STA 602 – Intro to Bayesian Statistics

Lecture 3

Li Ma

Last class

- Bayes' Theorem and the *ideal* approach to inference.
- Example: A political poll (a binomial experiment).

Political poll example revisited

An organization randomly selected 100 democrats and interviewed them about whether they support the incombent governer. Out of the 100 polled, 40 support the governer and 60 against. What can we say about the actual support rate of the governer θ?

Our model (or assumptions/hypotheses) are

(1) Given θ, X is Binomial(100,θ).

(2) θ is Uniform(0,1) *a priori*—representing our knowledge before data is observed.

Based on these two pieces, Bayes' Theorem allows us to use the observed data to update our knowledge about the parameter.

(3) θ is Beta(41,61) *a posteriori*—representing our updated knowledge after data is observed.

- Our analysis up to this point uses Uniform(0,1) to represent our prior knowledge about θ.
- More flexible choices for $p(\theta)$ can allow richer prior knowledge about the parameter to be incorporated into the analysis.

A richer class of priors for Binomial experiments

A flexible and convenient choice of the prior distribution for θ is the Beta(α,β) family of distributions.

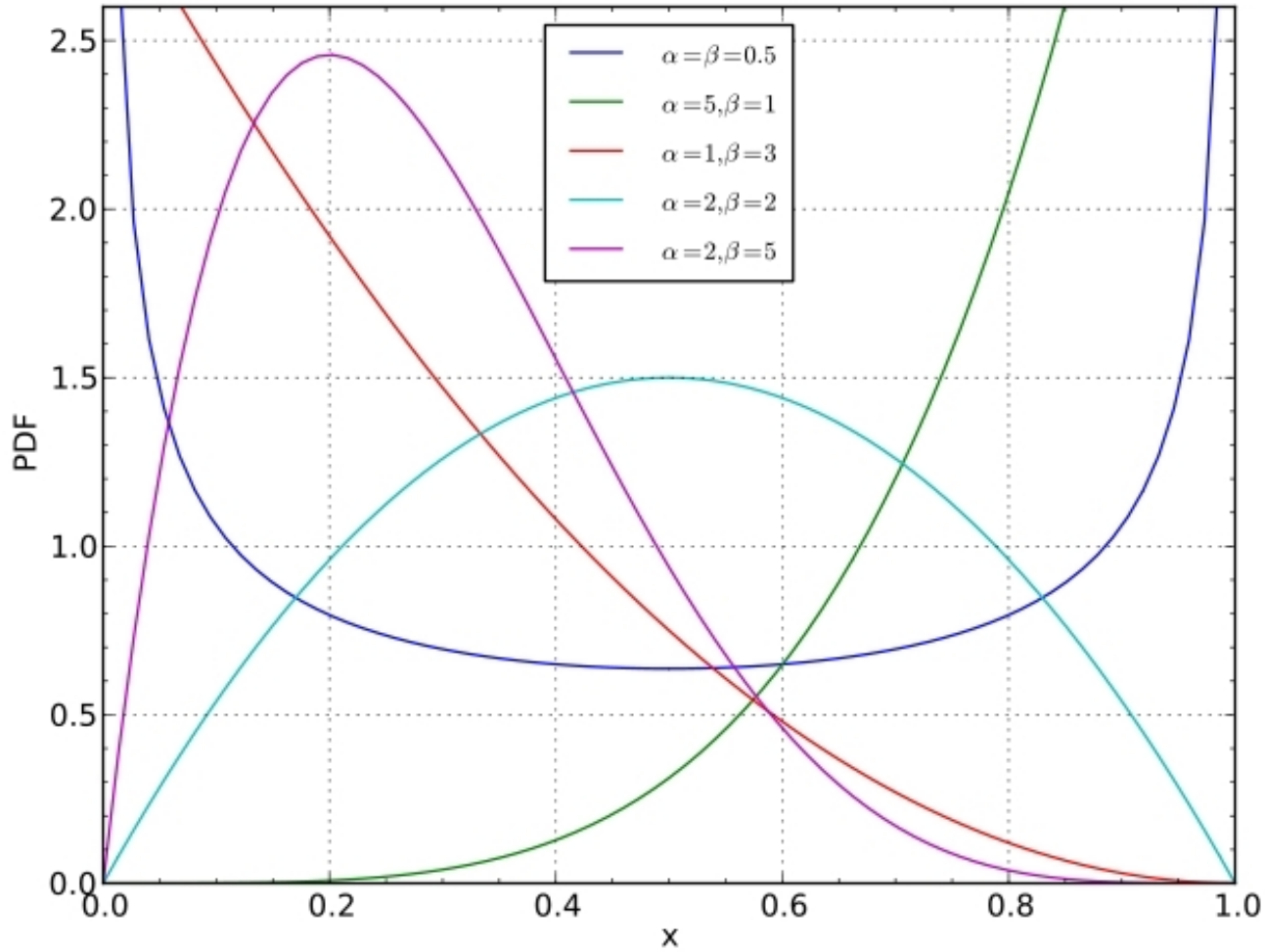
$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 < \theta < 1$$
$$= 0 \quad \text{otherwise}$$

where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$. (This is called the *Gamma* function.)

- Mean: $\mu_\theta = \frac{\alpha}{\alpha+\beta}$.
- Variance: $\sigma^2_\theta = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}$.

Parameters such as α and β that are used to characterize the prior distribution of θ are called *hyperparameters*.

Remark: When $\alpha = \beta = 1$, this is exactly the Uniform(0,1) distribution.



Source: Wikipedia.
`http://upload.wikimedia.org/wikipedia/commons/f/f3/Beta_distribution_pdf.svg`.

Bayesian inference using a Beta(α, β) prior

$$
\begin{aligned}
p(\theta|x) &\propto p(x|\theta)p(\theta) \\
&= \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= C(x)\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1 \\
&= 0 \quad \text{othewise.}
\end{aligned}
$$

To make this a density, we must have

$$
C(x) = \frac{1}{\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta}.
$$

But we don't need to calculate the integral in the denominator due to the following trick.

A trick to get the posterior density

The part that depends on θ

$$\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1$$

is the same as that in the p.d.f of a Beta$(x+\alpha, n-x+\beta)$ distribution. Thus the two density functions must be *identical*. So we must have

$$\begin{aligned} p(\theta|x) &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

A quick sum up

To make inference on a Binomial experiment using a Beta prior, we can proceed as follows

1. We model the distribution of the number of successes X conditional on the success probability θ as Binomial(n, θ).
2. We can then model the prior distribution $p(\theta)$ as a Beta(α, β) density. The value of α and β are chosen to represent our prior knowledge about the parameter.
3. After observing $X = x$, Bayes' theorem tells us that the posterior distribution of θ is Beta$(x + \alpha, n - x + \beta)$.

The three steps for obtaining posterior distribution

This example illustrates the general process of *all* Bayesian statistical analysis.

1. Model the distribution of data X conditional on a set of parameter θ.
2. Choose a prior distribution $p(\theta)$ for the parameters, and
3. Apply Bayes' Theorem to find the posterior distribution $p(\theta|x)$.

This posterior distribution is a *probabilistic* summary of our knowledge about the parameters given the data. It allows us to make probabilistic statements about the parameters such as follows.

- *Given the data*, the probability for θ to be in (0.7,0.8) is ....
- *Given the data*, the expected value of θ is ....

A conditional perspective versus the sampling approach, an unconditional perspective.

Choosing prior distribution to represent prior knowledge

prior elicitation { religious Bayesian / Empirical Bayesian / Frequentist

Suppose before the experiment is carried out, from historical background, we think that the actual proportion of supportors is about 0.5, "give or take" 0.1.

- Here by "give or take" I mean that we are willing to assume that prior standard deviation of θ is about 0.1. (Note that this is a probabilistic statement about θ!)

Which α and β values should we choose so that our prior Beta(α, β) will represent this knowledge?

We can choose the values for α and β so that

$$\mu_\theta = \frac{\alpha}{\alpha+\beta} = .5$$

and

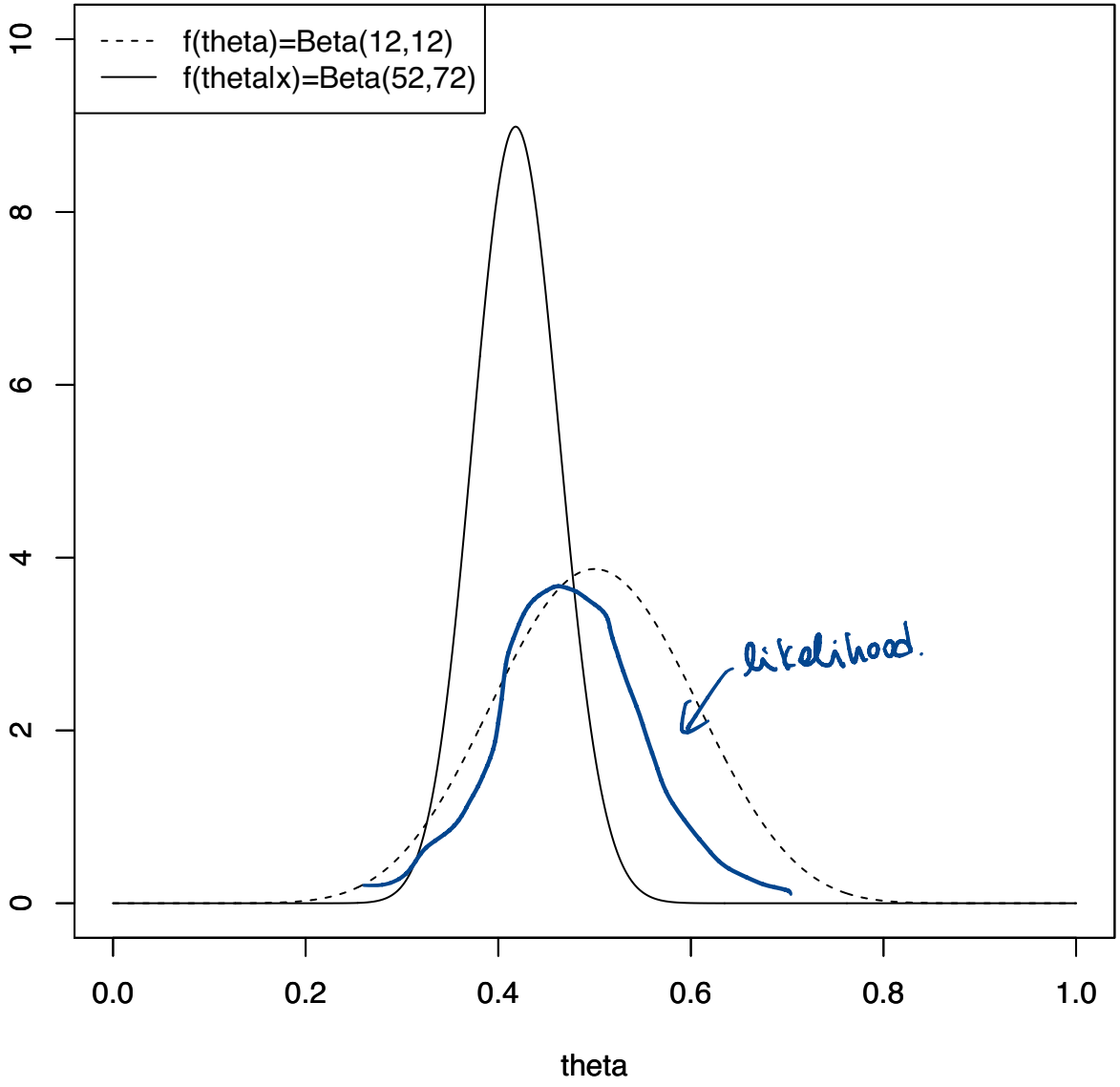
$$\sigma_\theta^2 = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1} = .1^2.$$

Solving these two equations we get

$$\alpha = \beta = 12.$$

Therefore we choose Beta(12,12) distribution as our prior distribution for θ.

From prior to posterior



Common devices for drawing inference from the posterior

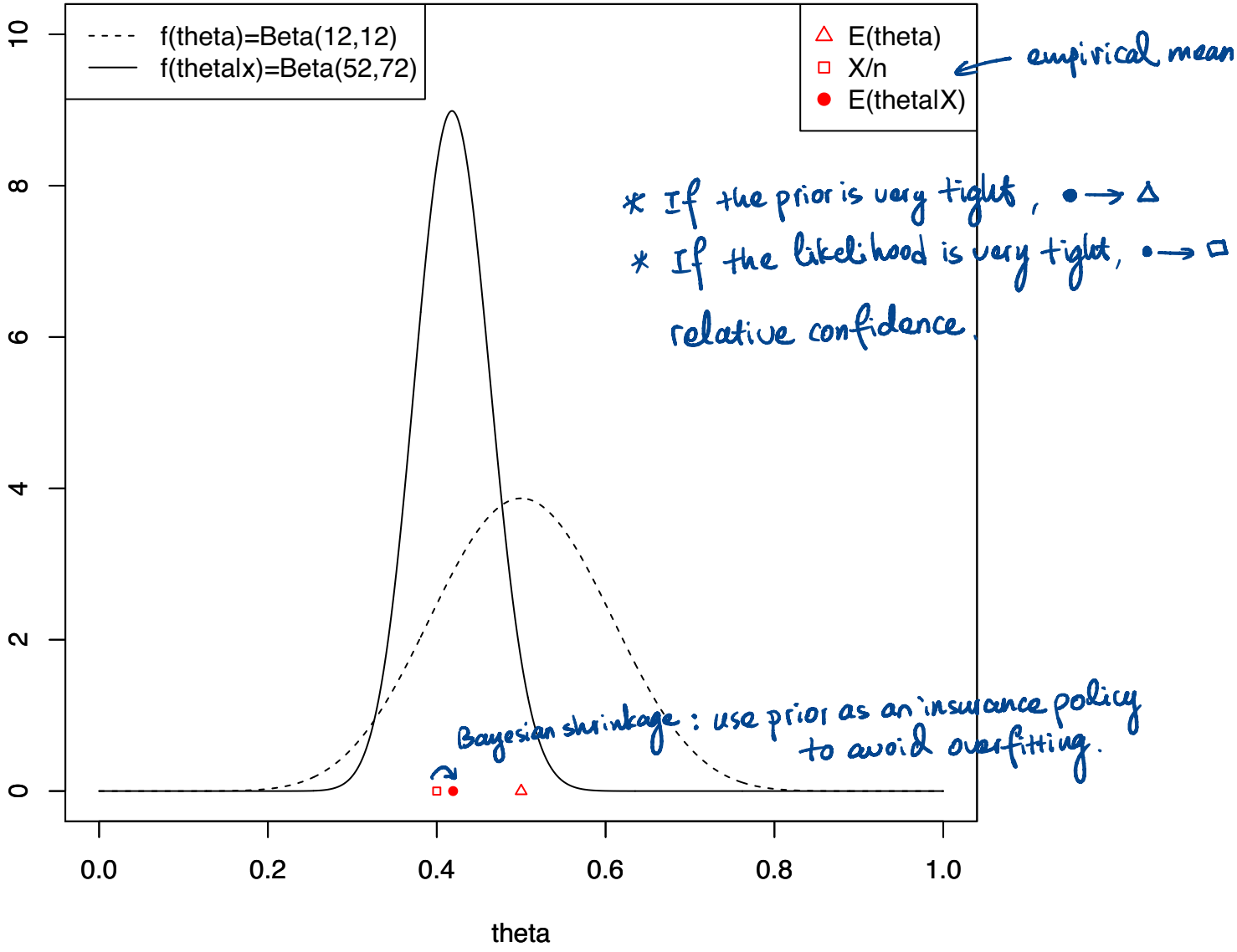
- Posterior summaries: e.g., posterior mean, mode, median, other quantiles, and higher moments.
  - Posterior mean and variance are examples.
- Posterior *credible intervals (or regions)*. equal-tail, HPD.
  - Intervals (or regions) in the parameter space such that the posterior probability for θ to be within is at a given level.
  - Contrast this with classical *confidence intervals*.
- Posterior *predictive distribution*.
  - The conditional distribution of a future observation given past observations, *marginalizing out* all unknown parameters.
  - It takes into account the uncertainty in our knowledge about the model in making predictions.
  - This is a unique Bayesian device, unavailable in the frequentist setting. Why?
- Making "optimal" decisions in a formal *decision-theoretic* framework. (More on this later.)

Using posterior summaries

- Characterize key features of the posterior distribution.
- In the political poll example:
  - Posterior mean: the average of the distribution is shifted.
  - Posterior variance: the spread of the distribution is also changed.

Let us quantify these two changes relative to the prior summaries.

From prior to posterior



Before observing data, the prior distribution is Beta(α, β). So

$$\mathrm{E}(\theta) = \frac{\alpha}{\alpha+\beta} = \mu_\theta$$

$$\mathrm{Var}(\theta) = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}.$$

After observing the data, $X = x$, the posterior distribution is Beta$(\alpha + x, \beta + n - x)$, and so

$$\mathrm{E}(\theta|X=x) = \frac{\alpha+x}{\alpha+\beta+n} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\cdot\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\cdot\frac{x}{n}$$

$$= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n} = \mu_{\theta|x}$$

and

$$\mathrm{Var}(\theta|X=x) = \frac{\mu_{\theta|x}(1-\mu_{\theta|x})}{\alpha+\beta+n+1}.$$

An interpretation of the posterior mean

* when $\alpha+\beta \to \infty$, $\mathbb{E}(\theta|X=x) \to \mu_\theta$
* when $n \to \infty$, $\mathbb{E}(\theta|X=x) \to \frac{x}{n}$

$$\mathrm{E}(\theta|X=x) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n}$$

- This is a weighted average of the prior mean μ_θ, and the observed average x/n.
- This is a compromise between our *prior expectation* (when there is no data) and the *observed average*.
- When we have a lot of data, i.e. n is very large compared to $\alpha+\beta$, the observed average x/n dominates $\mathrm{E}(\theta|X=x) \approx x/n$.

An interpretation of the posterior mean

prior #(successes)
prior #(failures)
Beta(α, β)
Beta($\alpha+x$, $\beta+n-x$)

$$\mathrm{E}(\theta|X=x) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n}$$

- When $\alpha+\beta$ is much larger than n, then the prior expectation μ_θ donminates $\mathrm{E}(\theta|X=x) \approx \mu_\theta$.
- $\alpha+\beta$ is the prior "sample size"—it measures the amount of information we have about θ.
- n is the observed sample size—it measures the amount of information we have in the data.

As for the variance

Before the data are observed,

$$\mathrm{Var}(\theta) = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}$$

- For the same μ_θ, the larger $\alpha+\beta$ is, the smaller $\mathrm{Var}(\theta)$ is—we are more certain about the value of θ.

After the data are observed,

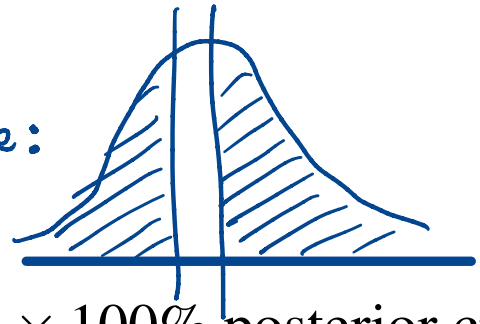
$$\mathrm{Var}(\theta|X=x) = \frac{\mu_{\theta|x}(1-\mu_{\theta|x})}{\alpha+\beta+n+1}.$$

either confident $(\alpha+\beta)$, or big sample set would decrease Var.

- For the same $\mu_{\theta|x}$, the larger $\alpha+\beta+n$—total amount of information, the smaller $\mathrm{Var}(\theta|X=x)$.

Credible intervals

could be:



- More generally, a level $(1-\alpha)\times 100\%$ posterior credible region is a subset $\mathscr{R}(\boldsymbol{x})$ in the parameter space such that

$$\mathrm{P}(\theta \in \mathscr{R}(\boldsymbol{x}) \,|\, \mathbf{X} = \boldsymbol{x}) = 1-\alpha.$$

the idea is general could be a region.

- The notation $\mathscr{R}(\boldsymbol{x})$ is to emphasize the fact that in practice usually the CR is chosen based on the observed data $\boldsymbol{x}$.
- In particular, when θ is one-dimensional, it is often convenient to choose $\mathscr{R}(\boldsymbol{x})$ to be an interval $[l(\boldsymbol{x}), u(\boldsymbol{x})]$, in which case, the lower and upper bounds are such that

$$\mathrm{P}(l(\boldsymbol{x}) \le \theta \le u(\boldsymbol{x}) \,|\, \mathbf{X} = \boldsymbol{x}) = 1-\alpha.$$

Here x is data not random variable. The only randomness is in θ.

Confidence interval: $P(l(\underline{X}) \le \theta \le u(\underline{X}) \,|\, \theta) = 1-\alpha$. The only randomness in $\underline{X}$.

↑ repeated experiment.

Choices of credible regions

Under repeated experiments, credible interval → confidence interval.

- $\mathscr{R}(\boldsymbol{x})$ are not unique. We need additional constraints to pin it down.
- Some common choices include:
  - *Highest posterior density (HPD)* regions. Intuitively, this is the "smallest" CR to achieve a given coverage probability.
    - The additional constraint is for any $\theta \in \mathscr{R}$ and $\theta' \notin \mathscr{R}$, the posterior density at θ and θ' satisfies

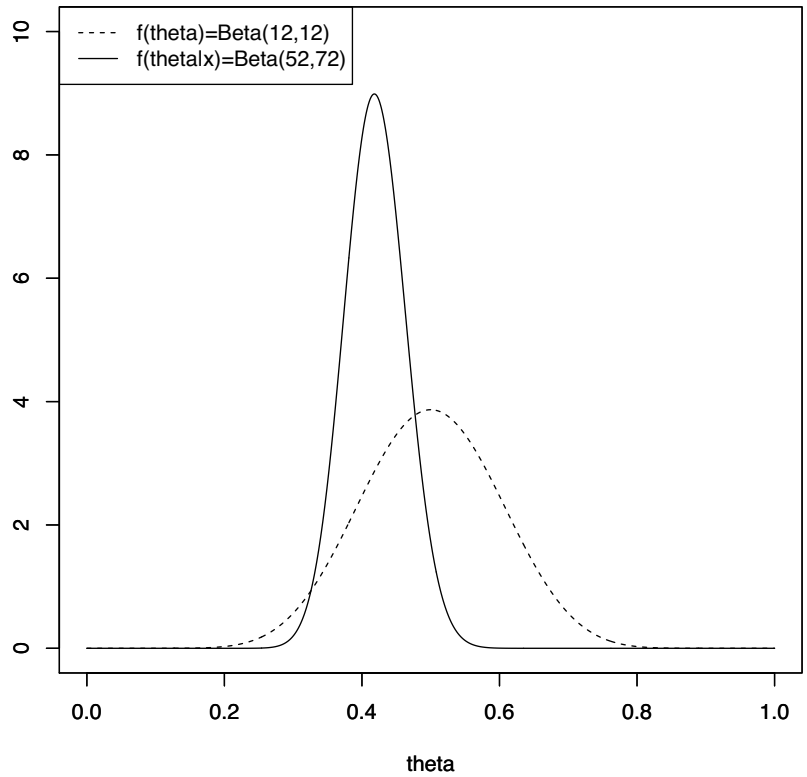
$$p(\theta \,|\, \mathbf{X} = \boldsymbol{x}) \geq p(\theta' \,|\, \mathbf{X} = \boldsymbol{x}).$$

  - *Equal-tailed credible intervals* (for one-dim parameters), aka *quantile-based credible intervals*.
    - Simply given by the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the posterior.
    - In other words, the additional constraint is

$$P(\theta < l(\boldsymbol{x}) \,|\, \mathbf{X} = \boldsymbol{x}) = P(\theta > u(\boldsymbol{x}) \,|\, \mathbf{X} = \boldsymbol{x}) = \alpha/2.$$

Example: A political poll

- In our political poll example, what is 95% HPD credible interval (CI) and 95% equal-tailed CI for θ?



Predictive distribution Bayesian Model Average.

- Now suppose we are interested in *predicting* the value of a future observation X_{n+1}, given n observations $\mathbf{X}_n$.
- What would a frequentist do?
  - Predict using the conditional distribution of X_{n+1} given $\mathbf{X}_n$ and a point estimate for $\hat{\theta}$ for θ:

$$\mathrm{P}(X_{n+1} \,|\, \mathbf{X}_n, \hat{\theta})$$

  - Uncertainty quantification is based on repeated experiments—each giving a different *training set* $\mathbf{X}_n$ and $\hat{\theta}$.
- The Bayesian approach:
  - Consider $(X_{n+1}, \mathbf{X}_n, \theta)$ as a joint random vector.
  - Use the conditional distribution of X_{n+1} given $\mathbf{X}_n$ for prediction

$$\begin{aligned}
\mathrm{P}(X_{n+1} = x \,|\, \mathbf{X}_n) &= \int \mathrm{P}(X_{n+1} = x, \theta \,|\, \mathbf{X}_n) d\theta \\
&= \int \mathrm{P}(X_{n+1} = x \,|\, \theta, \mathbf{X}_n) p(\theta \,|\, \mathbf{X}_n) d\theta \\
&= \int \mathrm{P}(X_{n+1} = x \,|\, \theta) p(\theta \,|\, \mathbf{X}_n) d\theta.
\end{aligned}$$

$X_1, \ldots, X_n, X_{n+1} \,|\, \theta \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(\theta)$ — i.i.d. — This can be any distri.

  - Uncertainty quantification is *conditional on* the training set $\mathbf{X}_n$.

Example: a political poll

frequentist : $p(X_{n+1} = x \mid \hat{\theta})$. plug in one point estimate
Bayesian : Model average. taking weights as $p(\theta \mid X_n)$.

- An additional person polled will respond $X_{n+1} \in \{0, 1\}$.
- Its predictive distribution is given by

$$\begin{aligned} \mathrm{P}(X_{n+1} = 1 \mid \mathbf{X}_n) &= \int \mathrm{P}(X_{n+1} = x \mid \theta) p(\theta \mid \mathbf{X}_n) d\theta \\ &= \int \theta \, p(\theta \mid \mathbf{X}_n) d\theta \\ &= \mathrm{E}(\theta \mid \mathbf{X}_n) = \frac{\alpha + x_n}{\alpha + \beta + n} = \frac{52}{52 + 72}. \end{aligned}$$

  In this case, the predictive probability for a future observation to be 1 happens to be equal to the posterior mean of θ. This is a feature of the binomial sampling model.

- Exercise: What's the predictive distribution of a future pair of observations (X_{n+1}, X_{n+2}).