Simplest Possible Guide to Belief Updating

Paul Johnson[1]

Center for Research Methods and Data Analysis

2018

KU

Outline

1. Why consider the Bayesian Way?

2. Bayesian Belief Updating
   - Two state model
   - How Beliefs are Updated
   - Visualize Updating Process

3. Updating in Statistics

4. Simulation: MCMC

Outline

This is Life

- I believe
- I observe
- I revise my belief

That's what I thought statistics would be about, but it wasn't

Instead, textbook "frequentist" statistics taught me

- The truth is 0
- The data says the truth is "statistically significantly different" from 0
- Therefore 0 is not the truth
- I can say only that 0 was wrong, not that anything else is correct
- "Asymptotically, ...". Many of the "more advanced" models (maximum likelihood estimates) are delivered with the warning that the standard errors and hypothesis tests are correct only for infinitely large sample sizes
    - Giggle about your dissertation project with 150 patients.

KU

Not informative, also not flexible

- The "usual" playbook is straight-jacketed with assumptions. This is $N(\mu, \sigma^2)$, that's χ^2 .
- Consider hierarchical linear random effect model for individual i in a geographical unit j

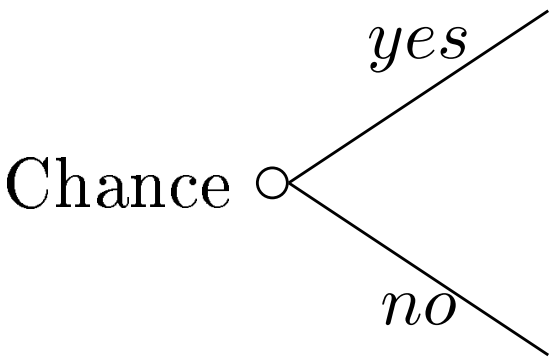
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + u_j + \varepsilon_i$$

  - We can do that if u_j is normally distributed across units and ε_i is normally distributed among individuals.
  - Otherwise not.

KU

Outline

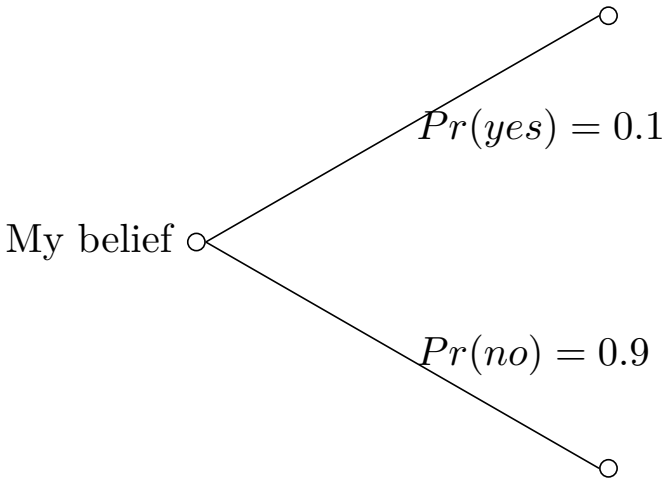
Uncertainty and the "State of Nature"



- Does Jennifer like me?
- Are we pregnant?
- Do I have cancer?
- Does the President use fake tan?

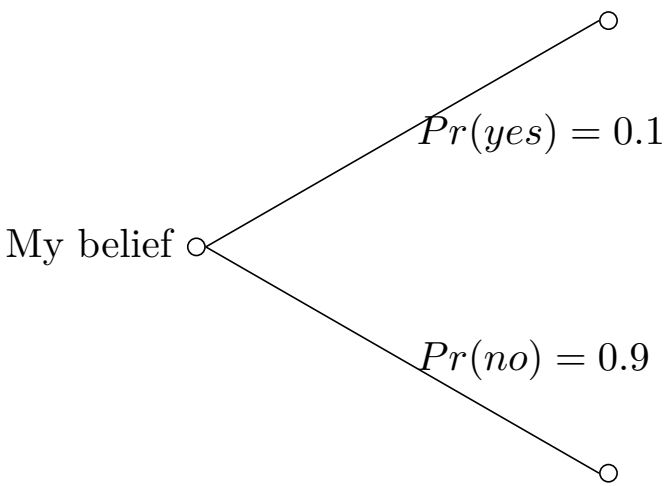
Those are my "beliefs"

Nature's probabilities are unknown. Let's refer to a belief.



- $Pr(yes)$ is the probability that *yes* is the "true state of nature" (in my humble opinion)
- It is my *subjective* belief, not a known fact (if such things exist)

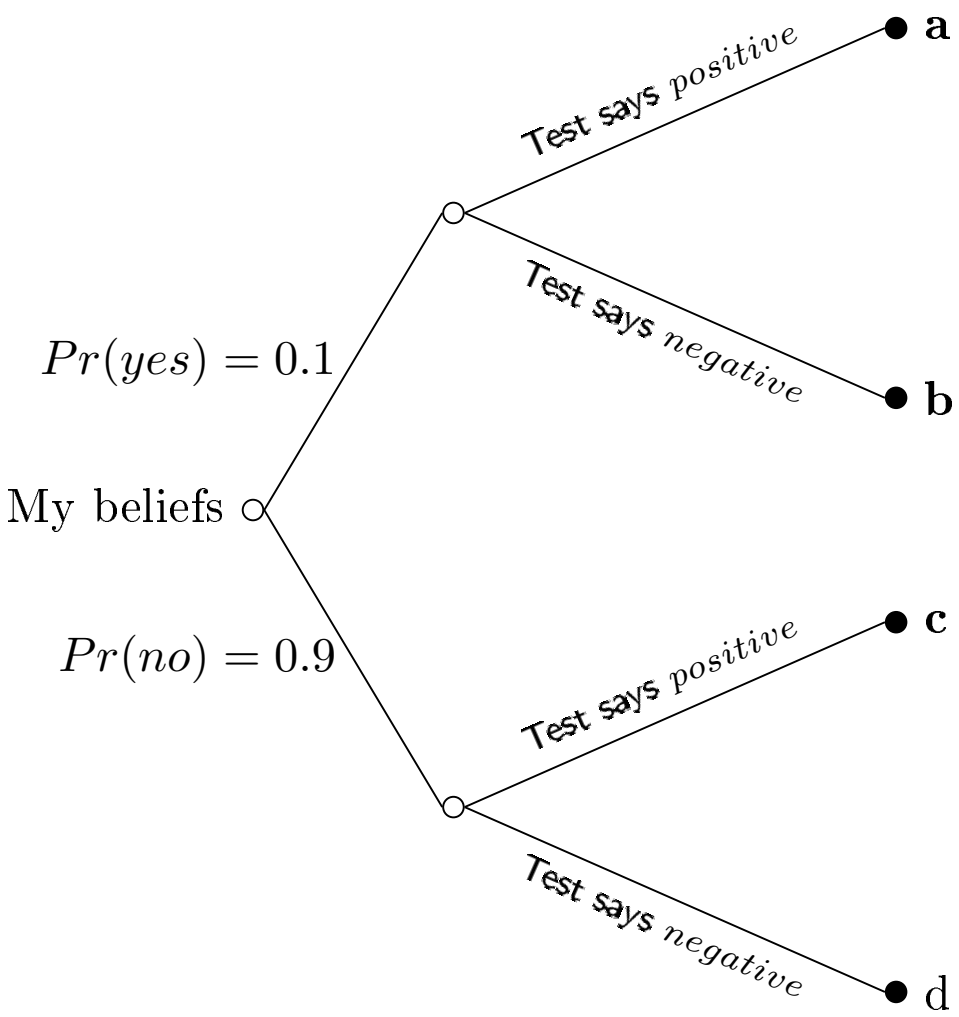
Those are my "beliefs"



My beliefs must be mathematically coherent:

- $Pr(yes) \geq 0$
- $Pr(no) \geq 0$
- $Pr(yes) + Pr(no) = 1.0$ (The probabilities sum to 1.0)

A Diagnostic Test is applied

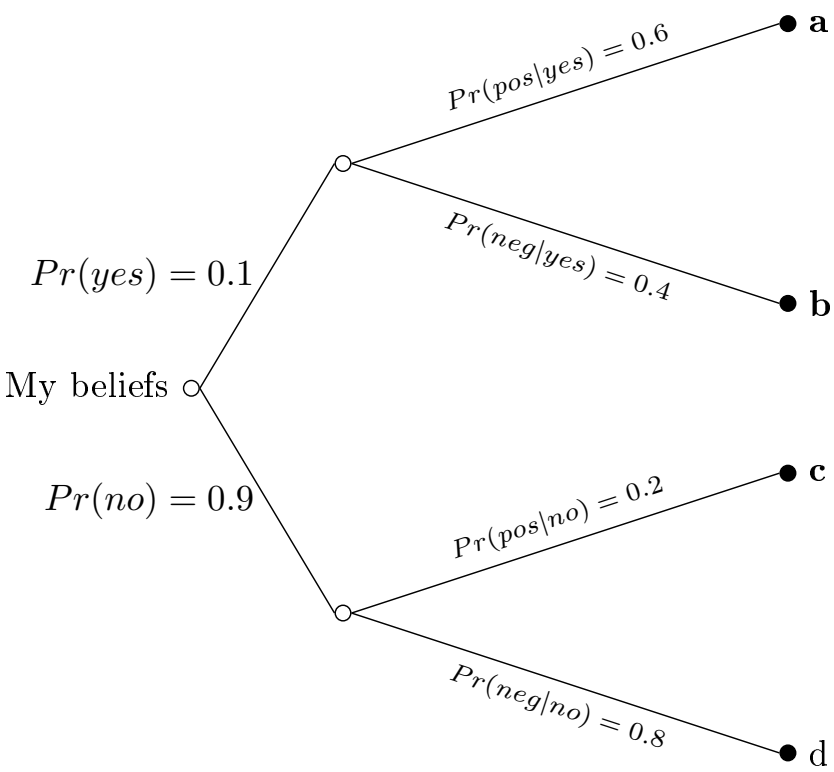


Here "*positive*" means the test says "*yes*".

- Ask Jennifer to go to movies. Does she say yes? "*positive*".
- Pee on a stick. Got +? "*positive*".
- Does orange stain rub off on a handkerchief? "*positive*".

Note I abbreviate labels "*pos*" and "*neg*" in the following.

Probabilities for the diagnostic test

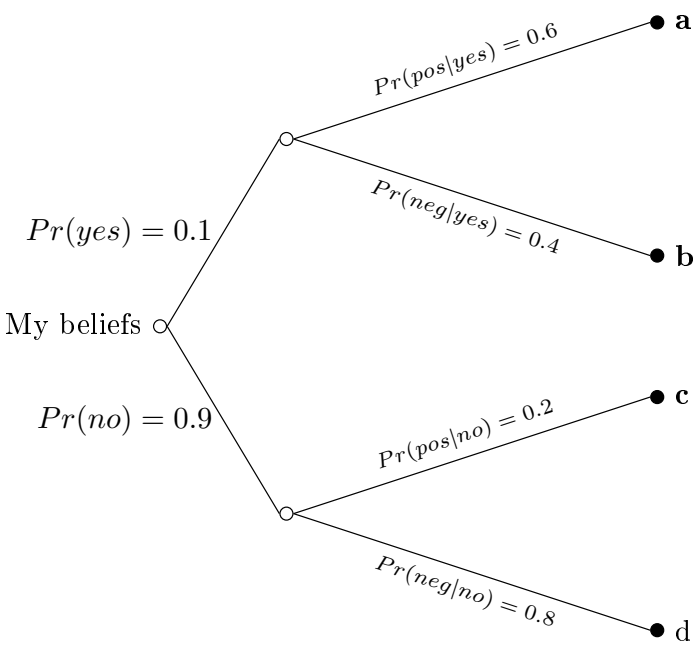


This is "conditional probability" notation.

The chance of one outcome, *given that the previous has happened*.

- $Pr(pos|yes)$ Truth is "yes" and test says "positive". Test is "accurate".
- $Pr(neg|no)$ Truth is "no" and test says "negative".
- $Pr(neg|yes)$ Truth is "yes", but test says "no". **Jargon**: False negative.
- $Pr(pos|no)$ Truth is "no", but test says "yes". **Jargon**: False positive.

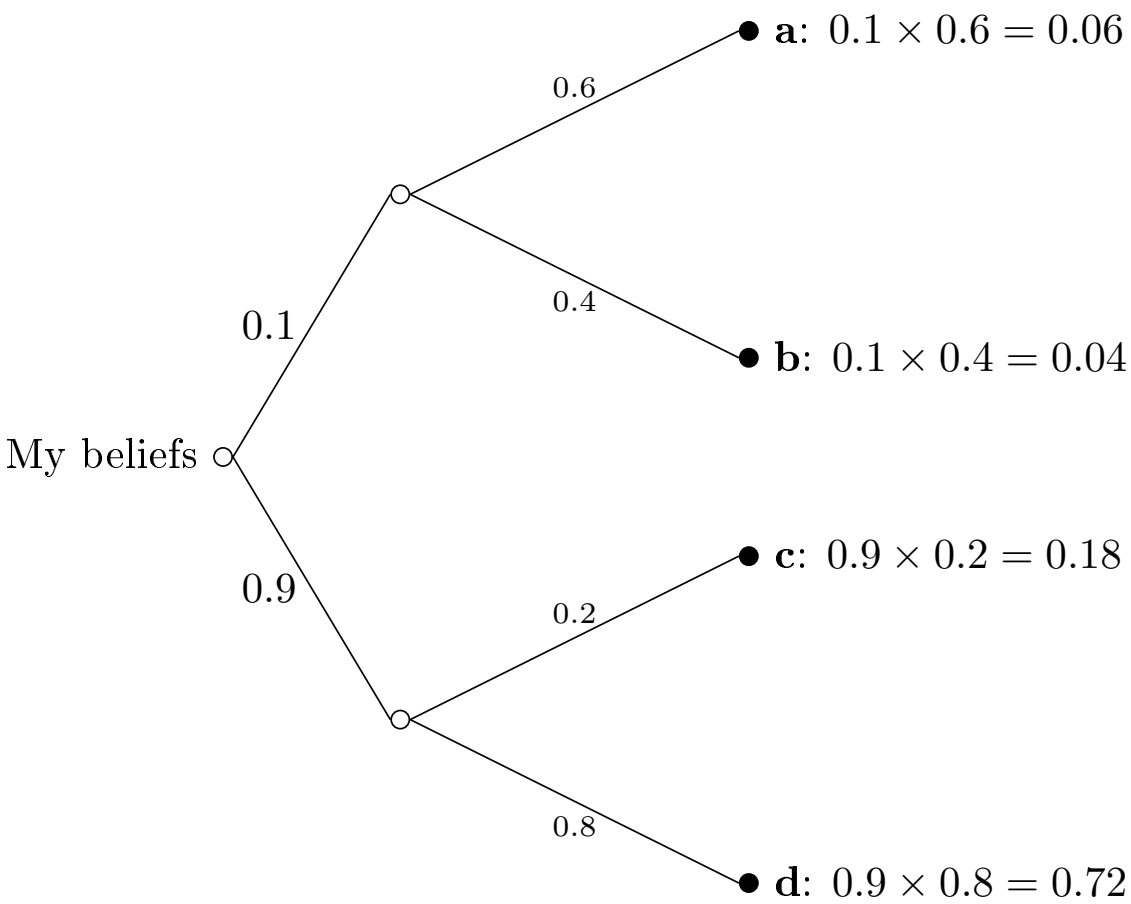
The Test is not Perfectly Accurate



The key part of this story. The test is not perfect

- Ask Jennifer to movies. She says *yes*! Does that mean she likes me?
  - Maybe not. *Maybe she just needs a ride to the theater.*
  - The chance of "pos" is $Pr(pos|yes) + Pr(pos|no)$
- She declines. Does that means she does not like me?
  - Maybe not!
  - Perhaps her parents say "No, you can't go with him"
  - Perhaps, she *really* does need to shampoo on Saturday.

Total probability of ending up at each of the 4 outcomes



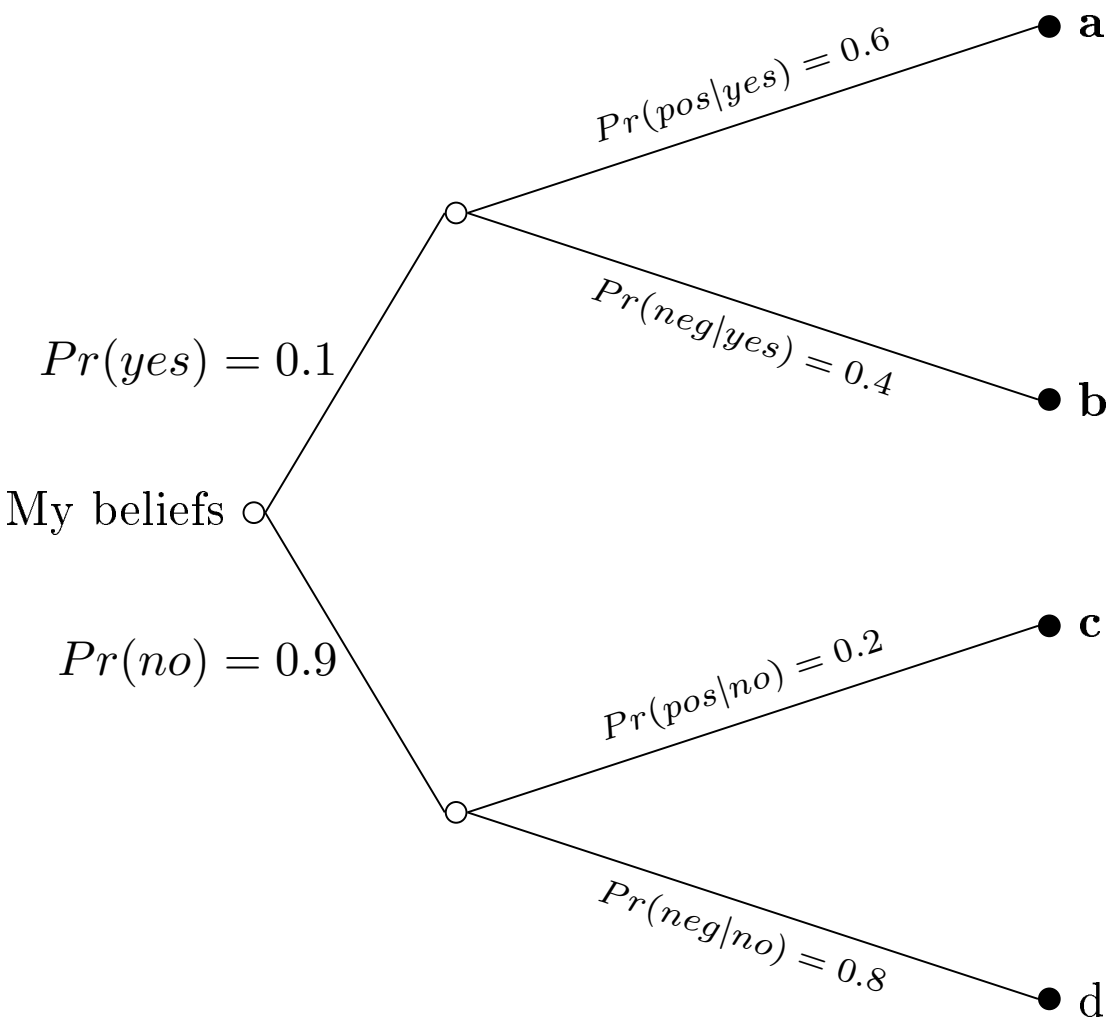
This is the "multiplication" property of probabilities:

Multiplication rule

The chance of ending up at a point is the product of probabilities for the branches leading to it.

- If we end up at node **a**, then the sequence was "yes" and "pos", $0.1 \times 0.6 = 0.06$
- Note $Pr(a) + Pr(b) + Pr(c) + Pr(d) = 1.0$

If "pos", we might be at either a and c?



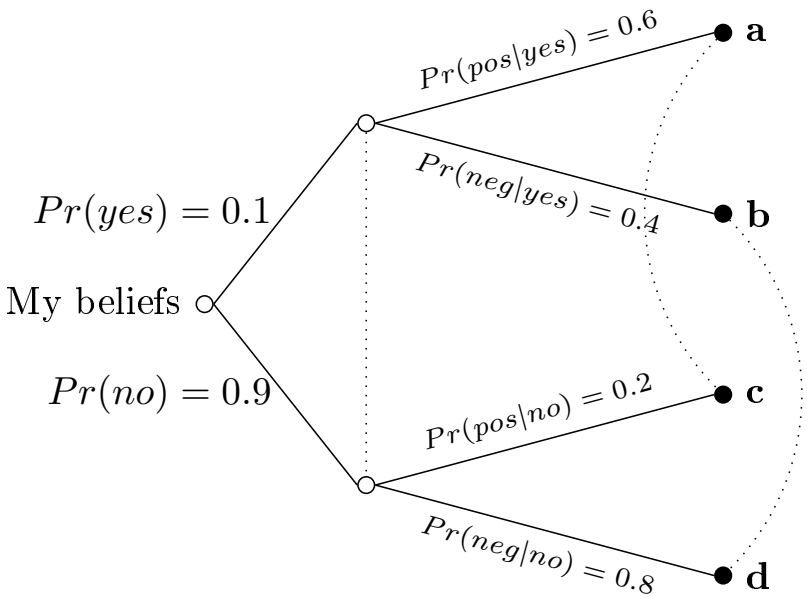
The test said "positive"

- If you end up at **a**, the truth is "yes"! ☺

If you end up at **c**, the truth is "no" ☹

Information is Limited

We don't directly observe yes or no at stage 1.

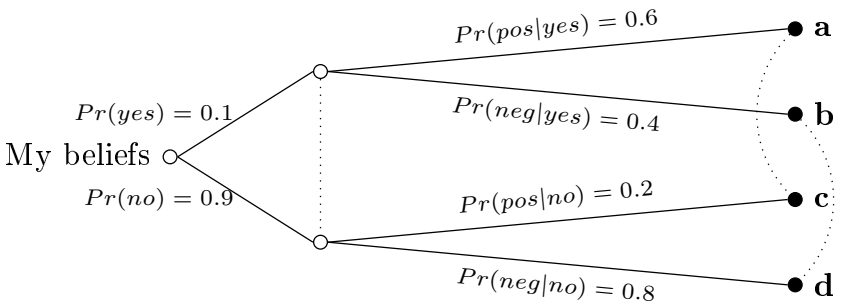


Information Sets.

The dotted lines are called "information sets". They group together nodes among which we cannot distinguish.

- If test says "positive", we know the outcome is either **a** or **c**
- If test says "negative", we know the outcome is either **b** or **d**

Impact of information revelation



- Multiplication property: with no additional information, the chance of **a** is just the proportion of the time **a** happens

$$\frac{Pr(a)}{Pr(a) + Pr(b) + Pr(c) + Pr(d)} = \frac{Pr(yes) \cdot Pr(pos|yes)}{1} = 0.06$$

- New information: test is "pos", then I can exclude **b** and **d** from that denominator

$$\frac{Pr(a)}{Pr(a) + Pr(c)} = \frac{Pr(yes) \cdot Pr(pos|yes)}{Pr(yes) \cdot Pr(pos|yes) + Pr(no) \cdot Pr(pos|no)}$$

- In this case, "Belief Updating" means excluding outcomes from the denominator!

Interpret each piece

- Can simplify the denominator. Note that denominator is same as the chance of a positive test, $Pr(pos)$:

$$Pr(pos) = Pr(yes) \cdot Pr(pos|yes) + Pr(no) \cdot Pr(pos|no)$$

- So rewrite right hand side as

$$\frac{Pr(a)}{Pr(a) + Pr(c)} = \frac{Pr(yes) \cdot Pr(pos|yes)}{Pr(pos)}$$

- Now Revise left hand side. It is a "**posterior belief**": the probability of yes given $positive$ test result.

$$Pr(yes|pos) = \frac{Pr(a)}{Pr(a) + Pr(c)}$$

Interpret each piece ...

- Putting left and right sides together, we have Bayes's Rule for updating beliefs:

$$Pr(yes|pos) = \frac{Pr(yes) \cdot Pr(pos|yes)}{Pr(pos)}$$

$$Pr(yes|pos) = \frac{prior\,probability\,of\,yes \times probability\,of\,positive\,if\,yes}{probability\,of\,positive\,test}$$

Insert the numbers to make this more tangible

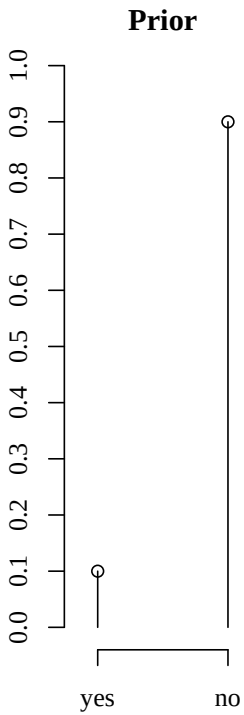
$$Pr(yes|pos) = \frac{Pr(yes) \cdot Pr(pos|yes)}{Pr(pos)} = \frac{0.1 \cdot 0.6}{0.1 \cdot 0.6 + 0.9 \cdot 0.2} = \frac{0.06}{0.06 + 0.18} = 0.25$$

Similarly, posterior chance of no given a positive test is

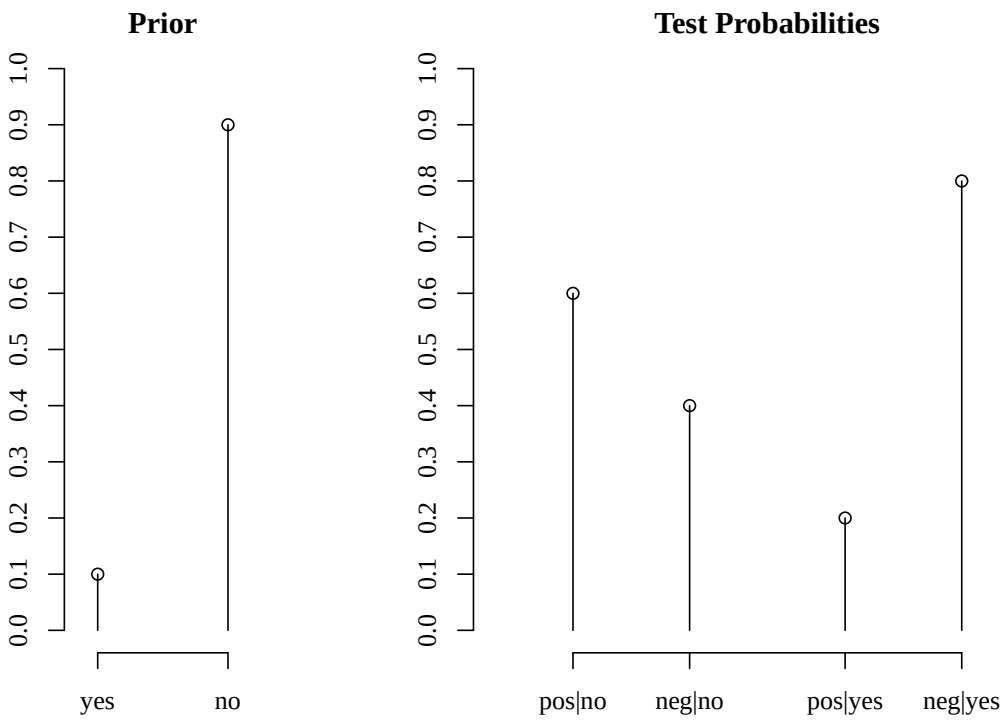
$$Pr(no|pos) = \frac{Pr(no) \cdot Pr(pos|no)}{Pr(pos)} = \frac{0.9 \cdot 0.2}{0.1 \cdot 0.6 + 0.9 \cdot 0.2} = 0.75$$

Note that the posterior beliefs sum to 1.0

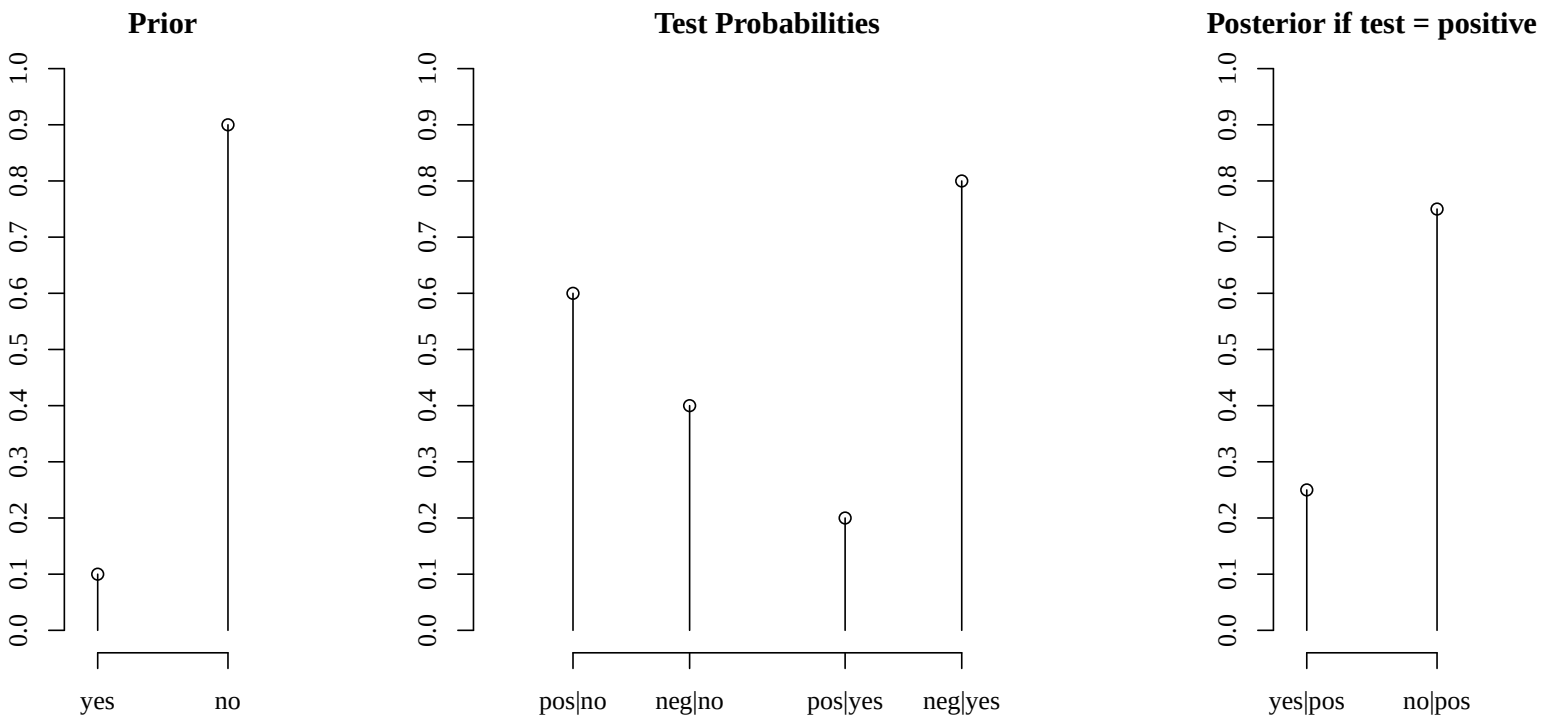
The Prior, likelihood, and posterior graphs



The Prior, likelihood, and posterior graphs



The Prior, likelihood, and posterior graphs



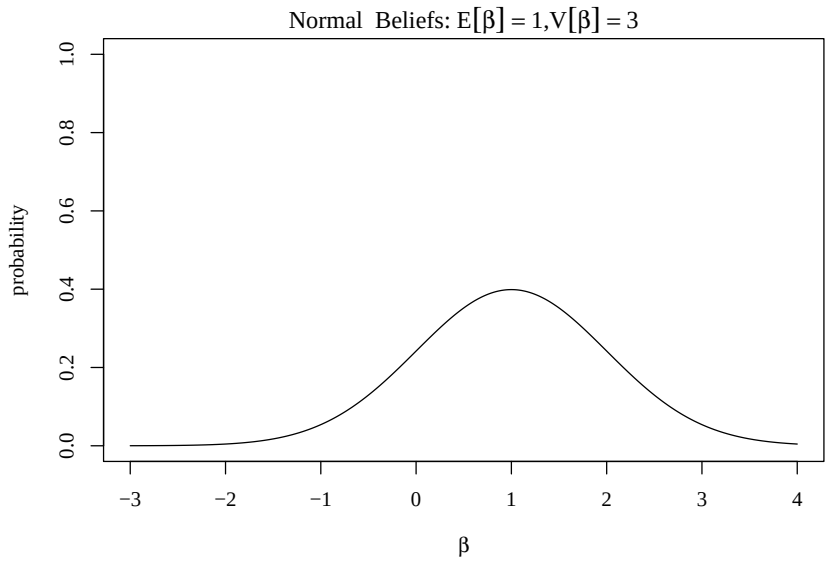
Summary of this simple example

- We began with beliefs, probabilities assigned on a discrete space, {"yes", "no"}
- We drew just 1 data point, {"pos", "neg"}
- The updated probabilities were simple to calculate! We know $Pr(yes|pos)$, $Pr(no|pos)$

Outline

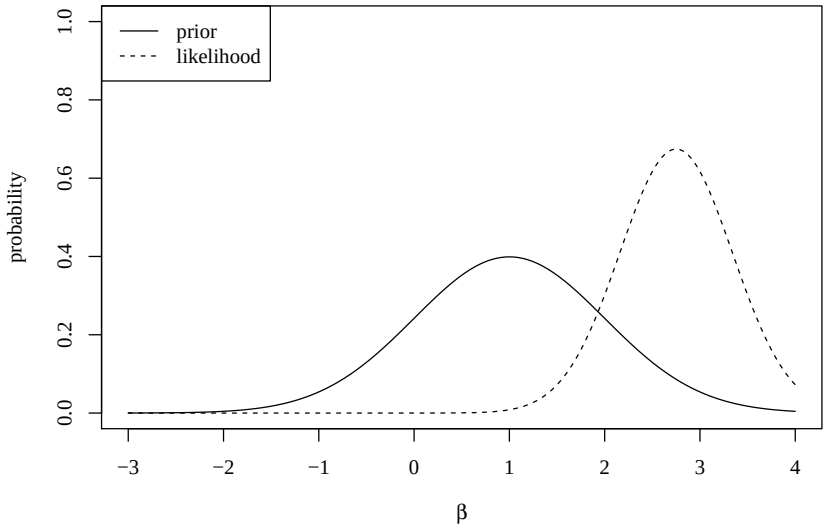
Statistical Problems: Prior is Usually a Continuum

- The prior space is usually a continuum (not on a discrete list)
- $p(\beta)$ is a probability density function



The Likelihood Function is continuous

- The likelihood is the chance that the data would be observed, if the truth is β
- In maximum likelihood analysis, we are only interested in the maximum, $\hat{\beta} = 2.75$



- A posterior probability $p(\beta|data)$ requires $p(data|\beta)$ for all values of β (across a continuum of possible β).

Bayesian updating

- What is the probability of each possible β?

$$p(\beta|data) = \frac{p_{prior}(\beta)p_{likelihood}(data|\beta)}{p_{marginal}(data)}$$

- The denominator was referred to as the "data marginal probability", Gelman et al BDA3 suggest "prior predictive distribution". The chance of observing each set of N data points.
- To calculate a denominator, we have to calculate

$$p_{marginal}(data) = \int prob(data)p(data|\beta)d\beta$$

Forget the denominator

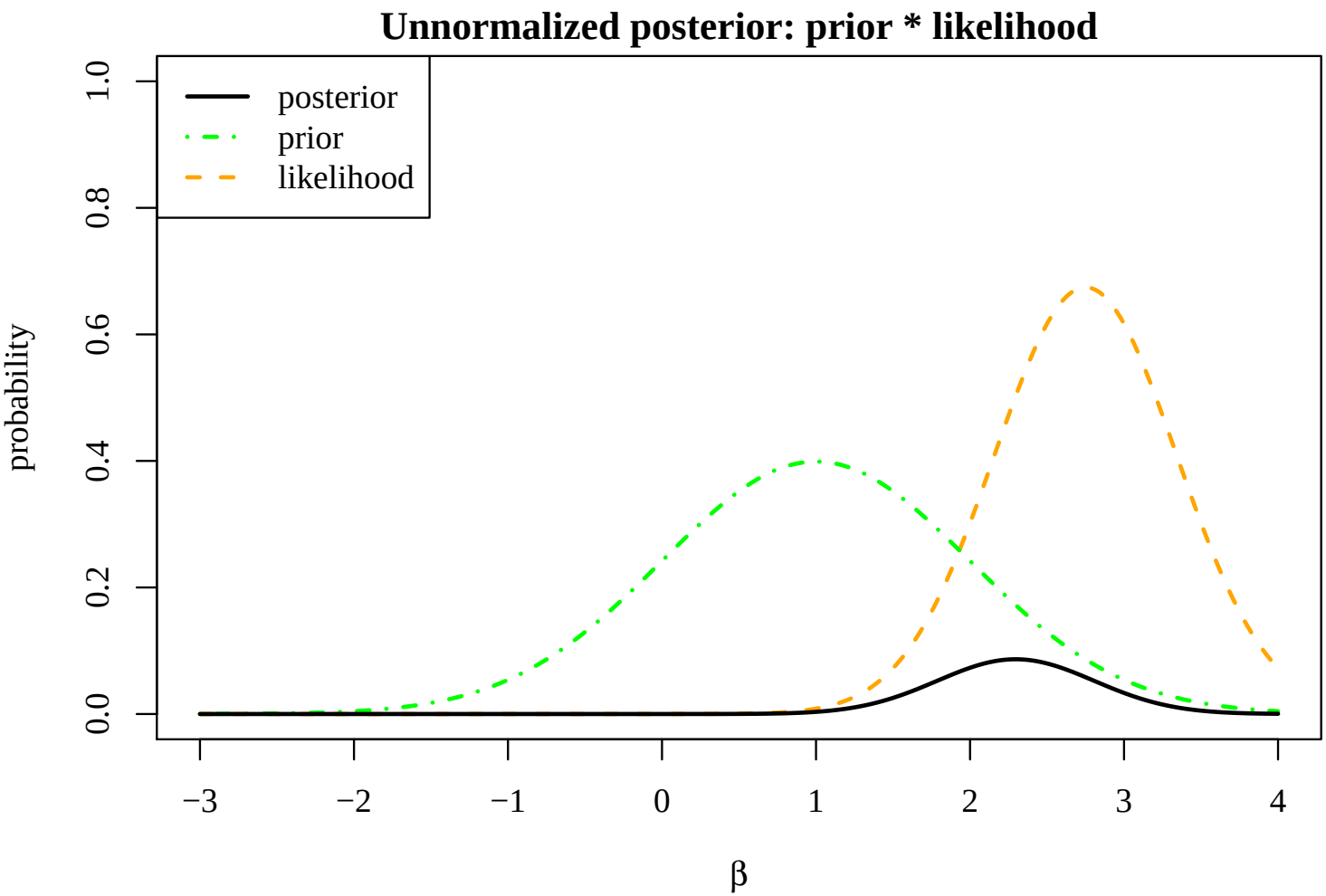
$$p(\beta|data) = \frac{p_{prior}(\beta)p_{likelhood}(data|\beta)}{p_{marginal}(data)}$$

- After data is gathered, $p_{marginal}(data)$ is a "fixed" "unknown" value. It is not relevant to finding the shape of the posterior
- So omit the denominator and think of the posterior as

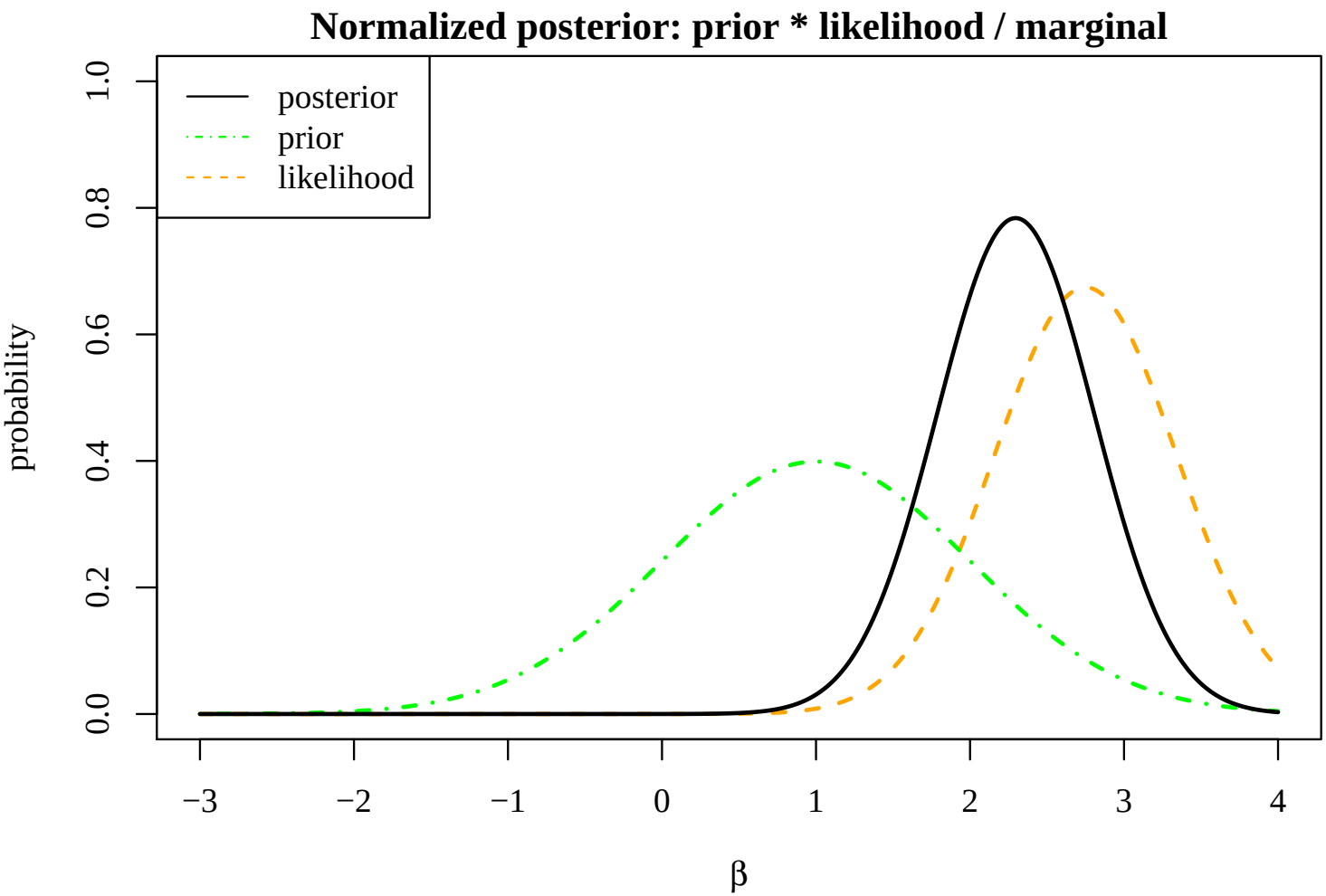
$$p(\beta|data) \propto p_{prior}(\beta) \times p_{likelhood}(data|\beta)$$

  $\propto$ means "is proportional to".
- $p(\beta|data)$ is not a true probability density, it is "unnormalized"

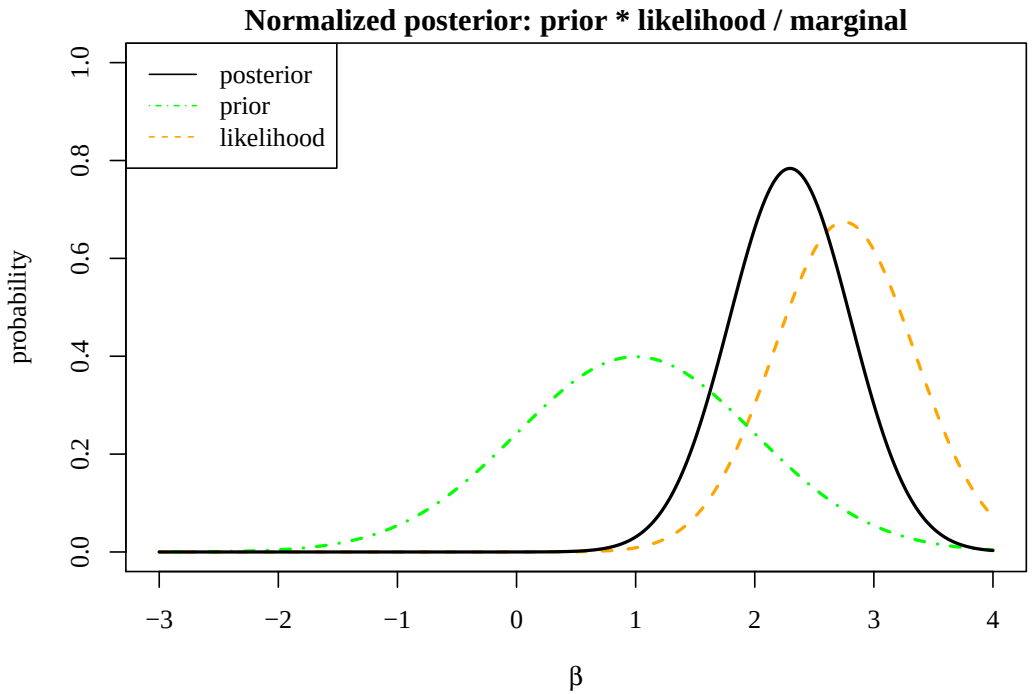
Calculating the Posterior: prior×likelihood (unnormalized)



Calculating the Posterior: prior×likelihood (normalized)

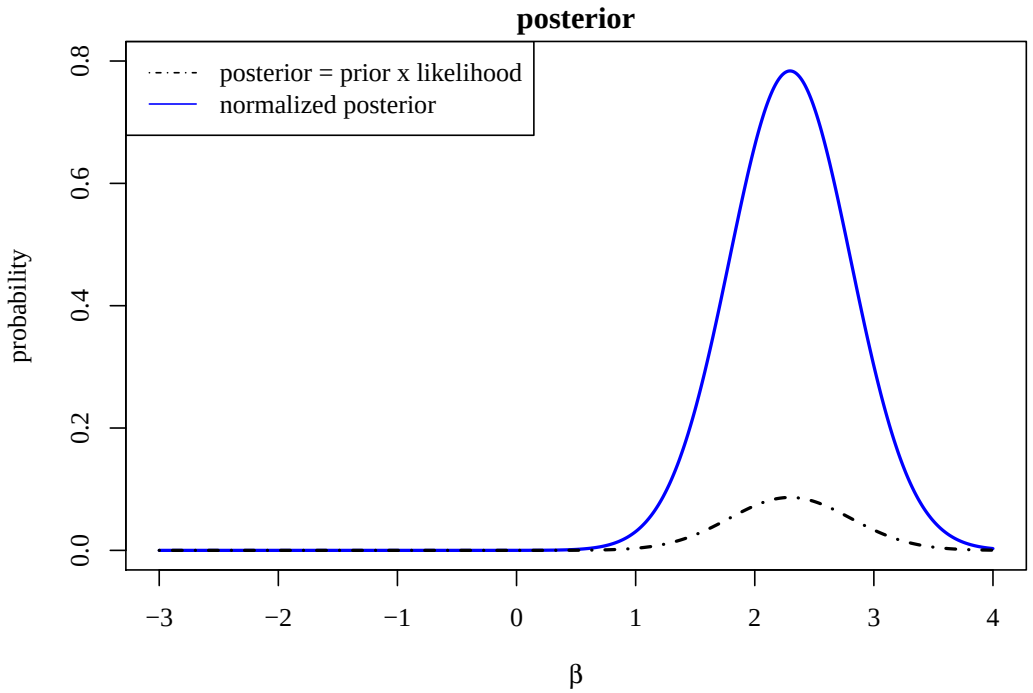


Calculating the Posterior



- Posterior is product of $p(\beta)p(data|\beta)$, for each β
- Analysts might summarize that by:
  - "most likely value", the mode of the posterior
  - "expected value", the mean of the posterior
  - "range of likely values", aka "HPD", say 95% range

Normalization is not usually necessary



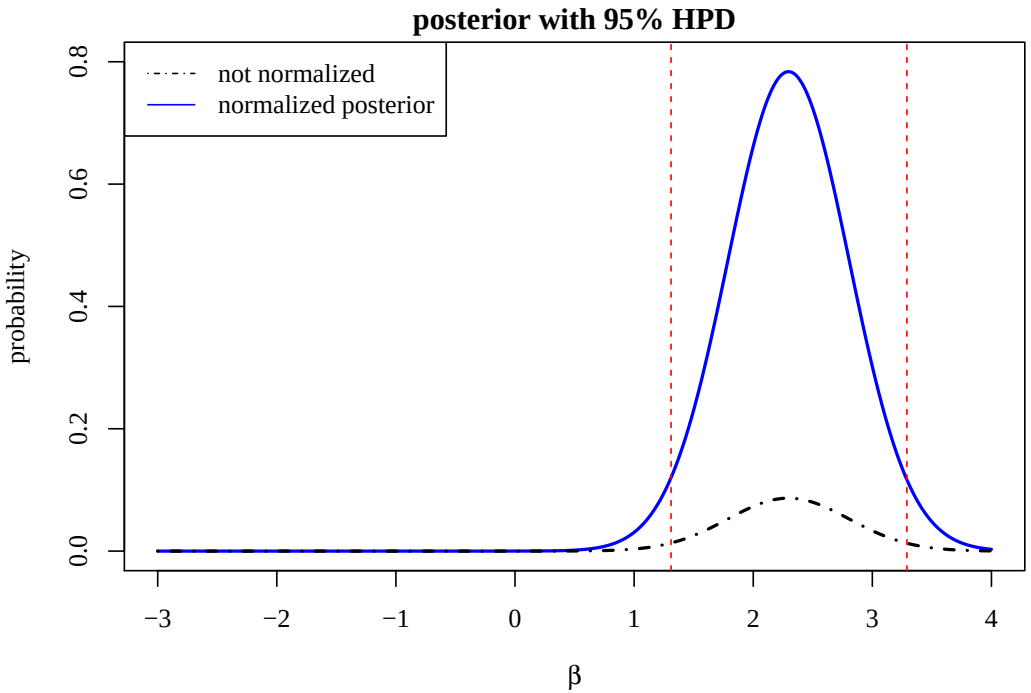
The Normalized and Unnormalized posteriors give the same information

- The mode of β is the same, 2.3
- The "HPD", the middle 95%, range. *I think* it is (1.31, 3.29)

Need normalized posterior to calculate expected value (area under the curve must equal 1.0)

Mode and HPD are same in Normalized/Unnormalized

Demonstrate 95% HPD in the previous graph

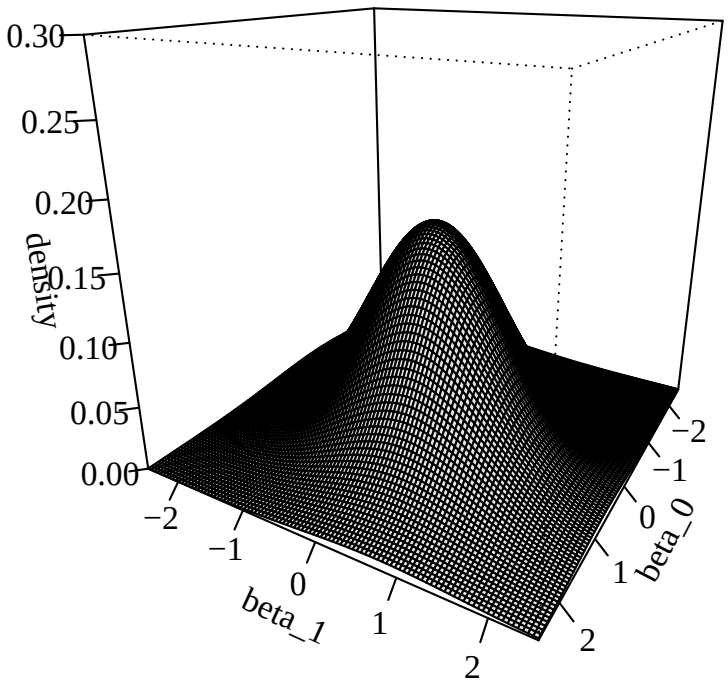


Calculating the Posterior

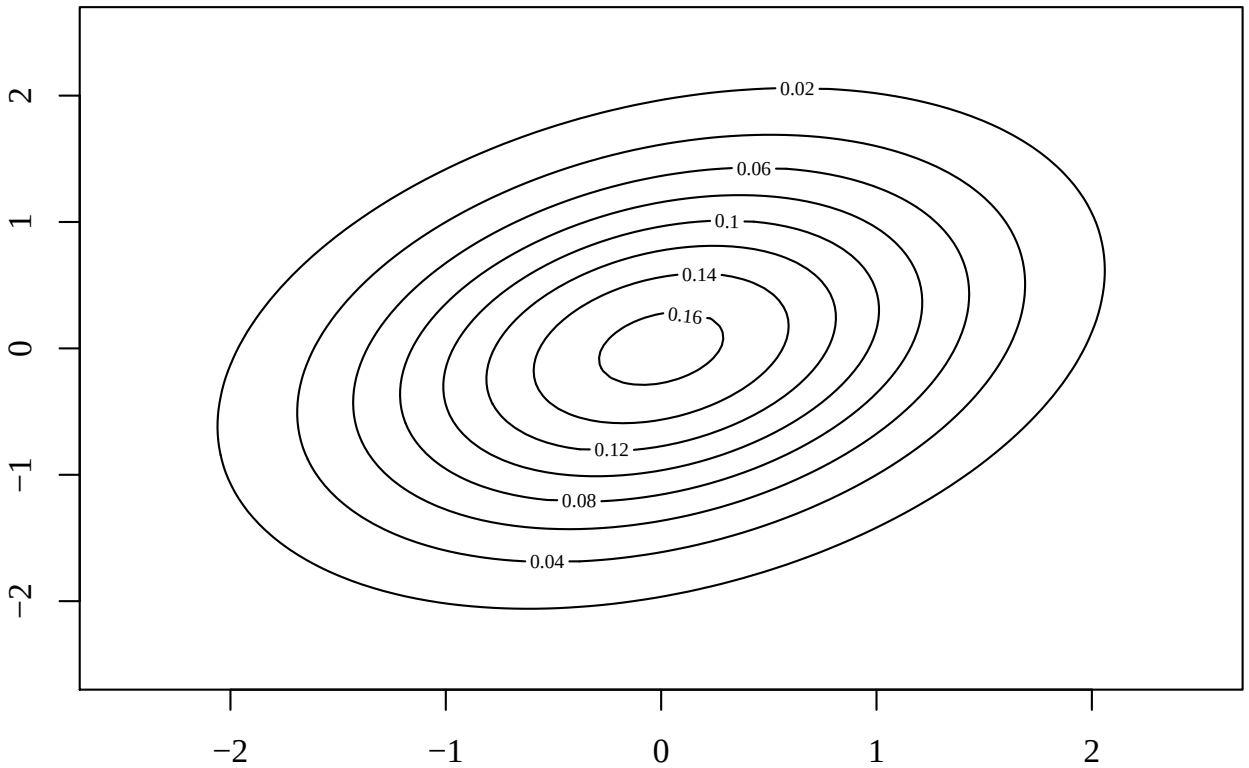
- Deriving $p(\beta|data)$ requires challenging derivations
- In pen & paper math, emphasis is on clever choice of functions
  - "conjugate priors": prior is chosen so that a prior and posterior have same formula
  - e.g., if prior model is $\beta \sim N(\mu, \sigma^2)$, then the posterior will also be normal, $\beta^{post} \sim N(\mu_{post}, \sigma^2_{post})$ Tractable math guides us updating $\mu \rightarrow \mu_{post}$ and $\sigma^2 \rightarrow \sigma^2_{post}$.
- Many problems are not tractable in that way.

KU

Beliefs in 2 dimensions



Beliefs in 2 dimensions ...



Imagine the math needed

- The 2-D belief requires
- A 2-D likelihood function
- Multiply $prior \times likelihood$ to get posterior across the whole 2-D continuum
- If problem has more coefficients, lots of data, it is not practical to approach the calculation in that way
- The "curse of dimensionality". Can't approximate continuum by discrete approximation in an m-dimensional parameter space

KU

Computer Simulation to Approximate the Posterior

- Possible now to "simulate" the posterior distribution (without actually "solving")
  - hardware breakthroughs (PCs, multi-core CPU, etc)
  - math/theory development since 1950 (Manhattan project and after)

Outline

Chi Feng created a Javascript program that allows us to explore/illustrate in a web browser

http://crmda.dept.ku.edu/StatsCamp2018/bayes/mcmc-demo/index.html

Metropolis-Hastings

- Idea originated in Los Alamos Labs (Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines. J. Chem. Phys., 21, 1087-1091.)
- Explore the posterior by making jumps
  - Look for improvements by "random walk"
    - accept proposals that "go down the hill" sometimes in order to explore

Standard MH with MV Normal posterior

- Random walk spends most of time wandering
- Converges in distribution to actual probability model after a while ("burnin phase")
- Draw samples from repeatedly that to form a posterior

Metropolis-Hastings Concerns

- "wasted" calculations
- If "parameter space" is not "round"–correlated parameters-then sampling is even less efficient
- banana-shaped posterior is bad scenario
  MH with banana shaped posterior

Competing Algorithms

- Gibbs sampling developed in 1990s
  - Basis of "BUGS" project (Bayesian Updating with Gibbs Sampling), WinBUGS, OpenBUGS
  - JAGS (Plummer) "Just another Gibbs Sampler"
- Hamiltonian No U-Turn Sampler (NUTS): Gelman & Columbia-based software Stan
- These are "free standing" programs and we can also interact with them via other software, such as R R Core Team (2017) or Python

Technical details

Lots of effort focused on finding out if

1. The Markov Chain has "converged" to a stable probability pattern

References

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Session

```
sessionInfo()
```

```
R version 3.5.1 (2018-07-02)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.10

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.8.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.8.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
     LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8
     LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8
    LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] mvtnorm_1.0-8
```

Session ...

```
20
loaded via a namespace (and not attached):
[1] compiler_3.5.1 tools_3.5.1
```