Frequentist versus Bayesian statistics

FW8051 Statistics for Ecologists

Department of Fisheries, Wildlife and Conservation Biology

Learning Objectives

- Understand differences in how probability is defined in Frequentist and Bayesian statistics
- Understand how to estimate parameters and their uncertainty using Bayesian methods
- Compare Bayesian and Frequentist inference, starting with a simple problem that we can solve analytically.

Frequentist Statistics

- Probability = relative frequency of some event across an infinite sequence of random, repeatable experiments or trials

Frequentist Statistics

- Probability = relative frequency of some event across an infinite sequence of random, repeatable experiments or trials
- Data are random and parameters are fixed and estimated using:
  - Maximum Likelihood: P(data; parameters)

Frequentist Statistics

- Probability = relative frequency of some event across an infinite sequence of random, repeatable experiments or trials
- Data are random and parameters are fixed and estimated using:
  - Maximum Likelihood: P(data; parameters)
- Data are used to test hypotheses which are either True or False.
  - p-value = P(getting data/statistic as or more extreme as our sample data/statistic | Null hypothesis is true)

Frequentist Statistics

- Probability = relative frequency of some event across an infinite sequence of random, repeatable experiments or trials
- Data are random and parameters are fixed and estimated using:
  - Maximum Likelihood: P(data; parameters)
- Data are used to test hypotheses which are either True or False.
  - p-value = P(getting data/statistic as or more extreme as our sample data/statistic | Null hypothesis is true)

Goal: make 'good' decisions with high probability (across potential repeated experiments)

Bayesian Statistics

- Probability reflects “belief’ ’ about the system, taking into account prior expectations and data

Bayesian Statistics

- Probability reflects "belief'' about the system, taking into account prior expectations and data
- Random variables are used to model all sources of uncertainty
  - Parameters have a distribution

Bayesian Statistics

- Probability reflects “belief’ ’ about the system, taking into account prior expectations and data
- Random variables are used to model all sources of uncertainty
  - Parameters have a distribution
- Data are fixed. Inference is performed, conditional on the data.

Bayesian Statistics

- Probability reflects “belief’ ’ about the system, taking into account prior expectations and data
- Random variables are used to model all sources of uncertainty
  - Parameters have a distribution
- Data are fixed. Inference is performed, conditional on the data.
- P(Hypothesis | data), ranges from 0 to 1

Bayesian Statistics

- Probability reflects "belief' ' about the system, taking into account prior expectations and data
- Random variables are used to model all sources of uncertainty
  - Parameters have a distribution
- Data are fixed. Inference is performed, conditional on the data.
- P(Hypothesis | data), ranges from 0 to 1

Still want to make 'good' decisions with high probability (across potential repeated experiments)…*calibrated Bayes*!

Key Difference: Probability

Frequentist: relative frequency of events

Bayesian: belief about the system

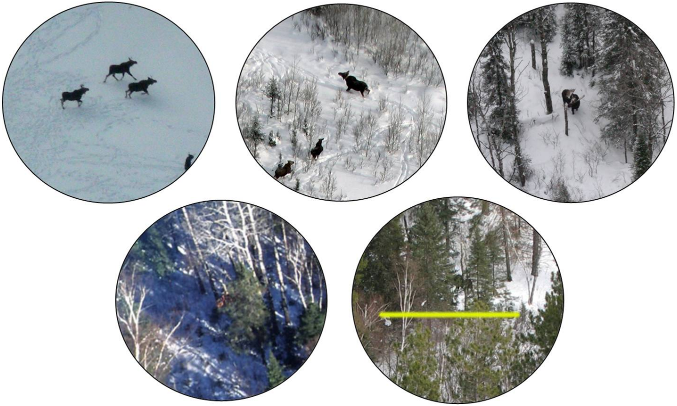
Key Difference: Probability

Frequentist: relative frequency of events

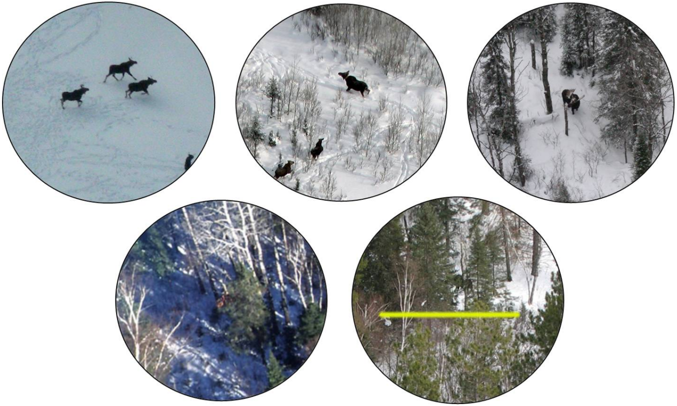
Bayesian: belief about the system

Lets compare inference from the two methods with a simple example

What is the probability, p, that a MNDNR biologist will detect a moose when flying in a helicopter?



What is the probability, p, that a MNDNR biologist will detect a moose when flying in a helicopter?



Goal: estimate p and characterize uncertainty wrt p.

Maximum Likelihood Using Binomial Distribution

1. Collect data: $n = 124$ trials, $y = 59$ moose observed

Maximum Likelihood Using Binomial Distribution

1. Collect data: $n = 124$ trials, $y = 59$ moose observed
2. Make assumptions about the data generating process:
   - Each trial is independent with constant probability of detection. What distribution would you use?

Maximum Likelihood Using Binomial Distribution

1. Collect data: $n = 124$ trials, $y = 59$ moose observed
2. Make assumptions about the data generating process:
   - Each trial is independent with constant probability of detection. What distribution would you use?
   - $y \sim$ Binomial(124, p), with p = Probability of a success (i.e., probability of seeing a moose)

Maximum Likelihood Using Binomial Distribution

1. Collect data: $n = 124$ trials, $y = 59$ moose observed
2. Make assumptions about the data generating process:
   - Each trial is independent with constant probability of detection. What distribution would you use?
   - $y \sim \text{Binomial}(124, p)$, with p = Probability of a success (i.e., probability of seeing a moose)
3. Estimate p using Maximum Likelihood

Maximum Likelihood

If $y \sim$ Binomial(n, p), then:

$L(p|y, n) = \frac{n!}{y!(n-y!)} p^y (1-p)^{n-y}$

Maximum Likelihood

If $y \sim$ Binomial(n, p), then:

$L(p|y,n) = \frac{n!}{y!(n-y!)} p^y (1-p)^{n-y}$

$\log(L) = log(\frac{n!}{y!(n-y!)}) + y\log(p) + (n-y)\log(1-p)$

Maximum Likelihood

If $y \sim$ Binomial(n, p), then:

$L(p|y,n) = \frac{n!}{y!(n-y!)} p^y (1-p)^{n-y}$

$\log(L) = log(\frac{n!}{y!(n-y!)}) + y\log(p) + (n-y)\log(1-p)$

Maximize $log[L(p|y)]$ with respect to p (take derivatives, set equal to 0, and solve) [On Board]:

Maximum Likelihood

If $y \sim$ Binomial(n, p), then:

$L(p|y,n) = \frac{n!}{y!(n-y!)} p^y (1-p)^{n-y}$

$\log(L) = log(\frac{n!}{y!(n-y!)}) + y\log(p) + (n-y)\log(1-p)$

Maximize $log[L(p|y)]$ with respect to p (take derivatives, set equal to 0, and solve) [On Board]:

$\hat{p} = y/n$

Maximum Likelihood

If $y \sim$ Binomial(n, p), then:

$L(p|y,n) = \frac{n!}{y!(n-y!)} p^y (1-p)^{n-y}$

$\log(L) = log(\frac{n!}{y!(n-y!)}) + y\log(p) + (n-y)\log(1-p)$

Maximize $log[L(p|y)]$ with respect to p (take derivatives, set equal to 0, and solve) [On Board]:

$\hat{p} = y/n$

$var(\hat{p}) = var(y/n) = var(y)/n^2 = p(1-p)/n$

$= I^{-1}(p),$ where $I(p) = E\left(-\frac{\partial^2 logL(p)}{\partial p^2}\right)$

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

- For large n, $\hat{p} \sim N(p, I^{-1}(p))$. This is the approximate sampling distribution of $\hat{p}$!

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

- For large n, $\hat{p} \sim N(p, I^{-1}(p))$. This is the approximate sampling distribution of $\hat{p}$!
- $I^{-1}(p) = \frac{p(1-p)}{n}$

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

- For large n, $\hat{p} \sim N(p, I^{-1}(p))$. This is the approximate sampling distribution of $\hat{p}$!
- $I^{-1}(p) = \frac{p(1-p)}{n}$

If $\hat{p} \sim N(p, I^{-1}(p))$, then $\frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \sim N(0,1)$

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

- For large n, $\hat{p} \sim N(p, I^{-1}(p))$. This is the approximate sampling distribution of $\hat{p}$!
- $I^{-1}(p) = \frac{p(1-p)}{n}$

If $\hat{p} \sim N(p, I^{-1}(p))$, then $\frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \sim N(0,1)$

$P(-z_{1-\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \leq z_{\alpha/2}) = \alpha$ where z is a standard normal random variable.

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

- For large n, $\hat{p} \sim N(p, I^{-1}(p))$. This is the approximate sampling distribution of $\hat{p}$!
- $I^{-1}(p) = \frac{p(1-p)}{n}$

If $\hat{p} \sim N(p, I^{-1}(p))$, then $\frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \sim N(0, 1)$

$P(-z_{1-\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \leq z_{\alpha/2}) = \alpha$ where z is a standard normal random variable.

To get a (1-α) CI, solve the above expression for p.

Frequentist Inference for p

For large n, a 95% CI = $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where does this come from?

- For large n, $\hat{p} \sim N(p, I^{-1}(p))$. This is the approximate sampling distribution of $\hat{p}$!
- $I^{-1}(p) = \frac{p(1-p)}{n}$

If $\hat{p} \sim N(p, I^{-1}(p))$, then $\frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \sim N(0,1)$

$P(-z_{1-\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{var(\hat{p})}} \leq z_{\alpha/2}) = \alpha$ where z is a standard normal random variable.

To get a (1-α) CI, solve the above expression for p.

$$P(\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > p > \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 0.95$$

Frequentist Inference

Calculate the large sample, normal-based confidence interval in R. $\hat{p} = 59/124 = 0.48$.

Frequentist Inference

Calculate the large sample, normal-based confidence interval in R. $\hat{p} = 59/124 = 0.48$.

```
# Estimate and SE
(theta.hat<-59/124)
```

```
## [1] 0.4758065
```

```
(se.theta.hat<-sqrt(theta.hat*(1-theta.hat)/124))
```

```
## [1] 0.04484873
```

```
# Confidence Interval
round(rep(theta.hat,2)+ c(-1.96,1.96)*se.theta.hat,2)
```

```
## [1] 0.39 0.56
```

How well do these work?

- Simulate 10,000 binomial random variables using: `x<-rbinom(10,000, size = n, p)` with `n = 15, 30, 100; p=0.1, 0.5, 0.9`
- Estimate 10,000 $\hat{p}$ values (`x/n`) and 10,000 95% CIs
- Determine how many of these CIs include p.

How well do these work?

- Simulate 10,000 binomial random variables using: `x<-rbinom(10,000, size = n, p)` with `n = 15, 30, 100; p=0.1, 0.5, 0.9`
- Estimate 10,000 $\hat{p}$ values (`x/n`) and 10,000 95% CIs
- Determine how many of these CIs include p.

```
# 10,000 repeated samples of size 124 and theta = theta.hat
ys<-rbinom(10000,size=124,prob=59/124)

# Calculate 10,000 theta^'s, SE(theta^)'s, CI's
theta.hats<-ys/124
se.theta.hats<-sqrt(theta.hats*(1-theta.hats)/124)
up.CIs<-theta.hats+1.96*se.theta.hats
low.CIs<-theta.hats-1.96*se.theta.hats

# Determine coverage
inCI<-I(low.CIs < 59/124 & up.CIs > 59/124)  # true theta is in the in
sum(inCI)/10000
```

```
[1] 0.9396
```

Frequentist Interpretation

In reality, we get 1 data set. We ended up with $\hat{p} = 0.48$, with 95% CI = (0.39, 0.56)

Frequentist Interpretation

In reality, we get 1 data set. We ended up with $\hat{p} = 0.48$, with 95% CI = (0.39, 0.56)

How do we interpret this CI?

Frequentist Interpretation

In reality, we get 1 data set. We ended up with $\hat{p}$ = 0.48, with 95% CI = (0.39, 0.56)

How do we interpret this CI?

p is either in the confidence interval or not! $P(p \in CI)$ = 0 or 1.

Frequentist Interpretation

In reality, we get 1 data set. We ended up with $\hat{p} = 0.48$, with 95% CI = (0.39, 0.56)

How do we interpret this CI?

p is either in the confidence interval or not! $P(p \in CI) = 0$ or 1.

The procedure we used should result in an interval that contains the true parameter 95% of the time.

Bayesian Inference

How does it differ? What are the steps?

1. Specify a likelihood for the data, $L(y|p)$

Bayesian Inference

How does it differ? What are the steps?

1. Specify a likelihood for the data, $L(y|p)$
2. Specify a prior distribution for the parameters, $\pi(p)$, reflecting our *a prior* belief about p

Bayesian Inference

How does it differ? What are the steps?

1. Specify a likelihood for the data, $L(y|p)$
2. Specify a prior distribution for the parameters, $\pi(p)$, reflecting our *a prior* belief about p
3. Use Bayes rule to determine the posterior distribution of p given the data, $p(p|y)$:

$$p(p|y) = \frac{L(y|p)\pi(p)}{p(y)} = \frac{L(y|p)\pi(p)}{\int L(y|p)\pi(p)dp}$$

Bayesian Inference

How does it differ? What are the steps?

1. Specify a likelihood for the data, $L(y|p)$
2. Specify a prior distribution for the parameters, $\pi(p)$, reflecting our *a prior* belief about p
3. Use Bayes rule to determine the posterior distribution of p given the data, $p(p|y)$:

$$p(p|y) = \frac{L(y|p)\pi(p)}{p(y)} = \frac{L(y|p)\pi(p)}{\int L(y|p)\pi(p)dp}$$

The posterior distribution captures our belief about the parameters after having collected data!

Bayes Theorem

$$p(p|y) = \frac{L(y|p)\pi(p)}{p(y)} = \frac{L(y|p)\pi(p)}{\int L(y|p)\pi(p)dp}$$

$p(y) = \int L(y|p)\pi(p)dp$ is the *marginal* distribution of y which requires integrating over p.

- This is a continuous version of the *total law of probability* formula we saw previously
- This integral is often difficult to solve, so Markov Chain Monte Carlo (MCMC) is used generate summaries of the posterior distribution, $p(p|y)$

Bayes Theorem

$$p(p|y) = \frac{L(y|p)\pi(p)}{p(y)} = \frac{L(y|p)\pi(p)}{\int L(y|p)\pi(p)dp}$$

$p(y) = \int L(y|p)\pi(p)dp$ is the *marginal* distribution of y which requires integrating over p.

- This is a continuous version of the *total law of probability* formula we saw previously
- This integral is often difficult to solve, so Markov Chain Monte Carlo (MCMC) is used generate summaries of the posterior distribution, $p(p|y)$

Keep focus on:

Posterior distribution $\propto$ Likelihood x prior distribution

Bayesian Inference

Likelihood (from binomial): $p(y|p) \propto p^{59}(1-p)^{124-59}$

Bayesian Inference

Likelihood (from binomial): $p(y|p) \propto p^{59}(1-p)^{124-59}$

Prior probability distribution for p?

Bayesian Inference

Likelihood (from binomial): $p(y|p) \propto p^{59}(1-p)^{124-59}$

Prior probability distribution for p?

- $\pi(p)$: $0 \leq p \leq 1$.
- $\pi(p)$: Beta distribution$(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

Bayesian Inference

Likelihood (from binomial): $p(y|p) \propto p^{59}(1-p)^{124-59}$

Prior probability distribution for p?

- $\pi(p)$: $0 \leq p \leq 1$.
- $\pi(p)$: Beta distribution$(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
- $\pi(p) \sim$ beta$(1, 1)$ (equivalent to a Uniform(0,1))

[Plot this using `curve(dbeta(x, 1, 1), from=0, to=1)`]

Bayesian Inference

Likelihood (from binomial): $p(y|p) \propto p^{59}(1-p)^{124-59}$

Prior probability distribution for p?

- $\pi(p)$: $0 \leq p \leq 1$.
- $\pi(p)$: Beta distribution$(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
- $\pi(p) \sim$ beta$(1, 1)$ (equivalent to a Uniform(0,1))

[Plot this using `curve(dbeta(x, 1, 1), from=0, to=1)`]

Use $\pi(p)$ and $p(y|p)$ and Bayes Theorem to calculate $p(p|y)$, the posterior distribution.

Bayes Theorem

$$p(p|y) = \frac{p(y|p)\pi(p)}{p(y)} = \frac{p(y|p)\pi(p)}{\int_{-\infty}^{\infty} p(y|p)\pi(p)d(p)}$$

$$p(p|y) \propto p(y|p)\pi(p)$$

Bayes Theorem

$$p(p|y) = \frac{p(y|p)\pi(p)}{p(y)} = \frac{p(y|p)\pi(p)}{\int_{-\infty}^{\infty} p(y|p)\pi(p)d(p)}$$

$$p(p|y) \propto p(y|p)\pi(p)$$

$$p(p|y) \propto p^{59}(1-p)^{124-59} \cdot 1$$

Bayes Theorem

$$p(p|y) = \frac{p(y|p)\pi(p)}{p(y)} = \frac{p(y|p)\pi(p)}{\int_{-\infty}^{\infty} p(y|p)\pi(p)d(p)}$$

$$p(p|y) \propto p(y|p)\pi(p)$$

$$p(p|y) \propto p^{59}(1-p)^{124-59} \cdot 1$$

$$p(p|y) \propto p^{60-1}(1-p)^{66-1}$$

Bayes Theorem

$$p(p|y) = \frac{p(y|p)\pi(p)}{p(y)} = \frac{p(y|p)\pi(p)}{\int_{-\infty}^{\infty} p(y|p)\pi(p)d(p)}$$

$$p(p|y) \propto p(y|p)\pi(p)$$

$$p(p|y) \propto p^{59}(1-p)^{124-59} \cdot 1$$

$$p(p|y) \propto p^{60-1}(1-p)^{66-1}$$

This is a beta distribution with parameters (60, 66).

Bayes Theorem

$$p(p|y) = \frac{p(y|p)\pi(p)}{p(y)} = \frac{p(y|p)\pi(p)}{\int_{-\infty}^{\infty} p(y|p)\pi(p)d(p)}$$

$$p(p|y) \propto p(y|p)\pi(p)$$

$$p(p|y) \propto p^{59}(1-p)^{124-59} \cdot 1$$

$$p(p|y) \propto p^{60-1}(1-p)^{66-1}$$

This is a beta distribution with parameters (60, 66).

The posterior distribution gives us the probability distribution of the parameter, given the data and our prior beliefs.

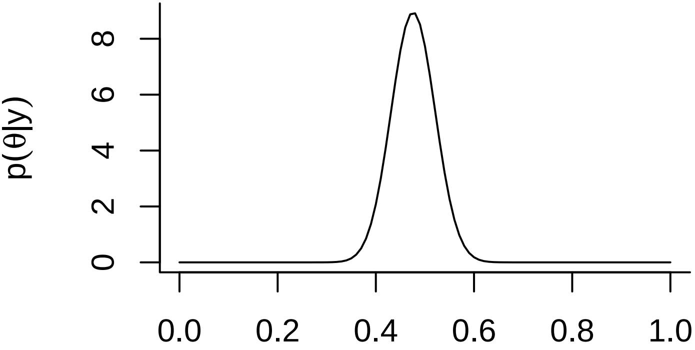
Bayesian Inference

Use `curve` to plot the posterior distribution = Beta(60,66).

Bayesian Inference

Use `curve` to plot the posterior distribution = Beta(60,66).

```
par(bty="L", mar=c(2,4.1,1,2.1))
# Plot the Posterior Distribution of theta
plot(curve(dbeta(x, 60, 66), from=0, to=1),
     type="l",xlab=expression(theta), ylab=c(expression(p(group("", theta, "|") *y))))
```



Credible interval

Find the endpoints, x_1 and x_2 such that $P(p \in (x_1, x_2)) = 0.95$.

Credible interval

Find the endpoints, x_1 and x_2 such that $P(p \in (x_1, x_2)) = 0.95$.

Use `qbeta` [remember, $\alpha = 60, \beta = 66$]

Credible interval

Find the endpoints, x_1 and x_2 such that $P(p \in (x_1, x_2)) = 0.95$.

Use `qbeta` [remember, $\alpha = 60, \beta = 66$]

```
# 95% credible interval
round(qbeta(c(0.025, 0.975), 60,66),2)
```

```
[1] 0.39 0.56
```

Same endpoints as Frequentist confidence interval, different interpretation!

Credible interval

Find the endpoints, x_1 and x_2 such that $P(p \in (x_1, x_2)) = 0.95$.

Use `qbeta` [remember, $\alpha = 60, \beta = 66$]

```
# 95% credible interval
round(qbeta(c(0.025, 0.975), 60,66),2)
```

```
[1] 0.39 0.56
```

Same endpoints as Frequentist confidence interval, different interpretation!

Interpretation: p has a 95% chance of being in the interval.

- Represents our *belief* based on *data* and our *prior* assumptions

Frequentist vs. Bayesian

Typical Conclusions

	Frequentist	Bayesian
Estimation	I have 95% confidence that the population mean is between 12.7 and 14.5 mcg/liter.	There is a 95% probability that the population mean is in the interval 136.2 g to 139.6 g.
Hypothesis Testing	If H_0 is true, we would get a result as extreme as the data we saw only 3.2% of the time. Since that is smaller than 5%, we would reject H_0 at the 5% level. These data provide significant evidence for the alternative hypothesis.	The odds in favor of H_0 against H_A are 1 to 3.

{Mary Parker, http://www.austincc.edu/mparker/stat/nov04/}

Advantages of Bayesian statistics

Advantages of Bayesian statistics

- Easier to fit complex models using Bayesian methods

Advantages of Bayesian statistics

- Easier to fit complex models using Bayesian methods
- Easy to characterize uncertainty for functions of the parameters

Advantages of Bayesian statistics

- Easier to fit complex models using Bayesian methods
- Easy to characterize uncertainty for functions of the parameters
- Intuitive appeal of credibility intervals (vs. confidence intervals)

Advantages of Bayesian statistics

- Easier to fit complex models using Bayesian methods
- Easy to characterize uncertainty for functions of the parameters
- Intuitive appeal of credibility intervals (vs. confidence intervals)
- Coherent philosophy of statistics
  - All inferences come from the posterior distribution
  - No separate theories for estimation, hypothesis testing, multiple comparisons, etc.

Disadvantages of Bayesian statistics

Disadvantages of Bayesian statistics

- With small samples, priors can make a big difference

Disadvantages of Bayesian statistics

- With small samples, priors can make a big difference
- Perceived subjectivity

Disadvantages of Bayesian statistics

- With small samples, priors can make a big difference
- Perceived subjectivity
- Computationally demanding when using MCMC

“Ecologists should be aware that Bayesian methods constitute a radically different way of doing science. Bayesian statistics is not just another tool to be added into ecologists’ repertoire of statistical methods. Instead, Bayesians categorically reject various tenets of statistics and the scientific method that are currently widely accepted in ecology and other sciences. The Bayesian approach has split the statistics world into warring factions (ecologists’ “density independence” vs “density dependence” debates of the 1950s pale by comparison), and it is fair to say that the Bayesian approach is growing rapidly in influence” - Brian Dennis (1996, Ecological Applications, p.1095-1103).

Pragmatic Statistician

Many do consider Bayesian methods another tool in the toolbox…

Pragmatic Statistician

Many do consider Bayesian methods another tool in the toolbox…

We will often fit models using both frequentist and Bayesian statistics (often, with similar answers)!

When is Bayesian Inference “Easier” or Prefered?

Dorazio 2016. Population Ecology 58:31-44

- “Hierarchical models of data that link a submodel of sampling processes with a submodel of ecological processes.”
- Inference for latent (i.e., unobserved) state variables
- Missing data problems
- Intractable likelihood functions
- Complex models of different sources and types of data (with shared parameters)