An Introduction to the Bayesian Approach to Statistical Inference

Alan Agresti, Distinguished Professor Emeritus
University of Florida, USA

Presented at La Sapienza Univ. and Univ. Cagliari

Outline

Focus of lecture

- Overview of the Bayesian approach for the most common statistical methods.
- Emphasis on examples of use, interpretations relative to the classical "frequentist" approach for interval estimation and significance testing, rather than theory, derivations, technical details.
- Lecture is based on *Foundations of Bayesian Statistics for Data Scientists, with R and Python* by A. Agresti, M. Kateri, R. Grove, and A. Mira (Taylor & Francis, 2026).
- Examples of analyses use `R` software. Our textbook has an appendix showing how also to conduct these statistical methods with `Python`.

INTRODUCTION TO BAYESIAN STATISTICS

- Classical (frequentist) versus Bayesian statistics
- Bayesian prior and posterior distributions
- Bayes' theorem and the Bayesian approach to statistical inference
- Bayesian point estimates and posterior intervals
- Bayesian significance testing and prediction
- Show inference about a proportion for a binary response variable to illustrate basic ideas

Classical (Frequentist) Approach to Statistical Inference

- With the classical (so-called *frequentist*) approach to statistical inference, probability statements refer to sample data that could be taken from the population when its parameters take particular values, but not to the parameters that describe the population.
- Classical statistical methods calculate probabilities about random variables and statistics such as test statistics that vary randomly from sample to sample, not about parameters.
- Probabilities cannot be stated for hypotheses, because hypotheses refer to parameter values. Likewise, probabilities do not apply to confidence intervals, once generated for sample data.
- Statistics have sampling distributions, parameters do not. Those sampling distributions have *frequentist* interpretations in terms of *what would happen in hypothetical repeated sampling of the same type*, but that sampling does not actually occur.

Bayesian vs Frequentist Approach to Statistical Inference

- The Bayesian statistical approach also treats parameters as having probability distributions. The Bayesian viewpoint is that *all* uncertainty can be modeled with probabilities, not just uncertainty of sample data but also uncertainty about parameter values.
- With the *frequentist* approach to statistical inference, once we obtain the data, probability statements (such as P-values) refer to other possible data from populations having particular values of parameters, but not to the parameters themselves.
- By contrast, given the data, the *Bayesian* approach applies probabilities to the parameters themselves, such as the probability that a particular hypothesis about a population mean is true or the probability that a particular interval contains the actual value of a population mean.

Bayesian vs Frequentist Approach (continued)

- Consider a comparison of population mean incomes for first job after receiving college degree (bachelor's), μ_1 for female graduates and μ_2 for male graduates. For the observed data, suppose the P-value in a t test of H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$ is 0.04. Then 0.04 is the probability the t test statistic takes value like the observed one or more extreme, in repeated samples of the same size, presuming that H_0 is true.
- That is, the P-value applies to potential samples, when H_0 is true. It is incorrect to interpret 0.04 as the probability that H_0 is true, because H_0 deals with parameter values. It is either true or not true.
- After observing the data, a Bayesian inference for estimating the population mean μ_1 might be "The probability is 0.95 that μ_1 falls between 22.8 and 24.6 thousand euros." A Bayesian inference for comparing population mean incomes μ_1 and μ_2 might be "The probability is 0.02 that $\mu_1 > \mu_2$ and 0.98 that $\mu_1 < \mu_2$."

Frequentist and Bayesian Approaches to Probability

Frequentist definition of probability: For an observation of a random phenomenon, the *probability* of a particular outcome (e.g., a *head* in flipping a coin) is the proportion of times that outcome would occur in an indefinitely long sequence of like observations under the same conditions.

Subjective definition of probability: The *probability* of a particular outcome is a person's assessment of the likelihood of that outcome, reflecting uncertainty based on the available information and experiences that influence that person's beliefs.

Bayesian statistics adopts the subjective definition of probability as its foundation, because it assumes probability distributions for parameters without relying on the notion of frequencies for long-run sequences of hypothetical observations.

Bayes' Theorem

Bayes' Theorem: For any two events A and B in a sample space,

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A \text{ and } B)}{P(A \text{ and } B) + P(A \text{ and } B^c)}$$

$$= \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}.$$

Regarding A as data and B as a parameter value, Bayes' Theorem enables us to find conditional probabilities about *parameter values, given the data*, using conditional probabilities for the *data, given parameter values*.

Bayes' Theorem and Posterior Distribution for Parameters

For a parameter θ, let $f(y \mid \theta)$ denote the *probability function* for an observation y, given a particular value of the parameter.

For independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)$, such as obtained with a simple random sample or an experiment employing randomization,

$$f(\boldsymbol{y} \mid \theta) = f(y_1 \mid \theta) f(y_2 \mid \theta) \cdots f(y_n \mid \theta).$$

The Bayesian approach assumes a **prior distribution** $p(\theta)$ for θ that reflects information available about θ before we observe the data $\boldsymbol{y}$.

Using Bayes' Theorem, the prior distribution combines with the information that the data provide to generate a **posterior distribution** $p(\theta \mid \boldsymbol{y})$ for θ,

$$p(\theta \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \theta) p(\theta)}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y} \mid \theta) p(\theta)}{\int_\Theta f(\boldsymbol{y} \mid \tilde{\theta}) p(\tilde{\theta}) d\tilde{\theta}}.$$

Bayesian statistical inferences are based on the posterior distribution.

Prior Distribution and Likelihood Function Determine Posterior Distribution

Likelihood function: Once we have observed $\boldsymbol{y}$, the **likelihood function** $L(\theta)$ is the joint probability function $f(\boldsymbol{y} \mid \theta)$ viewed as a function of the parameter θ when we substitute $\boldsymbol{y}$ in the formula for $f(\boldsymbol{y} \mid \theta)$.

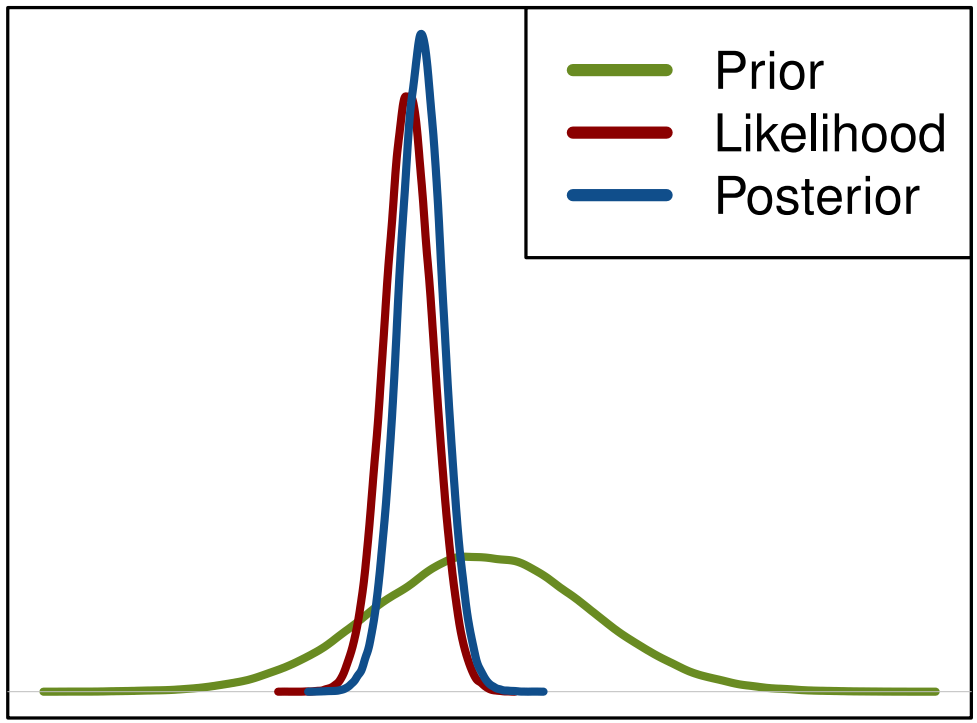
By Bayes' Theorem, the posterior probability function for θ is proportional to the product of the likelihood function with the probability function for the prior distribution of θ,

$$p(\theta \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \theta)p(\theta) = L(\theta)p(\theta).$$

When the prior distribution $p(\theta)$ is relatively flat, the posterior *pdf* $p(\theta \mid \boldsymbol{y})$ has a shape similar to the likelihood function $L(\theta)$.

Although interpretations differ, Bayesian and frequentist approaches usually lead to the same practical conclusions, because likelihood function is foundation of each.

Posterior Distribution Often Resembles Likelihood Function



Example: Posterior Distribution for Proportion

We estimate a population proportion π when data are n independent observations of a binary response variable (*success*, *failure* outcomes).

The number of successes y in the n observations is the outcome of a *binomial* random variable. Given y, the binomial likelihood function

$$L(\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \;\propto\; \pi^y(1-\pi)^{n-y}.$$

For a Bayesian prior distribution for π, lacking any prior information, a data scientist could use a *uniform distribution*

$$p(\pi) = 1,\ 0 \le \pi \le 1, \quad \text{so that}$$

$$p(\pi \mid y) \;\propto\; f(y \mid \pi)p(\pi) = L(\pi)p(\pi) \;\propto\; [\pi^y(1-\pi)^{n-y}] \cdot 1, \quad 0 \le \pi \le 1.$$

Beta Posterior Distribution

This is a special case of the *beta distribution*, which has *pdf*

$$p(\pi \mid \alpha_1, \alpha_2) \propto \pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}, \quad 0 \leq \pi \leq 1,$$

for *hyperparameters* $\alpha_1 > 0$ and $\alpha_2 > 0$.

- Uniform distribution is the beta distribution with $\alpha_1 = \alpha_2 = 1$.
- The mean of the beta distribution is $\alpha_1/(\alpha_1 + \alpha_2)$.
- The posterior *pdf* $p(\pi \mid y) \propto \pi^y(1-\pi)^{n-y}$ is the beta distribution with $\alpha_1 = y + 1$ and $\alpha_2 = n - y + 1$, having mean $= \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{y+1}{n+2}$.
- For large n, it is approximately bell-shaped with mean close to y/n, which is the sample proportion and maximum likelihood (ML) estimate.

Example: Proportion Supporting Legalized Abortion

A recent General Social Survey in the U.S. asked whether a woman should be able to get an abortion if she wants it for any reason. Of 1328 subjects who responded, 749 said *yes* and 579 said *no*.

Sample proportion $y/n = 749/1328 = 0.5640$

With uniform prior distribution, Bayesian posterior *pdf* $p(\pi \mid y)$ of π is beta with hyperparameters $\alpha_1 = y + 1 = 750$ and $\alpha_2 = n - y + 1 = 580$.

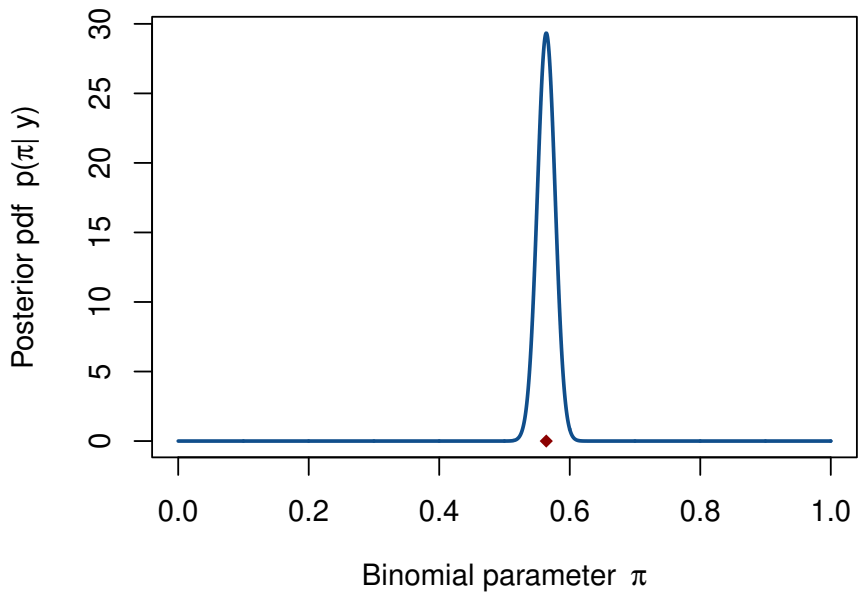
Mean $= (y + 1)/(n + 2) = 750/1330 = 0.5639$.

Range of plausible values for π is narrow, with 95% of the distribution between 0.537 and 0.590. The posterior $P(\pi < 0.50) = 0.0000015$.

```
------------------------------------------------------------------------------------------
> qbeta(0.025, 750, 580); qbeta(0.975, 750, 580)
[1] 0.537182            # 0.025 quantile of beta distribution with hyperparameters 750, 580
[1] 0.5904555           # 0.975 quantile of beta distribution
> pbeta(0.50, 750, 580) # cumulative probability at 0.50 for beta distribution
[1] 1.510934e-06        #   with hyperparameters 750, 580 is 0.0000015
------------------------------------------------------------------------------------------
```

Plot of Beta Posterior Distribution

Here is the beta posterior *pdf* with hyperparameters $\alpha_1 = 750$ and $\alpha_2 = 580$ for U.S. population proportion who believe a woman should be able to get an abortion for any reason (red dot = sample proportion):



Bayesian Point Estimates and Posterior Intervals

- **Posterior mean estimate of a parameter**: For a parameter θ in parameter space Θ, given the data $\boldsymbol{y}$ and the Bayesian posterior *pdf* $p(\theta \mid \boldsymbol{y})$, the *posterior mean* estimate of θ is

$$E(\theta \mid \boldsymbol{y}) = \int_{\Theta} \theta p(\theta \mid \boldsymbol{y}) d\theta.$$

- **Posterior equal-tail percentile interval estimate of a parameter** having probability $(1 - \alpha)$ is the region between the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution ($\alpha = 0.05$ in above example).
- **Highest posterior density (HPD) interval estimate of a parameter** has posterior *pdf* higher over all values in the interval than over all values not in it.

HPD interval is the shortest possible interval having chosen probability.

HPD interval and equal-tail percentile interval are identical when the posterior distribution is unimodal and symmetric.

Example: Proportion Supporting a Woman's Choice

The `binom` package in `R` can provide a wide variety of intervals for the binomial parameter. For the beta posterior with hyperparameters $\alpha_1 = 750$ and $\alpha_2 = 580$, the 95% posterior equal-tail percentile interval is (0.537, 0.590). The HPD interval is the default provided with the `binom.bayes` function in the `binom` package. The 95% HPD interval is (0.537, 0.591).

```
------------------------------------------------------------------------------------------------
> qbeta(c(0.025, 0.975), 750, 580) # 0.025 and 0.975 quantiles, values of beta hyperparameters
[1] 0.5371820 0.5904555            # 95% posterior equal-tail percentile interval
> library(binom)                   # requests use of binom package
>                                  # uniform prior is beta with alpha1 = alpha2 = 1
> binom.bayes(749, 1328, conf.level=0.95, prior.shape1=1, prior.shape2=1, type="central")
  method   x    n shape1 shape2      mean    lower    upper
1  bayes 749 1328    750    580 0.5639098 0.537182 0.590455 # equal-tail 95% percentile int.
> binom.bayes(749, 1328, conf.level=0.95, prior.shape1=1, prior.shape2=1)
  method   x    n shape1 shape2      mean    lower    upper
1  bayes 749 1328    750    580 0.5639098 0.537246 0.590519 # HPD 95% interval
> binom.confint(749, 1328, conf.level=0.95, method="asymptotic")
      method   x    n     mean     lower     upper
1 asymptotic 749 1328 0.564006 0.5373355 0.5906765 # frequentist 95% Wald confidence interval
------------------------------------------------------------------------------------------------
```

Frequentist (asymptotic "Wald") 95% CI, for ML estimate $\hat{\pi} = y/n$, is $\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}$.

Interpretation: Posterior Intervals vs Frequentist CIs

Although frequentist and Bayesian results are nearly identical, the interpretations are quite different.

- With frequentist approach, π is fixed, not a random variable. It either *is* or *is not* in the 95% confidence interval of (0.537, 0.591).
- Interpretation: If we used this method repeatedly with independent hypothetical samples, in the long run 95% of the confidence intervals would contain the actual value of π.
- With the Bayesian approach, π is itself a random variable that has a probability distribution. After observing the data and constructing the HPD posterior interval, we conclude that *the probability is 0.95 that π falls between 0.537 and 0.591*.
- The resemblance between frequentist and Bayesian results increases as n increases or as the variance of the prior distribution increases, because the maximum and shape of the posterior distribution is increasingly similar to the maximum and shape of the likelihood function.

Bayesian Significance Testing

With a continuous prior distribution, the posterior distribution is also continuous. Then, the posterior probability of any single value for a parameter such as a population proportion π is zero.

This accords with intuition in most applications that null hypothesis conditions such as $\pi = 0.50$ *exactly* are implausible. It is usually more relevant to summarize the evidence that $\pi < 0.50$ versus $\pi > 0.50$.

When neither posterior $P(\pi > 0.50 \mid y)$ nor $P(\pi < 0.50 \mid y)$ is close to 0, we regard 0.50 as a plausible value for π.

Example: Proportion Supporting a Woman's Choice

We find posterior $P(\pi > 0.50 \mid y)$ and $P(\pi < 0.50 \mid y)$.

```
-------------------------------------------------------------------------------
> pbeta(0.50, 750, 580) # cumulative probability at 0.50, from cdf of beta posterior
[1] 1.510934e-06        #    distribution with hyperparameter values 750 and 580
-------------------------------------------------------------------------------
```

Posterior $P(\pi < 0.50 \mid y) = 0.0000015$ and $P(\pi > 0.50 \mid y) = 0.9999985$, indicating extremely strong evidence that a majority of the population believe that a woman should have the right to an abortion for any reason. Corresponding P-value with frequentist inference for testing H_0: $\pi = 0.50$ (implicitly H_0: $\pi \leq 0.50$) against H_1: $\pi > 0.50$ is 0.0000017:

```
-------------------------------------------------------------------------------
> 1 - pbinom(748, 1328, 0.50) # one-sided (right-tail) P-value for binomial distribution
[1] 1.712424e-06              #    equals 1 - cumulative probability (from cdf)
-------------------------------------------------------------------------------
```

Interpretations: 0.0000015 is Bayesian posterior $P(\pi < 0.50)$, whereas frequentist P-value of 0.0000017 refers to hypothetical samples if H_0 were true; i.e., the probability of a sample like observed or more extreme if actually $\pi = 0.50$.

Bayesian Posterior Predictive Distribution

After observing the data, sometimes the main goal is to predict future observations. A Bayesian **posterior predictive distribution** is the probability distribution for a future observation Y_f,

$$f(y_f \mid \boldsymbol{y}) = \int_{\Theta} f(y_f \mid \theta) p(\theta \mid \boldsymbol{y}) d\theta.$$

We obtain $f(y_f \mid \boldsymbol{y})$ by averaging the probability function $f(y_f \mid \theta)$ for known θ with respect to the posterior distribution $p(\theta \mid \boldsymbol{y})$ of θ.

This is beyond our scope here, but Bayesian posterior predictive distributions are also useful for checking assumptions of the model. One such check compares simulated observations from the posterior predictive distribution with the observed data. The simulated data should look like the observed data, such as by having similar central tendency and variability.

BAYESIAN INFERENCE FOR PROPORTIONS

- Conjugate beta prior and posterior distributions
- Comparing proportions using simulation
- Bayesian posterior intervals and tail probabilities as alternatives to frequentist confidence intervals and P-values
- Generalization to multinomial parameters: Dirichlet prior and posterior distributions
- Brief mention of other methods for proportions presented in the textbook but not covered in this lecture

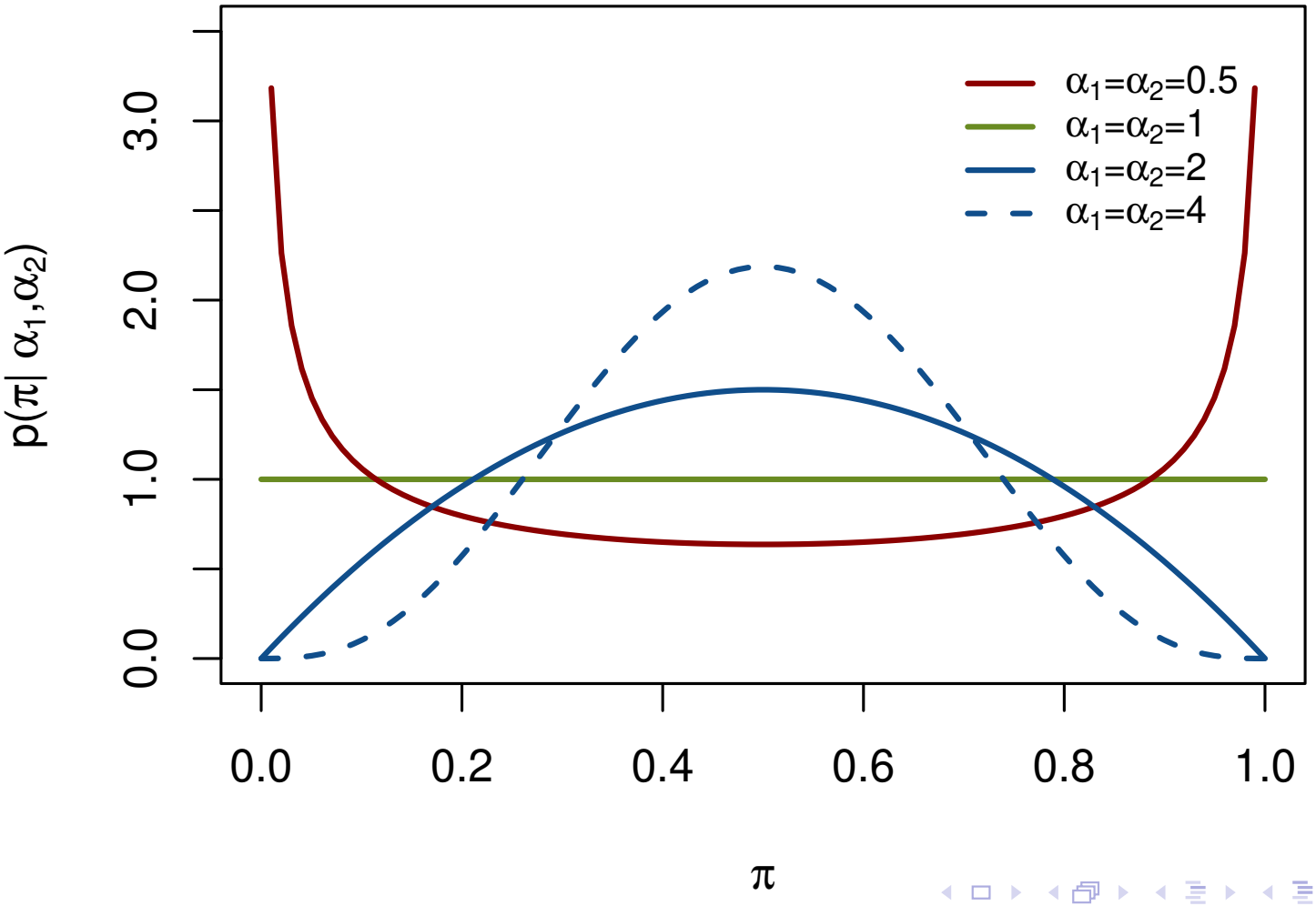
Beta Family of Prior Distributions

With hyperparameters $\alpha_1 > 0$ and $\alpha_2 > 0$, the *beta distribution* has *pdf*

$$p(\pi \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}, \quad 0 \leq \pi \leq 1,$$

where the gamma function $\Gamma(k) = \int_0^\infty e^{-x} x^{k-1} dx$ is the normalizing constant needed so that the *pdf* integrates to 1.

- $\mu = E(\pi) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$
- When $\alpha_1 = \alpha_2$, the beta *pdf* is symmetric around 0.50, with variance decreasing as common value increases.
- *Uniform distribution* over [0, 1] results when $\alpha_1 = \alpha_2 = 1$.
- Bimodal U-shape when $\alpha_1 = \alpha_2 < 1$ and bell shape when $\alpha_1 = \alpha_2 > 1$.
- Unimodal skewed to the left (i.e., longer tail pointing to the left) when $\alpha_1 > \alpha_2 > 1$; unimodal skewed to the right when $\alpha_2 > \alpha_1 > 1$.


$\alpha_1=\alpha_2=0.5$
$\alpha_1=\alpha_2=1$
$\alpha_1=\alpha_2=2$
$\alpha_1=\alpha_2=4$
$p(\pi|\ \alpha_1,\alpha_2)$
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
π

Conjugate Beta Prior and Posterior Distributions

For a binomial sample of y successes in n trials with parameter π and beta prior *pdf* $p(\pi \mid \alpha_1, \alpha_2)$, Bayes' theorem implies

$$\begin{aligned} p(\pi \mid y) \propto f(y \mid \pi; n) p(\pi \mid \alpha_1, \alpha_2) &\propto [\pi^y (1-\pi)^{n-y}][\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}] \\ &= \pi^{y+\alpha_1-1}(1-\pi)^{n-y+\alpha_2-1}, \quad 0 \leq \pi \leq \end{aligned}$$

also a beta distribution.

Posterior hyperparameter values $\alpha_1^* = y + \alpha_1$ and $\alpha_2^* = n - y + \alpha_2$.

Posterior distribution falls in same family of probability distributions as prior distribution, but hyperparameters are updated based on the data.

The beta prior distribution is a *conjugate prior distribution* for the binomial likelihood function.

Bayesian Posterior Mean Estimate of a Proportion

$$\tilde{\pi} = E(\pi \mid y) = \frac{\alpha_1^*}{\alpha_1^* + \alpha_2^*} = \frac{y + \alpha_1}{(y + \alpha_1) + (n - y + \alpha_2)} = \frac{y + \alpha_1}{n + \alpha_1 + \alpha_2}$$

$$= \left(\frac{n}{n + \alpha_1 + \alpha_2}\right)\frac{y}{n} + \left(\frac{\alpha_1 + \alpha_2}{n + \alpha_1 + \alpha_2}\right)\frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

- y/n is the sample proportion, which is also the ML estimate $\hat{\pi}$ of π.
- Posterior mean $\tilde{\pi} = E(\pi \mid y)$ is weighted average $w\hat{\pi} + (1 - w)E(\pi)$ of $\hat{\pi} = y/n$ and prior mean, $E(\pi) = \alpha_1/(\alpha_1 + \alpha_2)$, with $w = n/(n + \alpha_1 + \alpha_2)$.
- When $\alpha_1 = \alpha_2$, shrinks $\hat{\pi}$ toward prior mean of 0.50.
- Weight $w = n/(n + \alpha_1 + \alpha_2)$ given to $\hat{\pi}$ increases toward 1 as n increases.
- Amount of information in posterior mean estimate $\tilde{\pi}$ is summarized by n for data and $\alpha_1 + \alpha_2$ for beta prior distribution. Conceptually, effect of prior distribution is to add $\alpha_1 + \alpha_2$ *imaginary observations*, of which α_1 are successes.

Estimation: Comparing Bayesian and Frequentist (ML)

For point estimation of binomial parameter π, we compare Bayesian and frequentist estimators.

- ML estimator, sample proportion $\hat{\pi} = Y/n$, is unbiased (i.e., for any value of π, $E(\hat{\pi}) = \pi$) and has minimum variance among all the possible unbiased estimators of π.
- Bayesian posterior mean estimator $\tilde{\pi}$ of π for a beta prior distribution has expectation

$$\left(\frac{n}{n+\alpha_1+\alpha_2}\right)\pi + \left(\frac{\alpha_1+\alpha_2}{n+\alpha_1+\alpha_2}\right)\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right),$$

  a weighted average of π and the prior mean $\alpha_1/(\alpha_1+\alpha_2)$.
- The Bayes estimator is biased (i.e., $E(\tilde{\pi}) \neq \pi$), but its bias decreases toward 0 as n increases.
- *Mean squared error* (MSE) of estimator $\hat{\theta}$ summarizes how close $\hat{\theta}$ tends to be to parameter θ it estimates.

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta}-\theta)^2 = E\{[\hat{\theta}-E(\hat{\theta})]+[E(\hat{\theta})-\theta]\}^2 = \text{var}(\hat{\theta})+(\text{bias})^2.$$

Bayesian vs Frequentist and Bias/Variance Tradeoff

- If its variance decreases sufficiently, an estimator $\hat{\theta}$ that has some bias may have smaller MSE over a substantial range of θ values. This is called the *bias/variance tradeoff*.
- Although the Bayes estimator $\tilde{\pi}$ of π is biased, its variance

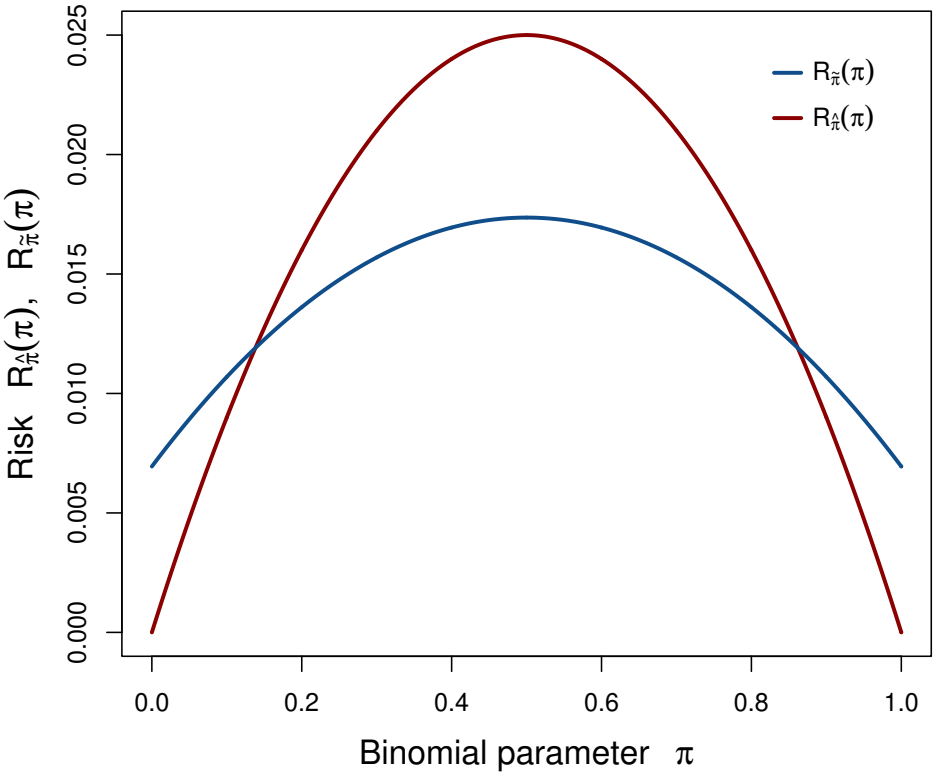
$$\text{var}(\tilde{\pi}) = \text{var}\left[\left(\frac{n}{n+\alpha_1+\alpha_2}\right)\frac{Y}{n}\right] = \left[\frac{n}{n+\alpha_1+\alpha_2}\right]^2 \text{var}\left(\frac{Y}{n}\right) = \left[\frac{n}{n+\alpha_1+\alpha_2}\right]^2 \frac{\pi(1-\pi)}{n},$$

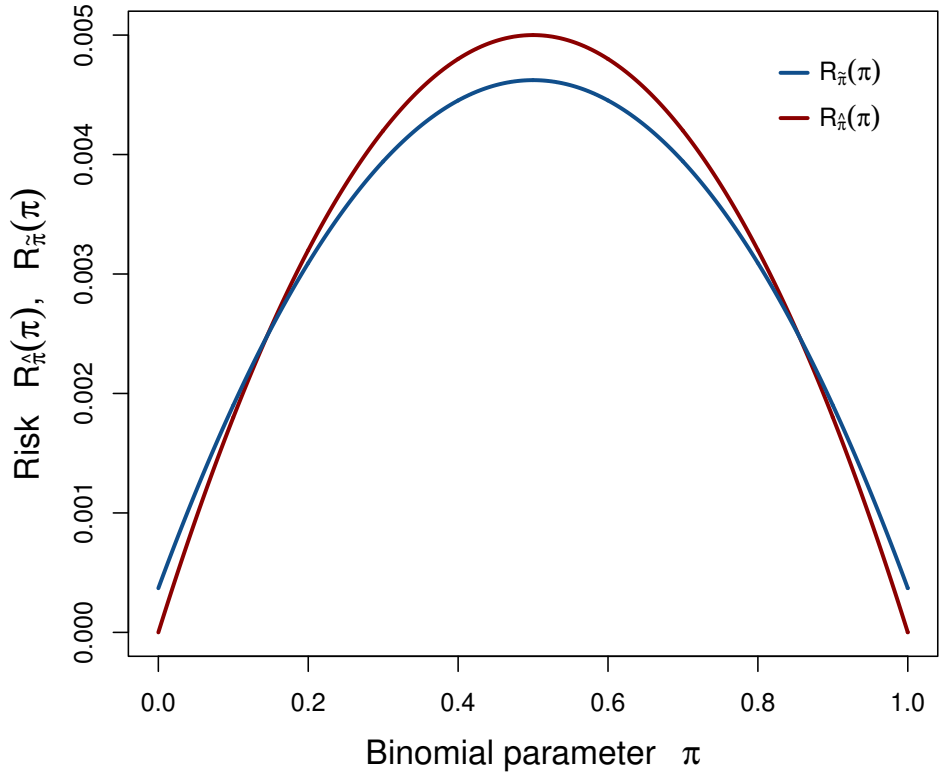
  is smaller than $\pi(1-\pi)/n$, the variance of the ML estimator.
- When $\alpha_1 = \alpha_2 = \alpha$, bias of $\tilde{\pi}$ is $\alpha(1-2\pi)/(n+2\alpha)$,

$$\text{MSE}(\tilde{\pi}) = \text{var}(\tilde{\pi}) + (\text{bias})^2 = \left(\frac{n}{n+2\alpha}\right)^2 \frac{\pi(1-\pi)}{n} + \frac{\alpha^2(1-2\pi)^2}{(n+2\alpha)^2}.$$

  If π is near the prior mean of 0.50, bias is relatively small, so MSE smaller for Bayes estimator than for ML estimator. But, if π is far from 0.50, such as close to 0 or 1, MSE smaller for ML estimator.

MSE is an example of a *risk function* $R(\pi)$, measuring expected distance of estimator from parameter for a squared-error *loss function*.





MSE for ML estimator $\hat{\pi} = Y/n$ (red) and Bayesian estimator $\tilde{\pi} = (Y+1)/(n+2)$ for uniform prior distribution (blue), for estimating binomial parameter π when $n = 10$ (left) and when $n = 50$ (right).

Bayesian Posterior Intervals for Comparing Proportions

Example: Comparing proportions of females and males who believe in hell (General Social Survey data)

$\hat{\pi}_1 = 498/674 = 0.739$ for females, $\hat{\pi}_2 = 316/468 = 0.675$ for males

- With independent binomial samples and independent beta prior distributions with $\alpha_1 = \alpha_2$ for π_1 and π_2, the prior distribution of $\pi_1 - \pi_2$ is symmetric around 0.
- With uniform prior distributions for π_1 and π_2, posterior distribution is Beta$(y_1 + 1, n_1 - y_1 + 1)$ = Beta(499, 177) for π_1 and Beta$(y_2 + 1, n_2 - y_2 + 1)$ = Beta(317, 153) for π_2.
- Simple way to construct a posterior interval for $\pi_1 - \pi_2$ is by simulating from the posterior distributions of π_1 and π_2 and using those values to approximate the posterior distribution of $\pi_1 - \pi_2$ and its quantiles.

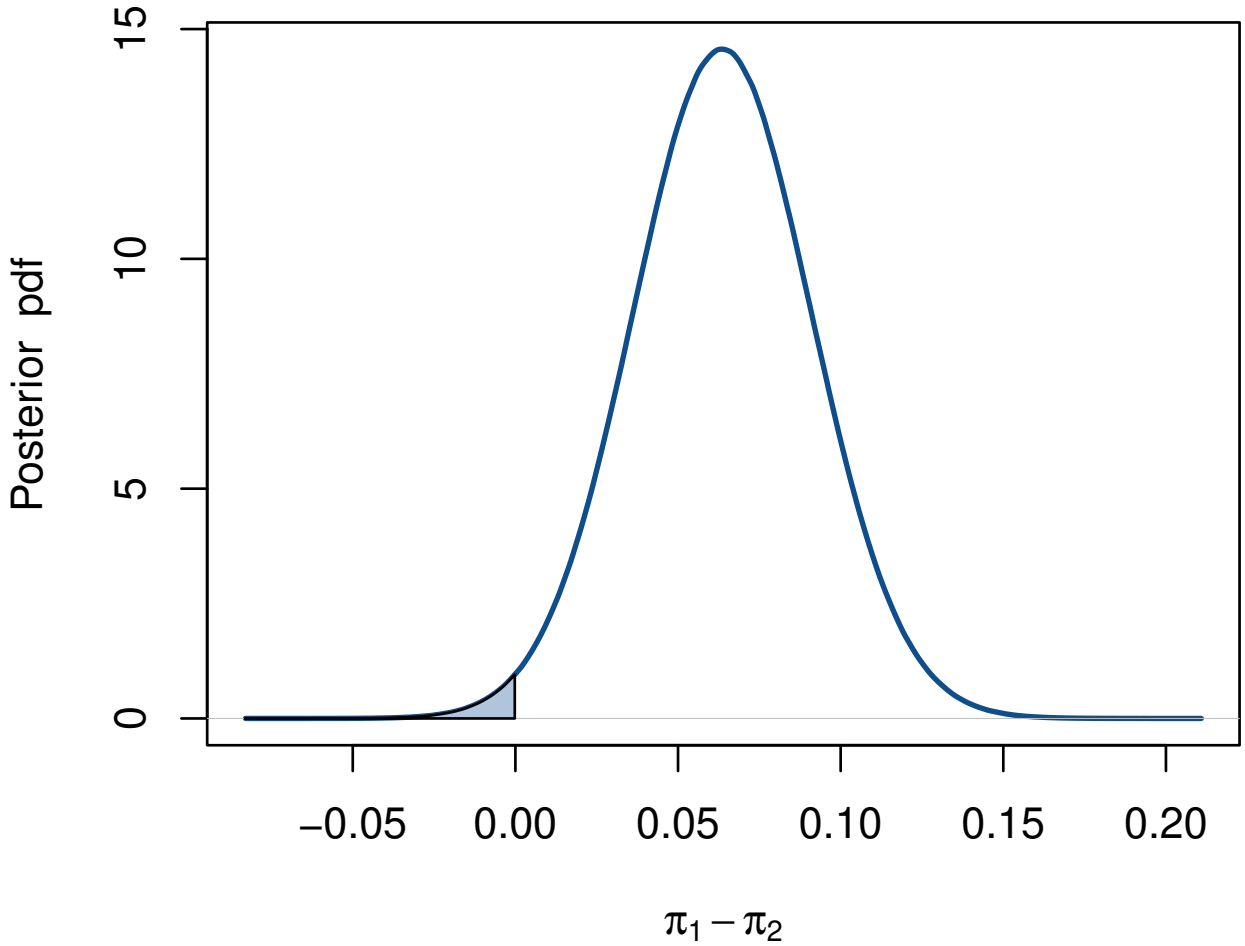
Simulation results using R

```
-------------------------------------------------------------------------------
> pi1 <- rbeta(50000000, 499, 177) # simulating from posterior distribution of pi1
> pi2 <- rbeta(50000000, 317, 153) # simulating from posterior distribution of pi2
> quantile(pi1 - pi2, c(0.025, 0.975)) # 0.025 and 0.975 quantiles of simulated difference
      2.5%      97.5%
0.01016046 0.11763070              # simulated 95% percentile interval for pi1 - pi2
> plot(density(pi1 - pi2), ylab="Density", xlab="Difference of Probabilities", main="")
-------------------------------------------------------------------------------
```

We also show results using R functions to simulate the HPD interval and to construct the frequentist 95% Wald confidence interval,

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2} se, \quad \text{i.e.} \quad (\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

```
-------------------------------------------------------------------------------
> library(HDInterval)
> hdi(pi1 - pi2, credMass=0.95)
    lower     upper
0.0099792 0.1174535                # simulated 95% HPD interval for pi1 - pi2
 > prop.test(c(498, 316), c(674, 468), correct=FALSE) # data c(y1, y2), c(n1, n2)
95 percent confidence interval:    # frequentist 95% Wald CI for pi1 - pi2
 0.009809589 0.117507867 # Conclude population cproportion believing in hell
                         #  between 0.010 and 0.117 higher for females than males
-------------------------------------------------------------------------------
```



Simulated posterior *pdf* of difference, between females and males, of probabilities of belief in hell. The shaded left-tail probability of 0.010 is the posterior probability that $\pi_1 - \pi_2 < 0$, that is, that females are less likely than males to believe in hell.

Bayesian Significance Tests for Comparing Proportions

We can precisely approximate $P(\pi_1 < \pi_2 \mid y_1, y_2)$ by proportion of simulated joint distribution for which $\pi_1 < \pi_2$, which is 0.01. It seems highly likely that $\pi_1 > \pi_2$.

```
------------------------------------------------------------------------------
> pi1 <- rbeta(50000000, 499, 177)
> pi2 <- rbeta(50000000, 317, 153)
> mean(pi1 < pi2); mean(pi1 > pi2) # simulated posterior P(pi1 < pi2), P(pi1 > pi2)
[1] 0.00978576
[1] 0.9902142
> prop.test(c(498, 316), c(674, 468), correct=FALSE) # data c(y1, y2), c(n1, n2)
X-squared = 5.4675, df = 1, p-value = 0.01937 # X-squared = chi-squared statistic
alternative hypothesis: two.sided # = square of z test stat; doesn't show direction
------------------------------------------------------------------------------
```

Frequentist chi-squared test of H_0: $\pi_1 = \pi_2$ does not utilize direction of effect, so corresponds to two-sided P-value of 0.0194. For H_1: $\pi_1 > \pi_2$, P-value is $0.0194/2 = 0.0097$; i.e., if H_0: $\pi_1 = \pi_2$ were true, probability would be 0.0097 of sample result or more extreme result in direction of H_1 (the direction of relatively more females than males believing in hell).

The Bayesian approach provides an explicit probability estimate of 0.0098 that males are more likely than females to believe in hell.

Bayesian Inference with Several Categories

When a categorical variable has more than two possible outcomes for each of n observations, the *multinomial distribution* generalizes the binomial to provide probabilities for the possible counts in the various categories.

Multinomial distribution For n independent observations with c possible outcome categories having probabilities $(\pi_1, \pi_2, \ldots, \pi_c)$ with $\sum_{j=1}^{c} \pi_j = 1$, the joint *pmf* for the numbers of outcomes $\{y_j\}$ in those categories, where $\sum_{j=1}^{c} y_j = n$, is

$$f(y_1, y_2, \ldots, y_c \mid \pi_1, \pi_2, \ldots, \pi_c; n) = \left(\frac{n!}{y_1! y_2! \cdots y_c!} \right) \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_c^{y_c}.$$

The beta distribution that serves as a conjugate prior distribution for a binomial likelihood function extends to the *Dirichlet distribution* for a multinomial likelihood function.

Dirichlet Conjugate Prior and Posterior Distribution

For c categories having probabilities $(\pi_1, \pi_2, \ldots, \pi_c)$, the *Dirichlet pdf* for the probabilities is

$$p(\pi_1, \ldots, \pi_c \mid \alpha_1, \ldots, \alpha_c) = \frac{\Gamma(\sum_{j=1}^{c} \alpha_j)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_c)} \pi_1^{\alpha_1 - 1} \cdots \pi_c^{\alpha_c - 1}, \quad 0 \leq \pi_j \leq 1,$$

where $\{\alpha_j > 0\}$ are hyperparameters.

- For a multinomial likelihood function having counts $\{y_j\}$, assuming a Dirichlet prior distribution for $(\pi_1, \ldots, \pi_c)$, posterior *pdf* is also Dirichlet, with hyperparameters $\{y_j + \alpha_j, j = 1, \ldots, c\}$.
- Impact of prior distribution is to add α_j *imaginary observations* to category j for all j before forming a sample proportion.
- Marginal posterior distribution of π_j is beta with hyperparameters $y_j + \alpha_j$ and $\sum_{k \neq j}(y_k + \alpha_k)$.

Example: What Percentage of People are Very Happy?

2021 GSS: Are you very happy, pretty happy, or not too happy?
Sample counts (778, 2304, 921), for percentages (19.4, 57.6, 23.0).

With uniform Dirichlet distribution over the probability simplex ($\alpha_1 = \alpha_2 = \alpha_3 = 1$), beta posterior distribution of π_1 for very happy category has hyperparameters $y_1 + \alpha_1 = 779$ and $(y_2 + \alpha_2) + (y_3 + \alpha_3) = (2304 + 1) + (921 + 1) = 3227$.

Hyperparameters are (2305, 1701) for pretty happy category.
Hyperparameters are (922, 3084) for not too happy category.

Here are 99% equal-tail percentile intervals for π_1, π_2, and π_3:

```
> qbeta(c(0.005, 0.995), 779, 3227)  # quantiles, beta hyperparameters for very happy
[1] 0.1786434 0.2108461              #   99% posterior percentile interval
> qbeta(c(0.005, 0.995), 2305, 1701) # quantiles, beta hyperparameters for pretty happy
[1] 0.5552075 0.5954250              #   99% posterior percentile interval
> qbeta(c(0.005, 0.995), 922, 3084)  # quantiles, beta hyperparameters for not too happy
[1] 0.2132832 0.2475323              #   99% posterior percentile interval
```

By Bonferroni method, posterior probability is at least $1 - 3(0.01) = 0.97$ that all three posterior intervals contain true population proportion values.

Be cautious with flat priors when c large!

The uniform Dirichlet prior having all $\{\alpha_j = 1\}$ seems non-informative, but it corresponds to adding c imaginary prior observations, which is a considerable effect when c is large relative to n.

Example: We observe which in a list of $c = 1000$ books is selected as the favorite book by each of $n = 100$ people.

- With all $\{\alpha_j = 1\}$, the posterior mean estimate of π_j is $(y_j + 1)/1100$.
- When book j receives 1 of the 100 observations, this is 0.002; but when it receives all 100 observations, this is 0.092.
- We could instead use a much more diffuse prior distribution. For Dirichlet hyperparameters $\{\alpha_j = 1/c\}$, the posterior mean is $(y_j + 1/c)/(n + 1)$, which takes values 0.015 and 0.995 for these two cases.

Other Methods (in text, beyond scope of this lecture)

- Another Bayesian method for summarizing evidence about hypotheses

$$H_0: \ \theta \in \Theta_0, \quad H_1: \ \theta \in \Theta_1$$

is the *Bayes factor*. When prior probabilities are equal for each hypothesis, this equals the *posterior odds*,

$$BF = \frac{P(\theta \in \Theta_1 \mid \boldsymbol{y})}{P(\theta \in \Theta_0 \mid \boldsymbol{y})} = \frac{\int_{\Theta_1} p(\theta \mid \boldsymbol{y}) d\theta}{\int_{\Theta_0} p(\theta \mid \boldsymbol{y}) d\theta}.$$

- Prior distributions $p(\theta \mid \boldsymbol{\lambda})$ often depend on unknown hyperparameters $\boldsymbol{\lambda}$. An *empirical Bayes* approach uses $\boldsymbol{y}$ to estimate $\boldsymbol{\lambda}$ by the values most consistent with the data in maximizing the marginal distribution $f(\boldsymbol{y} \mid \boldsymbol{\lambda})$ of $\mathbf{y}$.
- As alternative to beta, can use prior distribution (such as normal) for the *logit*, $\log[\pi/(1-\pi)]$, to relate to *logistic regression* modeling methods for binary response with explanatory variables.

BAYESIAN INFERENCE FOR MEANS

- Bayesian inference for a mean of a normal sampled distribution, assuming variance known
- Bayesian inference for a normal mean with variance unknown
- Bayesian inference for means using improper prior distributions
- Bayesian inference for comparing two means
- Bayesian inference for multiple means

Bayesian Inference for Normal Mean μ, Conditional on σ^2

Usual approach assumes independent $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from a $\mathcal{N}(\mu, \sigma^2)$ distribution, conditional on μ and σ^2, and a normal prior distribution for μ.

We first treat σ as known, so we do not also need a prior distribution for it. Unrealistic n practice, but relevant also for model with σ unknown.

Product of the normal *pdf*s $\{f(y_i \mid \mu), i = 1, \ldots, n\}$ yields joint *pdf*

$$f(\boldsymbol{y} \mid \mu) = \prod_{i=1}^{n} f(y_i \mid \mu) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma}\right) e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \quad \propto \quad \exp\left[-\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i-\mu)^2\right]$$

$$= \quad \exp\left\{-\left(\frac{1}{2\sigma^2}\right)\left[\sum_{i=1}^{n}[(y_i-\bar{y})+(\bar{y}-\mu)]^2\right.\right.$$

$$= \quad \exp\left\{-\left(\frac{1}{2\sigma^2}\right)\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)\right.\right.$$

Once we observe the data, the kernel of the likelihood function is

$$L(\mu) \propto e^{-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}}.$$

Conjugate Normal Prior and Posterior Distributions

With a $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution as the prior *pdf* $p(\mu \mid \mu_0, \sigma_0)$ for μ,

$$p(\mu \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \mu) p(\mu \mid \mu_0, \sigma_0) = L(\mu) p(\mu \mid \mu_0, \sigma_0) \propto e^{-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}.$$

Combining terms in a common exponent, keeping the terms involving μ, writing those terms as quadratic in μ and completing the square, this product is proportional to the *pdf* of another normal distribution,

$$p(\mu \mid \boldsymbol{y}) = \exp\left\{ -\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \left(\mu - \frac{\bar{y}\sigma_0^2 + \mu_0(\sigma^2/n)}{\sigma_0^2 + (\sigma^2/n)} \right)^2 \right] \right\}$$

In summary, posterior *pdf* is proportional to

$$e^{-\frac{(\mu-\tilde{\mu})^2}{2\tilde{\sigma}^2}} \quad \text{with} \quad \tilde{\mu} = \frac{\bar{y}\sigma_0^2 + \mu_0(\sigma^2/n)}{\sigma_0^2 + (\sigma^2/n)} \quad \text{and} \quad \tilde{\sigma}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1},$$

the kernel of a $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ *pdf*. Thus, normal distribution is *conjugate prior distribution* for normal likelihood function. The *precision* $1/\tilde{\sigma}^2$ equals sum of precision (n/σ^2) of data and precision of prior distribution.

Shrinkage of Sample Mean in Posterior Distribution of Mean

Mean $\tilde{\mu}$ of posterior normal distribution is weighted average of sample mean $\bar{y}$ and prior mean μ_0,

$$\tilde{\mu} = \frac{\bar{y}\sigma_0^2 + \mu_0(\sigma^2/n)}{\sigma_0^2 + (\sigma^2/n)} = \left(\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\right)\bar{y} + \left(\frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\right)\mu_0 = w\bar{y} + (1-w)\mu_0,$$

with weight $w = \sigma_0^2/[\sigma_0^2 + (\sigma^2/n)]$. Shrinks sample mean toward prior mean μ_0.

- Weight attached to $\bar{y}$ increases toward 1 as n increases.
- For fixed n, if we let σ_0^2 grow unboundedly,

  Posterior mean $\tilde{\mu}$ converges to $\bar{y}$.

  Posterior variance $\tilde{\sigma}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$ converges to σ^2/n.

  Posterior distribution of μ converges to $\mathcal{N}(\bar{y}, \sigma^2/n)$.

Inference for a Normal Mean with Variance Unknown

We again use a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution for μ

To obtain conjugacy, we let the precision $x = 1/\sigma^2$ have a *gamma* dist., $f(x \mid \alpha, \beta) \propto x^{\alpha-1}e^{-\beta x}$. Then, $z = 1/x = \sigma^2$ has an *inverse gamma* dist.,

$$f(z \mid \alpha, \beta) \propto z^{-(\alpha+1)}e^{-\beta/z}, \quad \text{for } z > 0,$$

for a *shape parameter* $\alpha > 0$ and a *scale parameter* $\beta > 0$.

- Skewed to right
- For fixed scale parameter β, it is increasingly peaked and bell-shaped as shape parameter α increases.
- For fixed shape parameter α, when the mean and standard deviation exist, standard deviation is proportional to mean, and both are proportional to scale parameter.
- Distribution is highly disperse when α and β are close to 0.

Conjugate Inverse Gamma Prior and Posterior for Variance

- Decompose joint prior distribution as $p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2)$, with $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution for $p(\mu \mid \sigma^2)$
- Likewise, for posterior, we factor $p(\mu, \sigma^2 \mid \boldsymbol{y})$ as

$$p(\mu, \sigma^2 \mid \boldsymbol{y}) = p(\mu \mid \sigma^2, \boldsymbol{y})p(\sigma^2 \mid \boldsymbol{y}).$$

  where $p(\mu \mid \sigma^2, \boldsymbol{y})$ is $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ distribution from treating σ^2 as known, and can show $p(\sigma^2 \mid \boldsymbol{y})$ is also inverse gamma.
- With modern computational methods, unnecessary to use conjugate prior distribution, and they usually are not available for most models. Another approach uses a prior distribution that is not conjugate but simpler to understand, such as exponential prior distribution for σ with rate β (special case of gamma distribution with $\alpha = 1$),

$$p(\sigma \mid \beta) = \beta e^{-\beta\sigma}, \quad \sigma \geq 0.$$

  Mean = standard deviation $= 1/\beta$, which is a scale parameter, median $= \log(2)/\beta = 0.693/\beta$.

Marginal Posterior Distribution of Mean and t Distribution

Marginal posterior distribution $p(\mu \mid \boldsymbol{y})$ of mean is

$$p(\mu \mid \boldsymbol{y}) = \int_0^\infty p(\mu, \sigma^2 \mid \boldsymbol{y}) d\sigma^2 = \int_0^\infty p(\mu \mid \sigma^2, \boldsymbol{y}) p(\sigma^2 \mid \boldsymbol{y}) d\sigma^2,$$

where $p(\mu \mid \sigma^2, \boldsymbol{y})$ and $p(\sigma^2 \mid \boldsymbol{y})$ are the normal and inverse gamma pdfs

This rather messy integration has the result that $p(\mu \mid \boldsymbol{y})$ is a generalized t distribution having location and scale parameters.

The textbook shows details,including the use of an alternative expression for $p(\mu \mid \sigma^2)$ as $\mathcal{N}(\mu_0, \sigma^2/n_0)$ for some real number $n_0 > 0$ that represents the number of *imaginary* prior observations.

In practice, software can approximate the posterior distribution and provide posterior intervals and tail probabilities.

Example: Bayesian Analysis for Anorexia Therapy

An experimental study compared three therapies for young girls suffering from anorexia. For each girl, weight was measured before and after a fixed period of treatment designed to aid weight gain.

For 29 girls undergoing cognitive behavioral (*cb*) therapy, weights at end of study minus weights at beginning were, in pounds:

$$1.7,\ 0.7,\ \ldots,\ 15.4,\ -0.7$$

Frequentist statistics for μ = population mean change in weight.

```
-------------------------------------------------------------------------------
> Anorexia <- read.table("http://bayes4ds.rwth-aachen.de/data/Anorexia.dat", header=TRUE)
> head(Anorexia, 2)               # first 2 lines in Anorexia data file
  subject therapy before after    # before and after therapy treatment
1       1      cb   80.5  82.2    # weight change = 82.2 - 80.5 = 1.7
2       2      cb   84.9  85.6    # cb is cognitive behavioral therapy
> y <- Anorexia$after[Anorexia$therapy=="cb"] - Anorexia$before[Anorexia$therapy=="cb"]
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -9.100  -0.700   1.400   3.007   3.900  20.900
> sd(y)      # standard deviation of weight changes for cb therapy
[1] 7.308504
> t.test(y)
t=2.2156, df=28, p-value=0.03502 # one-sided P-value = 0.035/2 = 0.0175 for H1: mu > 0
95 percent confidence interval: 0.2268902 5.7869029
-------------------------------------------------------------------------------
```

Bayesian Inference for Mean Weight Change

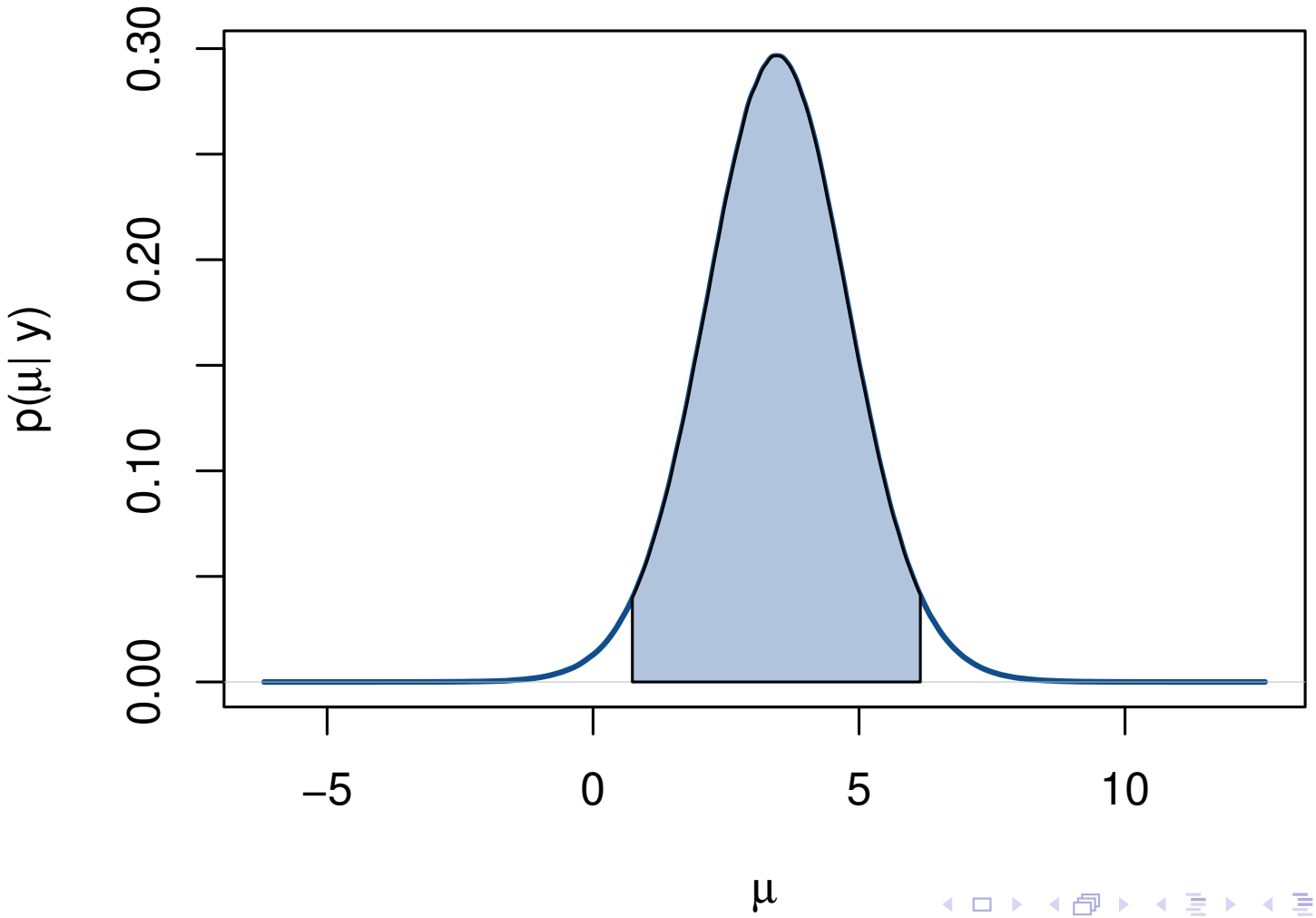
We use a highly disperse inverse gamma prior distribution for σ^2, with $\alpha = \beta = 0.5$ and a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution with $\mu_0 = 10$ and $\sigma_0 = 6$ for μ, based on a 95% prior belief that $-2 \leq \mu \leq 22$.

To obtain the Bayesian fit with software, applying MCMC = Markov chain Monte Carlo), we use an `R` package `MCMCpack` that can incorporate an inverse gamma prior distribution for σ^2:

```
> library(MCMCpack)
> y <- Anorexia$after[Anorexia$therapy=="cb"] - Anorexia$before[Anorexia$therapy=="cb"]
> fit.bayes <- MCMCregress(y ~ 1, mcmc = 20000000, b0=10, B0=1/6^2, c0=1, d0=1)
>   # mean has normal prior dist. with mean b0 = 10, precision B0 (std. dev. = 6)
>   # variance has inverse gamma prior dist. (c0/2 = shape, d0/2 = scale, both 0.5)
>   # mcmc = 20000000 uses huge number of iterations for MCMC fitting
>   # coding y ~ 1 fits linear model with intercept but no explanatory var's
> summary(fit.bayes)
1. Empirical mean and standard deviation for each variable
              Mean     SD
(Intercept)  3.359  1.348 # mean of posterior dist. for mu is 3.36
2. Quantiles for each variable:
               2.5%    25%    50%    75%  97.5%
(Intercept)  0.7327  2.461  3.347  4.243   6.05 # 95% posterior interval is (0.73, 6.05)
sigma2      32.7454 44.413 52.846 63.544  93.34 # quantiles for post. dist. of variance
> mean(fit.bayes[,1] <= 0) # <= represents "less than or equal to"
> # mean of indicators of whether values from post. dist. are less than or equal to 0
[1] 0.0066868              # provides posterior P(mu <= 0) = 0.0067
```

Figure for posterior distribution of mean weight change

Posterior distribution of mean weight change μ and 95% posterior interval for μ, based on normal and inverse gamma conjugate prior distributions, for cognitive behavioral therapy in anorexia study.



Inference for Means Using Improper Prior Distributions

To reduce the subjective element of selecting prior distributions for μ and σ^2, we can use an *improper prior distribution* for each of them, such as $p(\mu) = 1$ for $-\infty < \mu < \infty$, $p(\sigma^2) = 1/\sigma^2$ for $\sigma > 0$.

These are not true probability distributions, because their integrals over the possible parameter values are infinite and we cannot simulate from them.

Nonetheless, combining these improper prior distributions with likelihood function, posterior distribution of μ can be normalized to be proper. Posterior distribution of μ is characterized by

$$T = \frac{\mu - \bar{y}}{s/\sqrt{n}}$$

having t distribution with $df = n - 1$. Resulting Bayesian inferences identical to frequentist inferences using t distribution. e.g., 95% posterior interval for μ is $\bar{y} \pm t_{0.025,n-1}(s/\sqrt{n})$. For anorexia data:

$$\bar{y} \pm t_{0.025,28}(s/\sqrt{n}),\ 3.007 \pm 2.048(7.309/\sqrt{29}),\ (0.23, 5.79).$$

Inference for Comparing Two Means

For randomly selected Y_1 from group 1 and Y_2 from group 2, the standard model for the analysis assumes:

- Y_1 has a $\mathcal{N}(\mu_1, \sigma^2)$ distribution,
- Y_2 has a $\mathcal{N}(\mu_2, \sigma^2)$ distribution.

Data: Independent observations $\boldsymbol{y}_1 = (y_{11}, y_{21}, \ldots, y_{n_1 1})$ from group 1 and $\boldsymbol{y}_2 = (y_{12}, y_{22}, \ldots, y_{n_2 2})$ from group 2, with means $\overline{y}_1$ and $\overline{y}_2$ and standard deviations s_1 and s_2.

For frequentist analysis using *pooled estimate* s of common value σ,

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ with } s = \sqrt{\frac{\sum_i (y_{i1} - \overline{y}_1)^2 + \sum_i (y_{i2} - \overline{y}_2)^2}{n_1 + n_2 - 2}}$$

has the t distribution with $df = n_1 + n_2 - 2$. The $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2, n_1+n_2-2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Bayesian Inference for Comparing Two Means

Priors:

- We could treat μ_1 and μ_2 as independent from a $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution, for values provided of μ_0 and σ_0.
- We could use an inverse gamma prior distribution for σ^2 or a simpler prior distribution that is not conjugate, such as an exponential distribution for σ.
- Posterior inferences that agree with frequentist inferences result from using improper prior distributions $p(\mu_j) = 1$ for $-\infty < \mu_j < \infty$, $j = 1, 2$, and $p(\sigma^2) = 1/\sigma^2$ for $\sigma > 0$.

For large values of n_1 and n_2, marginal posterior distribution of $\mu_1 - \mu_2$ is close to

$$\mathcal{N}\left(\bar{y}_1 - \bar{y}_2, s^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right).$$

Example: Mean Housework Hours for Men and Women

GSS question in U.S.: "On average, how many hours a week do you personally spend on household work, not including childcare and leisure time activities?"

For 582 men and 694 women, sample distributions are skewed right, with means 8.31 for men and 11.87 for women, for difference of 3.56.

```
------------------------------------------------------------------------------
> Household <- read.table("http://bayes4ds.rwth-aachen.de/data/Household.dat",header=TRUE)
> tapply(Household$time, Household$gender, summary) # gender = 0 men, 1 women
$`0`  # men
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   6.000   8.309  10.000  72.000
$`1`  # women
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    4.00    8.00   11.87   15.00  100.00
> tapply(Household$time, Household$gender, sd)
        0         1
 9.437386 12.752392 # standard deviations of the observations for men and women
> y1 <- Household$time[Household$gender==1]
> y2 <- Household$time[Household$gender==0]
> t.test(y1, y2, var.equal=TRUE, conf.level=0.95) # var.equal=TRUE specifies equal var's
t = 5.5833, df = 1274, p-value = 2.88e-08
95 percent confidence interval: # 95% CI for difference between means for women and men
 2.312596 4.818127
------------------------------------------------------------------------------
```

Bayesian Modeling with brms R Package

We use a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution for each of μ_1 and μ_2 with $\mu_0 = 10$ and $\sigma_0 = 4$. This corresponds to a 0.95 prior probability that each of μ_1 and μ_2 falls between about 2 and 18.

Suppose prior probability is 0.99 that $\sigma \leq 100$; this is satisfied by an exponential prior distribution with rate parameter $\beta = 0.046$, which has mean = standard deviation $= 1/\beta = 1/0.046 = 21.7$.

```
------------------------------------------------------------------------------
> pexp(100, rate=0.046) # cumulative probability of exponential distribution at 100
[1] 0.9899482           #   equals 0.9899 when rate parameter = 0.046
> qexp(0.99, 0.046)     # cumulative probability of exponential distribution is 0.99
[1] 100.1124            #   at 100.1124 when rate parameter = 0.046
------------------------------------------------------------------------------
```

We use R package brms (Bayesian Regression Models using 'Stan'), as it can handle a wide variety of models and has similar form as the glm function that is popular for frequentist analyses.
To use this, install RTools by downloading it from https://cran.r-project.org/bin/windows/Rtools and then install rstan and brms from the R Console and load the packages:

Bayesian Comparison of Mean Housework Hours

```
-------------------------------------------------------------------------------------------------------
> install.packages("rstan", repos = "https://cloud.r-project.org/", dependencies = TRUE)
> library(rstan)
> options(mc.cores = parallel::detectCores()) # this line and next not required, but
> rstan_options(auto_write = TRUE)            #   recommended for good performance of brms
> install.packages("brms")
> Household$Women <- Household$gender   # the variable Women is 1 for women and 0 for men
> Household$Men <- 1 - Household$gender # the variable Men is 1 for men and 0 for women
> library(brms)
> fit.bayes <- brm(time ~ -1 + Men + Women, data=Household, family=gaussian, # gaussian = normal dist.
+                  prior = prior(normal(10, 4)) + prior(exponential(0.046), class=sigma))
> summary(fit.bayes)
Regression coefficients:
      Estimate Est.Error l-95% CI u-95% CI
Men       8.33      0.46     7.42     9.24 # posterior mean for men is 8.33
Women    11.85      0.44    10.97    12.69 # posterior mean for women is 11.85
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI
sigma    11.37      0.23    10.93    11.84 # mean of posterior dist. of sigma is 11.37
-------------------------------------------------------------------------------------------------------
```

Posterior means are very similar to sample means (11.87 and 8.31), because group sample sizes are large.

The model $E(Y_i \mid \boldsymbol{\beta}) = \beta_1 z_{i1} + \beta_2 z_{i2}$ uses indicators for the two groups:

$z_{i1} = 1$ if observation i is male, 0 otherwise

$z_{i2} = 1$ if observation i is female, 0 otherwise

Posterior Inference for Difference of Means

We can derive simulated values from the posterior distribution of the difference $\mu_1 - \mu_2$ by taking the difference between the simulated values from the posterior distributions of μ_1 and μ_2.

```
> posterior <- as.array(fit.bayes) # simulated values from posterior saved in an array
> diff_means <- posterior[,,2] - posterior[,,1] # mu1 - mu2 values from posterior array
            # posterior[,,2] has values for women and posterior[,,1] has values for men
> mean(diff_means <= 0) # Posterior probability that mu1 - mm2 less than or equal to 0,
[1] 0                   #   corresponds to P(H0) for H_1: mu_1 > mu_2
> quantile(diff_means, probs = c(0.025, 0.975)) # 95% posterior interval for mu1 - mu2
    2.5%    97.5%
2.264917 4.739305
```

With prior distributions that are flat relative to likelihood function, the posterior probability that $\mu_1 \leq \mu_2$, which is essentially 0, is very similar to frequentist one-sided P-value for testing H_0:$\mu_1 = \mu_2$ against H_1: $\mu_1 > \mu_2$, for which the implicit null hypothesis is H_0: $\mu_1 \leq \mu_2$.

Bayesian analysis: Probability is essentially 0 that population mean smaller for women than men. Frequentist analysis: If population means were equal, probability would be 1.4×10^{-8} that difference in sample means between women and mean would be at least as large as observed value of 3.56.

Bayesian Inference for Multiple Means

Bayesian inferences for comparing two means extend to set of means for c groups, with $c > 2$. We regard the c groups as categories of a categorical variable, referred to as a *factor*.

Common frequentist and Bayesian methods assume Y_{ij} for subject i in group j has $\mathcal{N}(\mu_j, \sigma^2)$ distribution, for $i = 1, \ldots, n_j$, $j = 1, \ldots, c$. The methods are special cases of linear models that may have multiple explanatory variables, categorical and/or quantitative.

Prior distributions treat $\{\mu_j\}$ as independent from some $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution, for particular values of μ_0 and σ_0, where σ_0^2 summarizes prior knowledge about between-groups variability.

Possible prior distributions for within-groups variability include inverse gamma distribution for σ^2 or exponential distribution for σ.

A more general *hierarchical* approach has hyperprior distributions for the hyperparameters of the prior distribution (Sec. 5.6.2 of textbook).

To show relative impacts on posterior means $\{\tilde{\mu}_j\}$ of within-groups variance σ^2 and between-groups variance σ_0^2, consider $\{\tilde{\mu}_j\}$ if σ^2 were known:

$$\tilde{\mu}_j = w_j \overline{y}_j + (1 - w_j)\mu_0, \text{ where } w_j = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n_j},$$

a weighted average of sample mean $\overline{y}_j$ for group j and prior mean μ_0,

- As n_j increases, w_j increases toward 1, and $\tilde{\mu}_j$ gets closer to $\overline{y}_j$.
- For any n_j, this also happens as σ_0 increases.
- By contrast, as σ_0 and/or $\{n_j\}$ decrease, the posterior mean estimates more greatly shrink the sample means toward μ_0.

Example: Comparing Mean Phone Holding Times

For its toll-free telephone number for making reservations, an airline conducted a randomized experiment to analyze whether subjects would remain on hold longer if they heard (a) an advertisement about the airline (group A), (b) Muzak (group M), or (c) classical music (group C).

For every 1000th call in a particular week, the experiment randomly selected one of three recordings and measured the number of minutes the caller remained on hold before hanging up.

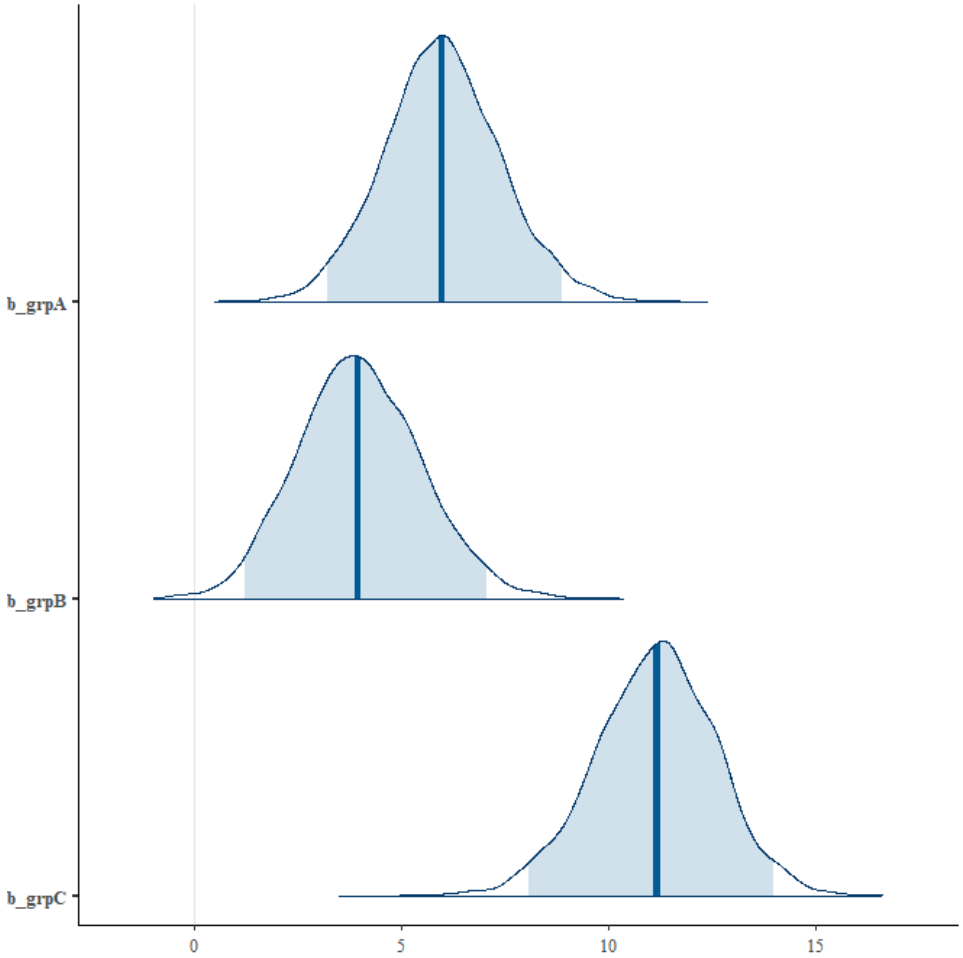
Because $\{\mu_1, \mu_2, \mu_3\}$ not expected to be large or have great variability but σ could be large, used $\mathcal{N}(8, 3^2)$ prior dist. for $\{\mu_j\}$ and exponential prior dist. for σ with mean $=$ standard deviation $= 10$.

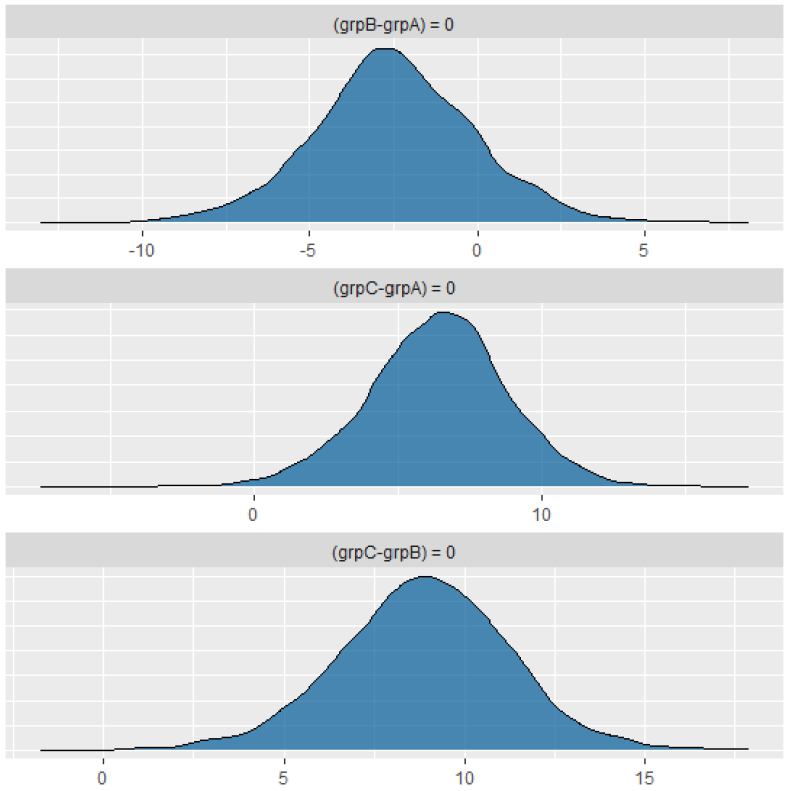
Recording	Observations	Mean	Standard deviation
Advertisement (A)	5, 1, 11, 2, 8	5.4	4.2
Muzak (M)	0, 1, 4, 6, 3	2.8	2.4
Classical (C)	13, 9, 8, 15, 15	12.0	3.3

Bayesian Comparison of Mean Holding Times

```
> time <- c(5, 1, 11, 2, 8, 0, 1, 4, 6, 3, 13, 9, 8, 15, 15)
> group <- c(rep("A", 5), rep("B", 5), rep("C", 5))
> Results <- data.frame(time, group)
> library(brms)
> Results$grp <- factor(Results$group) # optional here, but needed if group is numerical
> fit.bayes <- brm(time ~ -1 + grp, data=Results, family=gaussian, # gaussian = normal dist.
+         prior = prior(normal(8, 3)) + prior(exponential(0.10), class=sigma))
> summary(fit.bayes)
     Estimate Est.Error l-95% CI u-95% CI
grpA     6.01      1.47     3.14     9.01 # mean of posterior dist. of mu_A is 6.01
grpB     4.04      1.51     1.21     7.21 # mean of posterior dist. of mu_B is 4.04
grpC    11.05      1.49     7.94    13.80 # mean of posterior dist. of mu_C is 11.05
sigma    3.74      0.86     2.48     5.85 # mean of posterior dist. of sigma is 3.74
> library(bayesplot)                      # posterior distributions and histogram of them
> mcmc_areas(fit.bayes, pars = c("b_grpA", "b_grpB", "b_grpC"), prob=0.95)
> mcmc_hist(fit.bayes, pars=c("b_grpA", "b_grpB", "b_grpC"))
> hyp <- hypothesis(fit.bayes, c("grpB-grpA = 0", "grpC-grpA = 0", "grpC-grpB = 0"))
> hyp # use this to get posterior intervals for differences of means
Hypothesis Tests for class b:
       Hypothesis Estimate Est.Error CI.Lower CI.Upper
1 (grpB-grpA) = 0    -1.98      2.04    -6.01     2.27
2 (grpC-grpA) = 0     5.04      2.12     0.67     9.03
3 (grpC-grpB) = 0     7.02      2.19     2.39    11.00
> plot(hyp)
```

The posterior means of (6.0, 4.0, 11.0) shrink the sample means, which are (5.4, 2.8, 12.0), toward the prior mean of 8.





Posterior distributions of mean holding times for A (Advertisement), B (Muzak) and C (Classical) and for their pairwise differences (right).

MCMC BAYESIAN POSTERIOR COMPUTATION AND DIAGNOSTICS

- The Monte Carlo method
- Markov chains and the Markov property
- The idea behind Markov chain Monte Carlo (MCMC)
- Example of MCMC: Metropolis algorithm
- Diagnostics for MCMC: Effective sample size, R-hat, trace plots
- How close are reported posterior means to the true ones?

What Is the Computational Difficulty?

From Bayes' Theorem,

$$p(\theta \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \theta)p(\theta)}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y} \mid \theta)p(\theta)}{\int_{\Theta} f(\boldsymbol{y} \mid \tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

- Calculating the numerator is straightforward, once we select the probability function $f(y_i \mid \theta)$ and prior distribution $p(\theta)$.
- The denominator $f(\boldsymbol{y})$, the marginal probability function of the data, is a normalizing constant, not involving θ. However, we need to calculate it to obtain $p(\theta \mid \boldsymbol{y})$, and that integral may be complex.
- With multiple parameters, we also need integration to approximate the marginal posterior distribution of the parameter of special interest, by integrating out nuisance parameters, and to approximate summaries such as the posterior mean of that parameter. For most models, these integrals are not analytically available.

The Monte Carlo Method

Monte Carlo method: For independent random variables $Y_1, Y_2, \ldots, Y_T$ simulated from a probability mass function or density function $f(y)$, their empirical distribution forms a Monte Carlo approximation for $f(y)$. For a function $g(y)$, $\frac{1}{T}\sum_{i=1}^{T} g(Y_i)$ is a Monte Carlo approximation for $E[g(Y)]$.

- Monte Carlo methods use simulations incorporating repeated random sampling to solve problems that are difficult or impossible to solve analytically.
- Polish–American mathematician Stanislaw Ulam (for whom *Stan* programming language is named) devised Monte Carlo method in 1946 while working at Los Alamos National Laboratory on thermonuclear weapon development.
- Frequentist methods also sometimes use Monte Carlo methods, for instance to find ML estimates for non-standard distributions or in models with unobservable components, such as to adjust for some observations being missing.

Example of Monte Carlo

We used Monte Carlo to simulate the posterior distribution of $\pi_1 - \pi_2$ proportions of females and males who believe in hell. Let's get a posterior interval for the *risk ratio* π_1/π_2, using the posterior distributions,

Beta(499, 177) for π_1 and Beta(317, 153) for π_2:

```
-------------------------------------------------------------------------------
> pi1 <- rbeta(50000000, 499, 177) # simulating from posterior distribution of pi1
> pi2 <- rbeta(50000000, 317, 153) # simulating from posterior distribution of pi2
-------------------------------------------------------------------------------
> quantile(pi1/pi2, c(0.025, 0.975))
    2.5%    97.5%
1.014374 1.184076   # simulated 95% percentile interval for risk ratio
```

The posterior probability is 0.95 that the population proportion in the U.S. who believe in hell is between 1.01 and 1.18 times as high for women as for men.

Markov Chains and the Markov Property

Markov chain: A stochastic process such that conditional on observations at times $t, t-1, t-2, \ldots, 1$, the state of the process at time $t+1$ depends probabilistically only on the state at time t (called the *Markov property*).

Simplistic example (*random walk*): $x_0 = 0$, and then for $t = 0, 1, 2, \ldots$, X_{t+1} equals $x_t + 1$ with probability $\frac{1}{2}$ and equals $x_t - 1$ with probability $\frac{1}{2}$. It is a Markov chain because, given $\{x_0, x_1, \ldots, x_t\}$, the distribution for X_{t+1} depends only on x_t.

Markov chains are described in terms of whether they have certain properties. Under certain conditions (e.g., possible to move from any one state to any other, and starting from any state, guaranteed to return to it in a sufficiently long Markov chain) the Markov chain has a long-term *stationary distribution*.

Markov Chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) is a simulation method that uses the *Monte Carlo method* to simulate a long sequence of correlated random variables that are a *Markov chain*.

- For a random sequence of possible values $\theta^{(1)}, \theta^{(2)}, \ldots$ of the parameter θ, let $q(\theta^{(t+1)} \mid \theta^{(t)}, \theta^{(t-1)}, \ldots, \theta^{(1)})$ denote the *pdf* for $\theta^{(t+1)}$ conditional on previous θ values. The Markov property specifies that this satisfies

$$q(\theta^{(t+1)} \mid \theta^{(t)}, \theta^{(t-1)}, \ldots, \theta^{(1)}) = q(\theta^{(t+1)} \mid \theta^{(t)}).$$

- The MCMC algorithm uses the Monte Carlo method with some starting point $\theta^{(0)}$ to generate $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$ from conditional distributions $q(\theta^{(1)} \mid \theta^{(0)}), q(\theta^{(2)} \mid \theta^{(1)}), \ldots, q(\theta^{(T)} \mid \theta^{(T-1)})$.
- The distribution of this sequence of values approximates $p(\theta \mid \boldsymbol{y})$ because $p(\theta \mid \boldsymbol{y})$ is the stationary distribution of the Markov chain.

MCMC Iterations

- As the number of *iterations* T in the process increases, the empirical distribution of $\{\theta^{(t)}\}$ more closely approximates $p(\theta \mid \boldsymbol{y})$.
- In practice, $\theta^{(t)}$ when t is large better represents posterior distribution, as the Markov chain more closely approaches stationary distribution. To approximate posterior distribution better, most software does not use $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T_0)}\}$ from a *warm-up* period with T_0 much less than T, also referred to as the *burn-in*.
- Software also typically generates multiple Markov chains and uses values following the warm-up period for each to approximate the posterior distribution, combining them all to do this.
- By default, the `brm` function in `brms` `R` package uses four parallel Markov chains, with $T_0 = 1000$ and $T = 2000$.

Example of MCMC: The Metropolis Algorithm

If two possible values θ_a and θ_b for θ have $p(\theta_b \mid \boldsymbol{y})/p(\theta_a \mid \boldsymbol{y}) = r$, then the number of values in the Markov chain generated close to θ_b should be about r times the number of values generated close to θ_a.

Outline of the Metropolis MCMC algorithm:
For posterior distribution $p(\theta \mid \boldsymbol{y})$, given the value $\theta^{(t)}$ generated at step t of the algorithm:

1. Randomly generate θ^* from a symmetric probability distribution (called the *proposal distribution*) having mean $\theta^{(t)}$, such as a normal distribution having that mean.
2. If $r = p(\theta^* \mid \boldsymbol{y})/p(\theta^{(t)} \mid \boldsymbol{y}) \geq 1$, then take $\theta^{(t+1)} = \theta^*$ as the next value of the Markov chain.
3. If $r < 1$, then let $\theta^{(t+1)} = \theta^*$ with probability r and let $\theta^{(t+1)} = \theta^{(t)}$ with probability $1 - r$.

Metropolis Algorithm (continued)

How do we calculate r when we do not know the posterior distribution?

Bayes' Theorem tells us that

$$r = \frac{p(\theta^* \mid \boldsymbol{y})}{p(\theta^{(t)} \mid \boldsymbol{y})} = \frac{\frac{f(\boldsymbol{y}|\theta^*)p(\theta^*)}{f(\boldsymbol{y})}}{\frac{f(\boldsymbol{y}|\theta^{(t)})p(\theta^{(t)})}{f(\boldsymbol{y})}} = \frac{f(\boldsymbol{y} \mid \theta^*)p(\theta^*)}{f(\boldsymbol{y} \mid \theta^{(t)})p(\theta^{(t)})} = \frac{L(\theta^*)p(\theta^*)}{L(\theta^{(t)})p(\theta^{(t)})}.$$

We can calculate each term in the last fraction because we know the likelihood function $L(\theta)$ and the prior distribution $p(\theta)$.

In particular, we can find r without needing to know the normalizing constant $f(\boldsymbol{y})$, which cancels out in the second equality.

A Few More Details

- The particular choice of the family of distributions for the proposal distribution is not crucial, but its spread is important.
- With too small a standard deviation for the proposal distribution, the moves are very short and it takes a long time to move around the entire posterior distribution. With too large a standard deviation, a potential move will often be out into a tail, and the ratio r will be so small that with high probability the Markov chain stays where it is, not exploring the posterior *pdf* well.
- The Metropolis algorithm uses a symmetric proposal distribution, meaning that for two possible values θ_a and θ_b, $q(\theta_b \mid \theta_a) = q(\theta_a \mid \theta_b)$. The *Metropolis–Hastings algorithm* is a generalization for which the proposal distribution need not be symmetric, which can enable exploring the posterior distribution more efficiently, with the goal of faster convergence to the stationary distribution.

Other Methods and Generalizations

- *Gibbs sampling*: For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, each iteration cycles through the p parameters, using Monte Carlo to sample from the conditional distribution of one of them conditional on values for all the others. That is, for $j = 1, \ldots, p$, determine $p(\theta_j \mid \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$, and at iteration $t+1$ obtain $\theta_j^{(t+1)}$ by using Monte Carlo to sample from $p(\theta_j \mid \theta_1^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)})$.
- Most software (such as the `brms` package in `R`) uses modifications of Metropolis and Metropolis–Hastings algorithms that provide faster convergence to the stationary distribution.
- *Hamiltonian Monte Carlo*: Uses, with differential geometry, gradients of the log-posterior distribution to generate trajectories that efficiently explore regions of high probability mass.
- The `brms` package in `R` uses an extension of Hamiltonian Monte Carlo called the *No U-turn sampler* (NUTS) that stops when the simulated trajectory begins to double back and retrace its path.

Diagnostics for Checking a MCMC Process

When we implement an MCMC process, the precision of the approximation for the posterior distribution depends on the *number* of simulations in the process that the software uses.

Diagnostics help ensure the posterior distribution is adequately sampled and summary posterior results shown by software are trustworthy. We illustrate by viewing more output from the example comparing mean weekly household work times μ_1 for women and μ_2 for men.

```
------------------------------------------------------------------------------
> Household <- read.table("http://bayes4ds.rwth-aachen.de/data/Household.dat",header=TRUE)
>   # In Household, the variable gender is 0 for men and 1 for women
> Household$Women <- Household$gender   # indicator variable, 0 for men and 1 for women
> Household$Men <- 1 - Household$gender # indicator variable, 1 for men and 0 for women
> library(brms)
> fit.bayes <- brm(time ~ -1 + Women + Men, family=gaussian, data=Household,
+       prior = prior(normal(10, 4), class=b) + prior(exponential(0.046), class=sigma))
> print(fit.bayes, digits=3)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
Women   11.850     0.435   11.019   12.708 1.001     3793     2816
Men      8.329     0.470    7.419    9.260 1.001     4163     3025
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
sigma   11.370     0.217   10.957   11.799 1.000     3983     3253
------------------------------------------------------------------------------
```

Effective Sample Size

Effective sample size (ESS) indicates the equivalence of the correlated draws in terms of a number of independent draws.

`Bulk_ESS`: The *bulk effective sample size* for women of 3793 means for approximating the bulk of the posterior distribution, the MCMC process of $4(1000) = 4000$ values provides information equivalent to 3793 independent draws. Bulk ESS values above about 400 usually indicate that approximations for posterior means are decent.

`Tail_ESS`: The *tail effective sample size* gives effective number of independent draws for approximating tail characteristics of the posterior distribution. The tail ESS indicates how well we can trust reported posterior intervals.

By contrast, here is the output if we had used a length of only 100 (instead of 2000) for each of the four Markov chains, with a warm-up of length 50 for each:

```
-------------------------------------------------------------------------------------
> fit.bayes2 <- brm(time ~ -1 + Women + Men, family=gaussian, data=Household,
+   prior = prior(normal(10,4),class=b) + prior(exponential(0.046),class=sigma), iter=100)
> summary(fit.bayes2)                                          # iter specifies length of 100
  Draws: 4 chains, each with iter=100; warmup=50; thin=1; #   for each chain with warm-up
         total post-warmup draws=200                     #   of 50, so 4(100-50) = 200
Regression coefficients:                                 #   post warm-up draws
        Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Women      11.85      0.59    10.05    12.59 1.04       86       29
Men         8.22      0.62     6.23     9.10 1.07       58       28
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    11.37      0.22    11.02    11.84 1.01      212      141
-------------------------------------------------------------------------------------
```

Now, ESS values are small; compare posterior intervals of (10.0, 12.6) for women and (6.2, 9.1) for men to earlier ones of (11.0, 12.7) for women and (7.4, 9.3) for men.

Gauging Monte Carlo Standard Errors

fit.bayes reported Bayesian posterior means to three decimal places. But how much precision do we actually obtain? Here is R code to tell us:

```
> library(bayestestR)
> mcse(fit.bayes)
  Parameter        MCSE
1   b_Women 0.007287670    # Markov chain Monte Carlo standard errors for the
2     b_Men 0.007643984    #   approximations for the actual posterior means
```

For greater precision, we can be quite sure that the actual posterior mean for men falls within about $8.329 \pm 2(0.0076)$, or 8.314 to 8.344.

Alternatively, one can run MCMC for much longer.

```
> fit.bayes3 <- brm(time ~ -1 + Women + Men, family=gaussian, data=Household,
+        prior = prior(normal(10, 4), class=b) + prior(exponential(0.046), class=sigma),
+        iter = 2000000, warmup = 1000)
> print(fit.bayes3, digits=3)
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
Women   11.853     0.429   11.012   12.694 1.000  8268447  6179147
Men      8.333     0.468    7.414    9.250 1.000 10032898  6374529
sigma   11.370     0.225   10.939   11.822 1.000  9846006  6370968
> mcse(fit.bayes3)        # from bayestestR package
  Parameter         MCSE
1   b_Women 0.0001492393
2     b_Men 0.0001478554
```

R-hat Values for Comparing Parallel Markov Chains

When MCMC algorithm uses parallel Markov chains, we hope they all converge to a stationary distribution that is posterior distribution $p(\theta \mid \boldsymbol{y})$.

If not, the total variability in values observed when chains are mixed together, reflecting *between-chain variability* as well as *within-chain variability*, will tend to be greater than within-chain variability.

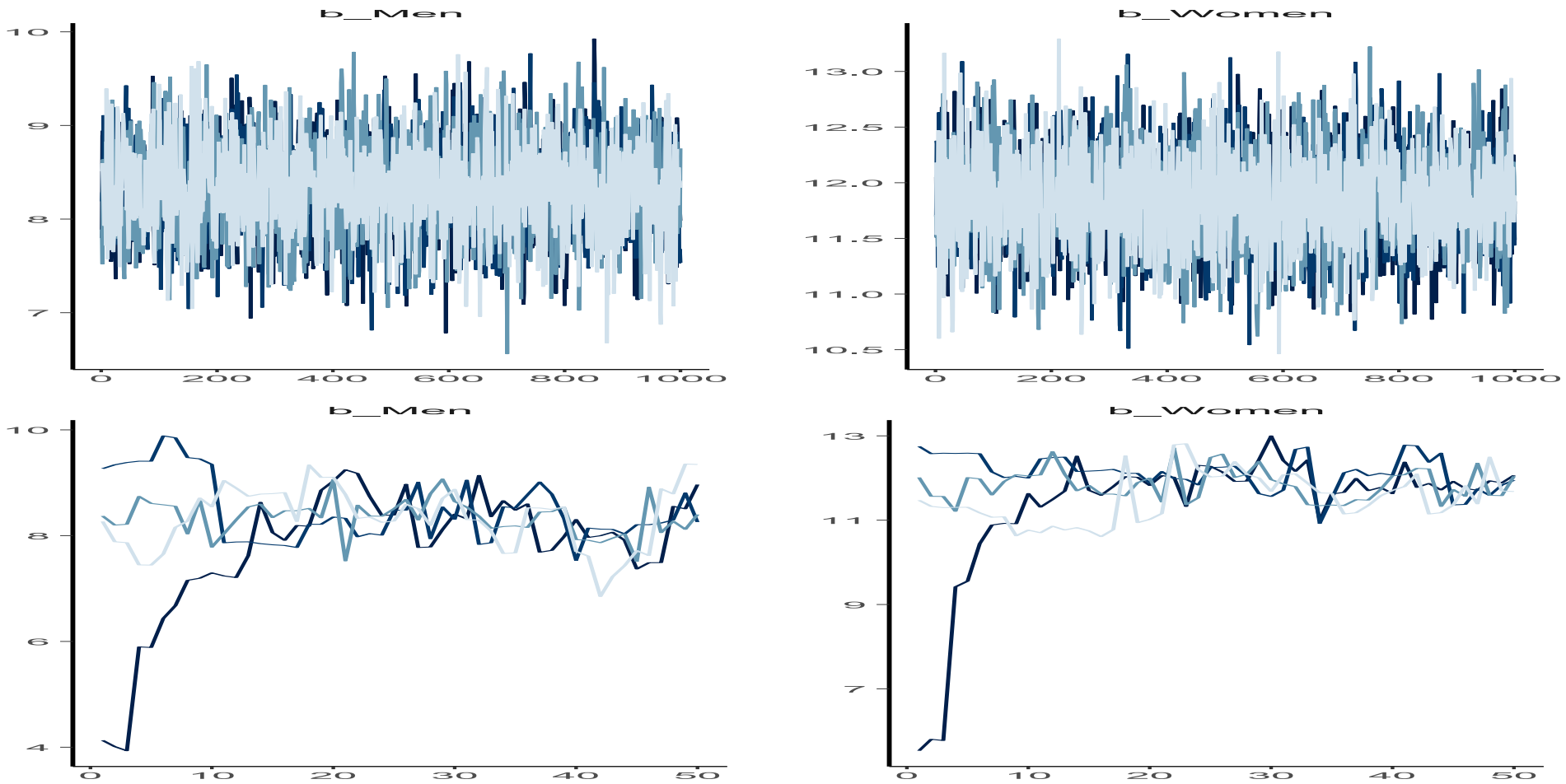
- The diagnostic index `Rhat`, also known as *Gelman-Rubin statistic*, summarizes consistency of convergence for the parallel chains, based on these two types of variability.
- *R-hat* $= 1.00$ indicates the MCMC process has likely converged adequately.
- When *R-hat* exceeds 1.00, this flags a situation where MCMC has failed to converge. You should then use a greater number of iterations.

Examining Trace Plots

A *trace plot* portrays chain values in sequential order, with a line connecting each pair of points in sequence, shown separately zig-zagging for each parallel Markov chain.

For any particular parameter, trace plot should have appearance of random variability, with good mixing and no particular trend or other behavior reflecting non-stationarity of a Markov chain.

Trace Plots for Household Work Example



Trace plots for mean household times for men and women using four Markov chains of post warm-up length 1000 each above, 50 each below.

Autocorrelation

For a sequence $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)})$ generated by an MCMC algorithm, the *autocorrelation* ρ_1 of lag 1 is the ordinary correlation computed for the pairs $(\theta^{(1)}, \theta^{(2)}), (\theta^{(2)}, \theta^{(3)}), \ldots, (\theta^{(T-1)}, \theta^{(T)})$.

The autocorrelation ρ_k of lag k for $k = 1, 2, \ldots$ is the correlation computed between pairs of observations $(\theta^{(1)}, \theta^{(1+k)}), (\theta^{(2)}, \theta^{(2+k)}), \ldots, (\theta^{(T-k)}, \theta^{(T)})$ that are k units apart.

Once the Markov chain is close to its stationary distribution, the effective sample size obtained with T iterations of the chain is approximately

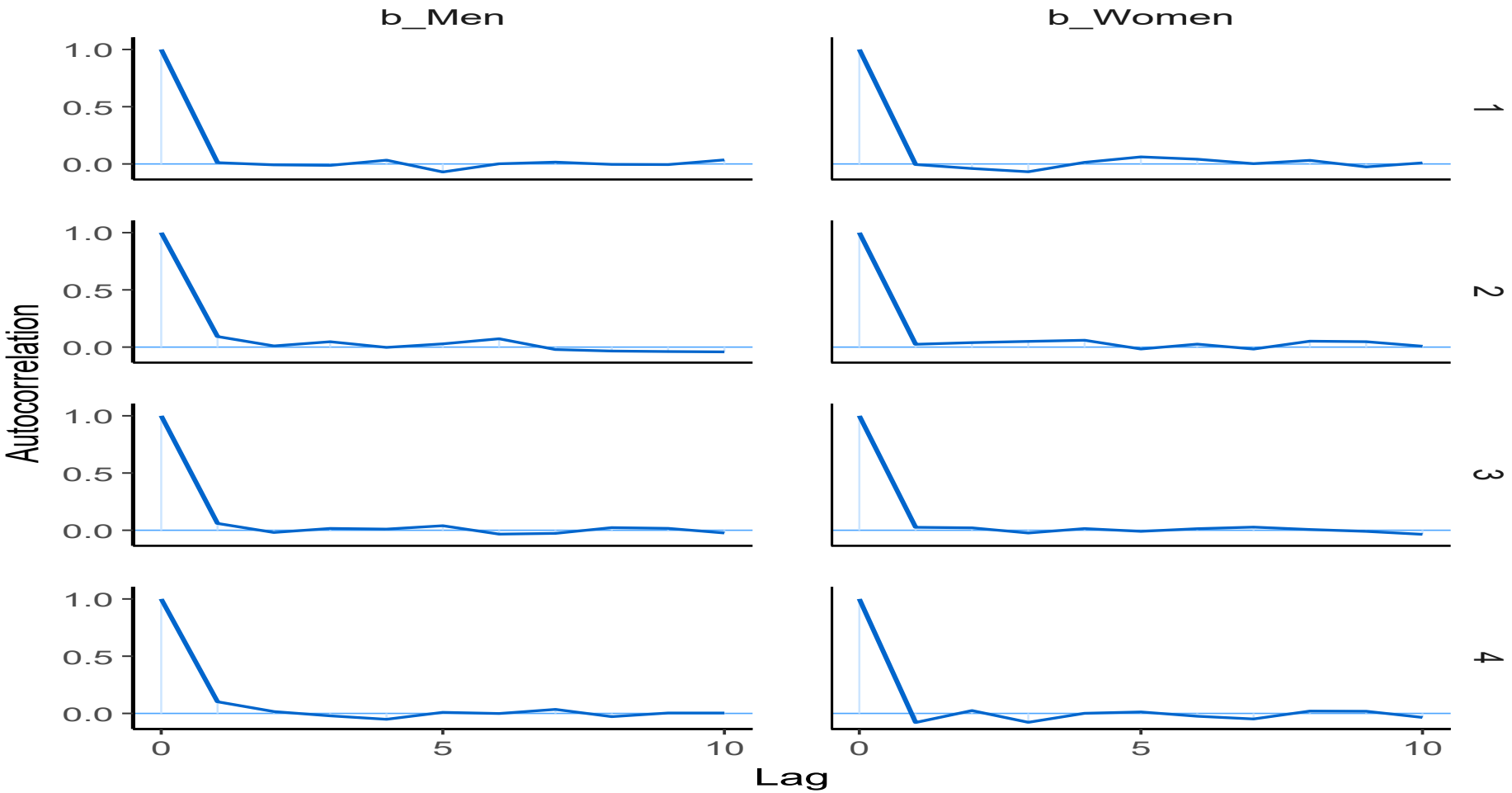
$$\text{ESS} = \frac{T}{1 + 2\sum_{k=1}^{\infty} \rho_k}.$$

As the autocorrelations in the chain increase further above zero, the denominator grows and the ESS becomes smaller.

Conversely, as the autocorrelations get closer to 0, the denominator approaches 1 and the ESS approaches T, and the estimators obtained from the MCMC process are more accurate and reliable.

Autocorrelations for Household Work

```
-------------------------------------------------------------------------------
> library(bayesplot)
> mcmc_acf(fit.bayes, pars=c("b_Men","b_Women"), lags=10) # autocorr's of lags 1,2,...,10
-------------------------------------------------------------------------------
```



BAYESIAN INFERENCE FOR LINEAR MODELS

- We summarize the normal linear model and show a Bayesian approach for fitting it
- Normal linear models can include categorical as well as quantitative explanatory variables
- Model building: How to determine which of p explanatory variables to include in model?
  - Bias/variance tradeoff: Allow parameter estimators to have some bias, to get benefit of smaller variance
  - BIC summary measure for model comparison has goal of keeping model from having more parameters than truly needed
  - Regularization (shrinking estimates toward 0) useful when p large and true effects expected to be small or non-existent

Bayesian Approach to Normal Linear Model

Model assumes independent observations such that for $i = 1, \ldots, n$,

$$Y_i \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mu_i, \sigma^2) \text{ with } \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- For prior, we can take $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$ where $\{\beta_j\}$ are independent, each with $\mathcal{N}(\lambda_0, \sigma_0^2)$ distribution, often with $\lambda_0 = 0$.
- With a common variance σ_0^2 for $p(\beta_j)$, prior effects are comparable in magnitude when we use *standardized* versions of explanatory variables.
- With σ^2 known, normal distribution is conjugate prior distribution with normal likelihood function. Posterior distribution of $\boldsymbol{\beta}$ is then multivariate normal distribution, and posterior mean estimate $\tilde{\beta}_j$ of β_j is weighted average of least squares estimate $\hat{\beta}_j$ and mean λ_0 of prior distribution for β_j.

Bayesian Normal Linear Model, Variance Unknown

- With unknown σ^2 and inverse gamma prior distribution for σ^2, $p(\sigma^2 \mid \boldsymbol{y})$ is also inverse gamma, and marginal posterior distribution of each $p(\beta_j \mid \boldsymbol{y})$ is a t distribution.
- With improper prior distributions $p(\boldsymbol{\beta} \mid \sigma^2) \propto 1$ and $p(\sigma^2) \propto 1/\sigma^2$, posterior distribution of

$$T = \frac{\beta_j - \hat{\beta}_j}{se_j},$$

is t distribution with $df = n - (p+1)$, where se_j is standard error of $\hat{\beta}_j$ from frequentist analysis.
- With highly disperse prior distributions for $\{\beta_j\}$ and for σ or σ^2, posterior means $\{\tilde{\beta}_j\}$ closely resemble least squares estimates, marginal posterior distribution of β_j around $\tilde{\beta}_j$ has same shape as frequentist sampling distribution of $\hat{\beta}_j$ around β_j, posterior interval for β_j similar to frequentist t confidence interval, posterior probabilities such as $P(\beta_j > 0 \mid \boldsymbol{y})$ are similar to frequentist P-values.

Example: Bayesian Linear Model for Mental Impairment

Study with $n = 40$ adults from a county in Florida investigated relationship between mental health and several explanatory variables.

Y = index of mental impairment; $\bar{y} = 27.30$ and $s_y = 5.46$.

Two useful explanatory variables:
x_1 = life events index; $\bar{x}_1 = 44.42$ and $s_{x_1} = 22.62$
x_2 = socioeconomic status (SES); $\bar{x}_2 = 56.60$ and $s_{x_2} = 25.28$.

```
-------------------------------------------------------------------------------
> Mental <- read.table("http://bayes4ds.rwth-aachen.de/data/Mental.dat", header=TRUE)
> pairs(~impair + life + ses, data=Mental) # scatterplot matrix for pairs (not shown)
> cor(Mental) # correlation matrix
           impair      life        ses
impair  1.0000000 0.3722206 -0.3985676
life    0.3722206 1.0000000  0.1233370
ses    -0.3985676 0.1233370  1.0000000
> summary(lm(impair ~ life + ses, data=Mental)) # least squares fit of linear model (lm)
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  28.22981     2.17422   12.984  2.38e-15
life          0.10326     0.03250    3.177   0.00300
ses          -0.09748     0.02908   -3.351   0.00186
---
Residual standard error: 4.556 on 37 degrees of freedom # estimates sigma = error std. dev.
Multiple R-squared:  0.3392,    Adjusted R-squared:  0.3034
-------------------------------------------------------------------------------
```

Bayesian Modeling of Mental Impairment Data

Based on previous studies, principal investigator felt 95% sure that $0 < \beta_1 < 0.4$ and 95% sure that $-0.4 < \beta_2 < 0$.

Suggests $\mathcal{N}(0.2, (0.1)^2)$ prior distribution for β_1, $\mathcal{N}(-0.2, (0.1)^2)$ prior distribution for β_2.

We use uninformative $\mathcal{N}(0, 100^2)$ prior distribution for intercept β_0 and exponential prior distribution for σ with rate $= 0.01$ (which has mean $=$ standard deviation $= 100$).

```
-------------------------------------------------------------------------------
> library(brms)
> fit.bayes <- brm(impair ~ life + ses, data=Mental, family=gaussian,
+   prior = prior(normal(0,100),class=Intercept) + prior(normal(0.2,0.1),class=b,coef=life)
+      + prior(normal(-0.2,0.1),class=b,coef=ses) + prior(exponential(0.01),class=sigma))
> print(fit.bayes, digits=3)
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI
Intercept   28.314     2.171   23.973   32.445
life         0.114     0.032    0.053    0.180 #  0.103 least squares
ses         -0.107     0.029   -0.166   -0.051 # -0.097 least squares
Further Distributional Parameters:
        Estimate Est.Error l-95% CI u-95% CI
> sigma    4.712     0.569    3.743    5.957   # 4.56 for least squares
-------------------------------------------------------------------------------
```

Bayesian Inferences about Effects

Posterior interval of (0.05, 0.18) for β_1 shifts the confidence interval (0.04, 0.17) slightly to right, because the prior mean for β_1 was 0.2.

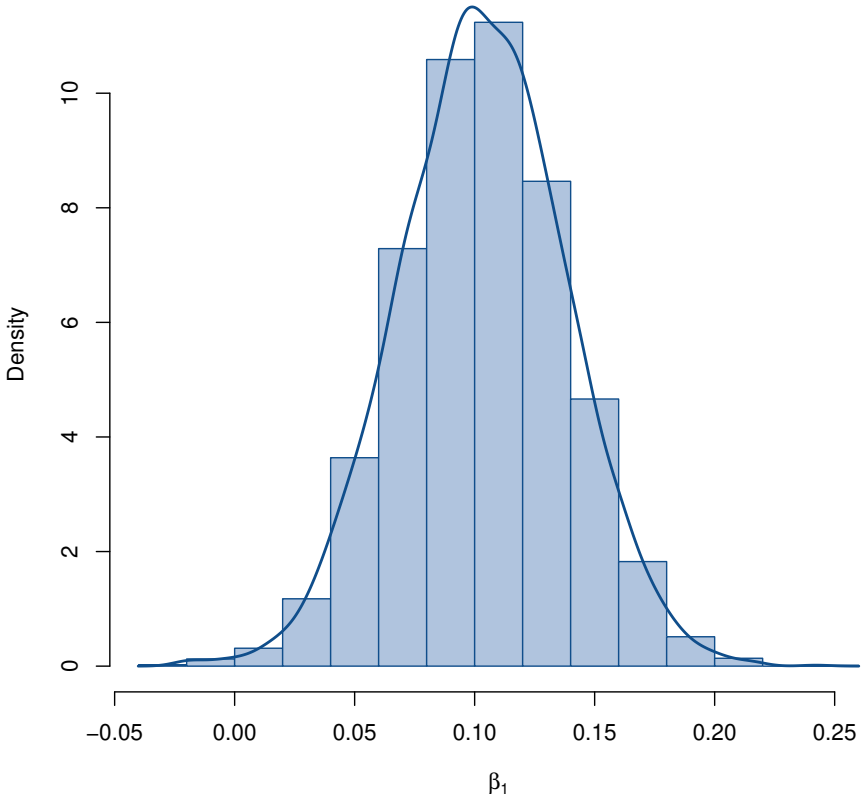
```
-------------------------------------------------------------------------------
> beta1 <- as_draws_df(fit.bayes)$b_life     # values from posterior dist. of beta1
> beta2 <- as_draws_df(fit.bayes)$b_ses      # values from posterior dist. of beta2
> mean(beta1 <= 0); mean(beta2 >= 0)
[1] 0.0005                # estimated posterior probability that beta1 <= 0
[1] 0.0000                # estimated posterior probability that beta2 >= 0
> hist(beta1, freq=FALSE, main = "", xlab = expression(beta[1]), ylab = "Density"))
                          # histogram of posterior distribution of life events effect
> dens1 <- density(beta1) # smoothed posterior distribution added on the histogram plot
> lines(dens1, lwd=2, lty=1)
-------------------------------------------------------------------------------
```

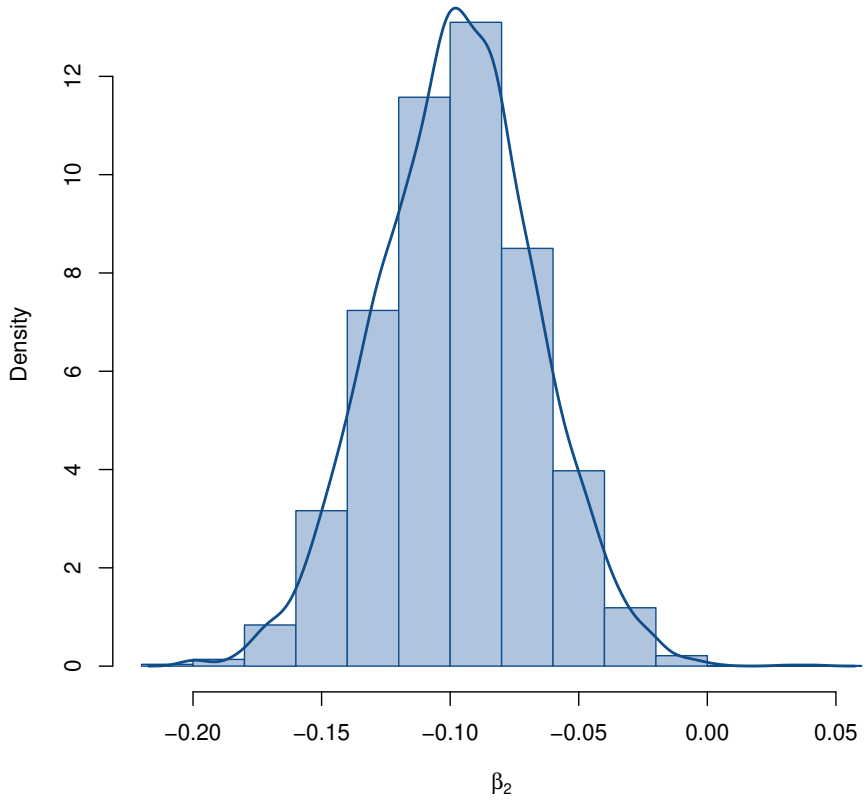
$P(\beta_1 \leq 0 \mid \boldsymbol{y}) = 0.0005$; that is, 0 is the 0.0005 quantile of the marginal posterior distribution of β_1, similar to one-sided P-value of 0.0015 for testing H_0: $\beta_1 = 0$ (and implicitly H_0: $\beta_1 \leq 0$) against H_a: $\beta_1 > 0$

$P(\beta_2 \geq 0) = 0.0000$ is similar to P-value of 0.0009 for testing H_0: $\beta_2 = 0$ (and implicitly H_0: $\beta_2 \geq 0$) against H_1: $\beta_2 < 0$.

Bayesian interpretation is simpler than frequentist P-value.

Histograms of Posterior Distributions, with Smoothings





Categorical Explanatory Variables: Indicators for Categories

Software sets up indicator variable but all except a reference category of categorical explanatory variable; e.g., with three categories, model

$$E(Y_i \mid \boldsymbol{\beta}) = \beta_0 + \beta_2 z_{i2} + \beta_3 z_{i3} \quad \text{has}$$

$z_{i2} = 1$ for observations in category 2, 0 otherwise
$z_{i3} = 1$ for observations in category 3, 0 otherwise,

where $\beta_0 = \mu_1$, $\beta_2 = \mu_2 - \mu_1$ and $\beta_3 = \mu_3 - \mu_1$.

An awkward aspect of no indicator for reference category is that mean for that category has different prior variance than the others; e.g., with $\mathcal{N}(0, \sigma_0^2)$ prior distribution for each β_j, variance of prior distribution for μ_1 is σ_0^2, variance of prior distribution for each other mean is $2\sigma_0^2$. When prior distributions are highly disperse, little effect.

A linear model can contain both categorical and quantitative predictors.

Example: Comparing Mean Fertility for Natives and Migrants

For sample of 52 married women above age of 45 in a Latin American city, comparison of fertility (number of children) for (RM = rural migrants, UM = urban migrants, UN = urban natives).
Quantitative explanatory variable: x = number of years of education

```
> Fertility <- read.table("http://bayes4ds.rwth-aachen.de/data/Fertility.dat", header=TRUE)
> tail(Fertility, 2)
   group education fertility
51    RM         9         7
52    RM        10         4 # 52 observations in Fertility data file
> library(brms)
> fit.bayes <- brm(fertility ~ education + group, data=Fertility, family=gaussian,
+              prior = prior(normal(0,100), class=Intercept) + prior(normal(0,100), class=b)
+              + prior(exponential(0.01), class=sigma))
> # if groups were instead (1, 2, 3), replace 'group' by 'factor(group)'
> summary(fit.bayes)
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept     7.16      0.37     6.43     7.88
education    -0.24      0.05    -0.33    -0.15 # effect of education on mean fertility
groupUM      -1.41      0.45    -2.29    -0.55 # compares UM with RM
groupUN      -1.64      0.47    -2.55    -0.72 # compares UN with RM
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI
sigma     1.35      0.14     1.10     1.66 # posterior interval for error standard dev.
```

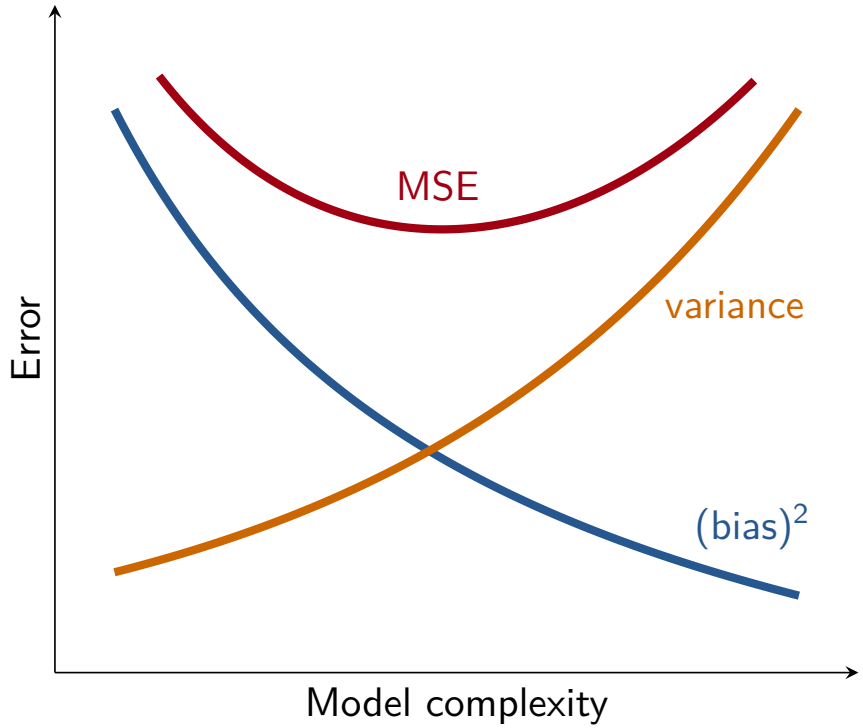
Model Building

- When the effect of x_1 on y changes as the value of x_2 changes, usual way of dealing with *interaction* between x_1 and x_2 in their effects adds cross-product term.
- Collinearity (strong correlations among explanatory variables) just as relevant with Bayesian as frequentist modeling
- Bias/variance tradeoff: As model complexity increases, estimator of $E(Y)$ at particular values of the explanatory variables has less bias but greater variance. Because the *mean squared error*

$$\text{MSE} = \text{variance} + (\text{bias})^2,$$

a Bayesian (or ML) estimator of true $E(Y)$ for a simpler model may tend to be *closer* than estimator for more complex model, if its variance is much smaller.

The bias/variance tradeoff in using a model to estimate a characteristic such as $E(Y)$:



Compared with an overly complex model, a simpler model having greater bias may have reduced variance and achieve MSE near the minimum.

BIC for Model Comparison

MSE takes different values at different settings of parameter values and of explanatory variables. So, for comparing two models according to MSE, one model is typically not always better than the other.

A summary measure for comparing models is the *Bayesian information criterion*, an analog of frequentist AIC. For a model M,

$$\text{BIC} = -2(\text{maximized log-likelihood}) + \log(n)(\text{number of parameters in } M).$$

BIC penalizes a model for having lots of parameters. Smaller BIC is better.

An argument motivating BIC is based on an approximation for the posterior probability $P(M \mid \boldsymbol{y})$ of model M in terms of prior probability $P(M)$. For large n,

$$P(M \mid \boldsymbol{y}) \approx P(M) \exp(-\text{BIC}/2).$$

Model Selection Strategies (e.g. using BIC)

- *Backward elimination* begins with a complex model and sequentially removes terms.
- *Forward selection* starts with the null model and adds terms sequentially until further additions do not improve the fit.

Such variable selection methods have no theoretical basis and need not yield a meaningful model. One possible strategy:

1. Construct initial main-effects model that includes key explanatory variables of interest as well as others known to be relevant or that show evidence of being so when used as sole predictors.
2. Conduct backward elimination using BIC as the criterion.
3. Check for plausible interactions among variables in model after (2).
4. For provisional model, use follow-up diagnostic investigations to check its adequacy (e.g., residuals, investigate influential observ's).

With large n, this process may yield an overly complex model, such as with a complex interaction term that improves the fit but not substantively. You might then exclude it to obtain a simpler model to interpret.

Example: Modeling Selling Price of a Home

Response = sale price (1000's of dollars) of 100 homes in Gainesville, FL

Explanatory variables:

size = size of house (in square feet)

taxes = annual tax bill (in dollars)

bedrooms = number of bedrooms

baths = number of bathrooms

new = indicator variable for whether home is new (1 = yes, 0 = no)

```
> Houses <- read.table("http://bayes4ds.rwth-aachen.de/data/Houses.dat", header=TRUE)
> library(brms)
> fit.bayes <- brm(price ~ size + taxes + new + bedrooms + baths, family=gaussian,
+      data=Houses, prior = prior(normal(0,1000), class=b)
+      + prior(normal(0,1000), class=Intercept) + prior(exponential(0.001), class=sigma))
> summary(fit.bayes)
          Estimate Est.Error l-95% CI u-95% CI
Intercept     6.33     37.49   -65.62    79.89
size          0.10      0.02     0.06     0.15
taxes         0.06      0.01     0.04     0.08
new          62.14     25.88    11.39   112.75
bedrooms    -16.90     13.78   -44.18     9.18
baths        -2.86     17.68   -38.30    31.71

> BIC(lm(price~size + taxes + new + bedrooms + baths, data=Houses))
[1] 1161.969
```

Results of Model Selection Processes

Backward elimination			Forward selection		
Step	Explanatory Variables	BIC	Step	Explanatory Variables	BIC
1	S, T, N, Be, Ba	1162.0	1	T	1177.8
2	S, T, N, Be	1157.4	2	S, T	1157.9
3*	S, T, N	1154.6	3*	S, T, N	1154.6
4	S, T	1157.9	4	S, T, N, Be	1157.4

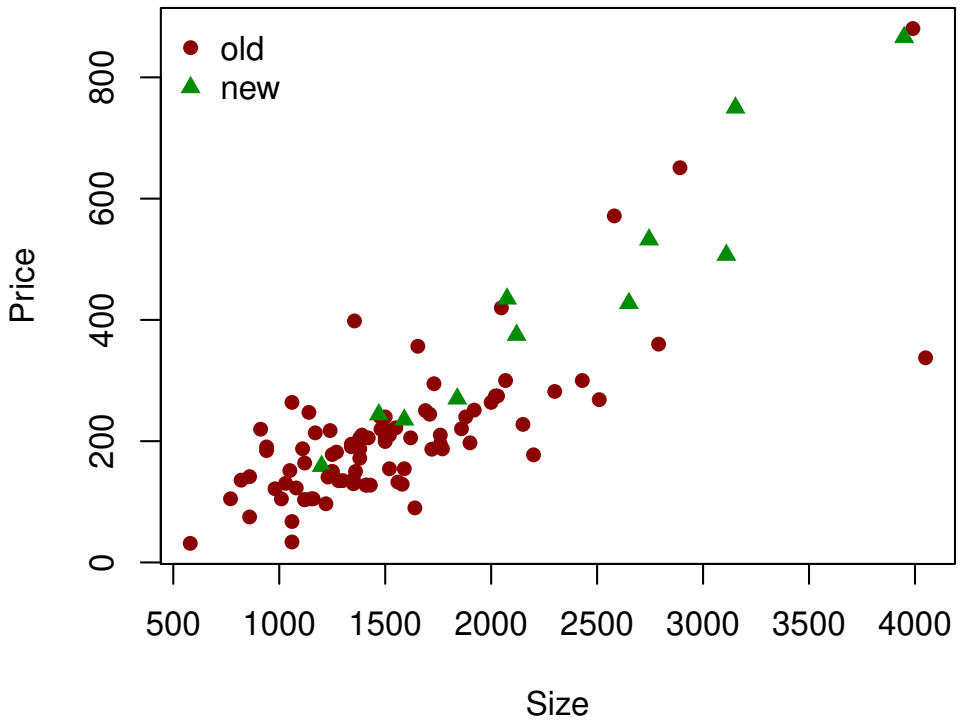
Fit of simpler model:

```
------------------------------------------------------------------------------------------
> fit.bayes2 <- brm(price ~ size + taxes + new, family=gaussian, data=Houses,
+       prior = prior(normal(0,1000), class=b) + prior(normal(0,1000), class=Intercept)
+       + prior(exponential(0.001), class=sigma))
> summary(fit.bayes2)
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept   -31.61     20.34   -70.83     9.41
size          0.09      0.02     0.05     0.13
taxes         0.06      0.01     0.04     0.08
new          69.82     25.15    20.52   118.78
------------------------------------------------------------------------------------------
```

The posterior means for effects are similar to more complex model. For the frequentist fits, adjusted $R^2 = 0.782$ for model with all explanatory variables and 0.783 for simpler model.

Further Modeling

At next stage, we could consider potential interaction terms. However, basic diagnostic analyses make additional model-fitting questionable, because of possible violations of linear modeling assumptions.



Scatterplot of house selling price (in thousands of dollars) by size (in square feet) and whether the house is new

Regularization Methods of Model-Fitting

When number of explanatory variables p is very large, often only a few $\{\beta_j\}$ are practically different from 0.

Unless n is very large, because of ordinary sampling variability $\tilde{\beta}_j$ tend to be much larger in absolute value than true values.

For frequentist methods, *regularization methods* use a *penalized likelihood* approach. Values that maximize the penalized-likelihood function shrink the ordinary ML estimates toward 0. For example, the *lasso* method finds the values of $\{\hat{\beta}_j\}$ that minimize

$$\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta} x_{ip})]^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| = SSE + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|,$$

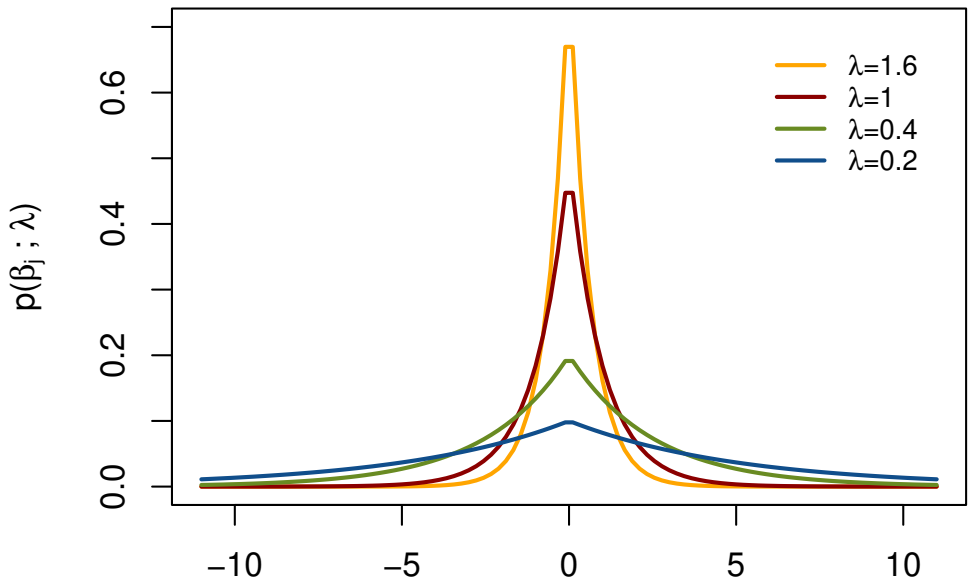
for some smoothing parameter $\lambda \geq 0$.
A lasso estimate corresponds to *mode* of Bayesian posterior distribution for normal linear model when independent prior distribution for each β_j is a *double-exponential (Laplace) distribution*.

Double-exponential prior distribution for β_j has *pdf*

$$p(\beta_j; \tau) = \frac{\tau}{2} \exp(-\tau|\beta_j|)$$

that is symmetric with mean = median = mode = 0.

τ is a a hyperparameter for which $1/\tau$ is a scale parameter and standard deviation $= \sqrt{2}/\tau$.



The double-exponential distribution is more highly-peaked at 0 and is more likely to yield a posterior mode of 0 for β_j as τ increases.

Example: Predicting GPA with Student Survey Data

Predict *cogpa* = college GPA using *hsgpa* = high school GPA (on a four-point scale), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *gender* (1 = female, 0 = male), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *age*, *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no).

When we use all as main effects in linear model and obtain Bayesian fit using highly disperse normal prior distributions, no explanatory variable has posterior mean more than 2 standard deviations from 0.

```
-------------------------------------------------------------------------------
> Students <- read.table("http://bayes4ds.rwth-aachen.de/data/Students.dat", header=TRUE)
> library(brms)
> fit.bayes <- brm(cogpa ~ hsgpa + relig + gender + ideol + age + abor + dhome + dres +
+                  tv + sport + news + aids + veg + affirm, family=gaussian, data=Students,
+                  prior = prior(normal(0,1000), class=Intercept) +
+                  prior(normal(0,1000), class=b) + prior(exponential(0.001), class=sigma))
> print(fit.bayes, digits=3)
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept    3.237     0.532    2.165    4.284
hsgpa        0.146     0.118   -0.088    0.377
relig       -0.120     0.066   -0.251    0.012
gender       0.180     0.119   -0.048    0.413
ideol        0.008     0.045   -0.080    0.096
age         -0.003     0.007   -0.016    0.011
abor        -0.043     0.157   -0.347    0.258
dhome       -0.000     0.000   -0.000    0.000
dres         0.003     0.014   -0.025    0.030
tv          -0.001     0.008   -0.017    0.015
sport       -0.011     0.014   -0.039    0.016
news         0.007     0.019   -0.029    0.045
aids         0.014     0.023   -0.030    0.059
veg         -0.031     0.153   -0.322    0.276
affirm      -0.164     0.153   -0.463    0.146
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI
sigma    0.361     0.040    0.293    0.449
-------------------------------------------------------------------------------
```

Since many true effects are likely essentially 0, it is sensible to instead use a double-exponential prior distribution highly peaked at 0.

Consider highly-disperse normal prior distribution for intercept but take standard deviation of prior distribution for each effect $= 0.1$.

Then, since the standard deviation $= \sqrt{2}/\tau = 0.1$,

scale parameter $= 1/\tau = 0.1/\sqrt{2} = 0.0707$.

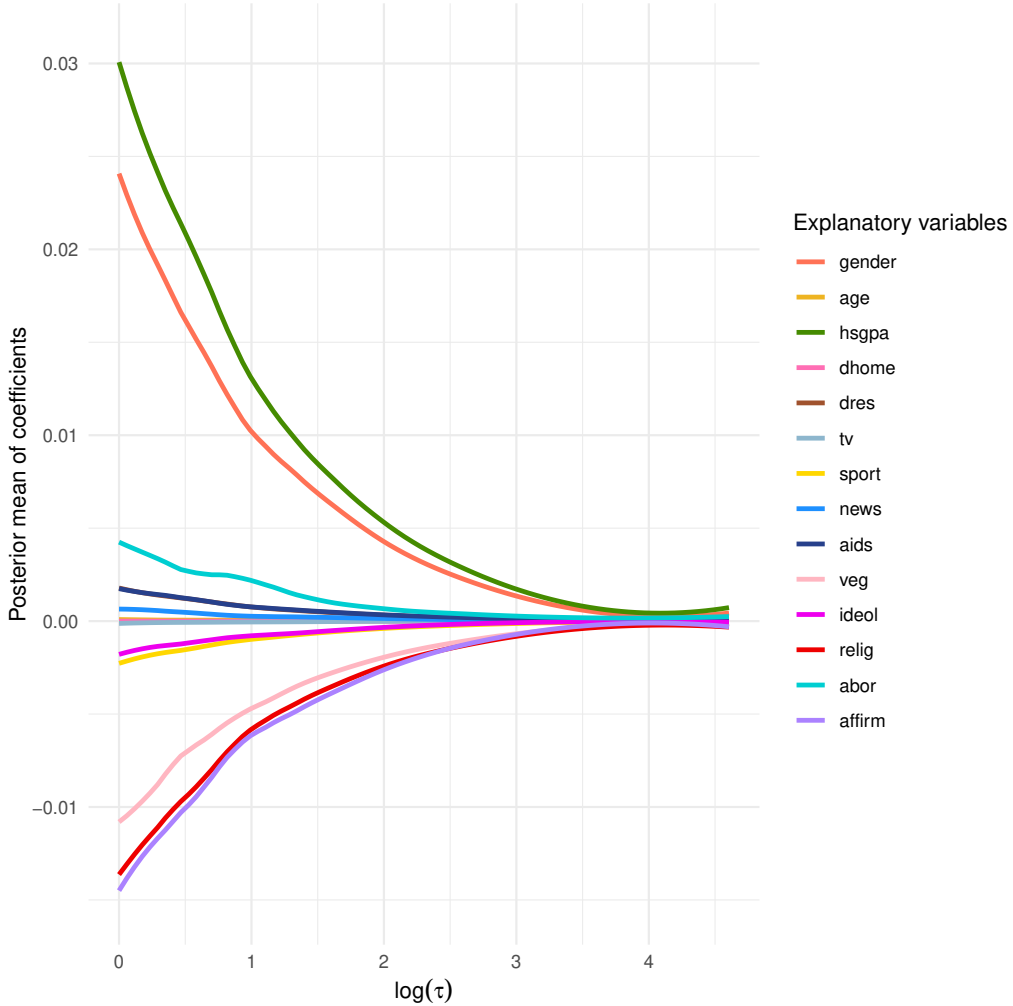
With such a highly-concentrated prior distribution, sensible to standardize the explanatory variables, but here we use the original ones to compare effect estimates to ones obtained with highly-disperse prior distributions:

```
-------------------------------------------------------------------------------------
fit.bayes2 <- brm(cogpa ~ hsgpa + relig + gender + ideol + age + abor + dhome + dres +
                          + tv + sport + news + aids + veg + affirm, family=gaussian,
                  data=Students, prior = prior(normal(0,1000), class=Intercept)
                  + prior(double_exponential(0, 0.0707), class=b) # double_exp. uses scale
                  + prior(exponential(0.001), class=sigma) )      #   parameter: 1/tau
print(fit.bayes2, digits=3)
          Estimate Est.Error l-95% CI u-95% CI
Intercept    3.332     0.345    2.628    3.975
hsgpa        0.069     0.079   -0.061    0.248 #  0.146 with disperse normal prior
relig       -0.066     0.051   -0.173    0.024 # -0.120 with disperse normal prior
gender       0.059     0.071   -0.057    0.221
ideol        0.000     0.030   -0.061    0.063
age         -0.000     0.006   -0.012    0.012
abor         0.001     0.075   -0.146    0.160
dhome       -0.000     0.000   -0.000    0.000
dres         0.010     0.012   -0.013    0.034
tv          -0.001     0.007   -0.016    0.013
sport       -0.014     0.012   -0.039    0.010
news         0.004     0.015   -0.027    0.035
aids         0.012     0.020   -0.025    0.053
veg         -0.027     0.074   -0.195    0.108
affirm      -0.040     0.075   -0.209    0.093
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
sigma    0.353     0.037    0.289    0.432 1.001     3498     2784
-------------------------------------------------------------------------------------
```

This choice of prior distributions more strongly shrinks the coefficients of all variables toward zero. A *Bayesian lasso* analog of traditional lasso, described in R Appendix of textbook, imposes double-exponential prior distribution and can shrink some posterior mean estimates exactly to zero.

Estimation Results

We plot the means of the posterior distributions of the coefficients against the logarithm of a smoothing parameter, τ, from the prior distribution, with greater values of $\log(\tau)$ showing greater regularization.



BAYESIAN GENERALIZED LINEAR MODELING

- Introduction to generalized linear models for possibly non-normal responses
- Bayesian approach for generalized linear models
- Generalized linear models assuming gamma distribution useful for positive-valued responses for which variance grows with mean
- Logistic regression for binary response variables enable modeling probability in a particular category
- Complete separation with binary responses results in infinite ML estimates in logistic regression modeling, whereas Bayesian estimates are finite
- Hierarchical modeling, such as multilevel models, requires *random effects* as well as *fixed effects* in a *generalized linear mixed model*

Introduction to Generalized Linear Models

Generalized linear models extend normal linear models to also encompass non-normal response distributions and to equate the linear predictor to nonlinear functions of mean.

Components of a GLM:

- *Response variable*: This component specifies Y and its probability distribution.
- *Explanatory variables*: This component specifies p explanatory variables for a *linear predictor* $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, where x_{ij} is the value of explanatory variable j for observation i.
- *Link function*: This component is a function g applied to $\mu_i = E(Y_i \mid x_{i1}, \ldots, x_{ip})$. The GLM relates $g(\mu_i)$ to the linear predictor,

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

GLMs for Normal, Binomial, and Poisson Responses

- For *continuous* responses, the most important GLM is the *normal linear model*, which assumes $Y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, ..., n$, and uses *identity link function*, for which $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.
- For *categorical* response Y, GLMs focus on category probabilities. For binary Y, with $1 =$ *success* and $0 =$ *failure*, $\mu_i = P(Y_i = 1)$, and *logistic regression model* with *logit link function* is

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \ldots, n,$$

- When Y has *counts* for outcomes, $E(Y)$ must be nonnegative, and its log can be any real number, like a linear predictor. A GLM using the *log link function*,

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \ldots, n,$$

is called a *loglinear model*. With assumption that Y has a *Poisson* distribution, it is called a *Poisson loglinear model*.

Bayesian Fitting of Generalized Linear Models

- A common structure for the prior distributions takes $\{\beta_j\}$ to be independent $\mathcal{N}(\lambda_0, \sigma_0^2)$ random variables, usually with $\lambda_0 = 0$.
- To enable a particular prior range of values for an effect but allow that the actual effect could be more extreme, we could instead use a *Cauchy* prior distribution, which has thicker tails than the normal distribution.
- As in linear modeling, standardizing explanatory variables makes the magnitudes of their effects comparable.
- As n increases, posterior joint distribution of $\{\beta_j\}$ more closely resembles the likelihood function, which is approximately a normal distribution around the ML estimate with covariance matrix the inverse of the information matrix.

GLM Assuming a Gamma Distribution for Y

When Y must be ≥ 0, var(Y) often grows as $E(Y)$ grows.
Example: We would expect $Y =$ annual income to be more variable when $E(Y) = 100,000$ euros than when $E(Y) = 10,000$.

The *gamma distribution*

$$f(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \; y \geq 0,$$

for shape parameter α and rate parameter β (scale parameter $1/\beta$) has mean $\mu = \alpha/\beta$, standard deviation $\sigma = \mu/\sqrt{\alpha}$ proportional to mean.

- The *pdf* becomes more bell-shaped as α increases, for fixed μ.
- Gamma GLMs usually assume that the shape parameter α is constant but unknown, like σ^2 in normal GLMs.
- Using the log link function ensures that the fitted values are all positive. (Fit similar to assuming *log-normal distribution* for Y)

Example: GLMs for House Selling Prices

We revisit data with Y = selling price, x_1 = size of house, and x_2 = whether the house is new (1 = yes, 0 = no). Scatterplot (page 97) showed variability in selling prices may be greater at larger house sizes. For Bayesian fitting of the *normal GLM*, we first include an interaction term to allow effect of size to differ for new homes and older homes.

```
-------------------------------------------------------------------------------------------
> Houses$size2 <- Houses$size - mean(Houses$size) # center size around its mean
> library(brms)
> fit.norm <- brm(price ~ size2 + new + size2:new, family=gaussian, data=Houses,
+    prior=prior(normal(0,1000), class=Intercept) + prior(normal(0,1000), class=b)
+    + prior(exponential(0.001), class=sigma))
> summary(fit.norm)
          Estimate Est.Error l-95% CI u-95% CI
Intercept   221.85      8.62   205.14   238.98
size2         0.16      0.01     0.13     0.18
new          34.16     33.32   -31.23    98.64 # estimated new effect = 34.16 at mean size
size2:new     0.09      0.03     0.03     0.16
      Estimate Est.Error l-95% CI u-95% CI
sigma    78.98      5.82    68.57    91.17
-------------------------------------------------------------------------------------------
```

The effect on $E(Y)$ of a square-foot increase in size is 0.16 (i.e., \$160) for older homes and $0.16 + 0.09 = 0.25$ (i.e., \$250) for newer homes.

Gamma GLM for Home Selling Price

For the Bayesian fit of GLM that instead assumes a gamma distribution for Y, we apply the log link function.

```
-------------------------------------------------------------------------------
> fit.gamma <- brm(price ~ size2 + new + size2:new, family=Gamma(link="log"), data=Houses,
+   prior = prior(normal(0,100), class=Intercept) + prior(normal(0,100), class=b)
+   + prior(exponential(0.01), class=shape), init=0) # needed initial value = 0 to run ok
> print(fit.gamma, digits = 5)
           Estimate Est.Error l-95% CI u-95% CI
Intercept  5.32313   0.03667   5.25267  5.39600
size2      0.00060   0.00006   0.00048  0.00073
new        0.23286   0.14017  -0.02876  0.52633
size2:new -0.00001   0.00015  -0.00029  0.00027
      Estimate Est.Error l-95% CI u-95% CI
shape  9.07915   1.25824  6.85544 11.74099
-------------------------------------------------------------------------------
```

Now, estimated size of the interaction term is trivial. The outlying observation is much less influential for the gamma GLM, because this model expects more variability in y when $E(Y)$ is larger.

The model without the interaction term but the same prior distributions gives essentially the same fit:

Gamma GLM without Interaction Term

```
-------------------------------------------------------------------------------------------
> fit.gamma2 <- brm(price ~ size2 + new, family=Gamma(link="log"), data=Houses,
+         prior = prior(normal(0,100), class=Intercept) + prior(normal(0,100), class=b)
+         + prior(exponential(0.01), class=shape), iter = 10000)
> # iter is number of iterations in MCMC fitting process (see Sec. 6.2.1; default = 2000)
> print(fit.gamma2, digits=5)
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept  5.32340   0.03518  5.25500  5.39361
size2      0.00059   0.00006  0.00048  0.00071
new        0.22253   0.11585 -0.00080  0.45016
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI
shape  9.20331   1.28656  6.84417 11.90852
> plot(fit.gamma2) # plots posterior distributions (not shown)
-------------------------------------------------------------------------------------------
```

$\tilde{\sigma} = \tilde{\mu}/\sqrt{\tilde{\alpha}} = \tilde{\mu}/\sqrt{9.2} = 0.33\tilde{\mu}.$

For example, when $\tilde{\mu} = \$100{,}000$, then $\tilde{\sigma} = \$33,000$, whereas when $\tilde{\mu} = \$500{,}000$, then $\tilde{\sigma} = 5(\$33,000) = \$165,000$.

By contrast, for normal GLM, $\tilde{\sigma}$ is constant at 78.98 (i.e., \$78,980).

For log link, estimated effects are *multiplicative*. For example, for a 1000 square foot increase in size, the estimated mean selling price multiplies by $\exp[1000(0.00059)] = 1.81$.

Model Comparison with BIC

```
-------------------------------------------------------------------------------
> BIC(lm(price ~ size2 + new + size2:new, data=Houses))
[1] 1174.066  # BIC for normal linear model

> BIC(glm(price ~ size2 + new + size2:new, family=Gamma(link=log), data=Houses))
[1] 1145.675  # BIC much better for gamma GLM than normal GLM

> BIC(glm(price ~ size2 + new, family=Gamma(link=log), data=Houses))
[1] 1141.075  # BIC slightly better for gamma GLM that excludes interaction term
-------------------------------------------------------------------------------
```

With interaction term, BIC = 1174.1 for the normal GLM and BIC = 1145.7 for the gamma GLM (much better).

Important note: In modeling, it is not sufficient to focus on how $E(Y)$ depends on the explanatory variables. The assumption about how var(Y) depends on $E(Y)$, through the choice of probability distribution for Y, can strongly affect standard errors and thus inferential conclusions.

Another Gamma Model

We do even a bit better by modeling mean (rather than its log) and removing intercept term, forcing estimated mean = 0 at size = 0 both for new and older homes:

```
-------------------------------------------------------------------------------------
> fit.gamma2 <- brm(price ~ -1 + size + size:new, family=Gamma(link=identity), data=Houses,
+ prior = prior(normal(0,100), class=b) + prior(exponential(0.01), class=shape))
> summary(fit.gamma2)
Regression Coefficients:
         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
size         0.13      0.00     0.13     0.14 1.00     2951     2720
size:new     0.05      0.02     0.01     0.09 1.00     2912     2446
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     9.42      1.33     7.04    12.16 1.00     2389     2424
> BIC(glm(price ~ -1 + size + size:new, family=Gamma(link=identity), data=Houses))
[1] 1134.883
-------------------------------------------------------------------------------------
```

Cauchy Distribution for Prior or Response Distribution

To enable an occasional rare event (black swan), we could instead use *Cauchy distribution*, which has thicker tails than normal distribution:

$$f(y \mid \beta, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{y-\beta}{\gamma}\right)^2\right]}, \quad -\infty < y < \infty,$$

bell-shaped around β = median = mode, with scale parameter $\gamma > 0$ that equals distance between the median and either quartile.

- Tails die out slowly; mean and variance do not exist.
- Standard version with $(\beta = 0, \gamma = 1)$ is t distribution with $df = 1$.
- To use the Cauchy prior distribution for μ, select median β as prior guess for μ and select γ to reflect $P(\beta - \gamma < \mu < \beta + \gamma) = 0.50$.

  To illustrate, suppose we believe $P(-2 < \mu < 2) \approx 0.50$. Then, $\mathcal{N}(0, 3^2)$ distribution and Cauchy distribution with $\beta = 0$ and $\gamma = 2$ satisfy this, but prior $P(|\mu| > 20)$ is 2.6×10^{-11} for normal prior and 0.063 for Cauchy prior.

Cauchy Model for a Continuous Response Variable

Example: modeling of house selling prices.

Cauchy not available as response distribution for linear model in `brms`, so we include code to incorporate a Cauchy log-likelihood function.

```
-------------------------------------------------------------------------------
\begin{code}
> cauchy_family <- custom_family("cauchy",                    # name of the custom family
+     dpars = c("mu", "scale"), # parameters for location (mu=median) and scale
+     links = c("identity", "log"),    # identity link for median and log link for scale
+     # Log-likelihood for Cauchy distribution
+       log_lik = function(i, prep, ...) {
+        mu <- prep$mu[i]
+        scale <- prep$scale[i]
+        return(d_cauchy(prep$y[i], location = mu, scale = scale, log = TRUE)) }  )
> library(brms)
> fit <- brm(price ~ size2 + new + size2:new, family = cauchy_family, data = Houses,
+            prior = prior(normal(0,100), class=Intercept) + prior(normal(0,100), class=b)
+                  + prior(exponential(0.001), class=scale))
> summary(fit)
          Estimate Est.Error l-95% CI u-95% CI
Intercept   206.17      5.95   194.18   217.22
size2         0.12      0.01     0.09     0.14
new          47.86     18.37    10.76    82.74
size2:new     0.13      0.03     0.06     0.18
scale    33.56      4.40    25.52    42.91
-------------------------------------------------------------------------------
```

Example: Model with Cauchy Prior Distributions

The Cauchy response model, like the normal linear model, assumes constant variability. To permit nonconstant variability, we could instead assume a gamma response distribution with log link function, now using a highly-disperse Cauchy prior distribution for model effect parameters with median = 0 and scale parameter = 100:

```
-------------------------------------------------------------------------------------
> fit.gamma <- brm(price ~ size2 + new + size2:new, family=Gamma(link="log"), data=Houses,
+   prior = prior(student_t(1,0,100), class=Intercept) + prior(student_t(1,0,100), class=b)
+         + prior(exponential(0.01),class=shape),init=0,iter=10000) # Cauchy = t with df=1
> # iter =  number of iterations in MCMC fitting process; inadequate with default iter=2000
> print(fit.gamma, digits=5)
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept  5.32299   0.03525  5.25553  5.39305
size2      0.00060   0.00006  0.00047  0.00073
new        0.23374   0.14384 -0.04360  0.52669
size2:new -0.00001   0.00015 -0.00030  0.00028
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI
shape  9.11032   1.27535  6.76924 11.76740
-------------------------------------------------------------------------------------
```

Results similar to those shown before for the gamma model. Both models indicate that interaction term not needed under gamma assumption.

Logistic Regression Modeling for Binary Response

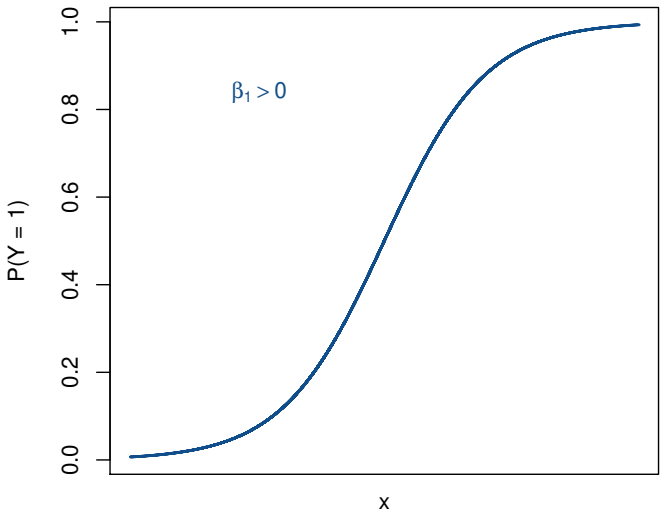
Logistic regression model

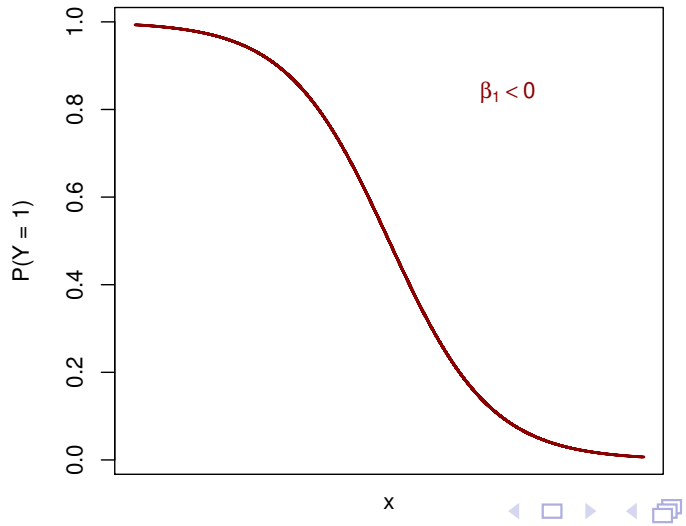
$$\text{logit}[P(Y_i = 1 \mid \boldsymbol{\beta})] = \log[\mu_i/(1-\mu_i)] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \ldots, n$$

corresponds to expression for $P(Y_i = 1)$,

$$P(Y_i = 1 \mid \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}.$$

For $p = 1$ explanatory variable, $P(Y = 1)$ follows a monotone S-shaped curve, for which β is a log odds ratio effect and e^{β} is an odds ratio.





Bayesian Inference for Logistic Regression

A $\mathcal{N}(0, \sigma_0^2)$ prior distribution for $\{\beta_j\}$ corresponds to *logit-normal* prior distribution for $P(Y_i = 1)$

- When $X \sim N(0, \tau^2)$, logit-normal *pdf* of $e^x/(1 + e^x)$ is symmetric about 0.5, unimodal when $\tau^2 \leq 2$, bimodal otherwise with modes closer to 0 and 1 as τ increases.
- Relatively non-informative (large τ) priors imply U-shaped priors on probability scale, with half probability close to 0 and half close to 1, like Beta(0.5, 0.5) prior, a special case of *Jeffreys prior*.
- In practical applications with logistic regression, using standardized explanatory variables, each $|\beta_j|$ rarely exceeds 10. With relatively large σ_0 in a $\mathcal{N}(0, \sigma_0^2)$ prior distribution for each β_j, such as $\sigma_0 = 10$, posterior distribution has same appearance as likelihood function.
- To permit unusually large effects, could instead use Cauchy prior distribution, such as with median $\beta = 0$ and scale parameter $\gamma = 2.5$, which has 84% of its density between -10 and 10 but has 0.01 and 0.99 quantiles at -80 and 80.

Example: Modeling of Endometrial Cancer Risk Factors

A study with 79 endometrial cancer patients analyzed how a histology grade that reflects the amount of the cancer tissue that is solid tumor growth (HG = 0, low; HG = 1, high) relates to three risk factors:

NV = neovasculation (1 = present in 13 cases, 0 = absent in 66 cases),

PI = pulsatility index; mean = 17.4, standard deviation = 10.0

EH = endometrium height; mean = 1.7, standard deviation = 0.7

PI and EH had very different units, so we use standardized versions (denoted PI2 and EH2) to enable comparing sizes of their effects.

We get a rather strange result for the NV effect estimate with the ML fit.

Results with Maximum Likelihood Fitting

```
------------------------------------------------------------------------------------
> Endo <- read.table("http://bayes4ds.rwth-aachen.de/data/Endometrial.dat", header=TRUE)
> Endo
   NV PI   EH HG   # HG is histology grade, the binary response variable
1   0 13 1.64  0
...
79  0 33 0.85  1
> Endo$PI2 <- scale(Endo$PI); Endo$EH2 <- scale(Endo$EH) # standardized variables
> fit <- glm(HG ~ NV + PI2 + EH2, family=binomial, data=Endo) # glm function gives ML fit
> summary(fit)                                                #   of GLMs; binomial family
            Estimate  Std. Error  z value  Pr(>|z|)          #   has logistic regression
(Intercept)  -1.2517      0.3688   -3.394  0.000688          #   as default
NV           18.1856   1715.7509    0.011  0.991543 # true ML estimate of NV effect is
PI2          -0.4217      0.4432   -0.952  0.341333 #   infinite: huge standard error
EH2          -1.9219      0.5599   -3.433  0.000597 #   for NV effect is a warning sign
------------------------------------------------------------------------------------
```

The estimated NV effect is unusually large yet has P-value $= 0.99$.

In fact, the true ML estimate of the NV effect is infinite.

Complete Separation and Quasi-Complete Separation

At least one infinite ML estimate occurs when the space of explanatory variable values exhibits *complete separation*.

For example, for $n = 6$, $y = 0$ at $x_1 = 1, 2, 3$ and $y = 1$ at $x_1 = 4, 5, 6$

Infinite ML estimates also occur under a weaker condition, *quasi-complete separation*, in which complete separation occurs except at a boundary point.

For example, if $y = 0$ at $x_1 = 1, 2, 3$ and $y = 1$ at $x_1 = 3, 4, 5$

```
-------------------------------------------------------------------------------------------
> xtabs(~NV + HG, data=Endo) # contingency table for NV and HG
    HG
NV    0   1
  0  49  17                  # quasi-complete separation:
  1   0  13                  #   when NV=1, no HG=0 cases occur

> library(detectseparation)
> detect_separation(x=cbind(Endo$NV, Endo$PI2, Endo$EH2), y=Endo$HG, family=binomial(link=logit))
Separation: TRUE  # detects complete or quasi-complete separation
Existence of maximum likelihood estimates
 X1  X2  X3
Inf   0   0       # 1st explanatory variable has infinite ML estimate
0: finite value, Inf: infinity, -Inf: -infinity
-------------------------------------------------------------------------------------------
```

Better Frequentist Inference and a Bayesian Approach

```
-------------------------------------------------------------------------------
> library(car)
> Anova(fit) # likelihood-ratio (LR) tests are valid even with infinite ML estimates
Response: HG
    LR Chisq  Df  Pr(>Chisq)
NV    9.3576   1    0.00222    # P-value = 0.002 gives strong evidence of NV effect,
PI2   0.9851   1    0.32093    #   unlike P-value of 0.99 in Wald test
EH2  19.7606   1    8.777e-06

> library(profileModel)        # use to get profile likelihood confidence intervals
> confintModel(fit, objective="ordinaryDeviance", method="zoom", endpoint.tolerance=1e-08)
               Lower    Upper  #   95% profile likelihood confidence intervals
NV           1.28411      Inf  #   NV effect on log odds has lower bound of 1.28
PI2         -1.37047  0.38176  #   Odds ratio at least exp(1.284) = 3.6
EH2         -3.16891 -0.95108  #   Wald CI not possible when ML estimate is infinite
-------------------------------------------------------------------------------
```

With complete or quasi-complete separation, Bayesian approach using proper prior distributions yields finite posterior mean estimates of all $\{\beta_j\}$.

With $\mathcal{N}(0, 10^2)$ priors, about 95% of prior distribution for odds ratio describing NV effect falls between $e^{-2(10)} = 2.1 \times 10^{-9}$ and $e^{2(10)} = 4.9 \times 10^8$, essentially entire real line.

Bayesian Fit of Logistic Regression for Endometrial Cancer

```
-------------------------------------------------------------------------------
> Endo$NV2 <- Endo$NV - 0.50 # so the prior variability is the same at each level of NV
> library(brms)
> fit.bayes <- brm(HG ~ NV2 + PI2 + EH2, family=bernoulli(link=logit), data=Endo,
+   prior=prior(normal(0,10), class=Intercept) + prior(normal(0, 10), class=b))
> summary(fit.bayes) # Bernoulli is binomial distribution for n=1 trial
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept     3.54      2.81    -0.30    10.29
NV2           9.77      5.60     2.20    23.21 # (1.28, Inf) for LR CI
PI2          -0.47      0.45    -1.41     0.37 # (-1.37, 0.38) for LR CI
EH2          -2.13      0.59    -3.38    -1.08 # (-3.17, -0.95) for LR CI
> plot(fit.bayes) # plots posterior distributions
-------------------------------------------------------------------------------
```

Estimated odds of higher-grade histology when neovasculation is present are $\exp(\tilde{\beta}_1) = \exp(9.77) \approx 17,500$ times estimated odds when neovasculation is absent, adjusting for PI and EH.

Posterior interval for β_1 of $(2.2, 23.2)$ indicates $\beta_1 > 0$ and that effect is strong, with odds ratio at least $e^{2.2} = 9.0$.

More about Bayes Fit for Endometrial Cancer Data

Because of the relatively flat log-likelihood function, posterior results for β_1 are highly dependent on value for σ_0 in $\mathcal{N}(0, \sigma_0^2)$ prior distribution for β_1.

Analysis	$\hat{\beta}_1$ (SD)	Interval[a] for β_1	$\hat{\beta}_2$ (SD)	$\hat{\beta}_3$ (SD)
ML	∞ (—)	(1.28, ∞)	−0.42 (0.44)	−1.92 (0.56)
Bayes, $\sigma_0 = 100$	80.7 (59.0)	(5.9, 222.3)	−0.49 (0.46)	−2.13 (0.60)
Bayes, $\sigma_0 = 10$	9.8 (5.6)	(2.2, 23.2)	−0.47 (0.45)	−2.13 (0.59)
Bayes, $\sigma_0 = 1$	1.6 (0.7)	(0.3, 3.1)	−0.22 (0.33)	−1.75 (0.43)

[a]Profile likelihood confidence interval and Bayes equal-tail posterior interval

The case $\sigma_0 = 1$ reflects strong prior belief that effects are not strong, Its posterior interval for β_1 is (0.3, 3.1), compared with (5.9, 222.3) for highly uninformative priors having $\sigma_0 = 100$.

With highly disperse prior distributions, estimated effect size is very imprecise, because log-likelihood function is so flat in β_1 dimension.

Although upper bound of Bayesian posterior interval for β_1 is not ∞, for practical purposes an upper bound such as $e^{23.2} = 1.2 \times 10^{10}$ for the corresponding odds ratio e^{β_1}, obtained when $\sigma_0 = 10$, *is* ∞.

Loglinear Models for Count Responses

Count responses have positive expected values. GLMs usually model *log* of $E(Y)$, which like linear predictor, can take any real-number value.

For a count observation Y_i for subject i, taking possible values $0, 1, 2, \ldots$,

$$\log[E(Y_i \mid \boldsymbol{\beta})] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \ldots, n,$$

has exponential effects of explanatory variables on expected value μ_i,

$$\mu_i = E(Y_i \mid \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}).$$

$E(Y_i)$ at $x_{ij} + 1$ equals $E(Y_i)$ at x_{ij} multiplied by e^{β_j}, adjusting for other explanatory variables.

For a count response, *Poisson distribution* $f(y \mid \mu) = e^{-\mu}\mu^y/y!$ for $y = 0, 1, 2, \ldots$ is simple, but forces $\text{var}(Y) = E(Y)$. *Negative binomial distribution* has two parameters, permits $\text{var}(Y) > E(Y)$ (*overdispersion*).

Modeling Rates for Count Data

Often expected value of a count response variable is proportional to an index t, such as *size* or *area* or *time*. A sample count y_i corresponds to a sample *rate* of y_i/t_i.

The loglinear model extends to focus on true rate $E(Y_i/t_i) = \mu_i/t_i$,

$$\log(\mu_i/t_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

$\log(\mu_i/t_i) = \log \mu_i - \log t_i$, and the adjustment term $\log t_i$ is called an *offset*.

With offset in linear predictor, expected response count satisfies

$$\mu_i = t_i \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}).$$

Mean has proportionality coefficient for t_i that depends on values of explanatory variables.

Example: Lung Cancer Survival Counts

For 539 males diagnosed with lung cancer, observe death counts according to stage of disease at seven two-month time intervals after diagnosis. Here, time at risk is number of months of observation of subjects still alive during that follow-up interval. (We show 3 intervals here.)

Follow-up time interval:	0–2 months			2-4 months			4-6 months		
Stage of disease:	1	2	3	1	2	3	1	2	3
Number of deaths:	15	17	89	5	11	56	13	13	29
Time at risk:	255	227	443	224	191	271	203	168	175

For μ_{ij} = expected number of deaths and t_{ij} = total time at risk for patients alive during follow-up time interval j with stage of disease i, we focus on effects of stage of disease on death rate.

Let (z_{i1}, z_{i2}, z_{i3}) denote indicator variables for the stages of disease. We find Bayesian fit of loglinear model (ignoring follow-up time interval) for effect of stage of disease on the death rate,

$$\log(\mu_{ij}/t_{ij}) = \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3}.$$

Bayesian Fit of Poisson Loglinear Model for Death Rates

```
-------------------------------------------------------------------------------
> Cancer <- read.table("http://bayes4ds.rwth-aachen.de/data/Cancer.dat", header=TRUE)
> Cancer    time stage deaths risktime
       1     1     1     15      255
       2     2     1      5      224
...
      21    7     3      6       46
> Cancer$logrisktime <- log(Cancer$risktime) # We use this as the offset in the model
> Cancer$stagef <- factor(Cancer$stage)
> fit.bayes <- brm(deaths ~ -1 + stagef + offset(logrisktime), family=poisson(link=log),
+              data=Cancer, prior = prior(normal(0,10), class=b))
> summary(fit.bayes)
Regression Coefficients:
        Estimate Est.Error l-95% CI u-95% CI
stagef1    -3.08      0.13    -3.35    -2.83
stagef2    -2.60      0.11    -2.82    -2.39
stagef3    -1.70      0.07    -1.84    -1.58
> hypothesis(fit.bayes, c("stagef3-stagef1=0","stagef3-stagef2=0","stagef2-stagef1=0"))
             Hypothesis Estimate Est.Error CI.Lower CI.Upper
1 (stagef3-stagef1) = 0     1.38      0.15     1.10     1.68 # compares stage 3 and stage 1
2 (stagef3-stagef2) = 0     0.89      0.13     0.64     1.16
3 (stagef2-stagef1) = 0     0.49      0.17     0.15     0.83
-------------------------------------------------------------------------------
```

Death rate progressively worsens as stage of disease advances. Estimated death rate at stage 3 is $\exp(\tilde{\beta}_3^S - \tilde{\beta}_1^S) = \exp[-1.70 - (-3.08)] = \exp(1.38) = 4.0$ times that at stage 1.

Bayesian Fit of Negative Binomial Loglinear Model

Do we get a better fit by assuming a negative binomial distribution for response?

```
------------------------------------------------------------------------------------
> fit.bayes2 <- brm(deaths ~ -1 + stagef + offset(logrisktime),
+                   family=negbinomial(link=log), data=Cancer,
+                   prior=prior(normal(0,10),class=b) + prior(exponential(0.001),class=shape))
> summary(fit.bayes2)
Regression coefficients:
        Estimate Est.Error l-95% CI u-95% CI
stagef1    -3.09      0.14    -3.36    -2.83
stagef2    -2.60      0.12    -2.83    -2.38
stagef3    -1.71      0.07    -1.87    -1.57
------------------------------------------------------------------------------------
```

Similar fit, because these data show no evidence of overdispersion relative to Poisson sampling. BIC values suggest that more complex models also having a follow-up time interval effect do not improve the fit.

The textbook has an example (number of male horseshoe crabs clustered around females during spawning season) with substantial overdispersion and better fits with negative binomial models.

Hierarchical Bayesian Modeling

Many Bayesian models have a *hierarchical* structure:

- In selecting a prior distribution determined by hyperparameters, one way to deal with uncertainty about the hyperparameters is a *hierarchical* approach that includes prior distributions for the hyperparameters themselves (in textbook, but not this lecture).
- *Multilevel models* apply to hierarchical data files in which the observations occur in clusters that are sampled from groups that themselves form clusters when sampled from groups at higher levels.
- Many studies observe Y for each subject repeatedly, such as when observations occur at several times in a *longitudinal study*. More generally, Y may be observed for *matched sets* of cases, such as children in families. Models with such clusters of observations needs to take into account the likely correlations among the responses within clusters (in text, but not this lecture).

Cluster-Specific Random Effects in Multilevel Models

The textbook considers each of these hierarchical types of analyses. Here, we illustrate multilevel models.

In frequentist modeling, a term in the linear predictor that applies to a particular subject or cluster in a sample from a population is called a *random effect*. These are *cluster-specific* effects, referred to as *subject-specific* when each cluster is a subject.

Random effects are distinguished from *parameters* that describe *population-averaged* effects of explanatory variables, which are called *fixed effects*.

Multilevel models contain a random effect for each cluster of observations at each level of the model. There terms account for the correlation that usually occurs between pairs of observations within clusters at each level.

Example: Smoking Prevention and Cessation Study

A study (from book by Hedeker and Gibbons) of efficacy of two programs for discouraging young people from smoking compared four groups, defined by a 2×2 factorial design according to whether a student was exposed to a school-based curriculum (SC; 1 = yes, 0 = no) and a television-based prevention program (TV; 1 = yes, 0 = no).

Subjects: 1600 seventh-grade students from 135 classrooms in 28 Los Angeles schools randomly assigned to the four intervention conditions.

Response variable: tobacco and health knowledge (THK) scale, measured at end of study; values between 0 and 7, with $\bar{y} = 2.66$ and $s_y = 1.38$. THK at beginning of study (PTHK = Pre-THK) used as covariate.

Student	School	Class	SC	TV	PTHK	THK
1	403	403101	1	0	2	3
2	403	403101	1	0	4	4
...						
1600	515	515113	0	0	3	3

Modeling with Random Effects for Smoking Prevention

Let y_{ijk} denote whether the follow-up THK score for student i within classroom j in school k exceeds 2 (1 = yes, 0 = no). Multilevel model

$$\text{logit}[P(Y_{ijk} = 1 \mid c_{jk}, s_k, \boldsymbol{\beta})] = c_{jk} + s_k + \beta_0 + \beta_1 \text{PTHK}_{ijk} + \beta_2 \text{SC}_k + \beta_3 \text{TV}_k.$$

Random effect c_{jk} for classroom j in school k has a $\mathcal{N}(0, \tau_c^2)$ distribution and random effect s_k for school k has a $\mathcal{N}(0, \tau_s^2)$ distribution.

We first find frequentist, ML fit. The term (1|class) represents random effect for classrooms and (1|school) represents random effect for schools.

```
> Smoking <- read.table("http://bayes4ds.rwth-aachen.de/data/Smoking.dat", header=TRUE)
> library(lme4)
> fit.ML <- glmer(y ~ (1|class) + (1|school) + PTHK + SC + TV, family=binomial, data=Smoking)
> summary(fit.ML) # Frequentist, ML fit of model
Random effects:
 Groups Name        Variance Std.Dev.
 class  (Intercept) 0.16728  0.4090 # estimated variability of classroom random effects
 school (Intercept) 0.06413  0.2532 # estimated variability of school random effects
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.13163    0.17827  -6.348 2.18e-10
PTHK         0.39512    0.04627   8.539  < 2e-16
SC           0.80014    0.16893   4.737 2.17e-06
TV           0.10786    0.16819   0.641    0.521
```

Bayesian Modeling for Smoking Prevention Study

Next we find a Bayesian fit for the multilevel model:

```
------------------------------------------------------------------------------------------
> fit.bayes <- brm(y ~ (1|class) + (1|school) + PTHK + SC + TV, family=bernoulli(link=logit),
+    prior = prior(normal(0,100), class=Intercept) + prior(normal(0,100), class=b)
+    + prior(exponential(0.001), class=sd, group="school")
+    + prior(exponential(0.001), class=sd, group="class"), data=Smoking)
> summary(fit.bayes)
Group-Level Effects:
~class (Number of levels: 135)
             Estimate Est.Error l-95% CI u-95% CI
sd(Intercept)    0.42      0.11     0.18     0.63 # est. std. dev. classroom random effects
~school (Number of levels: 28)
             Estimate Est.Error l-95% CI u-95% CI
sd(Intercept)    0.32      0.15     0.02     0.62 # est. std. dev. of school random effects
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept    -1.14      0.20    -1.53    -0.75
PTHK          0.40      0.05     0.31     0.49
SC            0.81      0.20     0.42     1.21
TV            0.11      0.20    -0.28     0.50
------------------------------------------------------------------------------------------
```

Results are similar with frequentist and Bayesian fitting of the multilevel model.

Suppose we ignore the clustering

The introduction of random effects in a multilevel model permits observations within a classroom to be positively correlated and observations within a school to be positively correlated.

Ignoring multilevel clustering in classrooms and schools and treating observations as independent, ordinary logistic regression model

$$\text{logit}[P(Y_{ijk} = 1 \mid \boldsymbol{\beta})] = \beta_0 + \beta_1 \text{PTHK}_{ijk} + \beta_2 \text{SC}_k + \beta_3 \text{TV}_k$$

does not contain any random effects. Here is a Bayesian fit:

```
-------------------------------------------------------------------------------------
> fit.bayes2 <- brm(y ~ PTHK + SC + TV, family=bernoulli(link=logit), data=Smoking,
+     prior = prior(normal(0,100), class=Intercept) + prior(normal(0,100), class=b))
> summary(fit.bayes2)
          Estimate Est.Error l-95% CI u-95% CI
Intercept    -1.12      0.13    -1.39    -0.88
PTHK          0.40      0.04     0.32     0.49
SC            0.77      0.11     0.56     0.98 # Est. Error = 0.20 for multilevel model
TV            0.13      0.10    -0.08     0.33 # Est. Error = 0.20 for multilevel model
-------------------------------------------------------------------------------------
```

Estimated fixed effects are similar to multilevel model, but standard errors are substantially underestimated for between-subjects effects (SC and TV).

Positively correlated observations do not impart as much information.

The end!

We hope this presentation has been useful, giving you an overview of the Bayesian approach.

Thanks very much for your time and attention!