VE414 Lecture 2

Jing Liu

UM-SJTU Joint Institute

May 16, 2019

- The updating or learning favour of Bayesian inference is particularly strong with Thomas Bayes' original study on the subject before Laplace refined the the idea of probability and formulated mathematical form of Bayes' theorem.

Thomas Bayes' original problem

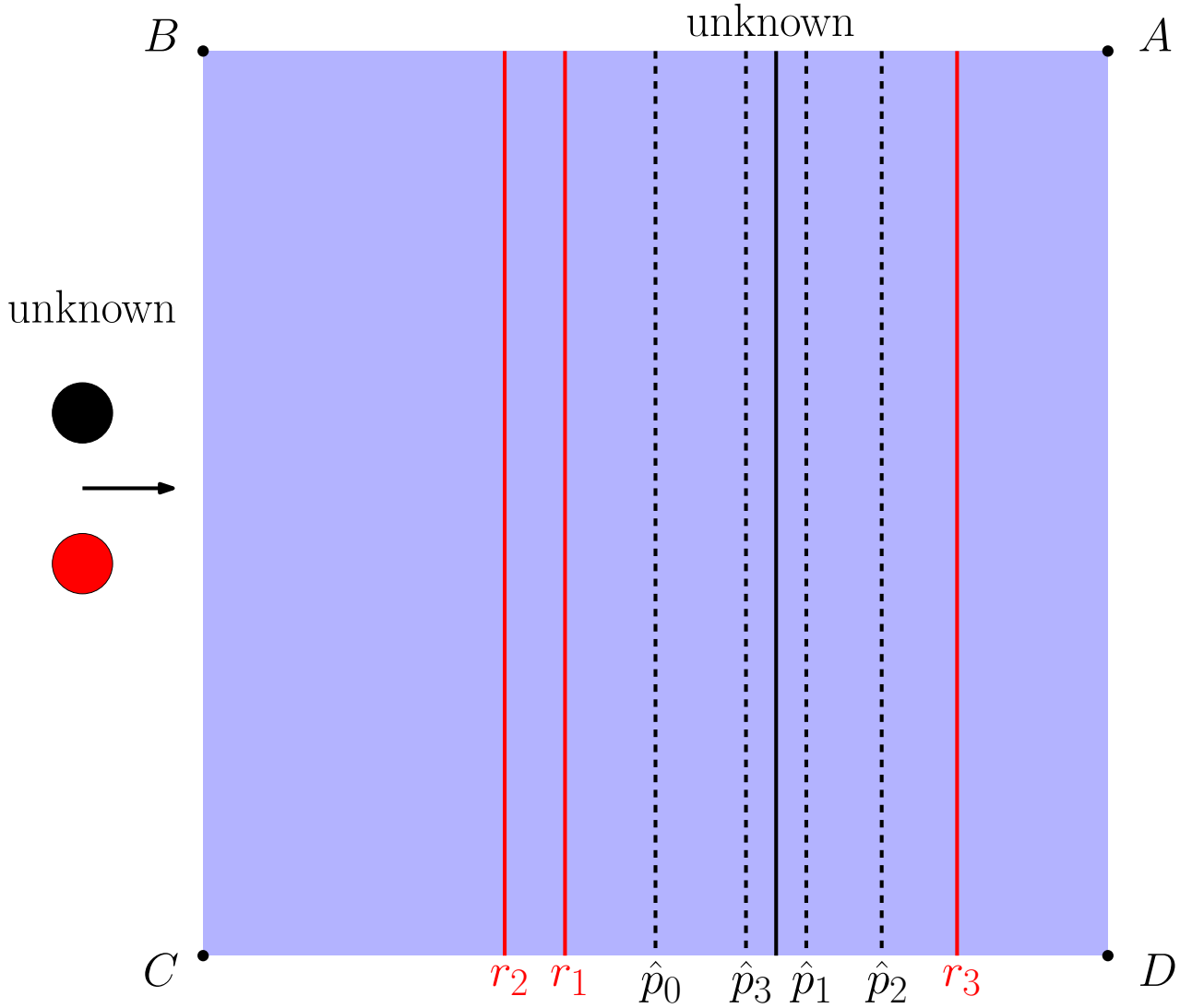
"*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named."

- In modern terms, it is about statistical inference on binomial proportion p,

$$X \sim \text{Binomial}\,(k, p)$$

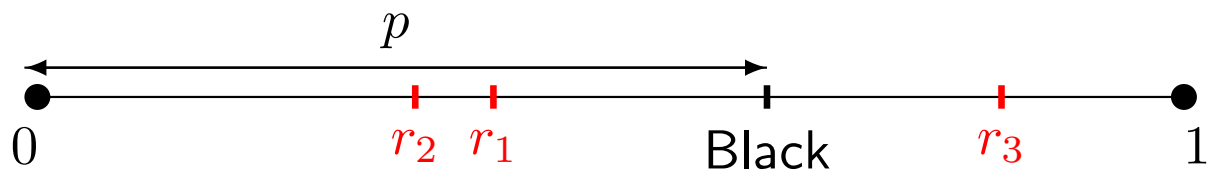
- **Bayes**[1] used a thought experiment to illustrate his process of inferencing.

[1] Thomas Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". . In: *Philosophical transactions of the Royal Society of London* 53 (1763), pp. 370–418.



Q: How would you estimate where the black ball was?

- The only information available is whether the succeeding red balls are on the left of where the black ball was, but not the actual positions of the red balls.



- Notice the position of the black ball determines the chance of whether a red ball is on the left of it, this connection allows us to estimate where it was.
- If X_k denotes the number of red balls out of k number of trials that are on the left of the black ball, with the assumption of independence, then

$$X_k \sim \text{Binomial}\,(k, p)$$

- In modern terms, we have identified the distribution from which the data are generated, thus the likelihood function, and the objective is to estimate p,

$$\mathcal{L}\,(p; x) = f_{X|p}\,(x \mid p) = \frac{k!}{x!(k-x)!} p^x (1-p)^{k-x}$$

- The maximum likelihood estimate is one of typical frequentist solutions

$$\tilde{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; x)$$

$$\Longrightarrow \tilde{p}_k = \frac{x}{k}$$

  where x denotes the number of success out of k number of trials.

- Having nice asymptotic properties is the main reason of using MLE; if there is very little data, it can be unreliable since it has no mechanism to take into account what could happen, it is entirely based on what has happened.
- Suppose $X_1 = 1$, $X_2 = 2$ and $X_3 = 2$ as depicted in the graph, then

$$\tilde{p}_1 = 1; \quad \tilde{p}_2 = 1 \quad \text{and} \quad \tilde{p}_3 = 2/3$$

Q: Do you see what I mean it does not take into account what could happen?

Q: Recall the Monty hall problem, can you guess what Bayes' estimates are?

Q: Have you ever questioned the definition of $f_{Y|X}$ when X is continuous?

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \qquad \text{where} \quad f_X(x) \neq 0$$

- The discrete case is clearly meaningful since PMF gives probability directly

$$f_{Y|X}(y \mid x) = \Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

  and it follows directly from the modern concept of conditional probability.

- However, in the continuous case, PDF does not link to probability directly

$$f_{Y|X}(y \mid x) \neq \Pr(Y = y \mid X = x)$$

  and the event to condition on has 0 probability in the continuous case

$$\Pr(Y = y \mid X = x) = 0$$

- That said, it is meaningful and should be understood in the following sense

$$
\begin{aligned}
F_{Y|X}(y \mid x) &= \lim_{\varepsilon \to 0^+} \Pr(Y \leq y \mid X \in (x, x+\varepsilon]) \\
&= \lim_{\varepsilon \to 0^+} \frac{\Pr(Y \leq y, X \in (x, x+\varepsilon])}{\Pr(x < X \leq x+\varepsilon)} \\
&= \lim_{\varepsilon \to 0^+} \frac{F_{X,Y}(x+\varepsilon, y) - F_{X,Y}(x, y)}{F_X(x+\varepsilon) - F_X(x)} = \frac{\partial F_{X,Y}(x,y)/\partial x}{f_X(x)}
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow \quad f_{Y|X}(y \mid x) &= \frac{\partial}{\partial y} \frac{\partial F_{X,Y}(x,y)/\partial x}{f_X(x)} \\
&= \frac{1}{f_X(x)} \frac{\partial^2 F_{X,Y}(x,y)}{\partial y \partial x} = \frac{f_{X,Y}(x,y)}{f_X(x)}
\end{aligned}
$$

Q: What does this modern understanding mean in terms of Bayes' theorem?

$$
\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)}
$$

- In additional to the Bayes' theorem for events, we have the following form

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x \mid Y = y)\Pr(Y = y)}{\Pr(X = x)}$$

for discrete random variables, and the following form

$$f_{Y|X=x}(y) = \frac{f_{X|Y}(x \mid y)\, f_Y(y)}{f_X(x)} \propto \mathcal{L}(y; x)\, f_Y(y)$$

for PDF as well as PMF given our understanding of conditional distributions.

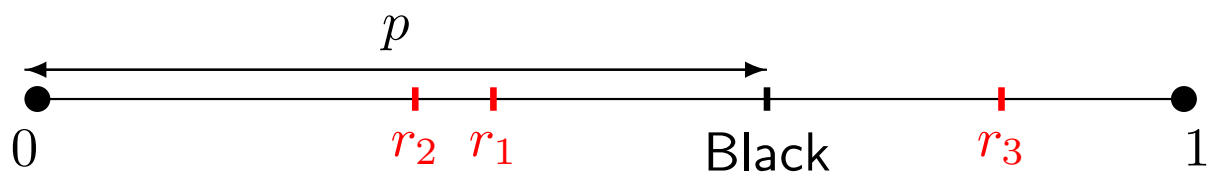
- Therefore, we often summarise various forms of Bayes' theorem simply as

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

since $f_X(x)$ depends only on the observed $X = x$, not on the unknown y.

- It is clear from this form, identifying the prior and the likelihood is crucial.

- Back to the Bayes' original problem,



for which we have identified that the following should be the likelihood

$$\mathcal{L}(p; x) = f_{X|P}(x \mid p) = \frac{k!}{x!(k-x)!} p^x (1-p)^{k-x}$$

Q: What should we use as our prior distribution before seeing any data?

- Although not in the following modern formulation, Bayes essentially used

uniform distribution $\mathrm{Unif}(0, 1)$: $$f_P(p) = \begin{cases} 1 & \text{for} \quad 0 \le p \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

and argued the prior should not prefer any particular $p \in [0, 1]$ over another.

Q: Given the data $X_3 = 2$, the uniform prior and the likelihood

$$\mathcal{L}(p; x) = \frac{k!}{x!(k-x)!} p^x (1-p)^{k-x}$$

what is the posterior density function based on the data $X_3 = 2$?

- As you have leaned from any elementary courses on estimation/prediction, if a single value is required to describe or predict a random variable, the mean of the random variable is one of the best choices. Without any data,

$$\hat{p}_0 = \mathbb{E}[P] = 1/2$$

is used as our estimate since the uniform distribution is our prior in this case.

Q: Given the arrival of new information in the form of each data point

$$X_1 = 1, \quad X_2 = 2, \quad X_3 = 2$$

what should we use as our estimate of p?

- For point estimates, Bayes naturally suggested the mean of the posterior

$$\hat{p} = \mathbb{E}\left[p \mid X = x\right] = \int_{-\infty}^{\infty} p f_{P|X=x}(p)\, dp$$

- Given the data $X_1 = 1$, $X_2 = 2$ and $X_3 = 2$, the corresponding estimate

$$\hat{p}_0 = \frac{1}{2}; \quad \hat{p}_1 = \frac{2}{3}; \quad \hat{p}_2 = \frac{3}{4} \quad \text{and} \quad \hat{p}_3 = \frac{3}{5}$$

alters with the information of the second ball in contrast to MLE

$$\tilde{p}_1 = 1; \quad \tilde{p}_2 = 1 \quad \text{and} \quad \tilde{p}_3 = 2/3$$

- Notice the Bayesian point estimator of p can be broken into two components

$$\hat{p}_k = w\frac{1}{2} + (1-w)\frac{x}{k} = w\hat{p}_0 + (1-w)\tilde{p}_k \qquad \text{where} \quad w = \frac{2}{2+k}$$

which reflect the prior knowledge and the information from the data.

- No surprise in seeing the two types of point estimates converge in some sense

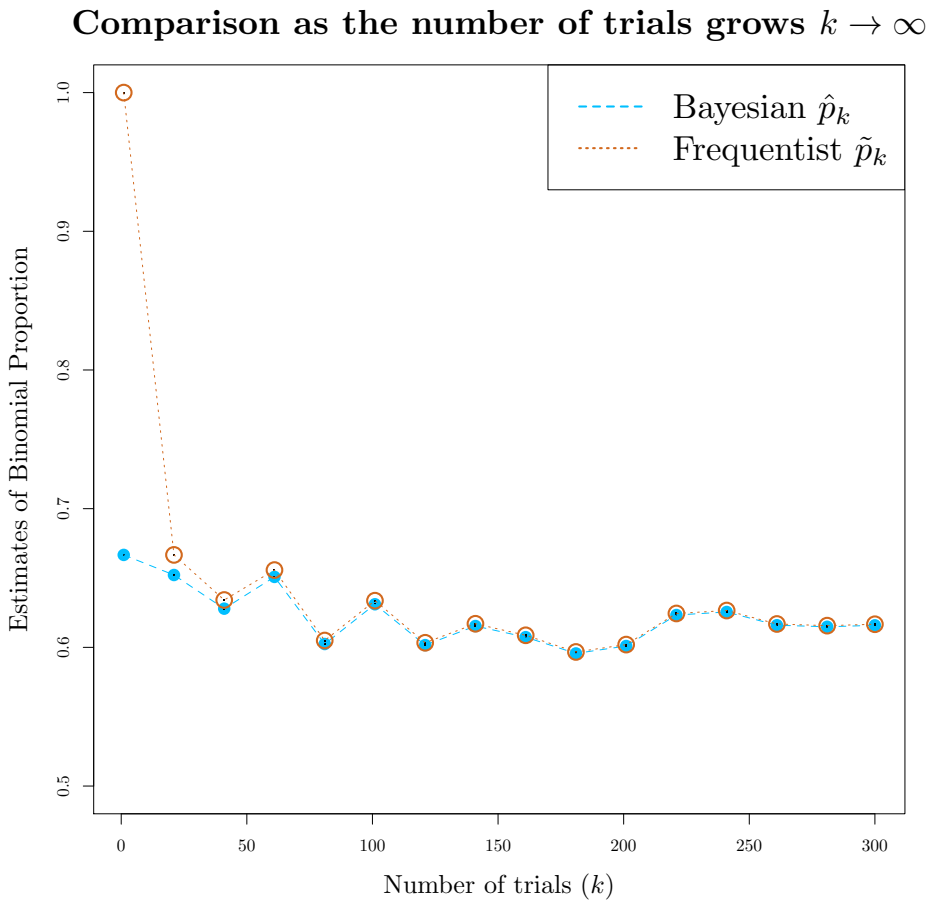


Figure: R Code: `freq_ci_414.R`

- However, Bayes did not focus on point estimates because he argued a point estimate fail to fully incorporate and reflect what we can learn from the data.

Q: What would you have as $\hat{p}_k$ and $\tilde{p}_k$ if you had the following data

$$X_2 = 2; \qquad X_{20} = 20; \qquad x_{200} = 200$$

Q: What should you report instead of point estimates?

Definition

In Bayesian analysis, a credible interval is an interval within which an unobserved parameter/variable value falls with a particular probability. Let $F_{Y|X}$ be the CDF,

$$F_{Y|X}(y_u \mid x) > F_{Y|X}(y_l \mid x)$$

then the interval $\mathcal{I} = [y_l, y_u]$ is a credible interval with coverage probability

$$F_{Y|X}(y_u \mid X) - F_{Y|X}(y_l \mid X)$$

If exactly $(\alpha/2)100\%$ of the posterior probability lies above and below $\mathcal{I}$, then $\mathcal{I}$ is known as the central credible interval with coverage probability of $1 - \alpha$.

```
> # Prior is taken to be beta(1,1), i.e. uniform
> # Prior Hyperparameters
> a.prior = 1; b.prior = 1

> # Inverse of cdf, aka quantile function
> qbeta(0.5, a.prior, b.prior)
```

```
[1] 0.5
```

```
> # Lower and upper limits
> l = qbeta(0.025, a.prior, b.prior)
> u = qbeta(0.975, a.prior, b.prior)

> # Print out the prior central credible interval
> cat(paste("(",l, sep = ""),
+     paste(u,")", sep =""), sep = ",")
```

```
(0.025,0.975)
```

```
> # Extreme case
> k.vec = c(2, 20, 200)    # Number of trials
> x.vec = c(2, 20, 200)    # Data
> n = length(k.vec)        # Number of cases

> for (i in 1:n){
+   # Posterior hyperparameters
+   a.posterior = a.prior + x.vec[i]
+   b.posterior = b.prior + (k.vec[i]-x.vec[i])
+
+   l = qbeta(0.025, a.posterior, b.posterior)
+   u = qbeta(0.975, a.posterior, b.posterior)
+   cat(paste("(",l, sep = ""),
+       paste(u,")", sep =""),
+       "\n", sep = ",")  # New line
+ }
```

```
(0.292401773821287,0.991596241340387)
(0.83890238478092,0.998795116551636)
(0.981814749945614,0.99987404868883)
```

```
> # Simulation study of posterior CCI
> set.seed(414)
> n = 5000
> k.vec = 1:n
> true.prob = 0.65

> # n Bernoulli trials
> x.vec = rbinom(n, size = 1, prob = true.prob)
> head(x.vec)
```

```
[1] 0 1 0 1 1 0
```

```
> # forming simulated the data
> x.vec = cumsum(x.vec)
> head(x.vec)
```

```
[1] 0 1 1 2 3 3
```

- Notice how the posterior central credible interval shrinks as k grows.

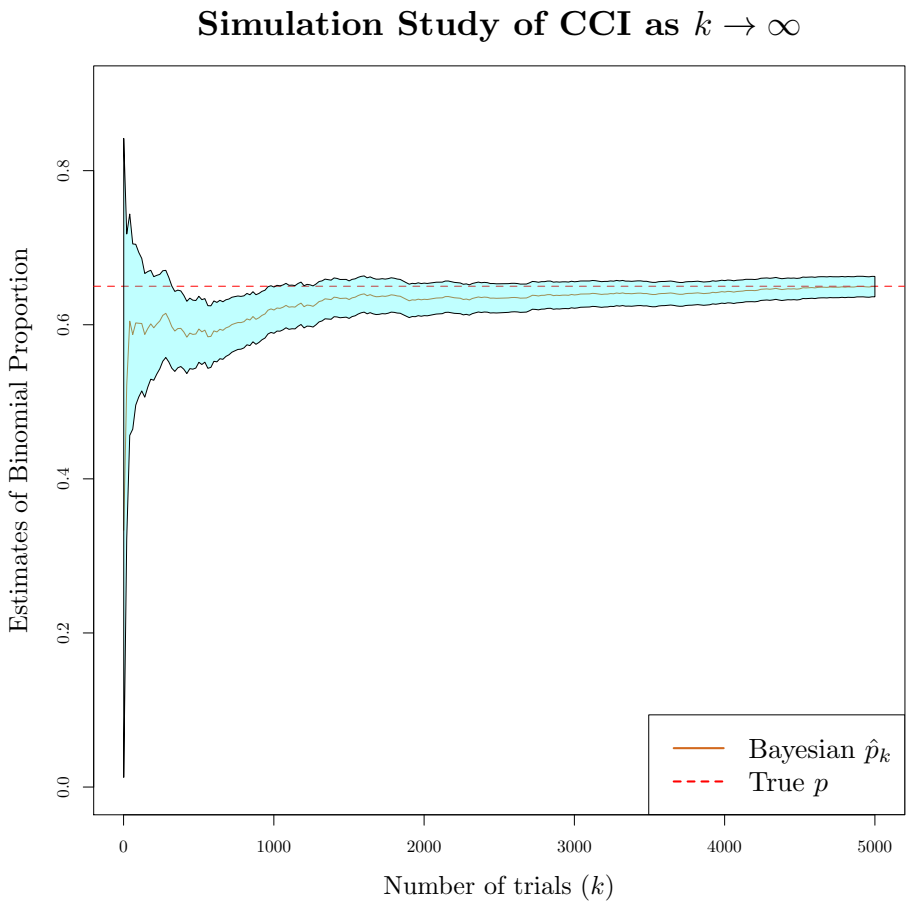
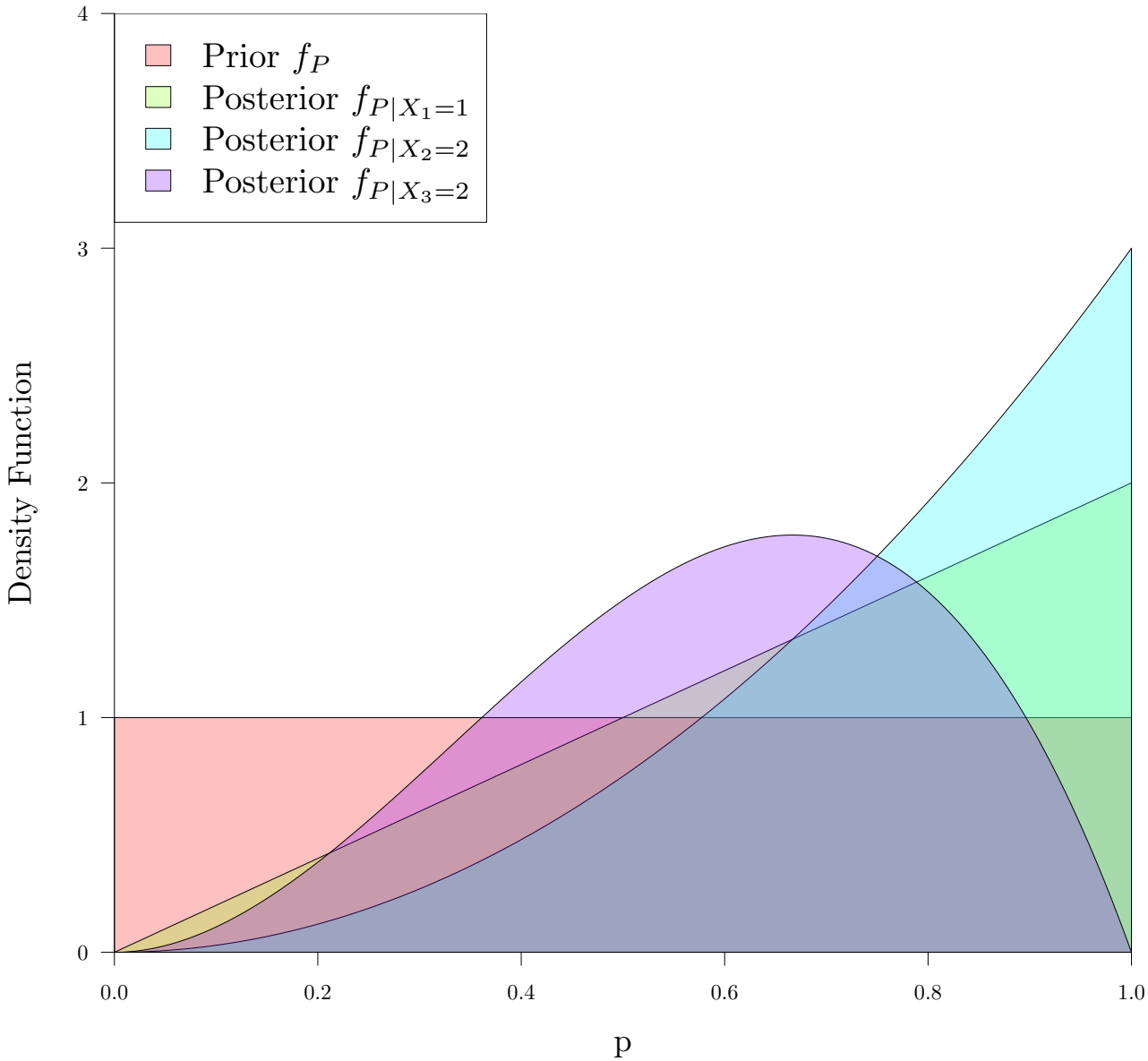


Figure: R Code: `central_credible_interval_binomial_beta.R`

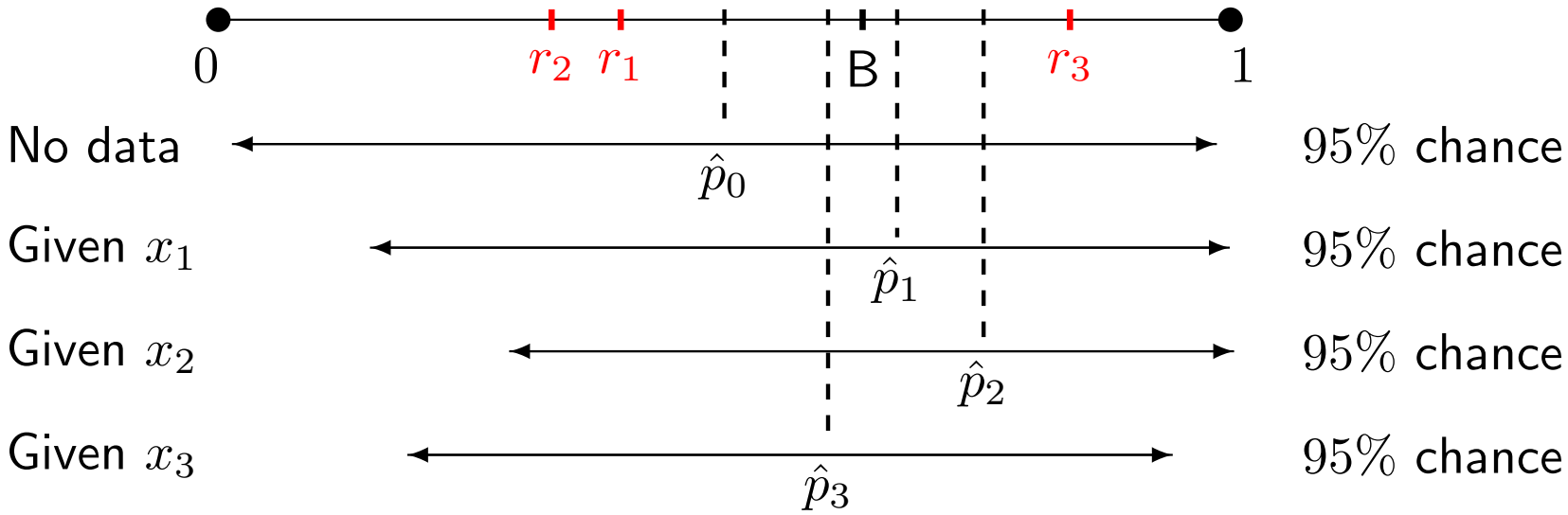
```
> # Back to Bayes' data
> k.vec = 1:3
> x.vec = c(1, 2, 2)
> n = length(k.vec)          # Number of cases

> for (i in 1:n){
+   # Posterior hyperparameters
+   a.posterior = a.prior + x.vec[i]
+   b.posterior = b.prior + (k.vec[i]-x.vec[i])
+   l = qbeta(0.025, a.posterior, b.posterior)
+   u = qbeta(0.975, a.posterior, b.posterior)
+   cat(paste("(",l, sep = ""),
+       paste(u,")", sep =""),
+       "\n", sep = ",")  # New line
+ }
```

```
(0.158113883008419,0.987420882906575)
(0.292401773821287,0.991596241340387)
(0.194120449683243,0.932414013511457)
```

Prior/Posterior Densities



- In the Bayes' study, 95% credible intervals have a 95% chance of correctly capturing the true p, thus the black ball, at various stage of the experiment.



- Notice those intervals are different from typical confidence intervals (CI).

Q: What is the difference? How can we compute a typical confidence interval?

- Recall confidence intervals for binomial proportion p rely on the central limit theorem, and it is unreliable for a small number of trials. Hence suppose the experiment continues, and we have 300 trials instead of just 3 trials.

- According to the central limit theorem, the following random variable

$$Z = (\tilde{p}_k - p) \Big/ \sqrt{\frac{\tilde{p}_k (1 - \tilde{p}_k)}{k}}$$

  follows a standard normal distribution as $k \to \infty$.

- Thus the following can be used as the $(1 - \alpha)100\%$ confidence interval of p.

$$\left(\tilde{p}_k - z_{\alpha/2} \sqrt{\frac{\tilde{p}_k (1 - \tilde{p}_k)}{k}}, \tilde{p}_k + z_{\alpha/2} \sqrt{\frac{\tilde{p}_k (1 - \tilde{p}_k)}{k}} \right)$$

  where $\alpha \in (0, 1)$ and $z_{\alpha/2}$ is a real number such that $\Pr\left(Z \leq -z_{\alpha/2}\right) = \frac{\alpha}{2}$.

- To understand a 95% confidence interval,
  1. consider $2^{16} = 65536$ samples of size 3000
  2. compute the 95% confidence interval for each sample
  3. check whether the true p is inside each of the confidence intervals

- Roughly 95% of the 95% confidence intervals contain the true p for large n

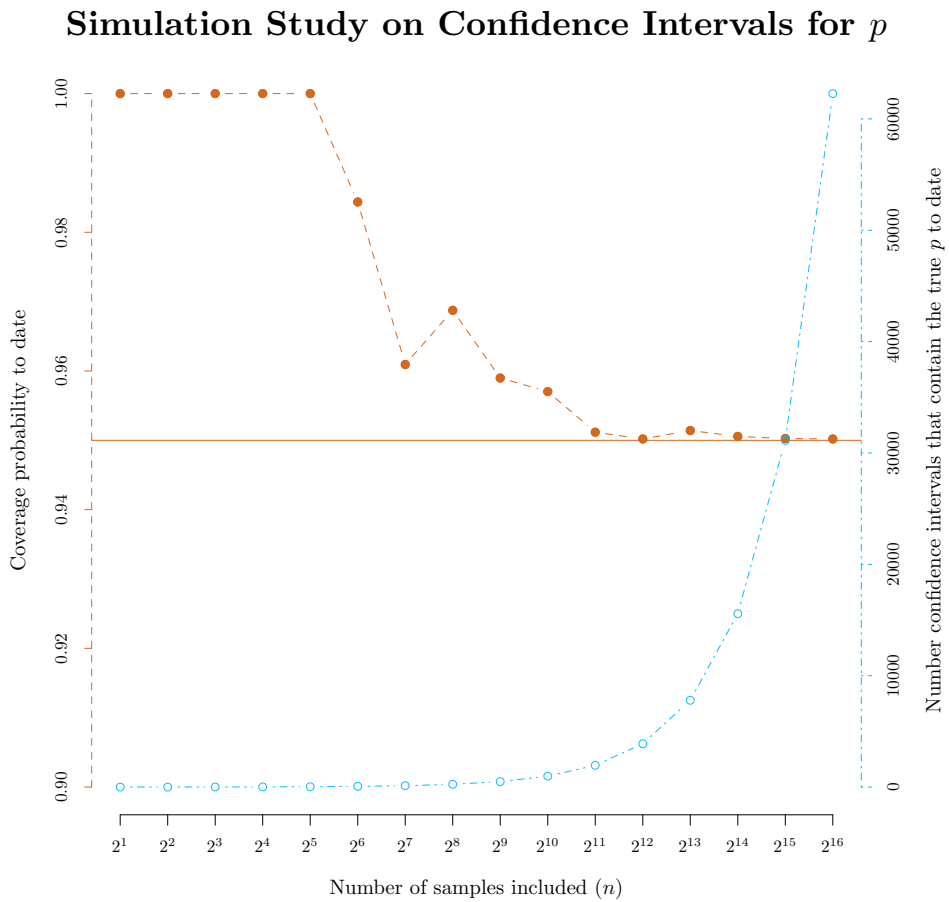


Figure: R Code: `bayes_ball_table_414.R`

- We can conduct something similar to a hypothesis test,

$$H_0 : p > 0.8$$

```
> a.prior = 1; b.prior = 1
> pbeta(0.8, a.prior, b.prior)        # cdf
```

```
[1] 0.8
```

```
> k.vec = 1:3; x.vec = c(1, 2, 2); n = length(k.vec)
> for (i in 1:n){
+   # Posterior hyperparameters
+   a.posterior = a.prior + x.vec[i]
+   b.posterior = b.prior + (k.vec[i]-x.vec[i])
+   cat(pbeta(0.8, a.posterior, b.posterior),
+       ";\t", sep = "")
+ }
```

```
0.64;   0.512;  0.8192;
```

Q: Is there any difference between Bayesian and Frequentist hypothesis testing?