Bayesian estimation of the parameters of a normal or Gaussian distribution

This program uses Stan to compute point and interval estimates of the parameters μ and σ of a normal distribution.

The data consists of a vector of N measurements for which we assume the normal distribution is a reasonable model.

Priors for the mean

A distinguishing feature of Bayesian analysis is that it requires a probability model for our initial state of knowledge about the parameters.

Your prior does not have to be very realistic, you just want to make sure that it does not effectively rule out any reasonable values for the parameter.

Critics of the Bayesian approach argue that this is subjective, and it is, but in scientific applications you almost always have some prior knowledge of the phenomenon that generated the data.

This may be in the form of physical laws that the process obeys, or it may come from previous studies.

A good approach is to use what are sometimes called "weakly informative" priors: the probability distribution you use to model your knowledge about the parameter has the mean your information tells you to expect, but double the standard deviation.

For example, if I am modeling ocean temperatures in the Caribbean in early September, we might have information that says it might be around 85 degrees, but probably not warmer than 95 or cooler than 75.

This suggests that a normal prior with a mean of 85 might be appropriate. For the standard deviation, suppose that the likely range is 4 standard deviations, so in this case the range is $95 - 75 = 20$ giving $\sigma = 5$, then double this estimate.

So we might consider a normal distribution with $\mu = 85$ and $\sigma = 10$ to be a weakly informative prior.

One of the advantages of the Bayesian approach is that it allows you to incorporate information not obtained from the data into the analysis, while the frequentist approach only uses the information in the data sample.

You can determine the interval containing 95% of the probability mass for a normal distribution using the `qnorm` function. For example, a standard normal ($N(0, 1)$) has about 95% of its probability mass between -1.96 and 1.96. We

can verify this with the `qnorm` function, which computes percentiles of a normal distribution:

```
c(qnorm(.025,0,1),qnorm(.975,0,1))

## [1] -1.959964  1.959964
```

By adjusting μ and σ, you can get the location and dispersion you feel is appropriate for μ for the data you have.

The default in the `normal.stan` model file is $N(0, 100)$, which has 95% of its probability mass between about -200 and 200. If the mean of your data is likely to lie outside this interval, you should recenter it near the likely value of μ.

Priors for the standard deviation

Andrew Gelman, one of the principal architects of Stan, favors using a half-cauchy distribution as a prior for standard deviations. The cauchy distribution has a similar shape to that of the normal, but much heavier tails (i.e., more observations that are relatively more distant from the mean).

You can use the R `qcauchy` function to obtain approximate percentiles of the half-cauchy. For example, using the `normal.stan` default prior of `cauchy(0,10)`,

```
qcauchy(.95,0,10)

## [1] 63.13752
```

tells us that 95% of the probability mass is to the left of 63, while for `cauchy(0,30)`,

```
qcauchy(.95,0,30)

## [1] 189.4125
```

95% is to the left of 189. If you need to accomodate larger standard deviations, increase the second parameter of the cauchy distribution.

Inference

The point estimates for μ and σ are the means of the posterior draw for those parameters.

The lower bound for the credible intervals is the 2.5^{th} percentile of the posterior draw, and the upper bound is the 97.5^{th} percentile.

Both the mean and percentiles are provided by the `print(stanfit)` command.

If you want to test further hypotheses about the parameters, you can use the `extract(stanfit)` command to extract the full posterior draw. With the default number of iterations, this will produce a draw of 4,000 values each for μ and σ.

Default model file `normal.stan`

```
//Estimate the parameters of a normal distribution
data {
  int N;                    //sample size is N
  real y[N];                //y consists of N real data values
}
parameters {
  real mu;                  //location parameter
  real<lower=0> sigma;      //dispersion parameter constrained to be nonegative
}
model {
  mu    ~ normal(0,100);    //normal prior for mu: centered at zero with sd=100
  sigma ~ cauchy(0,10);     //uniform prior for sigma

  y     ~ normal(mu,sigma); //normal likelihood given parameters (mu,sigma)
}
```