Why Bayesian approaches?
The average height of a rare plant

Estimation and comparison of averages is an important step in many ecological analyses and demographic models. In this demonstration you will be introduced to R and JAGS functions which will allow you to reach this goal. You will also be confronted with two alternative analysis frameworks to estimate averages: Frequentist and Bayesian. These two approaches represent different ways to evaluate information. We offer you a simple example that will help you to decide their relative advantages when obtaining inference from data. We use the *Hypericum cumulicola* data analyzed in the previous demo (Quintana-Ascencio *et al.* 2003), but we answer a different question. In this case we want to characterize the variation in height of *Hypericum cumulicola* by obtaining a robust estimate of its mean, and an assessment of the uncertainty of this estimate.

Figure 1. Studying *Hypericum cumulicola*

Getting ready. For this demo, you will need to download five files from the course website:
1) The *Hypericum cumulicola* individual height data (`hypericum_data_94_07.txt`).
2) The *Hypericum cumulicola* population height data (`popmeanHc.txt`).
3) The main R script for the analyses (`Averages in Jags.R`).
4) The script for the JAGS model with uninformed priors (`meanmodel_jags.R`)*.
5) The script for the JAGS model with informed priors (`meanmodel_informative_jags.R`)*.
6) You will also need to install the add-on package `rjags` (Plummer 2013) to communicate with JAGS**.

*Models for analysis with JAGS can either be defined in separate files (as we will do this week), or within the main R script (as we will do in the future).
** In the program below, the variables and settings of the model are specified in the R scripts and sent to JAGS, which then returns the output straight back to R, so there is no need to open JAGS directly (it does however need to be installed in the computer separately from R).

Part I. Plotting histograms and calculation of averages using a frequentist approach

Open the R script `Averages in Jags.R` . To see how this script works, let's take a look at the code. For this part we will concentrate in the upper portion of the script. The first line in the

program `rm(list=ls())`allows you to clear R's memory. This is a good practice when you start a new program. The second and third lines are used to obtain the current directory and set the directory that will be used in this session. You will need to modify this line to adapt it to your directory pathway. The function `read.table("file name.txt", header=T)` obtains the data from the txt files saved in your directory.

Since 1994, Quintana-Ascencio et al. (2003) have collected demographic data of *Hypericum cumulicola* in 14 populations at Archbold Biological Station, Florida USA. The line:

```
Height_data <-Hc_data$ht_init[Hc_data$stage !="sg" & Hc_data$ht_init<90]
```

filters the plant height data to include only plants older than one year (`Hc_data$stage !="sg"`) and eliminates a questionable point with an extreme size that we detected previously (height > 90 cm). The same line allocates the filtered data to the new variable "Height_data".

```
rm(list=ls())
getwd()
setwd("PATHWAY TO YOUR WORKING DIRECTORY")
Hc_data <- read.table("Hcdata.txt", header=T)
Hc_pop  <- read.table("popmeanHc.txt", header=T)
Height_data <-Hc_data$ht_init[Hc_data$stage !="sg" & Hc_data$ht_init<90]

hist(Height_data,100,main="Histogram of Hypericum cumulicola height (cm)")
abline(v=Hc_pop$pop_mean, col="blue")
abline(v=mean(Height_data), col="red")

mean(Height_data)
var(Height_data)
round(sqrt(var(Height_data))/sqrt(length(Height_data)),4)

summary(lm(Height_data~1))
Call:
lm(formula = Height_data ~ 1)
Residuals:
   Min     1Q Median     3Q    Max
-31.44  -8.74  -0.44   8.56  48.56
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.4401     0.1043   310.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 13.07 on 15686 degrees of freedom
```

It is always useful to check the distribution of the data. We will use the function `hist()` to create a histogram of adult plant heights, and the function `abline()` to add in blue the location of the 14 population means, and in red the location of the overall mean. Your plot should look like Figure 2, where the data seems to follow a truncated normal distribution.

There are alternative ways to obtain an average in R. We first use the functions `mean()` and `var()` to get an overall plant height mean of 32.4 cm (variance = 170.8). We also obtain the value of the standard error of the plant height mean (0.10) cm with a series of R functions (by calculating the

square root of the variance, dividing it by the square root of the sample size, and rounding the result to four digits). R also has a function that can let us accomplish these calculations in fewer steps.

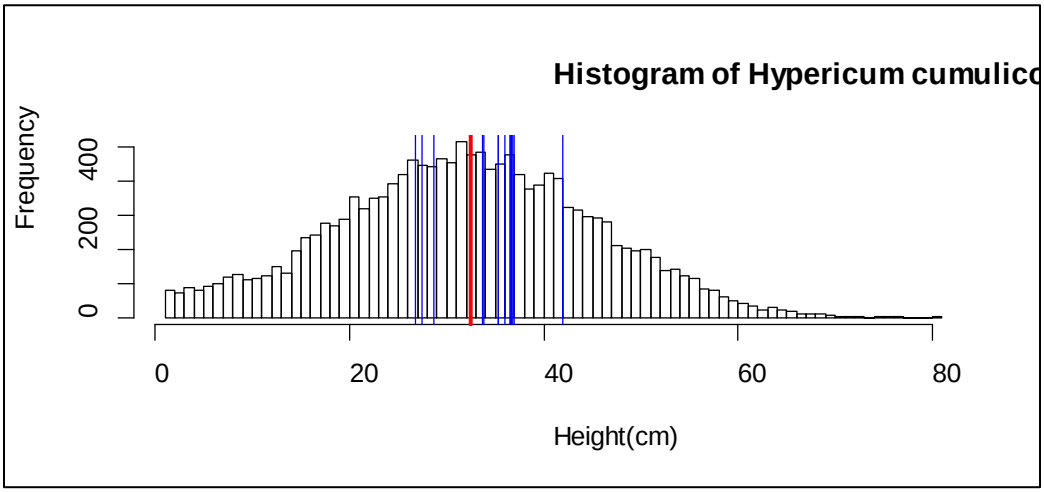

Figure 2. Histogram of individual heights of *Hypericum cumulicola* (1994-1997, n =15687). The red line represents the average height of the whole sample, while the blue lines represent the average of each of 14 populations in Archbold Biological Station.

We should also calculate and plot the mean height and variance when instead of using the data at the individual level, we use the means and variances calculated separately for each population (n=14 populations). This gives us different results (32.4 vs. 34.5, and 170.8 vs 160.6; Figure 3).

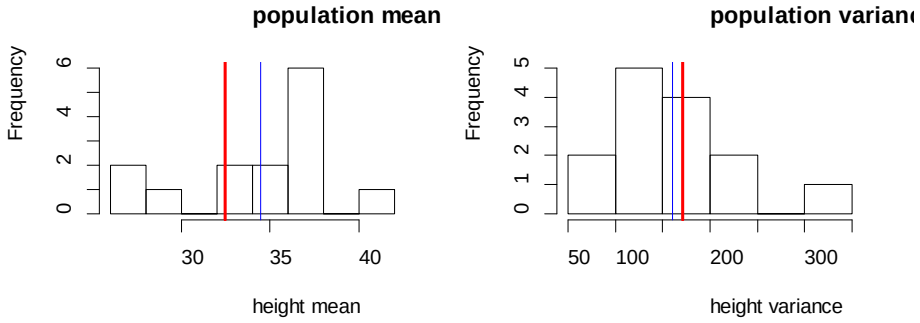

Figure 3. Histograms of mean and variance of height at the population level for *Hypericum cumulicola* (1994-1997, n=14 populations). Blue lines show the overall mean of the averages (34.4) and the overall mean of the variances (160.6) for each of the 14 populations. Red lines show the population mean and variance when calculated from data at the individual level.

Part II. Calculation of averages using a Bayesian approach

We can argue that with a sample of 15,687 individuals we have a reasonable estimate of the average *Hypericum cumulicola* height and its variance. We will now collect random samples of individuals from our data set and estimate their sample mean and uncertainty. We will compare

estimates obtained using Frequentist, and Bayesian approaches with informed and uniformed priors to the average of the whole data set. We use the function `sample( )` to obtain a sample of 10 random plant heights and allocated them to the variable `x`.

```
size <-10
x <- sample(Height_data,size)
n <- length(x)

pop.mean.mean <- mean(Hc_pop$pop_mean)
pop.mean.var <- var(Hc_pop$pop_mean)
log.pop.var.mean <- mean(log(Hc_pop$pop_sd^2))
log.pop.var.var <- var(log(Hc_pop$pop_sd^2))
```

We also create several variables that will be used in the model; n is the sample size and x is the vector that contains the random sample of the whole data. At the population level, pop.mean.mean is the mean of the means of each population, pop.mean.var is the variance of the means of each population, log.pop.var.mean is the mean of the logarithm of the variances, and log.pop.var.var is the variance of the logarithm of the variances.

We use the function `library(rjags)` to call the program that connects with JAGS, and define the models in separate scripts. For the **first model (**`meanmodel_jags.R`**)**, we define two uninformative priors. The first one is the uninformative prior for the mean `dnorm(0,1.0E-6)`. The second one is the uninformative prior for the variance `dgamma(0.001,0.001)`. Notice that the first prior describes a normal distribution while the second corresponds to a gamma distribution (in WinBUGS we would use a lognormal distribution). Because variance is indicated as precision (1/variance), our uninformative prior for the mean has a mean of zero and a large variance, implying our assumed lack of prior information.

```
library(rjags)
model =  "meanmodel_jags.R"

model{
## Priors
mean_height~dnorm(0,1.0E-6)     #uninformative prior for mean height
var_height~dgamma(0.001,0.001)  #uninformative prior for variance of the height
prec <- 1/var_height            #precision = 1 / variance
## Likelihood
for(i in 1:nobs){               #for each plant
Y[i]~dnorm(mean_height,prec)    #assume the height comes from normal distribution
}}
```

In the model, we also specify the calculation of the likelihood for each element of the sample. For more information read McCarthy (2007). Back in our main script, we allocate the data to mean.data, assign the initial values for the Markov chain, specify the parameters that will be obtained, and enter the settings for the Monte Carlo Markov chain (MCMC): **`n.chains`** =

number of chains, **`n.iter`** = number of iterations, **`n.adapt`** = number of chains discarded, **`n.iter (inside coda.samples)`** = the frequency for extracting values.

```
# Inits function
  inits=list(
  list(var_height=100,mean_height=100),
  list(var_height=80,mean_height=60),
  list(var_height=120,mean_height=90)
  )
# Bundle data
  mean.data=list(
  Y=x,
  nobs=n
  )
# Parameters to estimate
  params <- c("mean_height","var_height")
# MCMC settings, start Gibbs sampler and plot results and diagnostics
  jm = jags.model (model, data=mean.data, inits=inits, n.chains=length(inits),
       n.adapt=5000)
  update(jm, n.iter=10000)
  zc=coda.samples(jm,variable.names=params,n.iter=10000)
  gelman.diag(zc)
  plot(zc)
  summary(zc)
  R11<- summary(zc)
```

The function `jm()` determines the procedure for the model and allocates the results into the variable `R11`. A realization of the program is presented below (your values will vary according to the sample, but also from one run of the model to another).

```
Iterations = 15001:25000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean     SD Naive SE Time-series SE
mean_height  30.89  3.771  0.02177        0.02177
var_height  143.22 87.094  0.50284        1.39658

2. Quantiles for each variable:

             2.5%   25%    50%    75%  97.5%
mean_height 23.43 28.55  30.86  33.23  38.45
var_height  53.45 88.57 120.44 170.37 375.55
```

For the implementation without priors, we obtained an estimate of the sample mean of 30.89 and variance of 143.22, together with their 95% credible intervals.

For the **second model (**`meanmodel_informative_jags.R`**)**, we use informed priors based on the data for the 14 separate populations, which we can easily argue represents relevant prior information. We use `dnorm(pop.mean.mean,1/pop.mean.var)[34.47,0.057]` as the prior for the mean, and `dlnorm(log.pop.var.mean,1/var(log(Hc_pop$pop_sd^2))[5.019,8.069]` for the variance (transformed to precision = 1/v). Compare the distributions in Figure 4. You can think of a scenario where two different people sampled *Hypericum cumulicola* independently. The first person conducts an extensive census of several populations, while the second one takes a smaller overall sample of 10 individuals. The second person uses as prior the information from the larger sample. The data available for the second person is the sample of 10 individuals but they incorporate the information from the first person. We need to replace the following lines of code in the R scripts part to calculate informed parameters for the prior:

```
## Priors
mean_height~dnorm(m,v)        #informative prior for the mean height
var_height~dlnorm(mv,vv)      #informative prior for the variance of the height
prec <- 1/var_height          #precision = 1 / variance
# Bundle data
mean.data=list(Y=x, nobs=n, m=pop.mean.mean, v=1/pop.mean.var,
v=log.pop.var.mean, vv=1/log.pop.var.var)
```

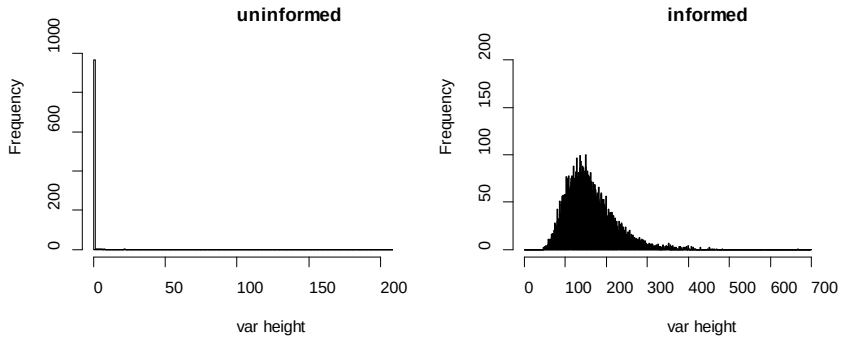

Figure 4. Distributions of priors for the Bayesian models of mean plant height.

A realization of the program is presented below (remember that the results will vary for each sample). For the implementation of the model with priors, we obtained an estimate of the sample mean of 32.47 and a variance of 143.18, together with their 95% credible intervals:

```
Iterations = 15001:25000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
              Mean     SD Naive SE Time-series SE
mean_height  32.47  2.779  0.01605        0.01619
var_height  143.18 42.904  0.24771        0.38093

2. Quantiles for each variable:
             2.5%   25%    50%   75%  97.5%
mean_height 27.02  30.6  32.44  34.3  38.02
var_height  78.54 112.4 136.78 166.7 246.58
```

Part III. Comparing results from all approaches

We now have to implement a frequentist approach to make estimations for the sample of 10 individuals (we used it at the beginning with all the data), and allocate all the results to a table (see code in the R script).

	mean	variance	M2.5%	M97.5	V2.5%	V97.5
Frequentist	30.90000	112.5444	23.31100	38.48900	NA	NA
B without priors	30.88961	143.2203	23.42798	38.45091	53.44854	375.5524
B with priors	32.46636	143.1764	27.01561	38.01692	78.53672	246.5807
Population	34.46857	160.6169	NA	NA	NA	NA
Overall	32.44015	170.7558	NA	NA	NA	NA

This realization of the Bayesian model with informed priors had closer values to the overall mean and variance and much narrower credible intervals (27.01-38.01) than those estimated with the uninformed Bayesian model (23.42-38.45) or the confidence interval of the frequentist approach (23.31-38.49). The last two were commensurate.

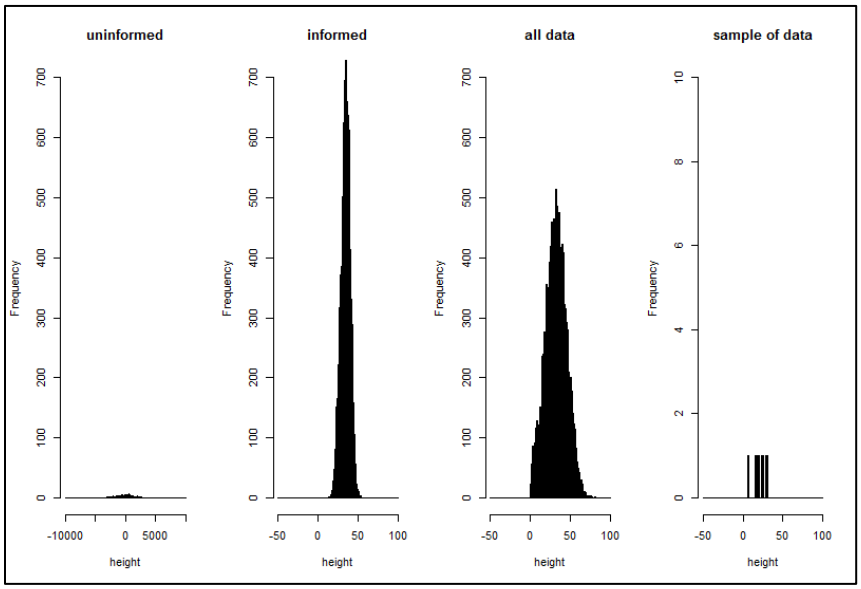

Figure 5. Distributions of the uninformed prior, informed prior, complete dataset and sample of the data used for the Bayesian models of mean plant height.

References.

Quintana-Ascencio, P. F., E. S. Menges, and C. Weekley. 2003. A fire-explicit population viability analysis of Hypericum cumulicola in Florida rosemary scrub. Conservation Biology 17: 433-449.

McCarthy. M.A. 2007. Bayesian Methods for Ecology. Cambridge University Press.

Plummer, M. 2013. JAGS Version 3.4.0 user manual.