Predictive Distributions

STAT 251, Prior-/Posterior- Predictive Distributions, Part 1

Overview

Goals in Statistics

Examples of Prior Predictive Distributions

Posterior Predictive Distributions

Final Remarks

Goals in Statistics

Two prevalent goals

Inference on parameters (population-level characteristics, such as the expected value of a response variable—μ—or the standard deviation—σ.)

Predictions of future data

The former (inference on parameters) relies on having a **prior distribution for the parameters** which is updated based upon observed data to form a **posterior distribution for the parameters**.

The latter (prediction of future data, y's) relies on blending the (prior/posterior) beliefs about the parameters with the assumed conditional distribution of the data to obtain the **(prior/posterior) predictive distribution**.

Predictive Distributions

The prior predictive distribution of a new y value, y_{new}, is denoted as $f(y_{new})$. The posterior predictive distribution of a new y value, y_{new}, given observed y's, $\boldsymbol{y}_{obs}$, is denoted as $f(y_{new}|\boldsymbol{y}_{obs})$

If θ is given a continuous prior distribution (e.g., Normal, Gamma, Inverse Gamma, Beta), then

Prior predictive distribution: $f(y_{new}) = \int_{-\infty}^{\infty} f(y_{new}|\theta)\pi(\theta)d\theta$

Posterior predictive distribution:

$f(y_{new}|\boldsymbol{y}_{obs}) = \int_{-\infty}^{\infty} f(y_{new}|\boldsymbol{y}_{obs}, \theta)\pi(\theta|\boldsymbol{y}_{obs})d\theta$

If θ is given a discrete prior distribution, then replace $\int$ with $\sum$.

Prior Predictive Distributions

If θ is given a continuous prior distribution (e.g., Normal, Gamma, Inverse Gamma, Beta), then

Prior predictive distribution: $f(y_{new}) = \int_{-\infty}^{\infty} f(y_{new}|\theta)\pi(\theta)d\theta$

If θ is given a discrete prior distribution (much less common), then

Prior predictive distribution: $f(y_{new}) = \sum_{\theta_k} f(y_{new}|\theta_k)\pi(\theta_k)$, where all possible θ_k values are considered.

Whether θ has a continuous or a discrete prior, this is interpreted as a weighted average of the *likelihood* function, where the weight is from the *prior* beliefs about θ. Different prior distributions for θ would yield different prior predictive distributions.

Connection between Marginal Likelihood and Prior Predictive Distribution

Recall that the marginal likelihood of an observed value of y, i.e. $f(y_{obs})$, is constructed as* $f(y_{obs}) = \int_{-\infty}^{\infty} f(y_{obs}|\theta)\pi(\theta)d\theta$.

However, the prior predictive distribution of a new y, i.e. $f(y_{new})$, is constructed as* $f(y_{new}) = \int_{-\infty}^{\infty} f(y_{new}|\theta)\pi(\theta)d\theta$

The functional forms of the marginal likelihood and the prior predictive distribution are IDENTICAL. The only difference is whether we are evaluating the functional form on the observed data or on potential future data.

* If θ has a discrete distribution, replace the integral with the sum

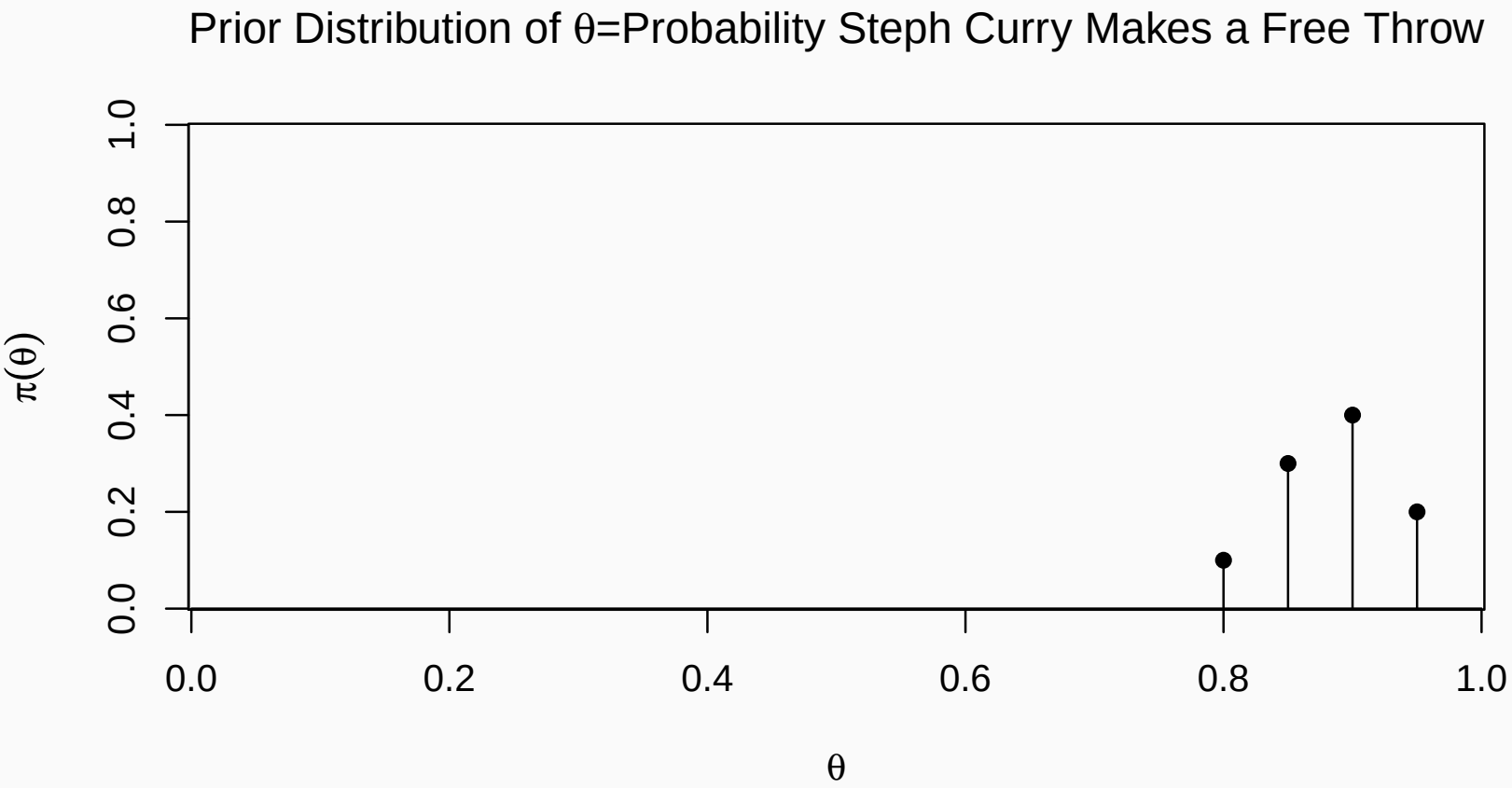
Examples of Prior Predictive Distributions

Toy Example with DISCRETE PRIOR

Consider Steph Curry's current ability to make free throws. First, we'll (unrealistically) have a **discrete prior** on $\theta \equiv$ Probability of making a free throw. For pedagogical purposes, I assume this prior:

$$\pi(\theta) = \begin{cases} 0.1, & \text{if } \theta = 0.80 \\ 0.3, & \text{if } \theta = 0.85 \\ 0.4, & \text{if } \theta = 0.90 \\ 0.2, & \text{if } \theta = 0.95 \\ 0, & \text{otherwise} \end{cases}$$

Toy Example with DISCRETE PRIOR (cont)



Toy Example with DISCRETE PRIOR (cont)

To get prior predictive distribution of y_{new} if n_{new}=1 free throw attempt is made, we calculate

$$
\begin{aligned}
f(y_{new}) &= \sum_{\theta_k \in \{0.8,0.85,.9,.95\}} f(y_{new}|\theta_k)\pi(\theta_k) \\
&= \sum_{\theta_k \in \{0.8,0.85,.9,.95\}} \binom{1}{y_{new}} \theta_k^{y_{new}} (1-\theta_k)^{1-y_{new}} \mathbb{1}_{(y_{new} \in \{0,1\})} \pi(\theta_k) \\
&= \begin{cases} .115, & \text{if } y_{new} = 0 \\ .885, & \text{if } y_{new} = 1 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

Toy Example with DISCRETE PRIOR (cont), Decomposing the prior predictive distribution

We assume (conditional on θ) that $Y_{new}|\theta \sim Binomial(n_{new}, \theta)$.

We have expressed the prior uncertainty that θ has each of these values via the **prior distribution**. (10% chance that $\theta = .8$, 30% chance that $\theta = .85$, etc.)

If $\theta = 0.80$ and $n_{new} = 1$, then $f(y_{new} = 1|\theta = 0.80) = 0.80$.

If $\theta = 0.85$ and $n_{new} = 1$, then $f(y_{new} = 1|\theta = 0.85) = 0.85$.

If $\theta = 0.90$ and $n_{new} = 1$, then $f(y_{new} = 1|\theta = 0.90) = 0.90$.

If $\theta = 0.95$ and $n_{new} = 1$, then $f(y_{new} = 1|\theta = 0.95) = 0.95$.

We can now use the law of total probability.

Toy Example with DISCRETE PRIOR (cont), Decomposing the prior predictive distribution

$$P(y_{new} = 1) = P(y_{new} = 1 \cap \theta = 0.80) + P(y_{new=1} \cap \theta = .85) + P(y_{new=1} \cap \theta = .9) + P(y_{new=1} \cap \theta = .95)$$

This equals

$$P(y_{new} = 1) = P(y_{new} = 1|\theta = 0.80)P(\theta = .8) + P(y_{new} = 1|\theta = .85)P(\theta = .85) + P(y_{new} = 1|\theta = .9)P(\theta = .9) + P(y_{new} = 1|\theta = .95)P(\theta = .95) = .8(.1) + .85(.3) + .9(.4) + .95(.2) = 0.885$$

Similarly,

$$P(y_{new} = 0) = P(y_{new} = 0|\theta = 0.80)P(\theta = .8) + P(y_{new} = 0|\theta = .85)P(\theta = .85) + P(y_{new} = 0|\theta = .9)P(\theta = .9) + P(y_{new} = 0|\theta = .95)P(\theta = .95) = .2(.1) + .15(.3) + .1(.4) + .05(.2) = 0.115$$

Toy Example with DISCRETE PRIOR (cont)

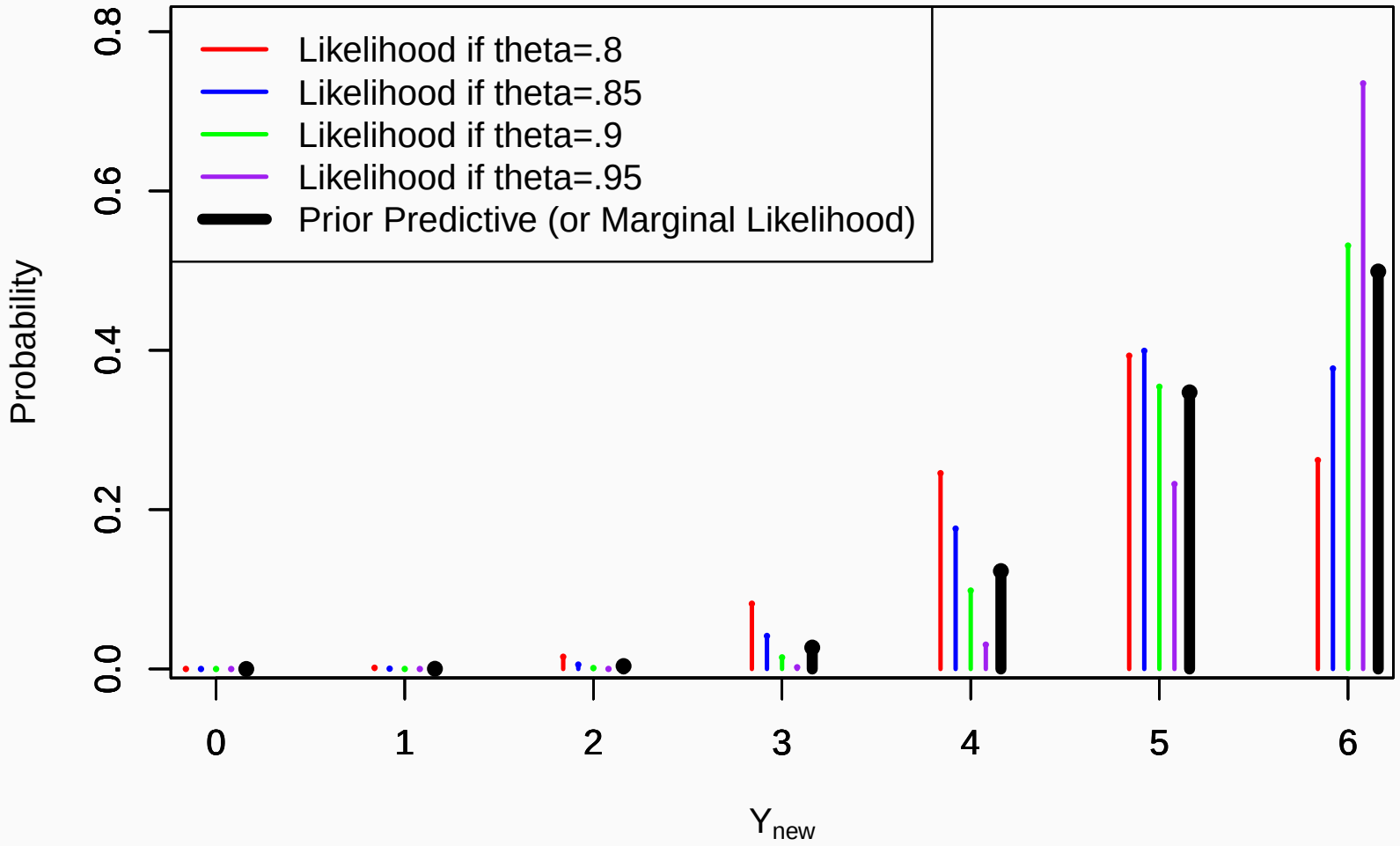
What if we want the prior predictive distribution of y_{new} assuming that Curry will shoot $n_{new} = 6$ free throws? To get prior predictive distribution of y_{new} if n_{new}=6 free throw attempts are made, we calculate

$$f(y_{new}) = \sum_{\theta_k \in \{0.8,0.85,.9,.95\}} f(y_{new}|\theta_k)\pi(\theta_k)$$

$$= \sum_{\theta_k \in \{0.8,0.85,.9,.95\}} \binom{6}{y_{new}} \theta_k^{y_{new}} (1-\theta_k)^{6-y_{new}} \mathbb{1}_{(y_{new} \in \{0,1,\ldots,6\})} \pi(\theta_k)$$

Consider the following plots of the likelihood for each given value of θ, and the marginal likelihood that results from weighting each likelihood relative to the prior beliefs re θ.

Likelihood and Prior Predictive Distributions for $Y_{new} \equiv$ number of made free throws in $n_{new} = 6$ attempts



Per the prior predictive distribution, what is more likely: that Steph would make 5 out of 6 or that he would make 6 out of 6?

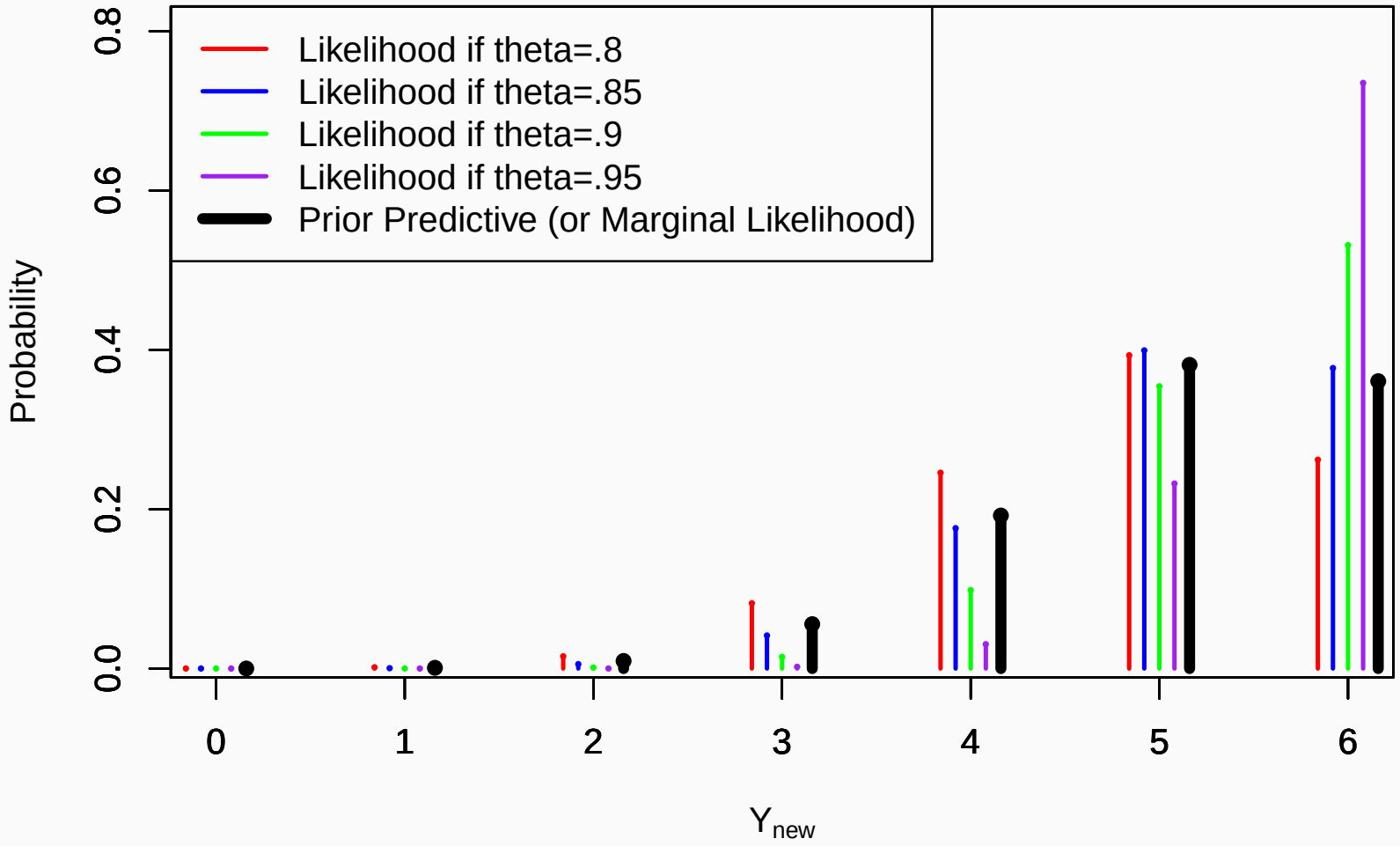
Toy Example with DISCRETE PRIOR (cont)

Consider the same situation (how many free throws would Steph make in 6 attempts), but now suppose that the prior distribution for θ had been:

$$\pi(\theta) = \begin{cases} 0.5, & \text{if } \theta = 0.80 \\ 0.3, & \text{if } \theta = 0.85 \\ 0.15, & \text{if } \theta = 0.90 \\ 0.05, & \text{if } \theta = 0.95 \\ 0, & \text{otherwise} \end{cases}$$

What would the prior predictive distribution be?

Likelihood and Prior Predictive Distributions for $Y_{new} \equiv$ number of made free throws in $n_{new} = 6$ attempts when using second choice for prior



Per **this** prior predictive distribution, which y_{new} is most likely?

Key Considerations

Prior predictive distribution is a weighted average of the (conditional) likelihoods for each possible θ.

Weights are from the prior belief about θ.

If you change the prior distribution of θ, you change the prior predictive distribution.

The prior predictive distribution can serve as a sanity check for the appropriateness of your prior distribution. If the prior and/or likelihood do not match what you actually believe, then the prior predictive distribution might not match the data you would expect to observe.

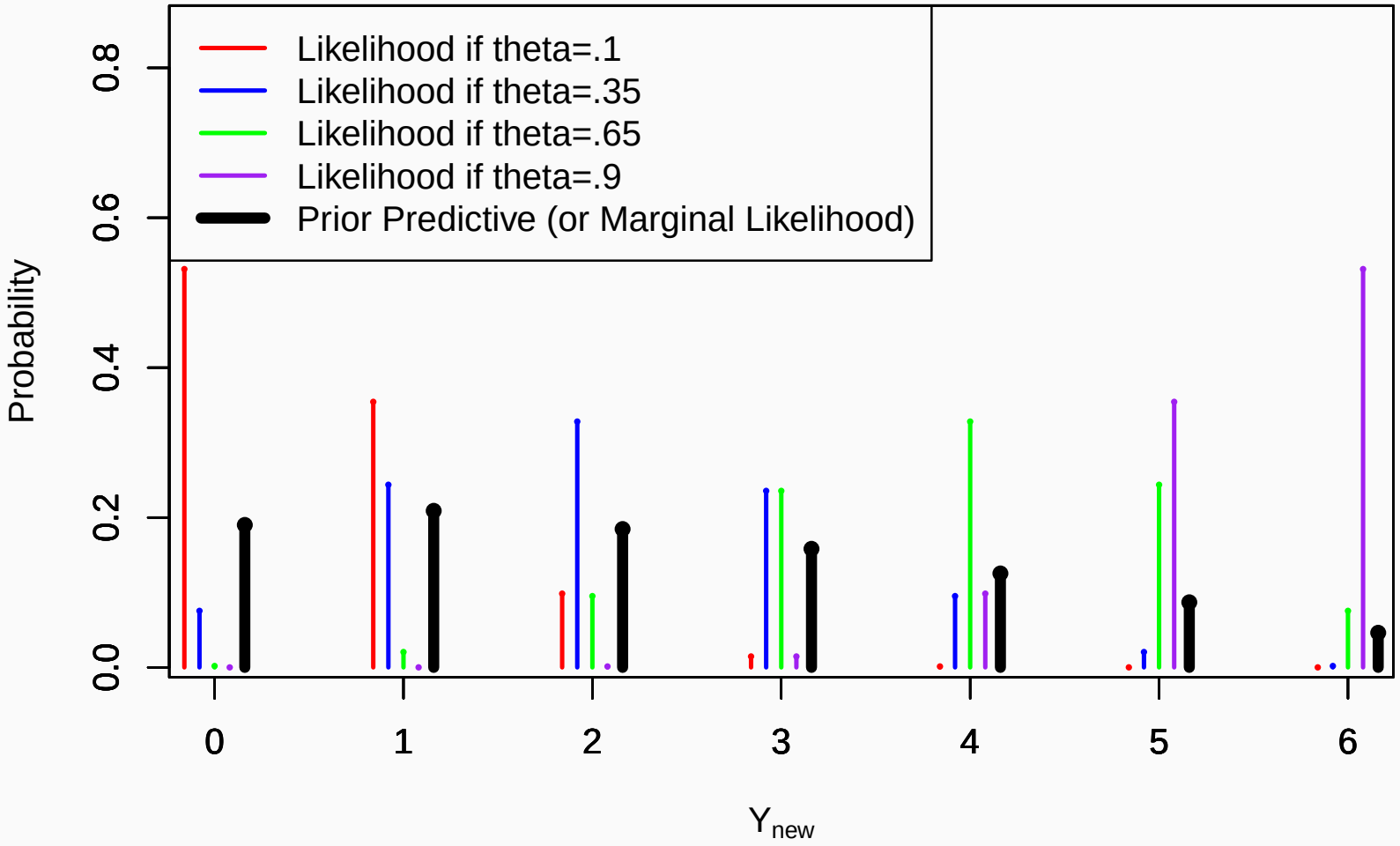
Example:

I would be shocked if Steph Curry made only 1 out of 6 free throw attempts. But consider the implication if I had assumed the following prior distribution for θ:

$$\pi(\theta) = \begin{cases} 0.30, & \text{if } \theta = 0.10 \\ 0.40, & \text{if } \theta = 0.35 \\ 0.25, & \text{if } \theta = 0.65 \\ 0.05, & \text{if } \theta = 0.90 \\ 0, & \text{otherwise} \end{cases}$$

What would the prior predictive distribution be?

Likelihood and Prior Predictive Distributions for $Y_{new} \equiv$ number of made free throws in $n_{new} = 6$ attempts when using a prior that doesn't reflect my beliefs



Per **this** prior predictive distribution, would $y_{new} = 1$ be shocking?

Objective Priors

The prior predictive distribution is often not indicative of the data we would expect to see if we have used an *objective prior*.

In fact, it is possible that the prior predictive distribution does not even exist (if the prior is too dispersed). (That is because $\int f(y_{new}|\theta)\pi(\theta)d\theta$ may not be finite.) This is called an *improper* distribution–it cannot be a valid statistical distribution if its integral over its entire support is not finite.

> Prior predictive distributions are generally not meaningful (and might not even exist) if the prior distribution of θ is *objective* (i.e., noninformative, or representing a high degree of vagueness in beliefs about what the parameter(s) might equal).

Now Example in Beta-Binomial Setting

Prior Predictive Distribution for y_{new} when
$y_{new}|\theta \sim Binomial(n_{new}, \theta)$ and when $\theta \sim Beta(a, b)$:

$$f(y_{new}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \binom{n_{new}}{y_{new}} \frac{\Gamma(a+y_{new})\Gamma(b+(n_{new}-y_{new}))}{\Gamma(a+b+n_{new})}$$
$$\times \quad \mathbb{1}_{(y_{new} \in \{0,1,\ldots,n_{new}\})}$$

Posterior Predictive Distributions

Posterior Predictive Distributions

The posterior predictive distribution of a new y value, y_{new}, given observed data, $\mathbf{y}_{obs}$, is denoted as $f(y_{new}|\mathbf{y_{obs}})$

If θ has a continuous posterior distribution*, then the posterior predictive distribution is

$$f(y_{new}|\mathbf{y_{obs}}) = \int_{-\infty}^{\infty} f(y_{new}|\mathbf{y_{obs}}, \theta)\pi(\theta|\mathbf{y_{obs}})d\theta.$$

When the data are assumed to be conditionally iid (as we have usually assumed in this course), then this simplifies slightly to

$$f(y_{new}|\mathbf{y_{obs}}) = \int_{-\infty}^{\infty} f(y_{new}|\theta)\pi(\theta|\mathbf{y_{obs}})d\theta.$$

*If θ has a discrete posterior distribution, then replace $\int$ with $\sum$.

Prior vs. Posterior Predictive Distributions

When the data are conditionally iid, the only difference in deriving the **posterior predictive distribution** (compared to deriving the **prior predictive distribution**) is that we are now weighting the likelihood based on the **posterior beliefs** about θ.

$$f(y_{new}) = \int_{-\infty}^{\infty} f(y_{new}|\theta)\pi(\theta)d\theta.$$

$$f(y_{new}|\mathbf{y_{obs}}) = \int_{-\infty}^{\infty} f(y_{new}|\theta)\pi(\theta|\mathbf{y_{obs}})d\theta.$$

Derivation

Sometimes it is straightforward to derive the prior/posterior predictive distribution. Consider the following scenario:

- Let y_{obs} represent the number of free throws that were made in n_{obs} (observed) attempts.

Derivation

Sometimes it is straightforward to derive the prior/posterior predictive distribution. Consider the following scenario:

- Let y_{obs} represent the number of free throws that were made in n_{obs} (observed) attempts.
- Let y_{new} represent the number of free throws that will be made in n_{new} (future) attempts.

Derivation

Sometimes it is straightforward to derive the prior/posterior predictive distribution. Consider the following scenario:

- Let y_{obs} represent the number of free throws that were made in n_{obs} (observed) attempts.
- Let y_{new} represent the number of free throws that will be made in n_{new} (future) attempts.
- Assume that given θ, every free throw is *iid* with probability θ of success. This implies, $y_{new}|\theta \sim Binomial(n_{new}, \theta)$ and that $y_{obs}|\theta \sim Binomial(n_{obs}, \theta)$, and that $y_{new}|\theta$ is independent of $y_{obs}|\theta$.

Derivation

Sometimes it is straightforward to derive the prior/posterior predictive distribution. Consider the following scenario:

- Let y_{obs} represent the number of free throws that were made in n_{obs} (observed) attempts.
- Let y_{new} represent the number of free throws that will be made in n_{new} (future) attempts.
- Assume that given θ, every free throw is *iid* with probability θ of success. This implies, $y_{new}|\theta \sim Binomial(n_{new}, \theta)$ and that $y_{obs}|\theta \sim Binomial(n_{obs}, \theta)$, and that $y_{new}|\theta$ is independent of $y_{obs}|\theta$.

Which would be a realistic family of priors to employ for $\pi(\theta)$? The Beta family!

Example in Beta-Binomial Setting

Because the posterior distribution of $\theta|y_{obs}$ is the $Beta(a^\star = a + y_{obs}, b^\star = b + (n_{obs} - y_{obs}))$ distribution, it follows that the posterior predictive distribution of $y_{new}|y_{obs}$ is

$$f(y_{new}|y_{obs}) = \frac{\Gamma(a^\star + b^\star)}{\Gamma(a^\star)\Gamma(b^\star)} \binom{n_{new}}{y_{new}} \frac{\Gamma(a^\star + y_{new})\Gamma(b^\star + (n_{new} - y_{new}))}{\Gamma(a^\star + b^\star + n_{new})}$$
$$\times \quad \mathbb{1}_{(y_{new} \in \{0,1,\ldots,n_{new}\})}$$

"Today's posterior [of θ] is tomorrow's prior" $\Rightarrow$ "Today's posterior predictive distribution [of y_{new}] is tomorrow's prior predictive distribution."

Example of prior, posterior predictive distributions

Let's consider a familiar scenario: the beta-binomial distribution.

We'll consider Steph Curry's ability to make free throws in NBA basketball games this season. We'll only use data from the 2016-2017 season. We'll look at the prior-predictive and posterior-predictive distributions for total made free throws in games 16–18 (y_{new}) where it is assumed we know he will shoot $n_{new} = 19$ attempts. For the prior-predictive, we will ignore the data from the first 15 games, whereas for the posterior-predictive, we'll use his first 15 games as our data, holding out his next three games to see how well our model would have predicted.

Notation

- y_{obs} is the number of free throws that Curry made in the first 15 games of the season.
- n_{obs} is the number of free throws that Curry attempted in those 15 games.
- θ, the parameter of interest, represents the probability he makes any given free throw attempt.

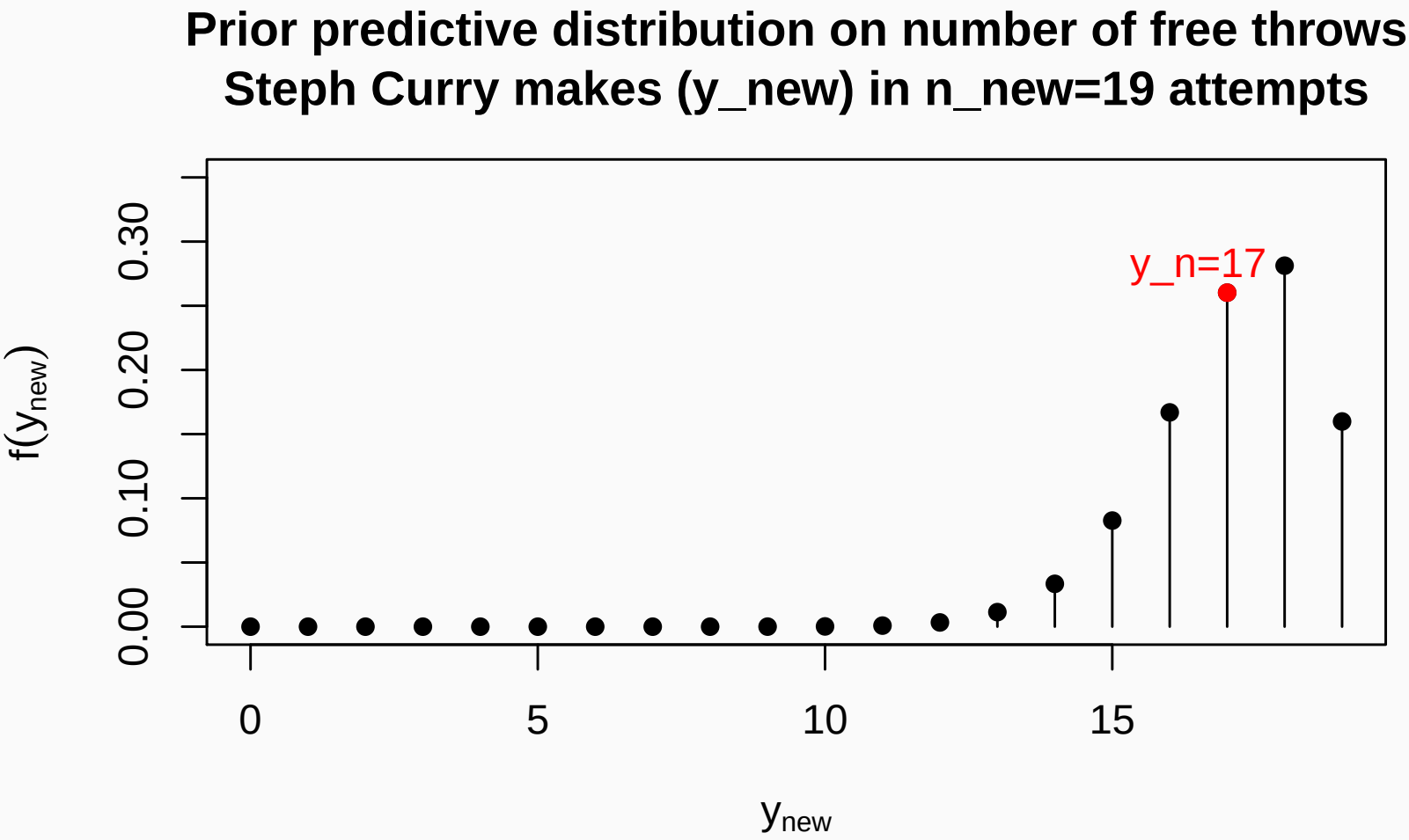
Model Specification

- Likelihood: $y_{obs}|\theta \sim Binomial(n_{obs}, \theta)$
- Prior: $\theta \sim Beta(a = 90, b = 10)$. (Strong prior belief that θ is close to 0.9.)
- Data*: $y_{obs} = 71$, $n_{obs} = 77$.
- Posterior Distribution: $\theta|y_{obs} \sim Beta(161, 16)$.
- n_{obs} is the number of free throws that Curry attempted in those 15 games.
- θ, the parameter of interest, represents the probability he makes any given free throw attempt.

In games 16–18, Steph Curry shot $n_{new} = 19$ free throws.*.

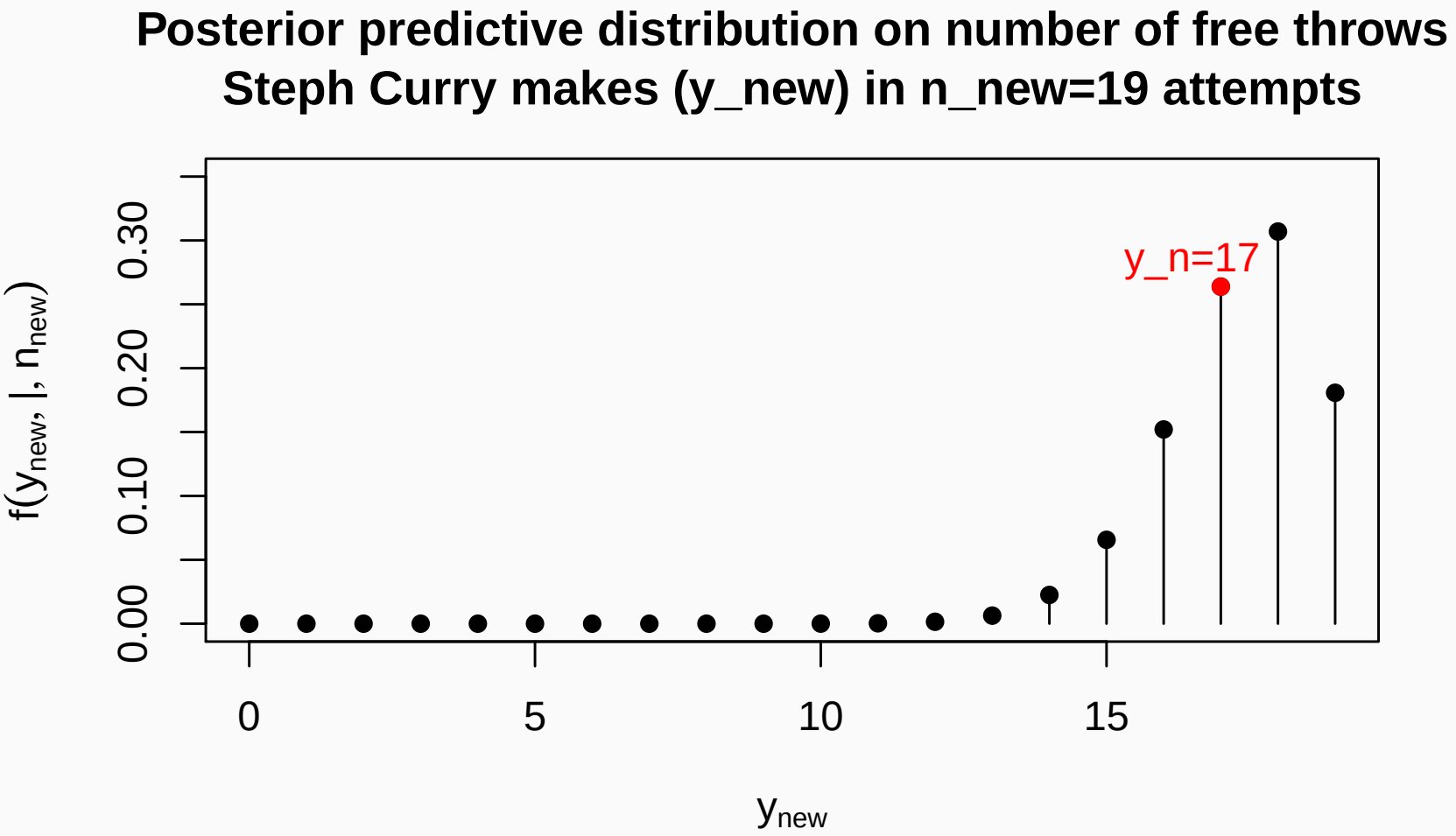
* Source:
`http://www.espn.com/nba/player/gamelog/_/id/3975/stephen-curry`,
accessed 29 November 2016

Prior Predictive Distribution



Curry actually made $y_{new} = 17$ of the $n_{new} = 19$ attempts. This result is certainly not inconsistent with what we would have expected per the **prior predictive distribution**.

Posterior Predictive Distribution



Curry actually made $y_{new} = 17$ of the $n_{new} = 19$ attempts. This result is certainly not inconsistent with what we would have expected per the **posterior predictive distribution**, either.

Comments

In the Free Throw Example, we can (and should) get the exact posterior predictive distribution analytically; Monte Carlo is NOT needed.

But in other settings, the exact derivation of the posterior predictive distribution is not always readily available.

How could we get a Monte Carlo estimate of the posterior predictive?

Monte Carlo Estimation of Posterior Predictive Distributions

Sometimes it is helpful to create a Monte-Carlo estimate of the posterior predictive distribution. The idea is that we simulate a (large) sample from the desired distribution. This is done in two stages:

1. Get a large sample of θ values from the posterior distribution, $\pi(\theta|y_{obs})$. Denote the jth value in the sample by $\theta^{(j)}$.
2. For each $\theta^{(j)}$, sample a value for y_{new} from the (conditional) distribution of y_{new}. That is, sample $y_{new}^{(j)}$ from $f(y_{new}|\theta = \theta^{(j)})$.

How could we get a Monte Carlo estimate of the prior predictive?

Final Remarks

Final Remarks, Part 1

The posterior predictive distribution can be analytically derived or estimated using Monte Carlo techniques. (Same goes for the prior predictive distribution)

In the Free Throw Example, we can (and therefore should) get the exact posterior predictive distribution analytically: Monte Carlo is NOT needed. But the exact derivation of the posterior predictive distribution is not always readily available.

Final Remarks, Part 2

The **posterior predictive distribution** is to predict a future **y**, whereas the **posterior distribution** is to express uncertainty in the **parameter(s)**. They both are conditional on observed data (i.e., $\mathbf{y}_{obs}$).

Final Remarks, Part 3

The **prior predictive distribution** may be used to check that our prior distribution for the parameter(s) would imply plausible **data** values.

- In my opinion, this is most helpful when we have informative (subjective) prior distributions for model parameters. If the prior distribution is very vague, the prior predictive distribution will generally be very disperse (in fact, it is possible that we don't even have a legitimate prior predictive distribution because it is so disperse, it has an infinite integral! We call this situation an *improper* distribution.)

Final Remarks, Part 4

The **marginal likelihood** (which is the denominator in the formula for the posterior distribution: $Posterior = \frac{Likelihood \times Prior}{Marginal\ Likelihood}$) is the same as the prior predictive distribution evaluated on the observed data (That is, plug in the actual data points, y_{obs}, in place of y_{new}.)

> The marginal likelihood expresses how likely the observed data were according to the prior predictive distribution.

Final Remarks, Part 5

The **posterior predictive distribution** may be used to

- Make predictions on what future data would look like—often, this is intrinsically important.
- Assess how well our model performs: we can make predictions on what future data would look like, and then collect future data and judge how well our predictions actually performed.