - 1.96\*sqrt(dat$var)  
dat$conf.high = dat$Nhat + 1.96\*sqrt(dat$var)  
  
# plot estimates of N over time  
par(mfrow = c(1,1))  
plot(dat$Nhat ~ dat$Year, pch = 16, cex = 1.5, xlab = "Time", ylab = "Estimated population size",   
 ylim = c(min(dat$conf.low), max(dat$conf.high)))  
lines(dat$Nhat ~ dat$Year, lwd = 1, lty = 2)  
arrows(x0 = dat$Year, x1 = dat$Year, y0 = dat$conf.low, y1 = dat$conf.high,   
 angle = 90, code = 3, length = 0.04, lwd =2)



# look at estimates  
dat

## Year M R C Nhat var conf.low conf.high  
## 1 1 20 10 20 39.09091 30.371901 28.28921 49.89261  
## 2 2 16 6 12 30.57143 33.826531 19.17195 41.97090  
## 3 3 16 11 16 23.08333 3.859509 19.23279 26.93388  
## 4 4 17 8 17 35.00000 32.400000 23.84348 46.15652  
## 5 5 21 11 19 35.66667 18.803419 27.16754 44.16580  
## 6 6 13 5 11 27.00000 32.000000 15.91257 38.08743  
## 7 7 10 4 11 25.40000 36.960000 13.48423 37.31577  
## 8 8 12 7 14 23.37500 11.848958 16.62823 30.12177  
## 9 9 17 8 17 35.00000 32.400000 23.84348 46.15652  
## 10 10 16 11 22 31.58333 11.487714 24.94020 38.22647  
## 11 11 20 11 18 32.25000 13.427885 25.06776 39.43224  
## 12 12 15 11 22 29.66667 8.649573 23.90228 35.43106  
## 13 13 17 12 20 28.07692 6.390533 23.12214 33.03171  
## 14 14 17 9 15 27.80000 12.567273 20.85173 34.74827  
## 15 15 16 9 14 24.50000 8.113636 18.91705 30.08295

# calculate lambda for each time step (need to add the NA on the end so it is the correct length)  
dat$lambda = c(dat$Nhat[2:nrow(dat)]/dat$Nhat[1:(nrow(dat)-1)], NA)  
  
# we can also calculate a variance on our estimate of lambda based on variance in estimates of N  
# first calculate the coefficent of variation (cv) for Nhat estimates  
cv0 = c(sqrt(dat$var[ 1:(nrow(dat)-1)])/dat$Nhat[1:(nrow(dat)-1)], 0)  
cv1 = c(0, sqrt(dat$var[2:nrow(dat)])/dat$Nhat[2:nrow(dat)])  
  
dat$lambda\_var = dat$lambda^2 \* (cv0 + cv1)  
  
# plot estimates of lambda over time  
plot(dat$lambda ~ dat$Year, pch = 16, cex = 1.5, xlab = "Time", ylab = "Estimated population growth rate", ylim = c(0.5, 2.1))  
lines(dat$lambda ~ dat$Year, lwd = 1, lty = 2)  
arrows(x0 = dat$Year, x1 = dat$Year, y0 = dat$lambda - dat$lambda\_var, y1 = dat$lambda + dat$lambda\_var,   
 angle = 90, code = 3, length = 0.04, lwd =2)  
abline(h = 1, lty = 3)

