Bayesian Statistics

Daniel B. Rowe, Ph.D.

Associate Professor
Department of Mathematics,
Statistics, and Computer Science
Outline

• Background

• Likelihood Distribution

• Prior Distribution

• Posterior Distribution

• Posterior Estimation
Bayesian Statistics - Background

We learned about the conditional probability of $B$ given $A$.

If $A$ and $B$ are events in $S$, and $P(A)>0$, then the *conditional probability of $B$ given $A*$ written is,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$
Bayesian Statistics - Background

We extended to more $A$ events, $A_1, A_2, \ldots$

Let $B_1, B_2, \ldots$ be a partition of the sample space, and let $B$ be any set.

Then for each $i=1,2,\ldots$,

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{\sum_{i=1}^{\infty} P(A | B_i)P(B_i)}$$

$$P(A) = \sum_{i=1}^{\infty} P(A | B_i)P(B_i) \quad S$$
Bayesian Statistics - Background

**Example:** Medical Test. $P(\text{have disease}|\text{test positive})$.

$T+$: The event that the test is positive.

$T−$: The event that the test is negative.

$D+$: The event that the person truly has disease.

$D−$: The event that the person truly does not have disease.

The sensitivity of test is $P(T+|D+) = .99$.

The specificity of test is $P(T−|D−) = .99$.

If the proportion of population that truly has disease is $10^{-6}$.

\[
P(D−|T+) = \frac{P(T+|D−)P(D−)}{P(T+)} = 0.99990101
\]

\[
P(T+) = P(T+|D+)P(D+) + P(T+|D−)P(D−)
\]
Bayesian Statistics - Likelihood

Assume that we have \( y_i = \mu + \epsilon_i \), where \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \) for \( i = 1, \ldots, n \).

This means that given \( \mu \) and \( \sigma^2 \), the PDF of \( y_i \) is

\[
f(y_i \mid \mu, \sigma^2) = (2\pi \sigma^2)^{-1/2} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right]
\]

and since these are independent observations, we wrote

\[
f(y_1, \ldots, y_n \mid \mu, \sigma^2) = (2\pi \sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right].
\]
Bayesian Statistics - Prior

In MLE, we sort of heuristically turned things around.

We took $f(y_1,\ldots, y_n \mid \mu, \sigma^2)$ which was a (probability) function of the data $y_1,\ldots, y_n$ given $\mu$ and $\sigma^2$ and changed it into a function $L(\mu, \sigma^2)$ of $\mu$ and $\sigma^2$ (given the data $y_1,\ldots, y_n$).

Why and how did this happen?

Truthfully $L$ is the probability of getting data $y_1,\ldots, y_n$ given $\mu$ and $\sigma^2$ and not probability of $\mu$ and $\sigma^2$ given data!
Bayesian Statistics - Prior

What happened to the rules of probability? i.e. Bayes’ Rule

\[ P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)} \]

Did we just throw out what we have learned?

To be correct, shouldn’t we instead write

\[ f(\mu, \sigma^2 \mid y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n \mid \mu, \sigma^2)f(\mu, \sigma^2)}{f(y_1, \ldots, y_n)} \]

\[ A \rightarrow y_1, \ldots, y_n \quad B \rightarrow \mu, \sigma^2 \]

?
Bayesian Statistics - Prior

What happened to the rules of probability? i.e. Bayes’ Rule

\[
P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)}
\]

Did we just throw out what we have learned?

To be correct, shouldn’t we instead write

\[
f(\mu, \sigma^2 \mid y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n \mid \mu, \sigma^2)f(\mu, \sigma^2)}{f(y_1, \ldots, y_n)}
\]

distribution of \(y\)'s given \(\mu\) and \(\sigma^2\)

distribution of \(\mu\) and \(\sigma^2\)

distribution of \(\mu\) and \(\sigma^2\) given \(y\)'s

\[A \rightarrow y_1, \ldots, y_n \quad B \rightarrow \mu, \sigma^2\]

A marginal distribution of \(y\)'s
Bayesian Statistics - Prior

We have \( f(y_1, \ldots, y_n \mid \mu, \sigma^2) \). The dist of RVs given \((\mu, \sigma^2)\).

We need \( f(\mu, \sigma^2) \), the dist of the parameters.

Given \( f(\mu, \sigma^2) \), we can get \( f(y_1, \ldots, y_n) \) by integration

\[
f(y_1, \ldots, y_n) = \int_{\sigma^2=0}^{\infty} \int_{\mu=-\infty}^{\infty} f(y_1, \ldots, y_n \mid \mu, \sigma^2) f(\mu, \sigma^2) \, d\mu \, d\sigma^2
\]

(but it is just a proportionality constant often neglected).
Bayesian Statistics - Prior

The distribution $f(\mu, \sigma^2)$ is called the prior distribution.

It is arrived at by quantifying expert opinion or using previous data.

There is a way of generating a distributional form for a prior distribution then all we need are its parameters.
Bayesian Statistics - Prior

Although any distribution that depends on certain parameters $\theta$ can be used as a prior distribution, we can obtain a “nice” one called a natural conjugate prior distribution. Then all we need to do is assess the parameters $\theta$ for this distribution either by expert opinion or from previous data.
Bayesian Statistics - Prior

A common joint distribution for the mean $\mu$ and variance $\sigma^2$ when data is normal is the natural conjugate prior distribution, $f(\mu, \sigma^2) = f(\mu | \sigma^2) f(\sigma^2)$

$$f(\mu | \sigma^2) = \left(\frac{2\pi\sigma^2}{\alpha}\right)^{-1/2} e^{-\frac{(\mu-\mu_0)^2}{2\sigma^2/\alpha}}$$

$$f(\sigma^2) = \frac{\kappa^{\nu/2}}{\Gamma\left(\frac{\nu-2}{2}\right)} \left(\frac{\nu}{2}\right)^{-\nu/2} \frac{\kappa}{2\sigma^2} e^{-\frac{\kappa}{2\sigma^2}}$$

$\mu_0, \alpha, \nu, \kappa$

Need to be assessed.

inverse gamma distribution
Bayesian Statistics - Prior

Parameters of prior are called hyperparameters.

The hyperparameters \((\mu_0, \alpha, \nu, \kappa)\) need to be assessed.

One way is from previous similar study data:

i.e. \(n_0\) observations with sample mean \(\bar{y}_0\) and sample variance \(s_0^2\) use

\[
\begin{align*}
\mu_0 &= \bar{y}_0 \\
\nu &= n_0 + 1 \\
\alpha &= n_0 \\
\kappa &= (n_0 - 1)s_0^2
\end{align*}
\]
Bayesian Statistics - Prior

The likelihood of the observations is

\[
f(y_1, \ldots, y_{n_0} \mid \mu, \sigma^2) = (2\pi \sigma^2)^{-n_0/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n_0} (y_i - \mu)^2}
\]

\[
f(\overline{y}_0, s_0^2 \mid \mu, \sigma^2) = \left(\frac{n_0}{2\pi \sigma^2}\right)^{1/2} e^{-\frac{n_0}{2\sigma^2}(\mu - \overline{y}_0)^2} \cdot \frac{n_0^{-1/2}}{(2\pi \sigma^2)^{(n_0-1)/2}} e^{-\frac{(n_0-1)s_0^2}{2\sigma^2}}
\]

\[
f(\mu \mid \sigma^2) = \frac{1}{(2\pi \sigma^2/\alpha)^{1/2}} e^{-\frac{\alpha(\mu - \mu_0)^2}{2\sigma^2}}
\]

\[
f(\sigma^2) = \frac{\kappa^{\nu/2} (\sigma^2)^{-\nu/2}}{\Gamma\left(\frac{\nu-2}{2}\right) 2^{(\nu-2)/2}} e^{-\frac{\kappa}{2\sigma^2}}
\]
Bayesian Statistics - Posterior

We can now form the posterior distribution

\[
f (\mu, \sigma^2 \mid y_1, \ldots, y_n) = \frac{f (y_1, \ldots, y_n \mid \mu, \sigma^2) f (\mu, \sigma^2)}{f (y_1, \ldots, y_n)}
\]

\[
f (\mu, \sigma^2) = f (\mu \mid \sigma^2) f (\sigma^2)
\]

\[
f (y_1, \ldots, y_n \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2}
\]

\[
f (\mu, \sigma^2) = \left(2\pi\sigma^2/\alpha\right)^{-1/2} e^{-\frac{(\mu - \mu_0)^2}{2\sigma^2/\alpha}} \frac{\kappa^{\nu/2} (\sigma^2)^{-\nu/2}}{\Gamma\left(\frac{\nu-2}{2}\right) 2^{(\nu-2)/2}} e^{-\frac{\kappa}{2\sigma^2}}
\]
Bayesian Statistics - Posterior

We can neglect $f(y_1, \ldots, y_n)$ since doesn’t have $(\mu, \sigma^2)$ and other constants

$$f(\mu, \sigma^2 \mid y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n \mid \mu, \sigma^2) f(\mu, \sigma^2)}{f(y_1, \ldots, y_n)}$$

$$f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto f(y_1, \ldots, y_n \mid \mu, \sigma^2) f(\mu, \sigma^2)$$

$$f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2 \sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2} (\sigma^2)^{-\frac{1}{2}} e^{\frac{(\mu - \mu_0)^2}{2 \sigma^2 / \alpha}} (\sigma^2)^{-\frac{\nu}{2}} e^{-\frac{\kappa}{2 \sigma^2}}$$

$$f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto (\sigma^2)^{-\frac{(n+\nu+1)}{2}} e^{-\frac{1}{2 \sigma^2} \left[ \sum_{i=1}^{n} (y_i - \mu)^2 + \alpha (\mu - \mu_0)^2 + \kappa \right]}$$
Now that we have a distribution $f(\mu, \sigma^2 \mid y' \text{s})$, we need to estimate the $(\mu, \sigma^2)$ parameters from it.

We can obtain (marginal) means

$$E(\mu \mid y' \text{s}) = \int \int \mu f(\mu, \sigma^2 \mid y' \text{s}) \, d\sigma^2 \, d\mu$$

$$E(\sigma^2 \mid y' \text{s}) = \int \int \sigma^2 f(\mu, \sigma^2 \mid y' \text{s}) \, d\mu \, d\sigma^2$$

or modes

$$\frac{\partial f(\mu, \sigma^2 \mid y' \text{s})}{\partial \mu} \bigg|_{\hat{\mu}, \hat{\sigma}^2} = 0$$

$$\frac{\partial f(\mu, \sigma^2 \mid y' \text{s})}{\partial \sigma^2} \bigg|_{\hat{\mu}, \hat{\sigma}^2} = 0$$
Bayesian Statistics - Estimation

The \((\mu, \sigma^2)\) that maximize the posterior distribution are maximum \textit{a posteriori} (MAP) estimates

\[
f(\mu, \sigma^2 \mid y's, \mu_0, \alpha, \nu, \kappa) = C(\sigma^2)^{-\frac{(\nu+n+1)}{2}} e^{-\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} (y_i - \mu)^2 + \alpha (\mu - \mu_0)^2 + \kappa \right]}
\]

\[
\ln(f(\mu, \sigma^2 \mid y's, \mu_0, \alpha, \nu, \kappa)) = -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} (y_i - \mu)^2 + \alpha (\mu - \mu_0)^2 + \kappa \right] - \frac{(\nu+n+1)}{2} \ln(\sigma^2) + C
\]

\[
LP = \ln(f(\mu, \sigma^2 \mid y's, \mu_0, \alpha, \nu, \kappa))
\]
Bayesian Statistics - Estimation

Maximum \textit{a posteriori} (MAP) estimates

\[
LP(\mu, \sigma^2) = -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} (y_i - \mu)^2 + (\mu - \mu_0)^2 + \kappa \right] - \frac{(\nu+n+1)}{2} \ln(\sigma^2) + C
\]

\[
\frac{\partial LP(\mu, \sigma^2)}{\partial \mu} \bigg|_{\hat{\mu}, \hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2} \left[ \sum_{i=1}^{n} 2(y_i - \hat{\mu})(-1) + 2\alpha(\hat{\mu} - \mu_0) \right] = 0
\]

\[
\frac{\partial LP(\mu, \sigma^2)}{\partial \sigma^2} \bigg|_{\hat{\mu}, \hat{\sigma}^2} = -\frac{\nu + n + 1}{2} \frac{2}{\hat{\sigma}^2} - \frac{-1}{2(\hat{\sigma}^2)} \left[ \sum_{i=1}^{n} (y_i - \hat{\mu})^2 + \alpha(\hat{\mu} - \mu_0)^2 + \kappa \right] = 0
\]
Bayesian Statistics - Estimation

Solving for $\mu$ and $\sigma^2$ yields MAP estimates

\[
\frac{\partial \text{LP}(\mu, \sigma^2)}{\partial \mu} \bigg|_{\hat{\mu}, \hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2} \left[ \sum_{i=1}^{n} 2(y_i - \hat{\mu})(-1) + 2\alpha(\hat{\mu} - \mu_0) \right] = 0
\]

\[
\frac{\partial \text{LP}(\mu, \sigma^2)}{\partial \sigma^2} \bigg|_{\hat{\mu}, \hat{\sigma}^2} = -\frac{\nu + n + 1}{2} \frac{2}{\hat{\sigma}^2} - \frac{-1}{2(\hat{\sigma}^2)^2} \left[ \sum_{i=1}^{n} (y_i - \hat{\mu})^2 + \alpha(\hat{\mu} - \mu_0)^2 + \kappa \right] = 0
\]

\[
\hat{\mu} = \frac{n}{\alpha + n} \bar{y} + \frac{\alpha}{\alpha + n} \mu_0
\]

\[
\hat{\sigma}^2 = \frac{1}{\nu + n + 1} \sum_{i=1}^{n} \left[ (y_i - \hat{\mu})^2 + \alpha(\hat{\mu} - \mu_0)^2 + \kappa \right]
\]

Can simplify with algebra
Bayesian Statistics - Additional Models

Bayesian Regression
Bayesian Time Series
Bayesian ANOVA
Bayesian Classification
Bayesian Multivariate Regression
Bayesian Image Reconstruction
Homework:

1) NONE