Interval Estimation

I can't believe schools are still teaching kids about the null hypothesis.

I remember reading a big study that conclusively disproved it years ago.
Interval Estimation

- Rarely are we interested in just a single point estimate for a parameter.
- Confidence intervals are used to:
  - Express uncertainty in an estimate.
  - Determine whether a hypothesized value falls within the interval.
- Interval estimates on predicted values.
Frequentist Confidence Interval

- Def'n: The fraction of intervals calculated from a large number of data sets generated by the same process that would include the true parameter value.

Bayesian Credible Interval

- Def'n: Posterior probability that the parameter lies within the interval.
Credible Intervals

\[ \int_{-\infty}^{A} p(\theta|Y) \, d\theta + \int_{B}^{\infty} p(\theta|Y) \, d\theta = \alpha \]

- Analytically estimated from posterior CDF
- Numerically estimated from quantiles of sample
- **NOT** estimated based on standard deviation
- Not necessarily symmetric
- Equal tail interval: both tails have the same probability
- Highest posterior density: narrowest possible interval
Beta-Binomial
Analytical CI in R

> ## Normal 95% CI
> mu = 5
> sigma = 3
> qnorm(c(0.025,0.975),mu,sigma)
[1] -0.879892 10.879892
> qnorm(c(0.025,0.975),1,0)
[1] -1.959964  1.959964

> ## Beta 95% CI
> p = 3
> n = 13
> qbeta(c(0.025,0.975),p,n-p)
[1] 0.05486064 0.48413775
Numerical Credible Intervals
Numerical CI in R

\[
> \text{quantile}(\beta, c(0.025, 0.975))
\]

\[
2.5\% \quad 97.5\%
\]

\[
9.362793 \quad 10.779553
\]

- Why numerical estimates of quantiles take longer to converge
- e.g. from 10000 steps, most extreme 250 used
Model Credible Interval

- Is a transformation of random variables, $f(y'|\theta)$
  - $f(x)$ is our process model
  - We are interested in the PDF of some new point $y'$
  - the model parameters $\theta$ are random, $p(\theta|Y)$
- Formally this transformation is

$$f(y') = p(\theta|Y) \text{det} \left( \frac{d\theta}{dy'} \right)$$

- Easier to understand/solve numerically
Example: Regression

- Plotting $y = b_0 + b_1 \times x$
- For each $[b_0, b_1]$ in the MCMC
- Interested in distribution of $E[y|x]$ for each $x$
Monte Carlo Error Propagation

• Input: Sample from **model inputs**
  - Parameters
  - Covariate/driver/IC uncertainty
• Action: Run model for each sample
• Output: Samples of **model outputs**
Regression Credible Interval

- Constructed as CI at each x
- Accounts for parameter uncertainty
- Does not account for variability of the data model
- Jensen's Inequality $f(x|\theta) \neq f(x|\theta)$
Monte Carlo CI in R

```r
xpred <- 1:20
ycred <- matrix(NA, nrow=10000, ncol=20)

for(g in 1:10000){
  ycred[g,] <- b0[g] + b1[g] * xpred
}

ci <- apply(ycred, 2, quantile, c(0.025, 0.975))

lines(xpred, ci[1,], col=3, lty=2)
lines(xpred, ci[2,], col=3, lty=2)
```
Bayesian Prediction

• Consider an observed data set $Y$ and a model with parameters $\theta$

• Want to calculate the posterior PDF of some new data point $y'$

• Need to integrate over all values $\theta$ can take on for ALL the model parameters (including variances)

$$p(y' | y) = \int \underbrace{p(y' | \theta)}_{\text{Likelihood of new data}} \underbrace{p(\theta | Y)}_{\text{Posterior}} \, d\theta$$
Bayesian Prediction Intervals

- CI of $p(y'|y)$ for each $x$
- Includes both data and parameter uncertainty
Monte Carlo PI in R

```r
xpred <- 1:20
ypred <- matrix(NA,nrow=10000,ncol=20)

for(g in 1:10000){
  Ey = b0[g] + b1[g] * xpred
  ypred[g,] <- rnorm(1,Ey,sig[g])
}

pi <- apply(ypred,2,quantile,c(0.025,0.975))

lines(xpred,pi[1,],col=3,lty=2)
lines(xpred,pi[2,],col=3,lty=2)
```