Part I:
The Bayesian paradigm
• Broadly speaking, the goal of data analysis is to learn about some population using information in a sample

• Typically, only observe a single sample

• What if we had drawn a different sample?
  ★ presumably we would have seen different results
  ★ how different? different in what ways?

• One of the biostatisticians primary tasks is the quantification of uncertainty
  ★ that specifically associated with only observing a sub-sample from the population

• To do this we need help in the form of a formal theoretical framework
  ★ one that is coherent and rigorous
Suppose I present you with a coin and ask: *What is the probability of the result being a head?*

- assuming the coin is ‘fair’, most folks would say: $P(\text{Head}) = 0.5$

Why? Where does the value ‘0.5’ come from?

- answer depends on how they interpret the word ‘probability’

Many schools of thought on how to define probability

The *classical* definition is based on the notion of equally likely outcomes

- symmetry
- e.g, flipping a coin or rolling a die
- intuitive but not straightforward to apply in more general settings
In statistical practice there are two main schools of thought or paradigms:

- frequentist
- Bayesian

What are the differences? In a nutshell ...
- the underlying assumptions
- the use of different types of information
- the interpretation of results

Before we get into details, why should we care?
- specific views on the nature of probability may dictate the approach taken for estimation/inference
- certain criteria/assumptions in a statistical method imply a particular viewpoint on probability
Key points

- Decision between the two paradigms is primarily a *philosophical* one.

- Each has a rigorous mathematical basis and have at their cores two well-known theorems:
  
  (1)

  (2)

- ‘Validity’ of an analysis is application-specific
  
  ★ we never know the ‘truth’ but are the assumptions reasonable?
The frequentist statistical paradigm

• Suppose we started flipping the coin over and over again
  ★ for each instance of this ‘trial’, record the outcome
  ★ observe a sequence similar to: H, H, T, H, T, T, T, H, H, ...

• Define ‘P(Head)’ as the \textit{limiting relative frequency} of the outcome ‘H’ across the sequence of trials

\[ P(\text{Head}) = \lim_{N \to \infty} \frac{\#'\text{Heads'} }{N} \]

★ all trials are assumed to be identical to each other

• Straightforward to generalize this definition to trials for which there are more than two possible outcomes
• How do we make use of this definition of probability in statistical practice?
  ★ when quantifying uncertainty in our analyses

• North Carolina infants dataset
  ★ all caucasian and African-American births between 2003-2004
  ★ NorthCarolina_data.dat on the course website
  ★ includes the following covariates

race  0/1 = Caucasian/African-American
sex  0/1 = female/male
mage mothers’ age at the time of birth, years
smoker mother smoked during the pregnancy, 0/1 = no/yes
gained weight gained by the mother during the pregnancy, lbs
weight birth weight of the baby, kg
weeks gestational period, weeks
dTime time between birth and death, days
  ★ 999 indicates survival to at least 365 days
• Dataset has information on 225,152 births

• First 10 observations in the data:

```r
> load("NorthCarolina_data.dat")
> infants[1:10,]

          race sex mage smoker gained weight weeks dTime
1          0   1    26    0      37   3.232   39  999
2          0   0    32    0      22   3.856   42  999
3          0   1    21    0      40   3.941   41  999
4          0   1    19    1      35   3.203   39  999
5          0   1    19    1      25   3.600   39  999
6          0   0    27    0      27   4.196   39  999
7          0   0    38    0      60   3.487   40  999
8          0   1    24    0      45   3.912   38  999
9          0   1    22    0      46   4.139   39  999
10         0   0    23    0      20   2.835   41  999
```
Suppose interest lies in the mean birth weight among babies born in North Carolina between 2003-2004

\[ \mu \]

Resources are limited in that we can only ‘collect’ \( n = 10 \) birth weights

\[
\text{sample1} <- \text{sample}(\text{infants$weight}, 10)
\]
\[
\text{mean} (\text{sample1})
\]

\[
\text{sample2} <- \text{sample}(\text{infants$weight}, 10)
\]
\[
\text{mean} (\text{sample2})
\]
• Not surprisingly, we get different values of \( \hat{\mu} \) from the two samples

• Repeat this ‘trial’ a large number of times ...

```r
> samples <- matrix(NA, nrow=10000, ncol=10)
> for(i in 1:10000) samples[i,] <- sample(Infants$weight, 10)
> muhat <- apply(samples, 1, mean)
> hist(muhat,
       breaks=seq(from=2, to=4, by=0.1),
       xlab=expression("Estimated mean birth weight, " * hat(mu)),
       ylab="Relative frequency",
       main="",
       col="blue",
       freq=FALSE)
```
Estimated mean birth weight, $\mu$

Relative frequency

0.0 0.5 1.0 1.5 2.0

0 2.0 2.5 3.0 3.5 4.0

BIO 233, Spring 2015
Sampling distributions

- The histogram presents the relative frequencies of the possible values that $\hat{\mu}$ takes on across the sequence of repeated trials
  - here the ‘trial’ is the sampling of $n=10$ infants
  - the ‘outcome’ of the trial is the estimated mean birth weight, $\hat{\mu}$

- The collection of relative frequencies is referred to as the sampling distribution
  - the histogram is an empirical version since $N$ is only 10,000

- The sampling distribution is the basis for quantifying uncertainty in the frequentist paradigm

- All of the standard measures of uncertainty are derived from the sampling distribution
  - standard error, confidence interval and p-value
Interpreting each of these measures of uncertainty requires consideration of the interpretation of the sampling distribution.

⋆ see, for example, the supplementary notes on the course website

Q: How can we characterize the sampling distribution when we only observe a single dataset?

⋆ i.e., a single realization from the hypothetical sequence of trials

• The central limit theorem is the primary tool

⋆ assuming certain regulatory conditions, the sampling distribution converges asymptotically to a Normal distribution

⋆ in ‘small samples’, the story is not so clear

⋆ indication of non-Normality in the example when $n=10$
The Bayesian statistical paradigm

- In many instances, it is difficult to conceptualize a hypothetical sequence of identical trails

- An alternative is to consider what our experience/knowledge tells us
  - arguably, this reflects how people think of ‘probability’ in most circumstances

- We can say that this experience/knowledge forms a ‘belief structure’
  - structure that assigns relative weight or probability to all possible outcomes of the trial

- As people gain experience/knowledge they (somehow) incorporate it into their belief structures
Two key questions:

Q: How do we formally quantify ‘belief’?

Q: How do we formally incorporate additional experience/knowledge?

The Bayesian paradigm provides a coherent framework within which both of these questions are answered.

The basic steps of a Bayesian analysis are:

1. quantify current beliefs via a prior distribution
2. quantify information provided by new data via the likelihood
3. use Bayes’ Theorem to update beliefs and form the posterior distribution

As we’ll see, the posterior distribution is the basis for all statements of statistical estimation and inference.
Parameters as random variables

- Suppose you have a coin and want to learn about $\theta = P(\text{Head})$

- How do we quantify ‘belief’ about the parameter $\theta$?
  - how do we assign relative weight or mass or probability to any given value?

- General idea is to treat $\theta$ as if it were a random variable
  - can then assign $\theta$ a distribution
  - use this distribution to assign relative weight or probability to different values of $\theta$
    - just like we do with any random variable

- Useful to interpret this distribution as a means to characterize uncertainty in our knowledge about $\theta$
• Note, we don’t say that \( \theta \) is a random variable
  ★ the underlying parameter is, of course, constant

• However, treating \( \theta \) as a random variable provides a route to using distributions as a means to assign relative weight/probability

• Very convenient for a number of reasons:
  ★ distributions have nice properties
    ♦ assign non-negative mass to potential values
    ♦ integrate to 1
  ★ we know how to work with distributions
    ♦ calculate moments, quantiles etc
  ★ there are loads of them!
Prior distributions

- Prior distributions encode experience/knowledge \textit{a priori} before seeing the data.

- What would be a reasonable distribution for $\theta$?
  - Has to be a distribution that only assigns mass on the interval $(0,1)$.
  - Ideally, flexible but easy to work with and interpret.

- One choice is the $\text{Beta}(a,b)$ distribution:
  \[
  \pi(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \quad \text{for } \theta \in (0,1)
  \]

- The \textit{hyperparameters} $a$ and $b$ determine the shape of the density function.
  - Different values encode different belief structures.
Reasonably flexible, even though there are only two hyperparameters
permits a wide range of belief structures regarding the value of $\theta$
The likelihood

- Let $f(y|\theta)$ denote the joint pdf or pmf of the sample $Y = (Y_1, \ldots, Y_n)$

- Given $Y = y$ is observed, the function of $\theta$ defined by

$$\mathcal{L}(\theta|y) = f(y|\theta)$$

is the *likelihood function*

- Intuitively, the likelihood encodes all the information about the parameter $\theta$ provided by the data
  
  * loosely interpreted as providing ‘relative weight’ to different values of $\theta$, as dictated by the information in the observed data
Suppose we observed 7 ‘Heads’ on 10 flips

assuming independence between the flips, it’s reasonable to adopt the binomial distribution

\[ \mathcal{L}(\theta | y) = \binom{10}{7} \theta^7 (1 - \theta)^{10 - 7} \]
Informally, we update beliefs by combining the prior and likelihood to obtain the posterior.
• We see that there is some apparent sensitivity to the choice of prior

• Increase to \( n=100 \) flips \( \Rightarrow \) observe 70 ‘Heads’
  
  * likelihood ‘overwhelms’ the prior and there is much less sensitivity
Bayes’ Theorem

- For events $A$ and $B$, assuming $P(A) > 0$, we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Assuming $P(B) > 0$, we can also write

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- In its’ simplest form, Bayes’ Theorem provides the relationship between these two conditional probabilities

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
If the event space for $B$ can be partitioned into $\{B_1, \ldots, B_K\}$, then we can write Bayes’ Theorem as

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A)}$$

$$= \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{K} P(A|B_j)P(B_j)}$$

where the denominator is obtained via an application of the law of total probability.
Bayes’ Theorem

- Let $\theta$ be a parameter of interest and $y$ denote the observed data.

- Let $\pi(\theta)$ denote the prior distribution and $L(\theta|y)$ the likelihood based on the observed data.

- The posterior distribution is given by

$$\pi(\theta|y) = \frac{L(\theta|y)\pi(\theta)}{\int L(\theta|y)\pi(\theta)d\theta}$$

- $y$ corresponds to $A$
- $\theta$ corresponds to $B_k$
- Form of the denominator is obtained as an application of the law of total probability to the marginal density of $y$. 
Once you have the prior and likelihood, Bayes’ Theorem is automatic
- it gives the formula for obtaining the posterior distribution

The posterior is interpreted as representing updated beliefs regarding $\theta$
- updated relative weighting or probability scheme
- updated quantification of uncertainty

- LBW is defined as a birth weight $\leq 2.5$kg.

Let $Y = 0/1$ be a binary indicator of a LBW.

The ‘target’ parameter is $\theta = P(Y = 1)$.

Towards estimating $\theta$, suppose we collect a random sample of $n$ births.

- record LBW status for each birth
- $Y_i = 0/1, i = 1, \ldots, n$
- **Prior distribution:**
  - $\theta$ takes on values on the interval $(0, 1)$
  - consider a Beta($a$, $b$) distribution
    \[
    \pi(\theta|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}
    \]
  - hyperparameters $a$ and $b$ are fixed and conditioned upon throughout

- **Likelihood:**
  - observed data is $y = (y_1, \ldots, y_n)$
  - may be reasonable to adopt the following likelihood:
    \[
    \mathcal{L}(\theta|y) = \prod_{i=1}^{n} \theta^{y_i}(1 - \theta)^{1-y_i}
    \]
  - assumptions?
Posterior distribution:

using Bayes’ Theorem we find that

\[
\pi(\theta|a, b, y) = \frac{\mathcal{L}(\theta|y)\pi(\theta|a, b)}{\int \mathcal{L}(\theta|y)\pi(\theta|a, b)\,d\theta}
\]

\[
= \frac{\prod_{i=1}^{n} \theta y_i (1 - \theta)^{1-y_i} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(a)} \theta^{a-1}(1 - \theta)^{b-1}}{\int_{0}^{1} \prod_{i=1}^{n} \theta y_i (1 - \theta)^{1-y_i} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(a)} \theta^{a-1}(1 - \theta)^{b-1} \,d\theta}
\]

See that this is a function of \( \theta \)

assigns ‘mass’ to values of \( \theta \in (0, 1) \)

In its current form, however, the posterior distribution doesn’t seem to be particularly useful!
Letting $y_+ = \sum y_i$ and collecting terms, the numerator is

$$\mathcal{L}(\theta|y)\pi(\theta) = \theta^{y_+}(1 - \theta)^{n-y_+} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1 - \theta)^{b-1}$$

$$= \mathcal{K} \theta^{y_++a-1}(1 - \theta)^{n-y_++b-1}$$

where $\mathcal{K} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$.

Since $\mathcal{K}$ doesn’t depend on $\theta$ we can write the denominator as

$$\int \mathcal{L}(\theta|y)\pi(\theta|a,b)\partial\theta = \mathcal{K} \int_0^1 \theta^{y_++a-1}(1 - \theta)^{n-y_++b-1}\partial\theta$$
• To evaluate this integral note that for the Beta($a, b$) distribution we have:

\[
\int_0^1 \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1} d\theta = 1
\]

• Equivalently, we can write

\[
\int_0^1 \theta^{a-1}(1 - \theta)^{b-1} d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}.
\]

• Consequently the denominator in the expression for the posterior is

\[
\mathcal{K} \int_0^1 \theta^{y_+ + a - 1}(1 - \theta)^{n - y_+ + b - 1} d\theta = \mathcal{K} \frac{\Gamma(y_+ + a)\Gamma(n - y_+ + b)}{\Gamma(n + a + b)}
\]
• Putting things together, and noting that \( K \) in the numerator and denominator cancel, we have

\[
\pi(\theta|a, b, y) = \frac{\Gamma(n + a + b)}{\Gamma(y_+ + a)\Gamma(n - y_+ + b)} \theta^{y_+ + a - 1}(1 - \theta)^{n - y_+ + b - 1}
\]

• This is the density for a \( \text{Beta}(y_+ + a, n - y_+ + b) \) distribution
  \( \star \) the data has directly updated the belief structure by changing the values of the parameters indexing the Beta distribution

\[
\text{Beta}(a, b) \longrightarrow \text{Beta}(y_+ + a, n - y_+ + b)
\]

  \( \star \) if \( a \) and \( b \) encoded our prior beliefs, then the new values of the parameters encode our ‘new’ belief structure

• Example of \textit{conjugacy}
  \( \star \) arises because of the specific choice of prior and likelihood
Returning to the low birth weight example, suppose we collected a random sample of $n = 500$ births
- from the 225,152 births in the NorthCarolina_data.dat dataset
- observe $y_+ = 45$ low birth weight events

**Prior distribution:**
- $\theta \sim \text{Beta}(40, 460)$
- a prior mean of $\mathbb{E}[\theta] = 0.08$

**Posterior distribution:**
- $\theta \sim \text{Beta}(y_+ + a, n - y_+ + b) \equiv \text{Beta}(85, 915)$
• We update beliefs by combining the prior and likelihood to obtain the posterior

\[ \theta = P(\text{LBW}) \]

• Intuitively, we can see how the posterior is a ‘compromise’ between the two sources of information
The previous example highlights an important feature of Bayes’ Theorem

\[
\pi(\theta|y) = \frac{\mathcal{L}(\theta|y)\pi(\theta)}{\int \mathcal{L}(\theta|y)\pi(\theta) \partial \theta}
\]

★ the denominator is a normalizing constant and not a function of \(\theta\)
★ ensures that the posterior integrates to 1

An equivalent representation of the posterior distribution is therefore given by

\[
\pi(\theta|y) \propto \mathcal{L}(\theta|y)\pi(\theta)
\]

The RHS is referred to as the kernel of the distribution
★ component of the posterior pdf/pmf that solely depends on \(\theta\)
• The kernel of a distribution uniquely identifies the distribution
  ★ why?

• In many instances we can exploit this knowledge to avoid having to calculate the integral in the denominator of Bayes’ Theorem

• For the low birth weight example we can write

\[
\pi(\theta|a, b, y) \propto \theta^{y+}(1 - \theta)^{n-y+} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}
\]

\[
\propto \theta^{y+ + a-1}(1 - \theta)^{n-y+ + b-1}
\]

which is the kernel for the Beta\((y_+ + a, n - y_+ + b)\) distribution
Alternative parameterizations

- Suppose that, instead of $\theta$, scientific interest actually lies with the odds of low birth weight

$$\psi = \frac{P(Y = 1)}{P(Y = 0)} = \frac{\theta}{1 - \theta}$$

- How could we proceed with a Bayesian analysis?

1. Parameterize the problem in terms of $\theta$ and perform a change of variables on the posterior distribution
   - for a given prior on $\theta$, calculate the posterior

$$\pi_\theta(\theta|y) \propto \mathcal{L}(\theta|y)\pi_\theta(\theta)$$
perform a change of variables by noting that $\psi = g(\theta)$

$$\pi_\psi(\psi|y) = \pi_\theta(g^{-1}(\psi)|y) \left| \frac{\partial}{\partial \psi} g^{-1}(\psi) \right|$$

(2) Parameterize the problem directly in terms of $\psi$

* specify a prior for $\psi$ and use Bayes’ Theorem

$$\pi_\psi(\psi|y) \propto \mathcal{L}(y|\psi)\pi_\psi(\psi)$$

* note the likelihood is

$$\mathcal{L}(y|\psi) = \binom{n}{y} \left( \frac{\psi}{1+\psi} \right)^y \left( \frac{1}{1+\psi} \right)^{n-y}$$

Q: How would one choose between the two approaches?
Practical issues

Elicitation and quantification of prior beliefs

- Challenging and requires care
  - generally requires a biostatistician and subject-matter expert to work closely
  - often requires a lot of iteration and creativity
- Various schemes and software exist for eliciting substantive prior information from
  - scientific knowledge
  - previous studies
  - pilot data
For the low birth weight example, we adopted a Beta\((a, b)\) as a prior

flexible of class of distributions for random variables on \((0,1)\)

How can we specify values for \(a\) and \(b\)?

One way forward is to consider the prior mean and variance:

\[
E[\theta] = \frac{a}{a + b}
\]

\[
V[\theta] = \frac{ab}{(a + b)^2(a + b + 1)} = \frac{E[\theta](1 - E[\theta])}{(a + b + 1)}
\]

by considering \(E[\theta]\) and \(V[\theta]\), we might be able propose sensible values for \(a\) and \(b\)

Firstly, from the National Vital Statistics Report for 2001, the nationwide prevalence of low birth weight was 7.7%
• Secondly, in the frequentist paradigm, the MLE is \( \hat{\theta} = y/n \) and its' variance is

\[
V[\hat{\theta}] = \frac{\hat{\theta}(1 - \hat{\theta})}{n}
\]

★ similar expression to \( V[\theta] \) on the previous slide
★ so we might loosely interpret \((a + b + 1)\) as a prior ‘sample size’

• Consider a range of priors by choosing \( a \) and \( b \) such that
  ★ \( E[\theta] = 0.08 \)
  ★ \( a + b \) is 100, 500 and 1,000

• We get
  
  (1) \( \theta \sim \text{Beta}(8, 92) \)
  (2) \( \theta \sim \text{Beta}(40, 460) \)
  (3) \( \theta \sim \text{Beta}(80, 920) \)
Often useful to visualize the prior distribution

★ as the prior ‘sample size’ increases, the spread in the distribution decreases

★ reflects greater certainty in the range of plausible values

\[ \theta = P(\text{LBW}) \]
Sensitivity analysis

• As the statistical task gets more and more complex, the specification of priors becomes much harder
  ★ multidimensional $\theta$
  ★ parameters don’t always have a particularly intuitive interpretation
  ★ folks don’t always agree with your belief structure

• Sensitivity analysis is a powerful tool
  ★ specify a range of different priors
  ★ examine the impact on the results and conclusions

• Unfortunately, the process isn’t always prescriptive and it’s often challenging knowing when to stop
• When the results and conclusions are not sensitive, one might take a measure of confidence
  ★ often viewed as a good thing

• Sensitivity to the choice of prior is not always a bad thing!
  ★ an indicator of insufficient information in the likelihood
  ★ should think about collecting additional data
Non-informative priors

- Recall our adoption of a Beta($a, b$) distribution for the prior of $\theta$ in the low birth weight example.

- Suppose we took $a=b=1$
  - Beta($a, b$) $\equiv$ Uniform(0,1)
  - *a priori* we put equal weight on all values of $\theta$ on (0,1)
  - interpret Beta(1, 1) as a ‘non-informative’ prior

- The extend to which this is a good thing depends on the situation
  - some argue that you always want to be non-informative
  - another argument is that this ignores information that you already have
  - but if you don’t have any good prior information, a non-informative prior is reasonable
 Either way, care needs to be taken when one uses the term ‘non-informative’
   • generally, one can only be non-informative on one scale

 Consider the odds of low birth weight:
   \[ \psi = \frac{\theta}{1 - \theta} \]

 If we adopt a Beta(1,1) prior for \( \theta \), what is the induced prior for \( \psi \)?
   • using a change of variables it’s straightforward to show that the induced density is
   \[ f(\psi) = \left( \frac{1}{1 + \psi} \right)^2 \quad \psi \in (0, \infty) \]
   • \underline{does not} assign equal \textit{a priori} weight to all values of \( \psi \)

**Q:** On which scale should we specify a non-informative prior?
Summarizing the posterior distribution

- The posterior, \( \pi(\theta|y) \), is the basis for estimation and inference about \( \theta \)

- Practically, we need to summarize the information in the posterior
  - particularly important for multi-dimensional \( \theta \)

- Typically, we report a range of features of the posterior
  - mean, median
  - standard deviation, IQR, 95% credible interval
  - \( P(\theta < \text{value}) \)
  - \( P(\theta_1 < \text{value AND} \theta_2 > \text{value}) \)

- In most instances, what you report depends on the scientific question and also the audience
Returning to investigation of the probability of a low birth weight in North Carolina

- drew a ‘sample’ of size $n=500$ and ‘observed’ $y+=45$ low birth weights
- suppose we adopt a Beta$(40, 460)$ prior
- posterior distribution: Beta$(85, 915)$
• Start by providing an ‘estimate’ of $\theta$
  ★ e.g., some measure of centrality
  ★ posterior mean: $\hat{\theta} = 0.085$
  ★ could also report the posterior median

• Might report $P(\theta > 0.10)$

\[
1 - \text{pbeta}(0.1, 85, 915)
\]
[1] 0.04953876

• A common measure of uncertainty is the \textit{posterior credible interval}
  ★ typically taken as the interval defined by the 2.5\textsuperscript{th} and 97.5\textsuperscript{th} quantiles

\[
\text{qbeta(c(0.025, 0.975), 85, 915)}
\]
[1] 0.06852638 0.10304497

• How do we interpret this interval?
Contrast this with the interpretation of a confidence interval in the frequentist paradigm:

- sequence of hypothetical trials
- on average, 95% of constructed confidence intervals cover the true value

Interestingly, the interpretation of a Bayesian credible interval is consistent with how frequentist intervals are often erroneously interpreted!

Note, just as with frequentist confidence intervals, we could define many credible intervals that all have the same property that

\[ P(\theta \in (L, U)) = 0.95 \]

the narrowest such interval is called the highest posterior density interval.
Prediction

• Suppose interest lies in predicting some future observation, $\tilde{Y}$

• If we knew the true value of $\theta$, we could generate predictions directly from the pdf or pmf $f(\cdot|\theta)$

• The problem is, of course, that we don’t know $\theta$!

• But we have observed the data $y$ and can define the posterior predictive distribution:

$$f(\tilde{y}|y) = \int f(\tilde{y}|\theta)\pi(\theta|y)\,d\theta$$

★ intuitively, the posterior predictive is an average of conditional predictive densities, averaging over the posterior distribution

★ average over the uncertainty in $\theta$
Summary

- *Probability* is the basis for our ability to address questions about a population using information from a sample.

- Two main paradigms in statistical practice: *frequentist* and *Bayesian*.

- Key points:
  - both provide a coherent and rigorous theoretical basis.
  - both rely on the chosen model(s) being scientifically driven.
  - use different information and employ different analytic techniques.
  - yield differing interpretations of results.
  - neither will provide perfect truth.
• **Benefits** of the Bayesian paradigm
  
  ◆ formalizes how (most) people think
  ◆ incorporation of substantive prior information
  ◆ single recipe for coherent inference
  ◆ sensitivity to prior distribution
  ◆ coherent framework for prediction

• **Drawbacks** of the Bayesian paradigm
  
  ◆ individuals have different prior beliefs
  ◆ prior specification is often challenging
  ◆ sensitivity to prior distribution
  ◆ computation can be complex