Confidence intervals & credible intervals

Kevin Stadler

Wed Apr 1 17:21:31 2015

Hoekstra et al. (2014) show how nobody (not even statistics professors) know/understand what confidence intervals (CIs) are, how to interpret them or what they are supposed to tell you. Even if you know what they are, in the heat of reading a paper/glimpsing at a graph you are much more likely to interpret them intuitively as something they are not:

is: *if we were to repeat the experiment/data collection procedure* then the CIs of 95% of the samples will contain the true mean

is not: given the sample we obtained, the CI is the region that we can be 95% sure the true mean lies in

Confidence intervals are a *frequentist* concept, meaning that there is the assumption of one underlying true mean. From the frequentist viewpoint any particular CI you're looking at does either contain this (unknown) true mean or not. The CI is a 'tag' indicating the quality of your experimental procedure but, importantly, for any particular sample *the CI is not intended to indicate a region we think the true mean is likely to lie in*!

If you're interested in knowing the range of likely values of the true mean based on your sample (the intuitive reading that most people go for), you are actually thinking of the credible interval (aka 'Bayesian confidence interval') for that parameter. The difference between those two types of intervals isn't just philosophical, because the Bayesian credible interval will generally be *wider* than the confidence interval, so people's 'intuitive' reading of CI's is an underestimate of the measure that they think they're looking at![1]

One conclusion from the paper is that this is yet another reason we should all go Bayes, but the obvious next question was: how do you (easily) calculate a credible interval? Let's try:

```
# Take a small normally distributed sample
x <- rnorm(10)

# Use a one sample t-test to get the confidence interval
t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.2924, df = 9, p-value = 0.2284
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.316307  0.359122
## sample estimates:
##  mean of x
## -0.4785923
```

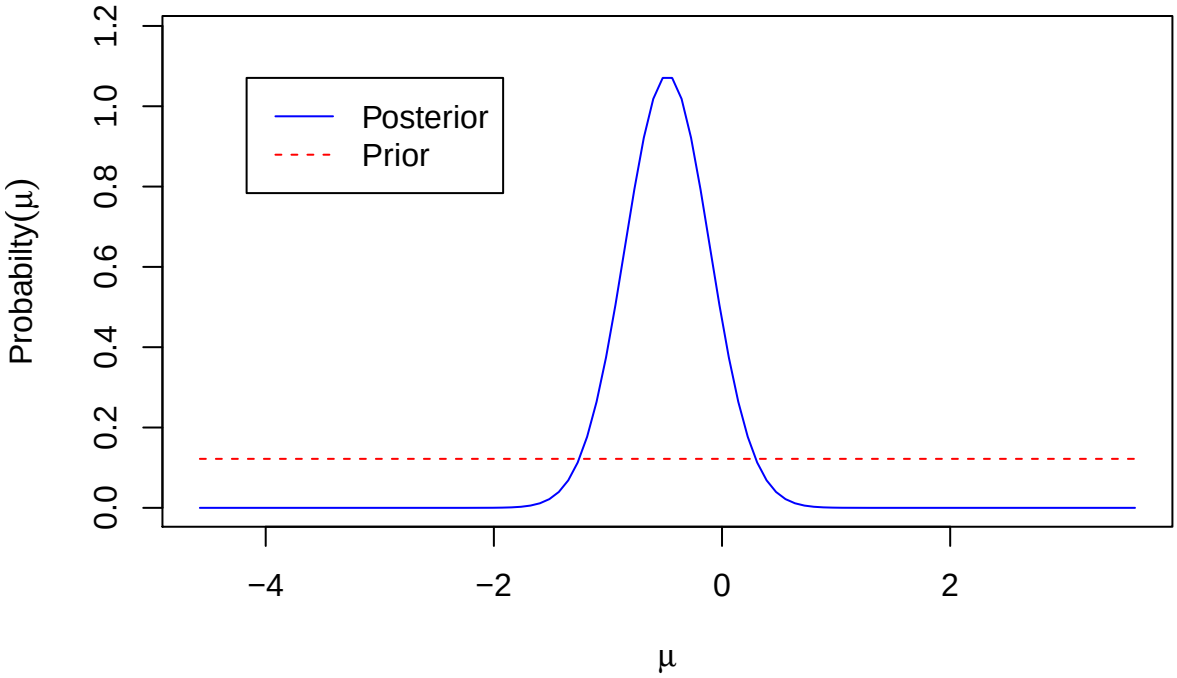
[1] http://stats.stackexchange.com/questions/5903/confidence-intervals-for-regression-parameters-bayesian-vs-classical just how much wider the credible interval is is mostly dependent on the size of your sample. See the final section of this document for a comparison.

Determining a credible interval for the mean of a sample

I first tried the Bolstad package, which has functions for inferring the most likely mean of a sample that's assumed to be normally distributed. Not making any assumptions about our sample, we use `normgcp` to infer the mean using a flat (uninformative) prior:

```
# the Bayesian inference occurs over a finite number of possible means. n.mu sets
# the number of means considered, which determines the resolution of the posterior
mu.mdl <- Bolstad::normgcp(x, n.mu=100)
```

```
## Standard deviation of the residuals :1.171
```



But how do we go from the posterior distribution to a credible interval? A first (naive) attempt at getting at a credible interval for the mean by cutting 2.5% off either side:

```
limit <- sum(mu.mdl$posterior)*0.025
lower <- which(cumsum(mu.mdl$posterior)>=limit)[1]
upper <- length(mu.mdl$posterior) - which(cumsum(rev(mu.mdl$posterior))>=limit)[1]
# naive credible interval assuming fixed (known) sigma:
mu.mdl$mu[c(lower,upper)]
```

```
## [1] -1.1824014  0.1424158
```

This interval is actually *smaller* than the confidence interval determined by the t-test above. What's going on? The issue can be seen in the output of `normgcp` above: the function only draws inferences about the *mean* of the distribution without considering the *standard deviation* which is simply assumed to be identical to the standard deviation of the sample. Known standard deviations can be passed as arguments to the function, but there is normally no reason to assume that we know the standard deviation. Really we will have to make inferences about the mean and standard deviation in tandem (and thus infer a credibility *region* within the distribution's multi-dimensional parameter space).

Bayesian inference over a (ostensibly) normally distributed sample

Here comes *BEST*: *Bayesian estimation supersedes the t test* (Kruschke 2013):

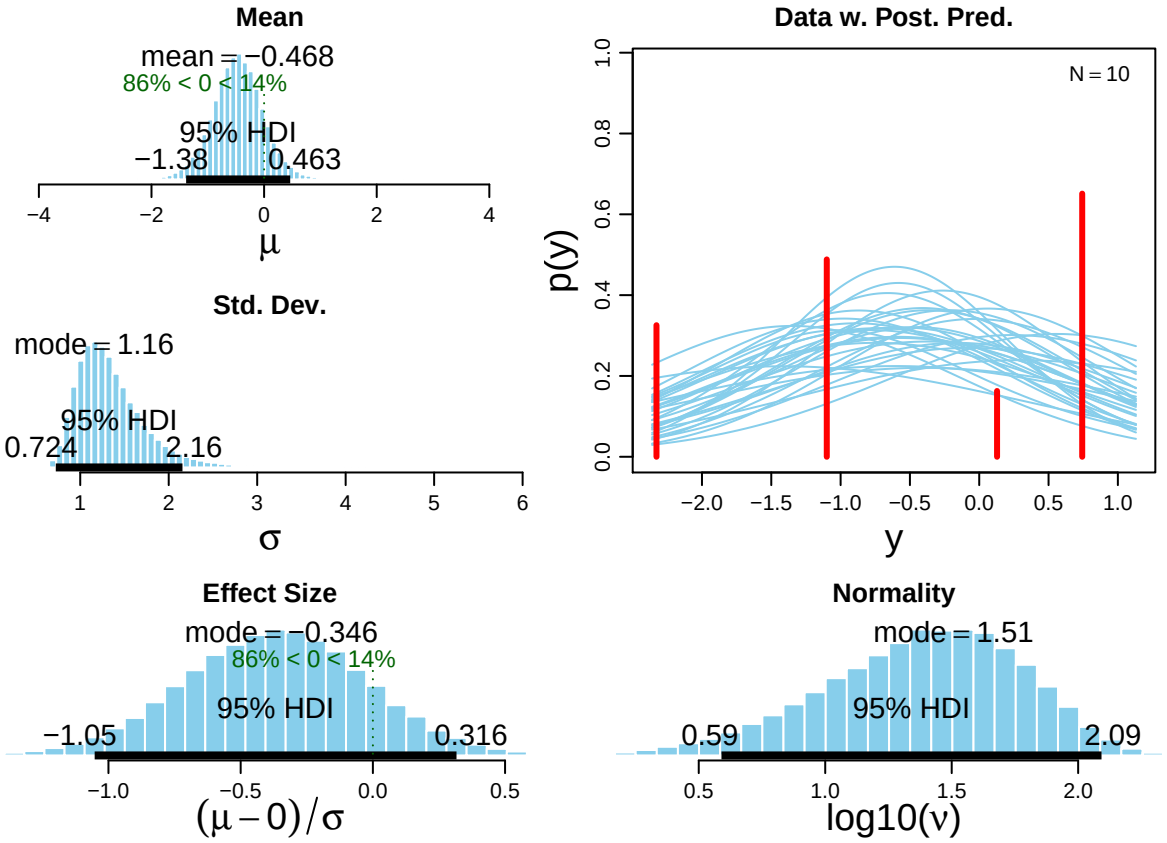
```
library(BEST)
best.mdl <- BESTmcmc(x, verbose=FALSE)
as.matrix(hdi(best.mdl))
```

```
##                mu        nu     sigma
## lower -1.3848890  1.018332 0.7236532
## upper  0.4632778 92.654112 2.1555796
## attr(,"credMass")
## [1] 0.95
```

Not only does this procedure give us a wider credible interval on μ (as we expected), it also gives estimates for standard deviation and ν, a measure of the sample distribution's normality.

The `hdi` function stands for *highest density interval* - since the posterior can have any (non-symmetric) shape, there are many 95% "credible intervals", i.e. many ways to carve out a region that covers 95% of the probability mass. Rather than simply cutting off 2.5% on either end of the posterior, you normally want to select the 95% region which has a greater probability density than any part outside the region. The highest density region of the different parameters is marked by the black bars below[2]:

```
plotAll(best.mdl)
```



[2]NB it is not necessarily always the case that the posterior distribution is unimodal, particularly if you're fitting a complex model. In this case the highest density interval might actually be *discontinuous*, i.e. it might actually consist of two disconnected regions at different locations in the parameter space!

Relation between the confidence and credible interval

The relative difference between the confidence intervals' and credible intervals' width depends strongly on the sample size, primarily because we cannot make any strong inferences about the true standard deviation of the underlying distribution from a small sample. Having to assume that the standard deviation could be very high means that very different settings of the mean don't actually affect the likelihood of the sample that much, leading to a flatter posterior distribution over possible mean values and consequently a wide credible interval.

```
get.cis <- function(x) {
  best.mdl <- BESTmcmc(x, verbose=FALSE)
  cbind(ci=t.test(x)$conf.int, hdi(best.mdl)[,c("mu", "sigma")])
}

compare.cis <- function(x) {
  cis <- get.cis(x)
  cat("Confidence interval:                ", cis[,"ci"], "\n")
  cat("Highest-density credible interval: ", cis[,"mu"], "(sd", cis[,"sigma"], ")\n")
  cat("credible:confidence interval ratio: ", diff(cis[,"mu"]) / diff(cis[,"ci"]))
}

compare.cis(rnorm(4))
```

```
## Confidence interval:                -1.381687 -0.7713592
## Highest-density credible interval:  -1.608579 -0.5860626 (sd 0.06778292 1.040264 )
## credible:confidence interval ratio:  1.675356
```

```
compare.cis(rnorm(5))
```

```
## Confidence interval:                -0.08079982 1.64629
## Highest-density credible interval:  -0.3386799 1.895844 (sd 0.3105238 2.335292 )
## credible:confidence interval ratio:  1.293808
```

```
compare.cis(rnorm(8))
```

```
## Confidence interval:                -1.054145 0.3596856
## Highest-density credible interval:  -1.137022 0.448539 (sd 0.4711624 1.773358 )
## credible:confidence interval ratio:  1.121465
```

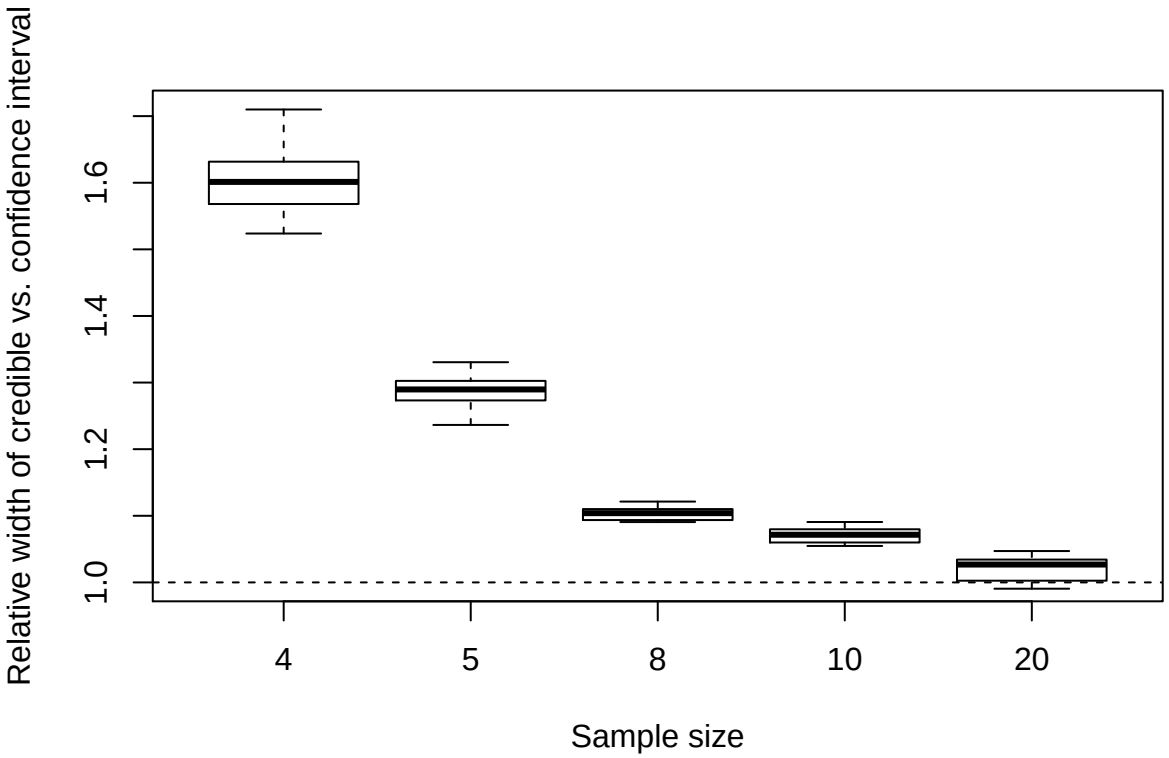
```
compare.cis(rnorm(10))
```

```
## Confidence interval:                -0.709105 0.450966
## Highest-density credible interval:  -0.7461504 0.5006663 (sd 0.4451786 1.486447 )
## credible:confidence interval ratio:  1.074776
```

```
compare.cis(rnorm(20))
```

```
## Confidence interval:                -0.5403241 0.2755756
## Highest-density credible interval:  -0.5341843 0.3134135 (sd 0.6069666 1.262597 )
## credible:confidence interval ratio:  1.03885
```

Below is a more exhaustive investigation of the credible:confidence interval ratio for different sample sizes, confirming that for $n = 4$ the credible interval is consistently about 60% wider than the CI. The boxplots summarise the credible:confidence interval width ratios for different sample sizes obtained from 10 replications each.



References

Hoekstra, Rink, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. 2014. "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin & Review* 21 (5): 1157–1164. doi:10.3758/s13423-013-0572-3. http://www.ncbi.nlm.nih.gov/pubmed/24420726.

Kruschke, John K. 2013. "Bayesian Estimation Supersedes the T Test." *Journal of Experimental Psychology: General* 142 (2): 573–603. doi:10.1037/a0029146. http://www.indiana.edu/~kruschke/BEST/.