Applied Bayesian Modeling module 3: **Bayesian inference for 1 continuous parameter "Everything's normal"**

Leontine Alkema, lalkema@umass.edu
Fall 2022

Recap from modules 1 and 2: introduction to Bayesian inference

- In Bayesian inference, parameters are considered random variables
- We draw statistical conclusions about parameters of interest using probability statements.
- General approach for some outcome of interest θ, i.e. regression coefficient, is based on learning via Bayes' rule
  - start off with *prior* probability distribution to quantify information related to θ
  - collect data, and use Bayes' rule to update the prior into the posterior distribution
  - use posterior distribution to draw conclusions

This module:

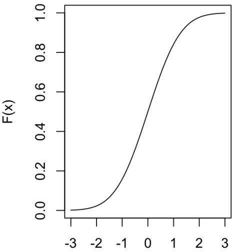
- Goal: do Bayesian inference for 1 continuous parameter
  - Set-up: estimate a mean parameter μ in a setting where data and prior are normal (to be discussed)
  - Steps: define likelihood function, define prior, obtain posterior using Bayes' rule, summarize the posterior
- To get there, to discuss:
  - Quick review of probability density functions and rules for continous outcomes (prereq, see readings on course webpage)
  - Defining notation, terminology
  - Inference using radon data

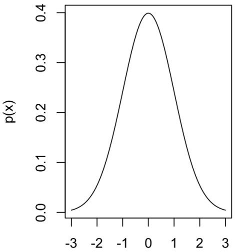
Uncountable possibilities: brief review of continuous RVs (Hoff 2.4.2)

- A probability distribution for a continuous random variable (RV) X can be defined using its cumulative distribution function (cdf) $F(x)$ and its probability density function (pdf) $p(x)$ with:

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} p(x')dx'.$$

- For continuous RV X with pdf $p(x)$:
  $0 \leq p(x)$ and $\int_{x \in \mathcal{X}} p(x)dx = 1$ where $\mathcal{X}$ is the sample space of X.





Main differences between discrete and continuous pdfs:

- $p(x)$ is NOT the probability that $X = x$,
- Sums are replaced by integrals

Brief review of pdfs for continuous RVs (ctd)

- A joint probability distribution for two continuous RVs X and Y can be defined using their joint cdf $F_{X,Y}(x,y)$ and joint pdf $p_{X,Y}(x,y)$:

  $$F_{X,Y}(x,y) = Pr(X \leq x \cap Y \leq y) = \int_{x'=-\infty}^{x} \int_{y'=-\infty}^{y} p_{X,Y}(x',y')dx'dy'.$$

  Subscripts are often left out (and will be left out in this class).
- The marginal pdf for Y can be obtained from the joint pdf: $p(y) = \int_{x' \in \mathcal{X}} p(x',y)dx'$ where $\mathcal{X}$ is the sample space of X (compare to rule of marginal probability for events).
- Conditional pdf is defined as $p(x|y) = p(x,y)/p(y)$
- From the definition of conditional pdf's, we can obtain Bayes' rule:

  $$p(x|y) = p(x,y)/p(y) = p(y|x)p(x)/p(y)$$

  (compare to Bayes' rule for events/discrete RV)

Usage of densities and steps in Bayesian inference

- Terminology regarding Bayesian inference about some parameter μ using data $\boldsymbol{y} = (y_1, \ldots, y_n)$:
  - Prior distribution $p(\mu)$: reflect knowledge about μ prior to observing data
  - Likelihood function or sampling distribution or data model or data distribution $p(\boldsymbol{y}|\mu)$: specifies the relation between data and μ, the hypothesized data generating mechanism
  - Posterior $p(\mu|\boldsymbol{y})$: prior is updated by conditioning on the data
- Steps for Bayesian inference about μ, using data $\boldsymbol{y}$:
  - Specify the likelihood function $p(\boldsymbol{y}|\mu)$.
  - Specify the prior $p(\mu)$.
  - Use Bayes' rule to obtain the posterior $p(\mu|\boldsymbol{y}) = \frac{p(\mu)p(\boldsymbol{y}|\mu)}{p(\boldsymbol{y})}$

Notation

- Notation in this course is generally aligned with BDA3 (p. 6)
  - Few exceptions for additional clarity, i.e., I aim to use **boldface** when referring to vectors or matrices.
- $p(\cdot)$ and $p(\cdot|\cdot)$ denote marginal and conditional distributions, with arguments determined by context (and subscripts left out)
- For discrete random variables or probability statements, we can also use $Pr(\cdot)$, i.e., Pr(A committed the crime).
- We use probability density and distribution exchangeably
- For standard distributions, we can use names or write out the density
  - i.e., for normal distribution:
    $\mu \sim N(m, s^2)$ is equivalent to $p(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2s^2}(\mu - m)^2\right)$

Inference using radon data

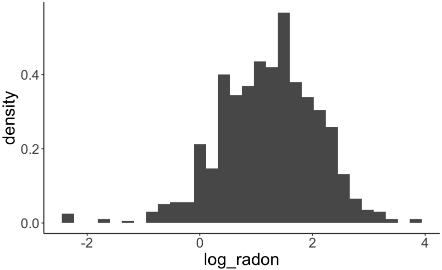
- From intro:
  - Radon is a naturally occurring radioactive gas. Its decay products are also radioactive; in high concentrations, they can cause lung cancer (several 1000 deaths/year in the USA).
  - Radon levels vary greatly across US homes.
  - Data:
    - Radon measurements in over 80K houses throughout the US (we focus on Minnesota)
    - Possible predictors: floor (basement or 1st floor) in the house, soil uranium level at county level.
  - Ultimate goal: predict radon levels for a non-sampled house in Minnesota (using a Bayesian hierarchical regression model).
- This module: estimate mean radon, assuming that all log-radon measurements are independent draws from a normal distribution.

Radon example: set up

- Let y_i denote log(radon) for house $i = 1, 2, \ldots, n$;
- We assume that all y_i are independent draws from a normal distribution; we can write this in different ways:

$$\begin{aligned} y_i|\mu,\sigma^2 &\sim N(\mu,\sigma^2), &(1)\\ p(y_i|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(\mu-y_i)^2\right). &(2)\end{aligned}$$

- Goal: estimate mean log-radon level μ, assume that σ^2 is known.



Bayesian inference about μ

- Bayesian inference about μ, using data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$:
  1. Specify the likelihood function $p(\boldsymbol{y}|\mu)$.
  2. Specify the prior $p(\mu)$.
  3. Use Bayes' rule to obtain the posterior $p(\mu|\boldsymbol{y}) = \frac{p(\mu)p(\boldsymbol{y}|\mu)}{p(\boldsymbol{y})}$

(1) Likelihood function: If $y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$ (independent), then

$$p(\boldsymbol{y}|\mu, \sigma^2) = \prod_{i=1}^{n} p(y_i|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2\sigma^2}(y_i - \mu)^2\right).$$

(2) We assume a normal prior on μ: $\mu \sim N(m_0, s_{\mu 0}^2)$,
where m_0 and $s_{\mu 0}$ refer to prior mean and standard deviation.

(3) It turns out that for this combination of prior and likelihood, with σ known, the posterior for μ is normal:

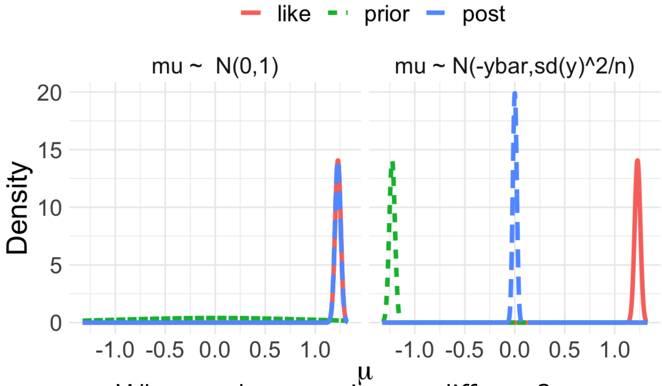
$$\mu|\boldsymbol{y}, \sigma^2 \sim N\left(\frac{m_0/s_{\mu 0}^2 + n \cdot \bar{y}/\sigma^2}{1/s_{\mu 0}^2 + n/\sigma^2}, \frac{1}{1/\sigma_{\mu 0}^2 + n/\sigma^2}\right).$$

- Details in next module, current focus is on seeing how posterior depends on prior and data, and what to do with it.

Bayesian inference for μ: Examples with different priors

$y_i|\mu,\sigma^2 \sim N(\mu,\sigma^2)$; $\mu \sim N(m_0, s^2_{\mu 0})$; $\mu|\boldsymbol{y},\sigma^2 \sim N\left(\frac{m_0/s^2_{\mu 0}+n\cdot\bar{y}/\sigma^2}{1/s^2_{\mu 0}+n/\sigma^2}, \frac{1}{1/s^2_{\mu 0}+n/\sigma^2}\right)$

- Use radon data and set $\sigma = s\{y\} = 0.86$, the st.dev. of the y_i's
- Results based on 2 different priors:
  $\mu \sim N(0,1)$ [LEFT] and $\mu \sim N(-\bar{y}, s\{y\}^2/n)$ [RIGHT]



Why are the posteriors so different?

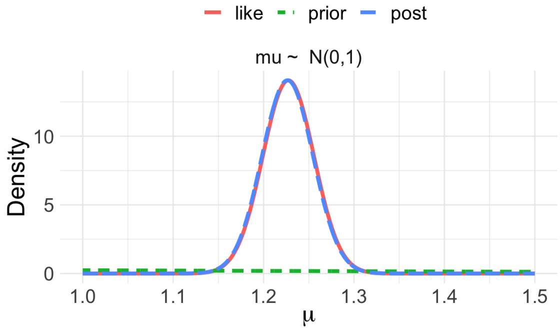
Conclusion?

- The posterior is a compromise between the likelihood and the prior
- This is to be expected, based on Bayes' rule $p(\mu|\boldsymbol{y}) = \frac{p(\mu)p(\boldsymbol{y}|\mu)}{p(\boldsymbol{y})}$,
- More details on this in the next module

We got the posterior, now what?

Inference based on posterior distribution

- In Bayesian inference, we use the posterior $p(\mu|\boldsymbol{y})$ to provide summaries of interest.
- Bayesian point estimates are often given by:
  - the posterior mean $E(\mu|\boldsymbol{y})$
  - or the posterior median μ^* with $P(\mu < \mu^*|\boldsymbol{y}) = 0.5$.



- Here $E(\mu|\boldsymbol{y}) =$ median $= 1.23$.

We got the posterior, now what? (ctd)

Inference based on posterior distribution

- Uncertainty can be quantified with credible intervals (CIs), definition (using 95% as an example) is as follows:
  - An interval is called a 95% Bayesian CI if the posterior probability that μ is contained in the interval is 0.95.
  - More formally, $(l(\boldsymbol{y}), u(\boldsymbol{y}))$ is called a 95% Bayesian CI if $P(l(\boldsymbol{y}) < \mu < u(\boldsymbol{y})|\boldsymbol{y}) = 0.95$.
  - This interpretation differs from a frequentist CI; it is a probability statement about the information about the location of μ.
- Interval options:
  - Quantile-based $100 \cdot (1-\alpha)$% CI is given by posterior quantiles $(\mu_{\alpha/2}, \mu_{1-\alpha/2})$, with $P(\mu < \mu_{\alpha/2}|\boldsymbol{y}) = P(\mu > \mu_{1-\alpha/2}|\boldsymbol{y}) = \alpha/2$.
  - Highest posterior density (HPD) intervals
- For the radon example with $\mu \sim N(0, 1)$, the quantile-based 95% CI is (1.17, 1.28).

Summary

- Bayesian inference about a parameter μ, using data $\boldsymbol{y}$:
  (1) relate data $\boldsymbol{y}$ to μ through a likelihood function $p(\boldsymbol{y}|\mu)$
  (2) set a prior distribution for μ, $p(\mu)$
  (3) use Bayes' rule to update the prior into the posterior distribution:

$$p(\mu|\boldsymbol{y}) = \frac{p(\mu)p(\boldsymbol{y}|\mu)}{p(\boldsymbol{y})},$$

  (4) use the posterior $p(\mu|\boldsymbol{y})$ to provide summaries of interest, e.g. point estimates and uncertainty intervals, called credible intervals (CIs).
- Model set-up in this module: everything's normal;
  When data and prior are normal, the posterior is normal too.
- Next module: derive the posterior using Bayes' rule, discuss the role of prior information, bias-variance trade-off.