# Bayesian Ingredients:
# A brief introduction

Hedibert Freitas Lopes[1]
hedibert.org

[1]Professor of Statistics and Econometrics at Insper, São Paulo.

# Outline

# Bayesian paradigm: an overview

- Combination of different sources/levels of information
- Sequential update of beliefs
- A single, coherent framework for
  - Statistical inference/learning
  - Model comparison/selection/criticism
  - Predictive analysis and decision making
- Drawback: Computationally challenging

# Example 1: Is Diego ill?

- Diego claims some discomfort and goes to his doctor.
- His doctor believes he might be ill (he may have the flu).
- $\theta = 1$: Diego is ill.
- $\theta = 0$: Diego is not ill.
- $\theta$ is the "state of nature" or "proposition"

# Adding some modeling

The doctor can take a binary and imperfect "test" $X$ in order to learn about $\theta$:

$$\begin{cases} P(X=1|\theta=0)=0.40, & \text{false positive} \\ P(X=0|\theta=1)=0.05, & \text{false negative} \end{cases}$$

These numbers might be based, say, on observed frequencies over the years and over several hospital in a given region.

# $X = 1$ is observed

Data collection

The doctor performs the test and observes $X = 1$.

# $X = 1$ is observed

Data collection

The doctor performs the test and observes $X = 1$.

Decision making

How should the doctor proceed?

# $X = 1$ is observed

**Data collection**

The doctor performs the test and observes $X = 1$.

**Decision making**

How should the doctor proceed?

**Maximum likelihood argument**

$X = 1$ is more likely from a ill patient than from a healthy one

$$\frac{P(X = 1|\theta = 1)}{P(X = 1|\theta = 0)} = \frac{0.95}{0.40} = 2.375$$

## $X = 1$ is observed

**Data collection**

The doctor performs the test and observes $X = 1$.

**Decision making**

How should the doctor proceed?

**Maximum likelihood argument**

$X = 1$ is more likely from a ill patient than from a healthy one

$$\frac{P(X = 1|\theta = 1)}{P(X = 1|\theta = 0)} = \frac{0.95}{0.40} = 2.375$$

The maximum likelihood estimator of $\theta$ is $\hat{\theta}_{MLE} = 1$.

# Bayesian learning

Suppose the doctor claims that

$$P(\theta = 1) = 0.70$$

# Bayesian learning

Suppose the doctor claims that

$$P(\theta = 1) = 0.70$$

This information can be based on the doctor's sole experience or based on existing health department summaries or any other piece of existing historical information.

# Bayesian learning

Suppose the doctor claims that

$$P(\theta = 1) = 0.70$$

This information can be based on the doctor's sole experience or based on existing health department summaries or any other piece of existing historical information.

Overall rate of positives

The doctor can anticipate the overall rate of positive tests:

$$\begin{aligned} P(X = 1) &= P(X = 1|\theta = 0)P(\theta = 0) \\ &+ P(X = 1|\theta = 1)P(\theta = 1) \\ &= (0.4)(0.3) + (0.95)(0.7) = 0.785 \end{aligned}$$

# Turning the Bayesian crank

Once $X = 1$ is observed, i.e. once Diego is submitted to the test $X$ and the outcome is $X = 1$, what is the probability that Diego is ill?

# Turning the Bayesian crank

Once $X = 1$ is observed, i.e. once Diego is submitted to the test $X$ and the outcome is $X = 1$, what is the probability that Diego is ill?

Common (and wrong!) answer: $P(X = 1|\theta = 1) = 0.95$

# Turning the Bayesian crank

Once $X = 1$ is observed, i.e. once Diego is submitted to the test $X$ and the outcome is $X = 1$, what is the probability that Diego is ill?

Common (and wrong!) answer: $P(X = 1|\theta = 1) = 0.95$

Correct answer: $P(\theta = 1|X = 1)$

# Turning the Bayesian crank

Once $X = 1$ is observed, i.e. once Diego is submitted to the test $X$ and the outcome is $X = 1$, what is the probability that Diego is ill?

Common (and wrong!) answer: $P(X = 1|\theta = 1) = 0.95$

Correct answer: $P(\theta = 1|X = 1)$

Simple probability identity (Bayes' rule):

$$\begin{aligned} P(\theta = 1|X = 1) &= P(\theta = 1)\left\{\frac{P(X = 1|\theta = 1)}{P(X = 1)}\right\} \\ &= 0.70 \times \frac{0.95}{0.785} \\ &= 0.70 \times 1.210191 \\ &= 0.8471338 \end{aligned}$$

# Combining both pieces of information

By combining

existing information (prior) $+$ model/data (likelihood)

the updated (posterior) probability that Diego is ill is 85%.

# Combining both pieces of information

By combining

existing information (prior) $\quad + \quad$ model/data (likelihood)

the updated (posterior) probability that Diego is ill is 85%.

More generally,

$$\texttt{posterior} = \frac{\texttt{prior} \times \texttt{likelihood}}{\texttt{predictive}}$$

# What if instead $X = 0$?

Maximum likelihood:

$X = 0$ is more likely from a healthy patient than from an ill one

$$\frac{P(X=0|\theta=0)}{Pr(X=0|\theta=1)} = \frac{0.60}{0.05} = 12,$$

so MLE of $\theta$ is $\hat{\theta}_{MLE} = 0$.

Bayes:

Similarly, it is easy to see that

$$\begin{aligned} P(\theta=0|X=0) &= P(\theta=0)\left\{\frac{P(X=0|\theta=0)}{P(X=0)}\right\} \\ &= 0.3 \times \frac{0.60}{0.215} \\ &= 0.3 \times 2.790698 \\ &= 0.8373093 \end{aligned}$$

# Sequential learning

The doctor is still not convinced and decides to perform a second more reliable test ($Y$):

$$P(Y = 0|\theta = 1) = 0.01 \quad \text{versus} \quad P(X = 0|\theta = 1) = 0.05$$
$$P(Y = 1|\theta = 0) = 0.04 \quad \text{versus} \quad P(X = 1|\theta = 0) = 0.40$$

# Sequential learning

The doctor is still not convinced and decides to perform a second more reliable test ($Y$):

$$P(Y=0|\theta=1)=0.01 \quad \text{versus} \quad P(X=0|\theta=1)=0.05$$
$$P(Y=1|\theta=0)=0.04 \quad \text{versus} \quad P(X=1|\theta=0)=0.40$$

## Overall rate of positives

Once again, the doctor can anticipate the overall rate of positive tests, but now conditioning on $X=1$:

$$\begin{aligned} P(Y=1|X=1) &= P(Y=1|\theta=0)P(\theta=0|X=1) \\ &+ P(Y=1|\theta=1)P(\theta=1|X=1) \\ &= (0.04)(0.1528662)+(0.99)(0.8471338) \\ &= 0.8447771 \end{aligned}$$

## $Y = 1$ is observed

Once again, Bayes rule leads to

$$
\begin{aligned}
P(\theta = 1|X = 1, Y = 1) &= P(\theta = 1|X = 1)\left\{\frac{P(Y = 1|\theta = 1)}{P(Y = 1|X = 1)}\right\} \\
&= 0.8471338 \times \frac{0.99}{0.8447771} \\
&= 0.8471338 \times 1.171907 \\
&= 0.992762
\end{aligned}
$$

# $Y = 1$ is observed

Once again, Bayes rule leads to

$$
\begin{aligned}
P(\theta = 1|X = 1, Y = 1) &= P(\theta = 1|X = 1)\left\{\frac{P(Y = 1|\theta = 1)}{P(Y = 1|X = 1)}\right\} \\
&= 0.8471338 \times \frac{0.99}{0.8447771} \\
&= 0.8471338 \times 1.171907 \\
&= 0.992762
\end{aligned}
$$

Bayesian sequential learning:

$$
P(\theta = 1|H) = \begin{cases} 70\% & , H\text{: before } X \text{ and } Y \\ 85\% & , H\text{: after } X = 1 \text{ and before } Y \\ 99\% & , H\text{: after } X = 1 \text{ and } Y = 1 \end{cases}
$$

## $Y = 1$ is observed

Once again, Bayes rule leads to

$$
\begin{aligned}
P(\theta = 1|X = 1, Y = 1) &= P(\theta = 1|X = 1)\left\{\frac{P(Y = 1|\theta = 1)}{P(Y = 1|X = 1)}\right\} \\
&= 0.8471338 \times \frac{0.99}{0.8447771} \\
&= 0.8471338 \times 1.171907 \\
&= 0.992762
\end{aligned}
$$

Bayesian sequential learning:

$$
P(\theta = 1|H) = \begin{cases} 70\% & , H\text{: before } X \text{ and } Y \\ 85\% & , H\text{: after } X = 1 \text{ and before } Y \\ 99\% & , H\text{: after } X = 1 \text{ and } Y = 1 \end{cases}
$$

Note: It is easy to see that $Pr(\theta = 1|Y = 1) = 98.2979\%$.

# $Y = 1$ is observed

Once again, Bayes rule leads to

$$
\begin{aligned}
P(\theta=1|X=1,Y=1) &= P(\theta=1|X=1)\left\{\frac{P(Y=1|\theta=1)}{P(Y=1|X=1)}\right\} \\
&= 0.8471338 \times \frac{0.99}{0.8447771} \\
&= 0.8471338 \times 1.171907 \\
&= 0.992762
\end{aligned}
$$

Bayesian sequential learning:

$$
P(\theta=1|H) = \begin{cases} 70\% & , H\text{: before } X \text{ and } Y \\ 85\% & , H\text{: after } X=1 \text{ and before } Y \\ 99\% & , H\text{: after } X=1 \text{ and } Y=1 \end{cases}
$$

Note: It is easy to see that $Pr(\theta=1|Y=1) = 98.2979\%$.
Conclusion: Don't consider test $X$, unless it is "cost" free.

# Example 2: Gaussian measurement error

Goal: Learn $\theta$, a physical quantity.

## Example 2: Gaussian measurement error

Goal: Learn $\theta$, a physical quantity.

Measurement: $X$

# Example 2: Gaussian measurement error

Goal: Learn $\theta$, a physical quantity.

Measurement: $X$

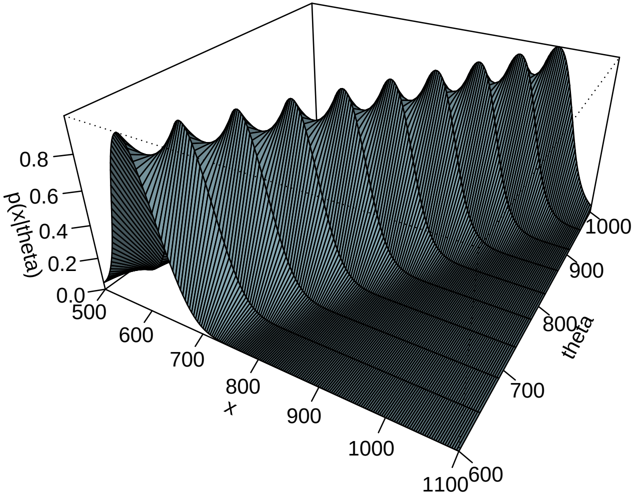
Model: $(X|\theta) \sim N(\theta, (40)^2)$

# Example 2: Gaussian measurement error

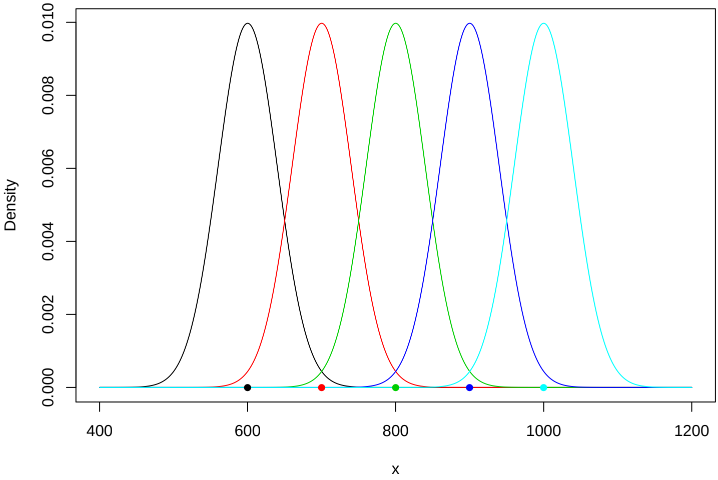
Goal: Learn $\theta$, a physical quantity.

Measurement: $X$

Model: $(X|\theta) \sim N(\theta, (40)^2)$



$p(x|\theta)$ for $\theta \in \{600, 700, \ldots, 1000\}$



# Large and small prior experience

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$

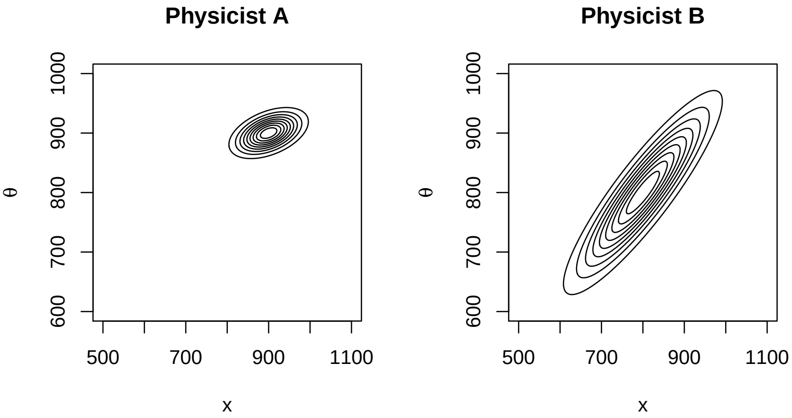
# Large and small prior experience

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, (80)^2)$

# Large and small prior experience

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$
Prior B: Physicist B (not so experienced): $\theta \sim N(800, (80)^2)$
Joint density: $p(x, \theta) = p(x|\theta)p(\theta)$



# Bayesian computation: predictive

Prior: $\theta \sim N(\theta_0, \tau_0^2)$ (Physicist A: $\theta_0 = 900$, $\tau_0 = 20$)
Model: $x|\theta \sim N(\theta, \sigma^2)$

Predictive:

$$p(x) = \int_{-\infty}^{\infty} p(x|\theta)p(\theta)d\theta$$

## Bayesian computation: predictive

Prior: $\theta \sim N(\theta_0, \tau_0^2)$ (Physicist A: $\theta_0 = 900$, $\tau_0 = 20$)

Model: $x|\theta \sim N(\theta, \sigma^2)$

Predictive:

$$p(x) = \int_{-\infty}^{\infty} p(x|\theta)p(\theta)d\theta$$

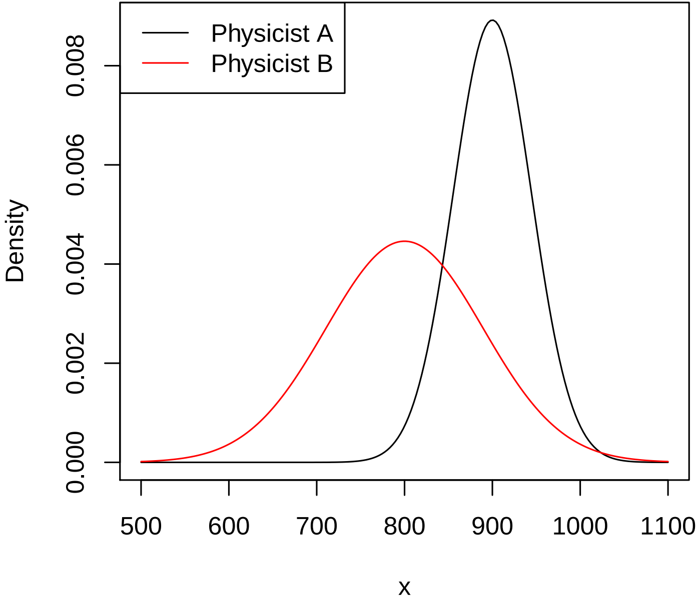
Therefore,

$$\begin{aligned} p(x) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\tau_0^2}} e^{-\frac{(\theta-\theta_0)^2}{2\tau_0^2}} d\theta \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau_0^2)}} e^{-\frac{(x-\theta)^2}{2(\sigma^2+\tau_0^2)}} \end{aligned}$$

or

$$x \sim N(\theta_0, \sigma^2 + \tau_0^2)$$

# Predictive densities



# Bayesian computation: posterior

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

such that

$$\begin{aligned}
p(\theta|x) &\propto (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\theta)^2}{2\sigma^2}} (2\pi\tau_0^2)^{-1/2} e^{-\frac{(\theta-\theta_0)^2}{2\tau_0^2}} \\
&\propto \exp\left\{-\frac{1}{2}\left[(\theta^2 - 2\theta x)/\sigma^2 + (\theta^2 - 2\theta\theta_0)/\tau_0^2)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau_1^2}(\theta-\theta_1)^2\right\}.
\end{aligned}$$

Therefore,

$$\theta|x \sim N(\theta_1, \tau_1^2)$$

where

$$\theta_1 = \left(\frac{\sigma^2}{\sigma^2+\tau_0^2}\right)\theta_0 + \left(\frac{\tau_0^2}{\sigma^2+\tau_0^2}\right)x \quad \text{and} \quad \tau_1^2 = \tau_0^2\left(\frac{\sigma^2}{\sigma^2+\tau_0^2}\right)$$

# Combination of information

Let

$$\pi = \frac{\sigma^2}{\sigma^2 + \tau_0^2} \in (0, 1)$$

Therefore,

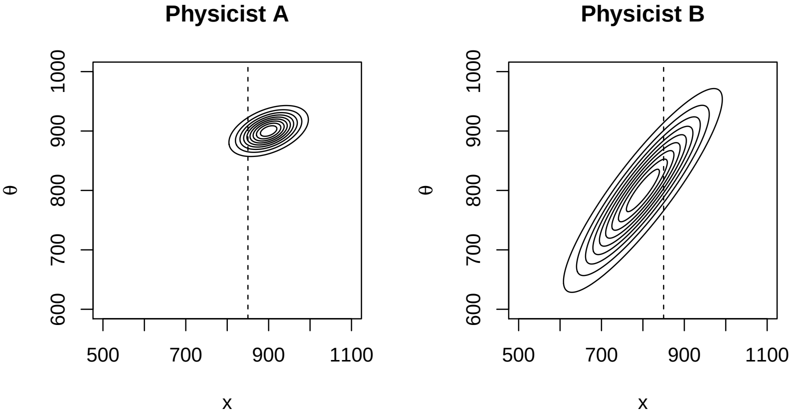
$$E(\theta|x) = \pi E(\theta) + (1 - \pi)x$$

and

$$V(\theta|x) = \pi V(\theta)$$

When $\tau_0^2$ is much larger than $\sigma^2$, $\pi \approx 0$ and the posterior collapses at the observed value $x$!

# Observation: $X = 850$



# Posterior (updated) densities

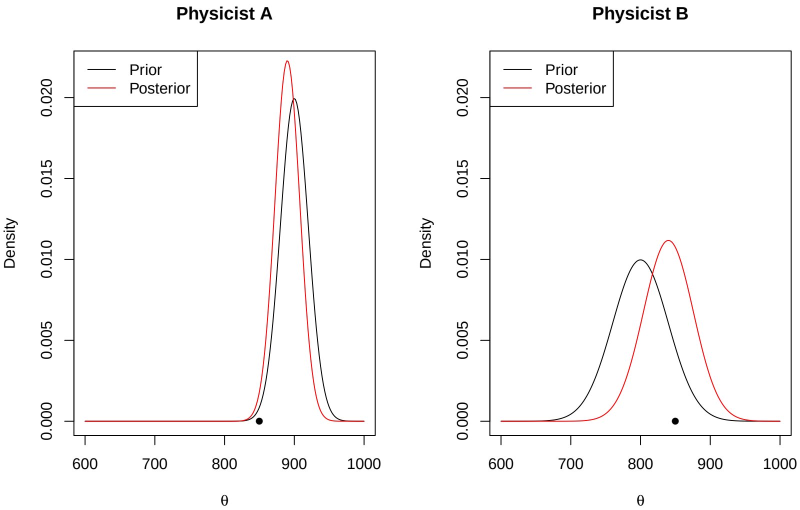
Physicist A

Prior: $\theta \sim N(900, (20)^2)$
Posterior: $(\theta | X = 850) \sim N(890, (17.9)^2)$

Physicist B

Prior: $\theta \sim N(800, (40)^2)$
Posterior: $(\theta | X = 850) \sim N(840, (35.7)^2)$

# Priors and posteriors



# Summary

Deriving the posterior (via Bayes rule)

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

and computing the predictive

$$p(x) = \int_{\Theta} p(x|\theta)p(\theta)d\theta$$

can become very challenging!

## Summary

Deriving the posterior (via Bayes rule)

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

and computing the predictive

$$p(x) = \int_{\Theta} p(x|\theta)p(\theta)d\theta$$

can become very challenging!

Bayesian computation was done on limited, unrealistic models until the Monte Carlo revolution (and the computing revolution) of the late 1980's and early 1990's.

# A more conservative physicist

Prior A: Physicist A (large experience): $\theta \sim N(900, 400)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, 1600)$

# A more conservative physicist

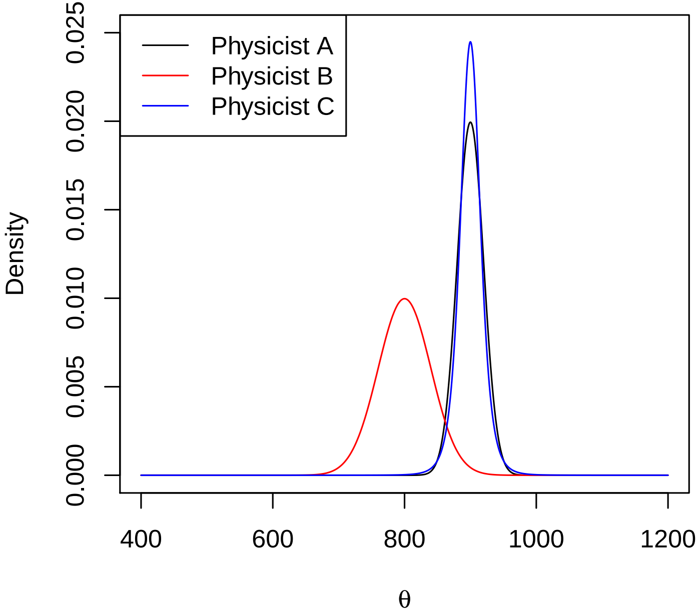
Prior A: Physicist A (large experience): $\theta \sim N(900, 400)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, 1600)$

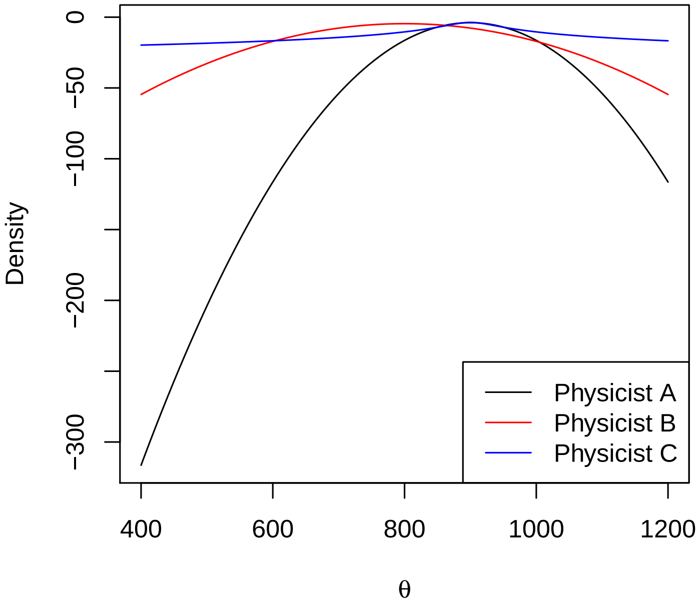
Prior C: Physicist C (largeR experience): $\theta \sim t_5(900, 240)$

$$V(\text{Prior C}) = \frac{5}{5-2} 240 = 400 = V(\text{Prior A})$$

# Prior densities



# Closer look at the tails



# Predictive and posterior of physicist C

For model $x|\theta \sim N(\theta, \sigma^2)$ and prior of $\theta \sim t_\nu(\theta_0, \tau^2)$,

$$p(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\tau_0^2}} \left(1 + \frac{1}{\nu}\left(\frac{\theta-\theta_0}{\tau_0}\right)^2\right)^{-\frac{\nu+1}{2}} d\theta$$

is not analytically available.

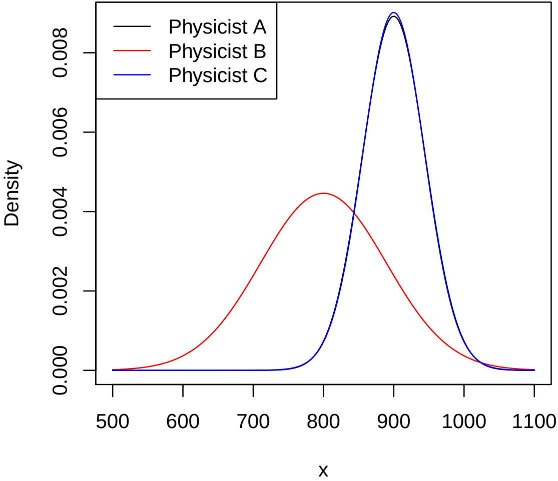
Similarly,

$$p(\theta|x) \propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}\left(1 + \frac{1}{\nu}\frac{(\theta-\theta_0)^2}{\tau_0^2}\right)^{-\frac{\nu+1}{2}}$$

is of no known form.

# Predictives

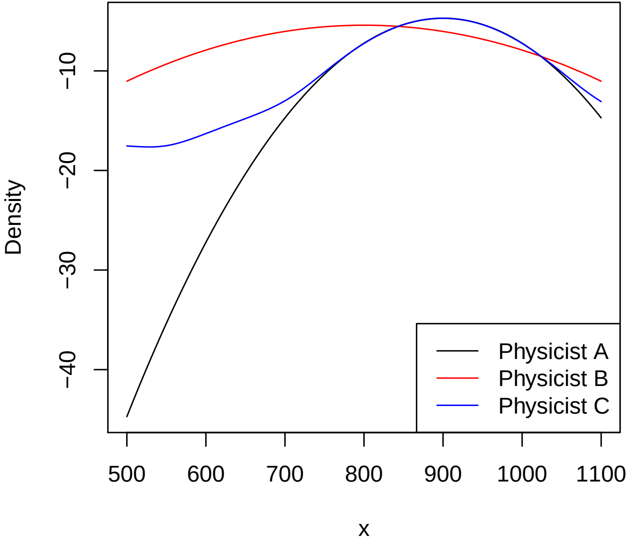
Monte Carlo approximation[2] to $p(x)$ for physicist C.




[2]Yet to be learned!

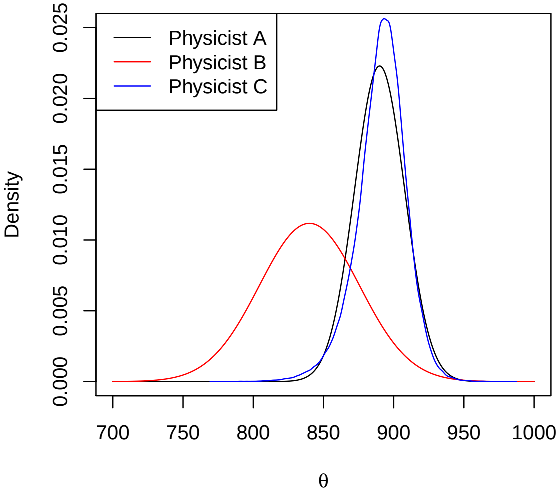
# Log predictives

Physicist C has similar knowledge as physicist A, but does not rule out smaller values for $x$.



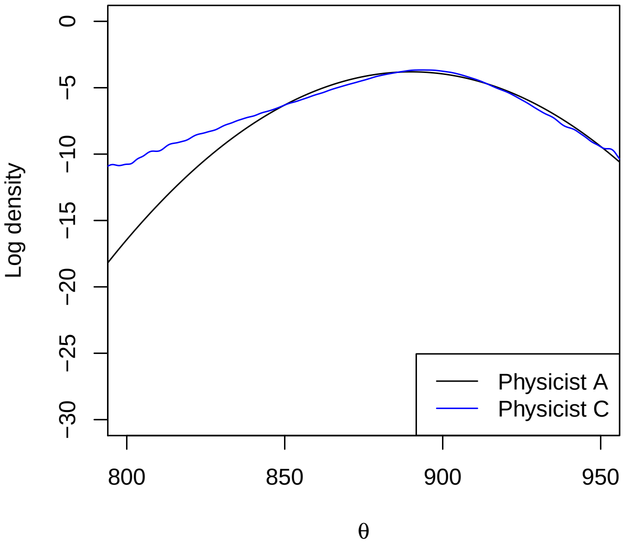
# Posteriors for $\theta$

Monte Carlo approximation[3] to $p(\theta|x)$ for physicist C.




[3]Yet to be learned!

# Log posteriors



# Monte Carlo: a toy example

In what follows, we will see how to approximate integrals and sample from unknown distributions via the well known *Monte Carlo* method.

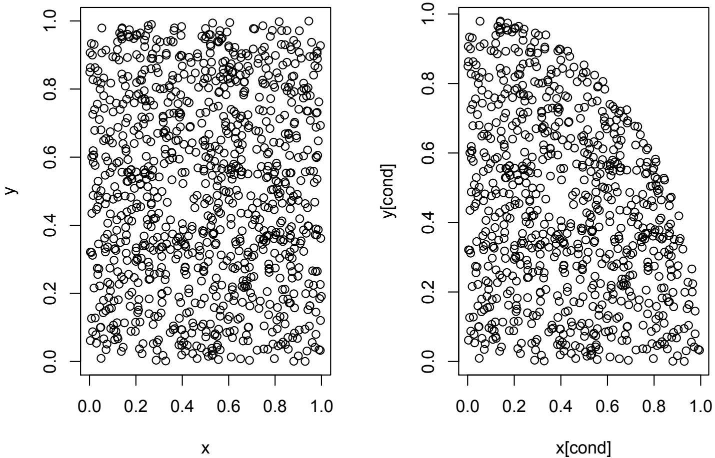
Let us think about calculating $\pi = 3.141593\ldots$

We could sample a bunch $(i = 1, \ldots, M)$ of pairs $(x_i, y_i)$ in the unit square $(0,1) \times (0,1)$ and compute the fraction $\alpha$ of those pairs satisfying the condition $x_i^2 + y_i^2 < 1$. In this case, $pi = 4\alpha$.

```
M = 1000
x = runif(M)
y = runif(M)
cond = (x^2+y^2)<1
par(mfrow=c(1,2))
plot(x,y)
plot(x[cond],y[cond])
pi.mc = 4*sum(cond)/M
```

## $\pi_{mc} = 3.196$

$$\frac{\pi}{4} = \int_0^1 \int_0^{\sqrt{1-x^2}} dydx$$



# Monte Carlo: Let us play with $M$

```
set.seed(12345)
M = 10000
x = runif(M)
y = runif(M)
cond = (x^2+y^2)<1
pi.mc = 4*cumsum(cond)/(1:M)
plot(1:M/1000,pi.mc,ylim=c(2.7,3.3),type="l",
     xlab="thousands of draws",ylab="pi approx.")
abline(h=pi,col=2)

for (i in 1:20){
  x = runif(M)
  y = runif(M)
  cond = (x^2+y^2)<1
  pi.mc = 4*cumsum(cond)/(1:M)
  lines(1:M/1000,pi.mc,col=i)
}
```

# MC error