10 Application: dose estimation, Bayes & non-Bayes

In this example, the data consist of bacterial counts per gram (Listeria in broiler legs: Aarnisalo et al 2007). But the data could represent generally any other applications where some concentrations are measured. The question of interest might be to estimate population average dose in broiler legs per gram, or quantify the probability of a specific high dose, based on this observed small set of data. These data represent sampled broiler legs from three different food companies, one measurement per leg. Denote the measurements as $X_1, \ldots, X_n$. Let us first look for an estimate of population average concentration per leg in Finland. To simplify further, it is ignored that the legs represent three companies (clustering). Hence, we assume they are randomly drawn 'from the market'. Denote the unknown population mean as $\mu$.

This example provides a demonstration of both bayesian and non-bayesian approaches, based on a simple model.

A very large proportion of measurements are zero. This is usually the case in food safety applications. A distribution that is very much concentrated to zero is said to be zero inflated. There are a few very large measurements too. Apparently, no standard distribution fits well these kind of data. Let’s demonstrate a few different approaches that can be computed easily by simulation or otherwise.

1. Pick a model.
2. Compute the estimate and quantify the uncertainty by simulations or otherwise, from the specified model.

A usual approach would be to compute a transformation of the original measurements, $X$, and hope that the transformed variables $Y = g(X)$ would show an empirical distribution that resembles the form of some known standard density. A usual transformation is logarithm, but since some of the
measurements are zero, a logarithm cannot be directly used. But we could try \( Y = \log(1 + X) \). Another approach would be to divide the data in two parts: zeros and non-zeros. This would lead to a mixture model. This might be a good approach here because after the transformation \( \log(1 + X) \) the data distribution is still very concentrated at zero.

If only population average is to be estimated, a simple model might be reasonable even though it does not fit all the extreme values well. But for estimating probabilities that an individual dose is higher than some tolerable limit, it becomes more important to model well the upper tail of the distribution. Let’s consider quantifying the probability

\[
P(X > 25).
\]

In a frequentist model this would represent the true fraction of broiler legs with \( X > 25 \) in an infinite population of broiler legs. In our small data, 1.21% of the measurements were higher than 25, so we get a point estimate

\[
P(X > 25) \approx 0.0121.
\]

Next, we would like to describe the uncertainty of this result by different methods of simulations and other calculations. To make the examples simple (and feasible for short calculations), we adopt a simple exponential distribution with only one parameter. The computations are shown as R code.

```r
cfu10<-c(0,0,0,0,0,20,0,0,0,3,0,3,0,0,8,0,0,0,0,0,0,...
0,5,10,13,3,0,3,28,14,5,211,5,0,0,3,18,0,0,5,8,0,10,100,...
0,0,0,0,0,0,0,0,64,0,55,0,70,0,35,0,0,0,0,0,5,27,0,...
0,20,84,0,68,0,52,0,0,34,0,0,5,0,0,148,0,1470,0,0,0,0,...
5,8,5,0,0,0,5,0,8,5,103,0,0,25,0,35,0,0,0,10,8,0,0,0,13,0,...
0,13,0,0,0,18,0,0,0,5,18,8,0,27,0,9,0,0,0,26,0,0,3,0,0,297,...
```

Figure 2: Empirical distribution of \( \log(1 + X) \).
0,5,0,5,0,0,0,0,27,0,0,15,0,0,0,0,0,0)
x<-cfu10/10
  - the concentrations per gram
y<-log(1+x)
  - the transformed measurements

10.1 Frequentist approach, exact result

The simple model for the transformed variables \( Y = \log(1 + X) \) is

\[
Y \sim \exp(\mu),
\]

which is not a perfect model since a large proportion of \( Y \)-values are zero, and exponential random variables are all positive. The expected value of exponential density is \( 1/\mu \), which gives a point estimate for parameter \( \mu \):

\[
\hat{\mu} = 1/\bar{Y}_n = 2.495269,
\]

This is also the maximum likelihood estimate, MLE. From the chosen model and MLE estimate, we can compute

\[
\hat{P}(X > 25) = \hat{P}(Y > \log(1 + 25))
= \int_{\log(1+25)}^{\infty} \hat{\mu} \exp(-\hat{\mu}t) dt = 1 - F(\log(1 + 25)) = \exp(-\hat{\mu} \log(1 + 25)) = 0.0002946,
\]

which is a point estimate of the required probability. The same could be computed by simulation:

```r
gsim <- rexp(10000,2.495269)
P25 <- sum(gsim >log(1+25))/length(gsim)
```

Frequentist confidence interval for \( \mu \) could be obtained by using pivot-technique. This is based on observing that if \( Y_i \sim \exp(\mu) \), then

\[
\frac{\sum Y_i}{n} = \frac{\bar{Y}_n}{n} \sim \text{Gamma}(n, n),
\]

Hence, the CI is obtained by solving the 95\% quantiles of a \( \text{Gamma}(n, n) \)-distribution

```r
lower <- qgamma(0.025, shape = length(y), rate = length(y))
upper <- qgamma(0.975, shape = length(y), rate = length(y))
```

and using them in the equation:

\[
\text{lower} < \mu \bar{Y}_n < \text{upper},
\]

so that:

\[
\frac{\text{lower}}{Y_n} < \mu < \frac{\text{upper}}{Y_n},
\]
which gives us:

\[
2.129053 = \frac{0.8532359}{0.4007584} < \mu < \frac{1.158240}{0.4007584} = 2.890120.
\]

The corresponding upper and lower bounds for the probability are:

\[
P_{\text{upper}}(X > 25) = 0.0009715093
\]

\[
P_{\text{lower}}(X > 25) = 8.138752e - 05
\]

Let's draw some density functions and empirical histograms:

\[
u <- (1:5000)/1000
\]

\[
\text{plot(u,dexp(u,1/mean(y)),'l',xlab="log(1+Y)",ylab="exponential density")}
\]

\[
\text{points(u,dexp(u,2.129053),'l',col="red")}
\]

\[
\text{points(u,dexp(u,2.890120),'l',col="red")}
\]

\[
\text{hist(y,100)}
\]

\[
\text{hist(rexp(165,1/mean(y)),100)}
\]

Figure 3: Densities based on estimates \( \hat{\mu}, \mu_{\text{lower}}, \mu_{\text{upper}} \). Black line: density based on MLE (\( \hat{\mu} \)).

### 10.2 Frequentist approach, asymptotic result

According to theory, the asymptotic distribution of the MLE of \( \mu \) is:

\[
\hat{\mu} = 1/\bar{Y}_n \sim N(\mu, 1/I(\mu)) = N(\mu, \mu^2/n).
\]

Therefore:

\[
\frac{1/\bar{Y}_n - \mu}{\mu/\sqrt{n}} \sim N(0, 1),
\]

which can also be written in this form:

\[
\sqrt{n}(1/(\bar{Y}_n\mu) - 1) \sim N(0, 1).
\]
Figure 4: Empirical distribution of simulated 165 values from Exp(\(\mu\))-density, compared with the actual empirical distribution of data. Is the estimate of \(P(X > 25) = P(Y > 3.258)\) reasonable?

Hence, the 95\% asymptotic CI can be obtained by writing:

\[-1.96 < \sqrt{n}(1/(\bar{Y}_n\mu) - 1) < 1.96,\]

And this is the same as:

\[\bar{Y}_n(-1.96/\sqrt{n} + 1) < \frac{1}{\mu} < \bar{Y}_n(1.96/\sqrt{n} + 1),\]

which provides the result:

\[0.3477038 < \frac{1}{\mu} < 0.4729192,\]

\[2.114526 < \mu < 2.876011,\]

This happens to be nearly the same as the exact solution.

10.3 Frequentist non-parametric approach: simple bootstrap

A CI for the probability \(P(X > 25)\) (proportion of legs with \(X > 25\) in an infinite population) could be obtained by a simple Bootstrap method.

1. Draw randomly a new artificial set of 165 observed values from the data, using sampling with replacement.
2. Compute the percentage of values that are \(> 25\) from this artificial data.
3. Repeat sufficiently many times (10000).
4. Draw the histogram of the percentages and find the 95\% interval of the histogram.

```
P25<-1:10000
   -this will define a vector with length 10000
```
for(i in 1:10000){P25[i] <- sum(sample(x, replace=TRUE)>25)/length(x)}
- this will generate 10000 bootstrap samples of x, and count the proportion of those >25

hist(P25,100)
- draws the histogram

quantile(P25,probs=c(0.025,0.975))
- computes the 95% bootstrap CI from the simulated distribution

As a result, we get CI95%=[0, 0.0303] which is a frequentist confidence interval for the proportion of legs with $X > 25$ cfu/g in an infinite population of legs, according to simple bootstrap method. Since (if?) we can assume that the infinite population proportion can be interpreted as the probability that a randomly chosen leg from this population has $X > 25$, we can also say that this is the CI for such sampling probability.

Figure 5: Bootstrap-distribution of the proportion of measurements with $X > 25$ cfu/g.

10.4 Bayesian approach, exact result

In bayesian approach, we aim to compute a posterior probability for some real quantity that is uncertain to us. The uncertain quantity here would be the proportion of legs with $X > 25$ in a large, or 'infinite', population. In contrast, it is not meaningful to speak of probabilities of probabilities because probability is by definition a degree of uncertainty concerning some real quantity that is of interest. The proportion in a large well defined population is such a quantity, but a proportion in infinite population is strictly speaking not. It is just a convenient approximation.

So, define the infinite population proportion as $P25$. The data consist of the same 165 values as before. We also assume the same exponential model as before. The posterior density of parameter $\mu$ is now, according to Bayes formula:

$$
\pi(\mu \mid Y_1, \ldots, Y_n) = \frac{\prod_{i=1}^{n} \pi(Y_i \mid \mu)\pi(\mu)}{\text{constant}},
$$

As a prior density, we can try to choose an uninformative flat prior among conjugate densities: Gamma($\alpha, \beta$). This gives the posterior as
\[ \text{Gamma}(\alpha + n, \sum_{i=1}^{n} Y_i + \beta). \]

If the prior parameters \( \alpha \) and \( \beta \) are very small, the result depends almost solely on data. The posterior mean is

\[
\frac{\alpha + n}{\sum_{i=1}^{n} Y_i + \beta} \approx 1/\bar{Y}_n,
\]

which gives an estimate for \( \mu \). The point estimate is the same as in the frequentist approach, although the interpretation is different. 95\% Credible Interval can be obtained by computing:

\[
\text{lower} \leftarrow \text{qgamma}(0.025, \text{shape} = \text{length}(y), \text{rate} = \text{sum}(y))
\]
\[
\text{upper} \leftarrow \text{qgamma}(0.975, \text{shape} = \text{length}(y), \text{rate} = \text{sum}(y))
\]

This gives the result: 2.129053 < \( \mu \) < 2.890120. Hence, the bayesian CI is numerically the same as the frequentist CI, although with different interpretation. The required probability \( P(X > 25) = P(Y > \log(1 + X)) \) depends on the unknown parameter, so that it is actually the conditional probability:

\[
P(X > 25 \mid \mu) = \exp(-\mu \log(1 + 25)).
\]

In bayesian approach, we aim to compute the posterior probability \( P(X > 25 \mid \text{data}) \), which is obtained by computing the marginal probability:

\[
P(X > 25 \mid Y_1, \ldots, Y_n) = \int_{0}^{\infty} P(X > 25 \mid \mu)\pi(\mu \mid Y_1, \ldots, Y_n)\,d\mu,
\]

In other words, the possible values of the unknown parameter are weighted by the posterior density while we integrate over \( \mu \). The posterior density was a Gamma density, so that we have:

\[
P(X > 25 \mid Y_1, \ldots, Y_n) = \int_{0}^{\infty} \exp(-\mu \log(1 + 25))\Gamma(n, \sum_{i=1}^{n} Y_i)\,d\mu
\]

And inserting the function of the Gamma density we obtain the mathematical expression for the integral:

\[
\int_{0}^{\infty} \exp(-\mu \log(1 + 25)) \frac{(\sum_{i=1}^{n} Y_i)^n}{\Gamma(n)} \mu^{n-1} \exp(-\mu \sum_{i=1}^{n} Y_i)\,d\mu.
\]

Again, the result follows by solving this integral by paper and pencil (result: \( (\sum_{i=1}^{n} Y_i/(\sum_{i=1}^{n} Y_i + \log(26)))^n \approx 0.0003576747 \)), or by computing the integral using numerical integration methods, or by simulation:

\[
\text{mu} \leftarrow \text{rgamma}(10000, \text{shape}=\text{length}(y), \text{rate}=\text{sum}(y))
\]
\[
\text{P25} \leftarrow \exp(-\text{mu} \times \log(1+25))
\]
\[
\text{P25result} \leftarrow \text{mean}(\text{P25})
\]
\[
\text{hist}(\text{P25}, 100)
\]
\[
\text{quantile}(\text{P25}, \text{probs}=c(0.025, 0.975))
\]

7
which gives: \(7.872829e - 05 < P25 < 9.533395e - 04\), and the mean (P25result) 0.0003587195. Alternatively, we could first simulate \(\mu\) from its posterior density, and then using each of the simulated values generate new predicted values ypred:

```r
mu <- rgamma(1000000, shape=length(y), rate=sum(y))
ypred <- rexp(1000000, mu)
P25result <- sum(ypred > log(1+25))/length(ypred)
```

resulting to (P25result) 0.000361. Note: this technique requires much larger simulation!

If we aim to estimate the true proportion in a ‘large population’, it makes sense to report the bayesian CI, but if we want to report the probability that an arbitrary broiler leg has \(X > 25\), then this (prior or posterior) probability is a single number, and that is \(\approx 0.00036\) which represents the whole uncertainty we still have after having these data and this model and this prior.

![Histogram of P25](image)

Figure 6: Posterior distribution of the proportion of legs with \(X > 25\) cfu/g in a large population, when the model is exponential.

### 10.5 Bayesian approach, approximate result

The exact posterior density could be approximated by a normal density

\[N(\hat{\mu}, I(\hat{\mu})^{-1}).\]

The posterior was a gamma-density, with some parameters \((A, B) = (n, \sum_{i=1}^{n} Y_i)\), so we know that the mode is \(\hat{\mu} = (A - 1)/B\). Otherwise, we would first have to find it by solving the point where the first derivative of the density is zero. The reciprocal of the observed information at \(\hat{\mu}\) is \((A - 1)/\hat{\mu}^2\) = \((A - 1)/B^2\). Hence, the approximation is

\[N\left(\frac{n - 1}{\sum_{i=1}^{n} Y_i}, \frac{n - 1}{(\sum_{i=1}^{n} Y_i)^2}\right) = N\left(\frac{164}{66.12514}, \frac{164}{66.12514^2}\right).\]

The 95\% interval from this normal density gives

\[2.100566 < \mu < 2.859726.\]

Finally, the required probability is obtained in a similar way as above

8
leading to $0.044699 < P_{25} < 1.084429 e - 03$, with mean $0.0003803768$. (Alternatively, we would need to solve the integral given above where the gamma-density is replaced by the normal density).

**Note:** the bayesian approach could be based on an informative prior that would represent some prior knowledge based on earlier data or expert opinion. But then the posterior might not have an easy solution in the form of a standard density. Also, the conditional distribution of data could be replaced by something more realistic than the exponential distribution. Even if the posterior would not be among any standard densities, we could still use MCMC methods to simulate it.

### 10.6 Bayesian non-parametric approach

Short intro to BNP, bayesian non-parametrics: In bayesian non-parametric modeling, the model could just as well be described as a model with infinite number of parameters. For example, in a change point problem, the number of change points could be an unknown parameter with possible values $0, 1, 2, 3, \ldots, \infty$. This would lead to a posterior distribution of the number of change points as well as their unknown locations, together with all other unknown parameters of the model. In other words, the assumed data model would then be a piecewise constant Poisson process, but with unknown number of 'pieces' with constant (but unknown) levels of intensity. The prior is then actually defined for functions: $\pi(f)$, where $f$ belongs to the set of piecewise constant functions. As a result, we then obtain the posterior distribution $\pi(f \mid data)$, and taking the posterior mean of such functions results to a smooth function (that is not any of the piecewise constant functions). In this way, more realistic models may be obtained, instead of arbitrarily choosing a single parametric model.

Also, in models with mixture distributions, the number of distributions in the mixture could be unknown. Hence, the data points could be probabilistically grouped into a number of groups, each with unknown group specific parameters, but without declaring a pre-fixed number of groups. All these examples have an unknown number of parameters which can also be estimated.

Here, we define the data $X'$ according to the original cfu/10g values. (These are all integers). Then we can use Poisson model, from which a mixture model can be constructed, using so called Dirichlet process prior:

$$X'_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i \sim \text{Gamma}(0.01, 0.01)$$

$$\alpha_i \sim DPP$$

Finally, we simulate predicted values from the model for computing the probability of exceeding 25. As a result of MCMC simulation we get:
\[ P(X > 25 \mid X_1, \ldots, X_{165}) = 0.018. \]

For the number of component distributions in the mixture, we obtained 95\% CI of [9-18], with mean 13.

A simplified and slightly different approach with fixed mixture distribution can be found in Aarnisalo et al [1].

References