CBCS
HTTP://AGROBIOL.SGGW.WAW.PL/CBCS

COMMUNICATIONS IN BIOMETRY AND CROP SCIENCE
VOL. 10, NO. 2, 2015, PP.58–64

INTERNATIONAL JOURNAL OF THE FACULTY OF AGRICULTURE AND BIOLOGY,
WARSAW UNIVERSITY OF LIFE SCIENCES – SGGW, POLAND

TEACHING CORNER

Confidence intervals: am I unconsciously Bayesian?

Andrea Onofri

Department of Agriculture, Food and Environmental Sciences, University of Perugia, Italy.
E-mail: andrea.onofri@unipg.it

CITATION: Onofri A. (2015). Confidence intervals: am I unconsciously Bayesian? *Communications in Biometry and Crop Science* 10, 58–64.

Received: 10 March 2015, Accepted: 30 May 2015, Published online: 17 August 2015

ABSTRACT

To most biologists, the exact meaning of confidence intervals is very difficult to grasp, though such intervals are shown in many of our papers as measures of data variability. One of the reasons lies in the fact that the traditional way of teaching confidence intervals suggests much more than they actually deliver. Therefore, working with biologists, statistics teachers need a convincing way of introducing this topic and, to my experience, Monte Carlo simulation offers some opportunities. However, understanding the crude meaning of frequentist confidence intervals may be disappointing for biologists, who might be seduced by the intuitive appeal of Bayesian credible intervals.

Key Words: *credible intervals; priors; Bayes' rule; R; RJAGS*

INTRODUCTION

A few months ago I made a survey among the students attending my course in 'Experimental Methods in Agriculture'. One of the questions was:

> QUESTION: „I sampled 5 maize plants from a field and found that their average height was $\bar{x}$ = 2. The confidence interval for the population mean was 1.8 - 2.2. What is the meaning of such a statement?".

The possible answers were (height is measured in meters and, for better clarity, measurement units will not be shown):

1. There is 95% probability that the 'true' population mean (μ) is between 1.8 - 2.2.
2. If we sample repeatedly from our population of maize, the estimated confidence intervals will contain the true mean in 95% of cases.
3. The true population mean is certainly between 1.8 - 2.2.
4. The true population mean can take any value between 1.8 - 2.2.

It is perhaps necessary to give some detail about the background for this survey: it came after the first half of the course, approximately one month after the lecture about point and interval estimation. At that stage of the course, I expected all the students to be familiar with the fundamental fact that whenever we make an experiment, even if it is a very simple one, our main interest is not in describing a sample, but in estimating the characteristics of the whole population from which the sample was taken. The whole survey consisted of 20 questions, and the students were asked to select the correct answers, without looking at textbooks or class notes, just using their memory and intuition.

In the end, 75% of my 36 students chose answer 1 while none of them chose answer 2. This came out quite as a shock to me: in a frequentist setting, the correct answer is clearly 2. Indeed, it should be intuitively clear that there is a 'true' (fixed) average height μ for my maize population, though I will never come to known it exactly, unless I measure all the plants in the population, clearly an impossible task. Therefore, I am forced into taking a sample from this population and measuring its average height ($\overline{x}$). My intuition suggests that further samples will show different average heights, some closer while others farther from the true μ, but this true μ will not change. Recalling the frequentist definition of probability (from Wikipedia: „the limit of the relative frequency of an event in a large number of trials"), it would seem pretty clear that it makes no sense to attach any sort of probability to the true value of μ as this is not going to change at all during my experiment! This is why answers 1, 3 and 4 do not make sense in a frequentist setting. Furthermore, the confidence interval (1.8 to 2.2) built from my sample may either contain μ or not, but I have no hints to favor one of the two situations. And the limits of the interval (1.8 - 2.2) are actually meaningless: when I repeat the sampling, they will very likely change.

The above reasoning seems pretty clear, so why did not students select answer 2? Why did they intuitively embrace the perspective of answer 1? It's clear that I did not deliver the correct message during my lecture! In the following days I asked my colleagues biologists about their point of view. I was surprised to note that most of them see confidence intervals very much like it is described in answer 1.

It has been stated that *„confidence intervals seductively suggest more than it is actually delivered. What is delivered is an interval. What is suggested is that the probability that the estimated parameter is in the interval is 0.95"* (Dennis 1996). We clearly see that answer 1 is strongly rooted in biology.

THE USUAL TEACHING APPROACH

Indeed, if we take a look at the milestones of biometry, such as Sokal and Rohlf (1981) or Snedecor and Cochran (1991), confidence intervals are always described starting from the sampling distribution of random variables. It is indeed pretty intuitive for all students that when we sample repeatedly from a normal population $N(\mu, \sigma)$, we obtain a new population of means ($\overline{x}$) that is normally distributed, with mean equal to μ and standard deviation equal to $\sigma_{\overline{x}}$. Considering the values $\overline{x}$ and s (estimates of μ and σ) obtained at each sampling, the following probabilistic statement clearly holds:

$$P\left(-t_{0.975,n-1} \leq \frac{\overline{x}-\mu}{s_{\overline{x}}} \leq t_{0.975,n-1}\right) = 0.95 \qquad (1)$$

where $t(0.975, n\!-1)$ is the 97.5th percentile of a t distribution with n – 1 degrees of freedom (where n is the number of sampled units). From there, with simple math it is derived that:

$$P(\overline{x} - t_{0.975,n-1}\, s_{\overline{x}} \leq \mu \leq \overline{x} + t_{0.975,n-1}\, s_{\overline{x}}) = 0.95 \qquad (2)$$

which, unfortunately, may lead us to answer 1. Indeed, it is just a matter of wrong interpretation: such an inequality holds on the long run, meaning that 95% of the intervals built during the repeated sampling process will contain μ. On the contrary, it does not hold for every single sampling effort, which is where the misinterpretation arises! Obviously, I do not intend to criticise the wonderful 'classical' books about biometry and their approach: a careful reading shows that all the above authors give a correct account of how the above equations should be interpreted. However, the risk of misinterpretation is very high.

A BETTER METHOD

My course is strongly rooted in frequentist statistics. So it is fundamental that I deliver the correct interpretation of confidence intervals. To this aim, I should better avoid all the above equations and use Monte Carlo simulation instead. Let's imagine that the true average height in the original population is μ = 2.0 m, while the true standard deviation is 0.2 m. In R we can mimick the experiment by:

```
set.seed(1234)
sample <- rnorm(5, 2.0, 0.2)
sample
## [1] 1.758587 2.055486 2.216888 1.530860 2.085825
mean(sample)
## [1] 1.929529
```

At this step, equation 2 turns out useful to calculate the two limits for the confidence interval:

```
mean(sample) - sd(sample)/sqrt(5) * qt(0.975, 4)
## [1] 1.583293
mean(sample) + sd(sample)/sqrt(5) * qt(0.975, 4)
## [1] 2.275766
```

What insights do we gain from such an interval?

1. A measure of precision: for a given sample size, the smaller the interval, the higher the sampling precision.
2. The confidence that if we repeat the sampling, in 95% of the cases our confidence interval contains μ. Indeed, Monte Carlo simulations allow us to demonstrate that this is true. In R, we just need to create an object to store the results (1 if our interval contains μ, 0 if it does not) and perform 100,000 simulations:

```
result <- rep(0, 100000)
for (i in 1:100000){
sample <- rnorm(5, 2.0, 0.2)
limInf<- mean(sample) - sd(sample)/sqrt(5) * qt(0.975, 4)
limSup<- mean(sample) + sd(sample)/sqrt(5) * qt(0.975, 4)
if (limInf<= 2.0 &limSup>= 2) result[i] = 1
}
sum(result)/100000
## [1] 0.95052
```

We need to note the following issues:

1. It is false that there is 95% probability that the 'true' population mean is between 1.583293 - 2.755766 m. In fact, the true population mean is always 2.0.
2. It is false that if we repeatedly sample the population, 95% of the observed means will be between 1.583293 - 2.275766 m. Indeed, we see that the all the observed means in 100,000 simulations are within that interval:

```
result <- rep(0, 100000)
for (i in 1:100000){
sample <- rnorm(5, 2.0, 0.2)
if (mean(sample) <= 2.755766 & mean(sample) >= 1.583293) result[i] = 1
}
sum(result)/100000
## [1] 1
```

3. It is false that there is 95% probability that the statement 1.583293 <μ< 2.755766 is true. In normal experimental conditions we do not know anything about this, while in this Monte Carlo simulation, where μ was known, we can say that our first confidence interval was able to 'capture' it.

THE BAYESIAN PERSPECTIVE

Indeed, answer 1 is wrong in a frequentist setting, but it is very close to the definition of a Bayesian credible interval. Like the frequentists, the Bayesians recognise that the experimental observations are driven by a 'stochastic' process, described by a certain likelihood function. In our case, observations are 'drawn' from a normal distribution with parameters μ and σ. However, the Bayesians attach to these parameters a 'prior' distribution, e.g. p(μ), which summarises their personal belief before making the experiment. They may have a very vague belief, for instance that the true mean height of our maize population is somewhere between 1 and 3 m, without favouring any particular values within this range; they may also have a very strong belief, for instance that the true mean height is normally distributed around 2, with a standard deviation of 0.2.

In all cases, the Bayesians use the observed sample to update their prior belief and determine a 'posterior' distribution, e.g. p(μ|x), which summarises their new belief about μ, after having seen the data x. And how is the prior belief updated? By using Bayes' rule, i.e. a very old (published in 1764) and simple rule of conditional probability. As we see, in the Bayesian perspective μ is not a fixed unknown quantity, but it has a distribution of probability 'attached' to itself. The existence of this distribution makes such statements as „there is 95% probability that μ is within the interval" or „I am 95% sure that the results of my experiment are correct" perfectly logic, while they are nonsensical in frequentist statistics.

I will not give any further details about Bayesian methods here, which would be far beyond the aims of this paper; I will just suggest to the interested biologists one of the good books on the topic, such as McCarthy (2007). However, it may be useful to show how a Bayesian credible interval looks like.

For a swift calculation, we will use R and the package *rjags* (Plummer 2014), that is, a good interface between R and JAGS, a wonderful free program for Bayesian analyses (JAGS stands for Just Another Gibbs Sampler). If you intend to reproduce the example below, you need to install them in your computer system.

Let's consider again the above-mentioned sample drawn from a normal distribution and submit it to Bayesian analysis with R + JAGS, to obtain a posterior distribution for μ. First of all, we need to create a model specification (JAGS code), which is written to a string of text (*modelSpec*) in R and finally stored to an external text file („firstModel.txt"), by using the function *writeLines()*. Such a specification should contain: (i) a likelihood function for the observations, and (ii) a prior distribution for μ and σ:

```
modelSpec<- „
model{
  #likelihood
  for(i in 1:N){
x[i] ~ dnorm(mu, 1/sigma2)
}
  #priors
  mu ~ dunif(1, 3)
  sigma2 ~ dunif(0, 0.3)
}
#end of model specification
„
writeLines(modelSpec, con=„firstModel.txt")
```

The likelihood function corresponds to a normal distribution, with mean equal to μ and precision equal to $1/\sigma^2$ (this parameterisation is specific to JAGS: the precision is the inverse of the variance σ^2).

Our prior beliefs for the parameters are rather vague: (i) μ ranges from 1 to 3, according to a uniform distribution; (ii) σ^2ranges from 0 to 0.3, according to a uniform distribution (uniform distribution means that all the values in the range are equally likely).

Successively, we create a list to host the data needed for the analysis (*dataset*) and a list of initial values for the parameters to be estimated (*init*). Finally, we send model specification and other data to JAGS, by using the function *jags.model()*, provided by the package *rjags*. This function returns samples from the posterior distribution, and the 95% credible interval is any region in this posterior that contains 95% of the values (Kruschke 2011).

```
library(rjags)
set.seed(1234)
sample <- rnorm(5, 2.0, 0.2)
dataset <- list(x = sample, N = length(sample))
init<- list(mu = mean(sample), sigma2 = var(sample))
mcmc<- jags.model(„firstModel.txt”, data = dataset, inits = init,
n.chains = 4, n.adapt = 100)
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
##    Graph Size: 13
##
## Initializing model
update(mcmc, 1000)
res <- jags.samples(mcmc, c(„mu”), 1000)
res$mu
## mcarray:
## [1] 1.93085
##
## Marginalizing over: iteration(1000),chain(4)
lims<- quantile(res$mu, c(0.025, 0.975))
lims
##     2.5%    97.5%
## 1.579337 2.283275
```

It is reassuring to see that the Bayesian credible interval (last line in the above code) and the frequentist confidence interval are very similar in this case. Notwithstanding, the conceptual difference is huge, and answer 1 becomes correct when it is referred to the Bayesian credible interval while still being wrong with the frequentist confidence interval.

We need to be careful, however. Indeed, if we change the vague prior into a more informative one, the results are going to change a lot. For example, if we use a normal prior for μ(with mean=2 and precision=1/0.07) and a narrower uniform prior distribution for σ [i.e. we change *mu ~ dunif(1, 3)* with *mu ~ dnorm(2, 1/0.07)* and *sigma2 ~ dunif(0, 0.3)* with *sigma2 ~ dunif(0, 0.1)* in the above code], we will obtain a much narrower credible interval.

```
modelSpec2 <- „
model{
  #likelihood
  for(i in 1:N){
x[i] ~ dnorm(mu, 1/sigma2)
}
  #priors
  mu ~ dnorm(2, 1/0.07)
  sigma2 ~ dunif(0, 0.1)
}
#end of model specification
„
writeLines(modelSpec2, con=„firstModel2.txt”)
mcmc2 <- jags.model(„firstModel2.txt”, data = dataset, inits = init,
n.chains = 4, n.adapt = 100)
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
##    Graph Size: 15
##
## Initializing model
update(mcmc2, 1000)
res2 <- jags.samples(mcmc2, c(„mu”), 1000)
res2$mu
## mcarray:
## [1] 1.945657
##
## Marginalizing over: iteration(1000),chain(4)
lims<- quantile(res2$mu, c(0.025, 0.975))
lims
##     2.5%    97.5%
## 1.738699 2.159008
```

This is perhaps the most controversial aspect of Bayesian methods: the priors may be used to inject arbitrary elements into the process of data analyses (Lele and Dennis 2009).

AM I SATISFIED WITH THIS?

As a teacher I am happy with the above: I am sure that if I use equation 2 only for the calculation of confidence intervals and Monte Carlo simulation to demonstrate their meaning, I can deliver a more correct message during my lectures. But as a biologist, I have to admit that, like my students and my colleagues, I feel somewhat disappointed by the crude meaning of frequentist confidence intervals. In other words, although the correct answer to my initial question is 2, I find myself wishing it were 1: that would really be satisfactory! However, answer 1 would only be appropriate for a Bayesian credible interval and, although confidence and credible intervals may be very close when a flat prior is used, I do not think that mixing the two concepts is a good thing for students.

A few weeks ago, I presented some of these feelings within the Bayesian Statistics community in Google+, by sharing a post with the same title as this paper. Afterwards, I received several comments where the intuitive appeal of credible intervals was used as an intrinsic advantage of the Bayesian approach over the frequentist approach. I do not intend to pursue the Bayesians/Frequentists war to any extent here, but I have to note that Bayesian analyses have recently become in fashion in the agricultural field (Che and Xu 2010, Cotes et al. 2006, Forkman and Piepho 2014, Josse et al. 2014, Mila and Carriquiry 2004). Kery (2010) lists six advantages of the Bayesian approach to statistics and wonders why not everyone is a Bayesian.

Considering the ironic claim of IJ Good („People who do not know they are Bayesians are called non-Bayesians”; cited in Kery, 2010), I am asking myself: am I one of those people who are Bayesians, but do not know, yet? Perhaps, time has come for me to inject some Bayesianism in my courses.

REFERENCES

Che X., Xu S. (2010). Bayesian data analysis for agricultural experiments. *Canadian Journal of Plant Science* 90, 575–603.

Cotes J.M., Crossa J., Sanches A., Cornelius P.L. (2006) A Bayesian approach for assessing the stability of genotypes. *Crop Science* 46, 2654–2665.

Dennis B. (1996). Discussion: should ecologists become Bayesians. *Ecological Applications* 6, 1095–1103.

Forkman J., Piepho H.-P. (2013). The performance of empirical BLUP and Bayesian prediction in small randomized complete block experiments. *Journal of Agricultural Sciences* 151, 381–395.

Josse J., van Eeuwijk F., Piepho H.-P., Denis J. (2014). Another look at Bayesian analysis of AMMI models for genotype-environment data. *Journal of Agricultural, Biological, and Environmental Statistics* 19, 240–257.

Kery M. (2010). *Introduction to winBUGS for ecologists. A Bayesian approach to regression, ANOVA, mixed models and related analyses*. Academic Press, Amsterdam.

Kruschke J.K. (2011). *Doing Bayesian data analyses. a tutorial with R and BUGS*. Academic Press, Amsterdam.

Lele S.R., Dennis B. (2009). Bayesian methods for hierarchical models: are ecologists making a faustian bargain? *Ecological Applications* 19, 581–584.

McCarthy M. (2007). *Bayesian methods for ecology*. Cambridge University Press, New York.

Mila A., Carriquiry A. (2004). Bayesian analysis in plant pathology. *Phytopathology* 94, 1027–1030.

Plummer M. (2014). *Rjags*: *Bayesian graphical models using MCMC. R package version 3-14. http://CRAN.R-project.org/package=rjags.*

Snedecor G., Cochran W. (1991). *Statistical methods*. Page 503. IOWA State University Press, 8th Edition, AMES (Iowa).

Sokal R., Rohlf F. (1981). *Biometry*. W.H. Freeman, New York.