Posterior Predictive Distribution

To make *predictions* using θ in our coin tossing example, we can use

$$P(heads \mid D) = \int_\theta P(heads, \theta \mid D) d\theta = \int_\theta P(heads \mid \theta, D) P(\theta \mid D) d\theta = \int_\theta \theta P(\theta \mid D) d\theta$$

Here, we used the fact that we defined $P(heads) = \theta$ and that $P(heads) = P(heads \mid D, \theta)$ (this is only the case for coin flipping - not in general).

In general, the posterior predictive distribution is

$$P(Y \mid D, X) = \int_\theta P(Y, \theta \mid D, X) d\theta = \int_\theta P(Y \mid \theta, D, X) P(\theta|D) d\theta$$

Unfortunately, the above is generally *intractable* in closed form and sampling techniques, such as Monte Carlo approximations, are used to approximate the distribution.

Machine Learning and estimation

In supervised Machine learning you are provided with training data D. You use this data to train a model, represented by its parameters θ. With this model you want to make predictions on a test point x_t.

- **MLE** Prediction: $P(y|x_t;\theta)$ Learning: $\theta = argmax_\theta P(D;\theta)$. Here θ is purely a model parameter.
- **MAP** Prediction: $P(y|x_t,\theta)$ Learning: $\theta = argmax_\theta P(\theta|D) \propto P(D \mid \theta) P(\theta)$. Here θ is a random variable.
- **"True Bayesian"** Prediction: $P(y|x_t, D) = \int_\theta P(y|\theta) P(\theta|D) d\theta$. Here θ is integrated out - our prediction takes all possible models into account.