An Introduction to the Bayesian Approach to Statistical Inference

Alan Agresti, Distinguished Professor Emeritus
University of Florida, USA

Presented at UF Statistics Club

© Alan Agresti, 2026

Focus of lecture

- Overview of the Bayesian approach for statistical inference.
- 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 Alan Agresti, Maria Kateri, Ranjini Grove, and Antonietta Mira (Taylor & Francis, 2026).
- Examples of data analyses use `R` software. The textbook has an appendix showing how also to conduct these analyses using `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 \$67,800 and \$69,600." 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 (Thomas Bayes, 1701-1761)

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*.

Note: Richard Price found this result in Bayes' papers when he passed away in 1761 and re-worked it and published it in 1764. Unaware of this, Pierre-Simon Laplace re-discovered the theorem in 1774 and showed how to apply it to data analysis as explained on the next page.

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 $f(\boldsymbol{y} \mid \theta)$.

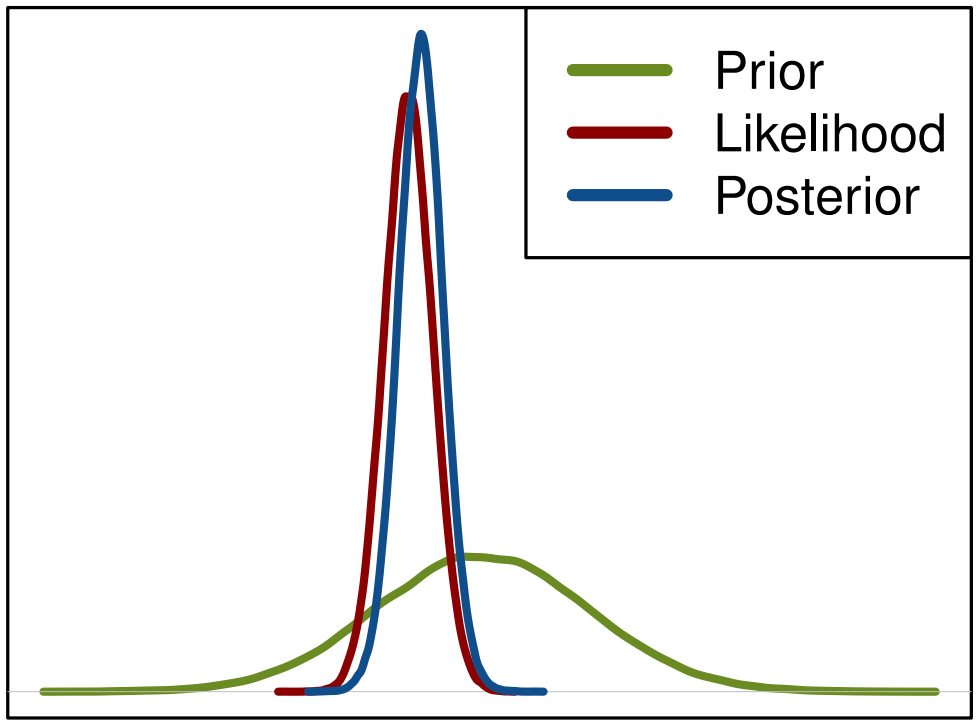
By Bayes' Theorem, the posterior probability function for θ is proportional to the product of the likelihood function with the probability function for the prior distribution of θ,

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

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*

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

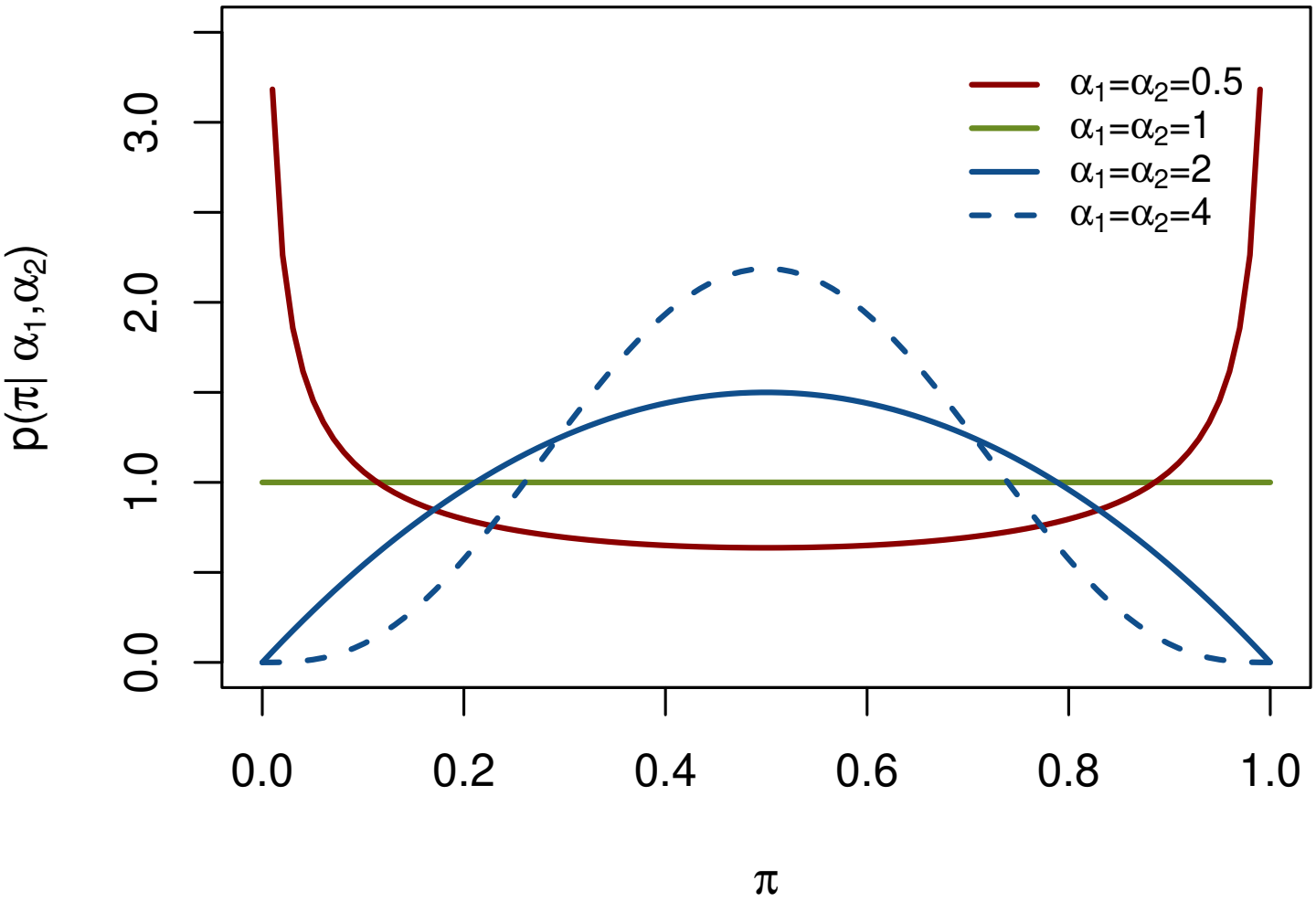
for *hyperparameters* $\alpha_1 > 0$ and $\alpha_2 > 0$.

- 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

  $$\text{mean} = \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{y+1}{n+2} = \frac{n}{n+2}\left(\frac{y}{n}\right) + \frac{2}{n+2}\left(\frac{1}{2}\right),$$

  a weighted average of sample proportion $\frac{y}{n}$ and prior mean $\frac{1}{2}$, corresponding to sample proportion after adding 2 *imaginary* observations, 1 of each type.
- For large n, it is approximately bell-shaped with mean close to sample proportion $\frac{y}{n}$, which is the maximum likelihood (ML) estimate.

Beta distributions



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 $\alpha_2 = n - y + 1 = 580$.

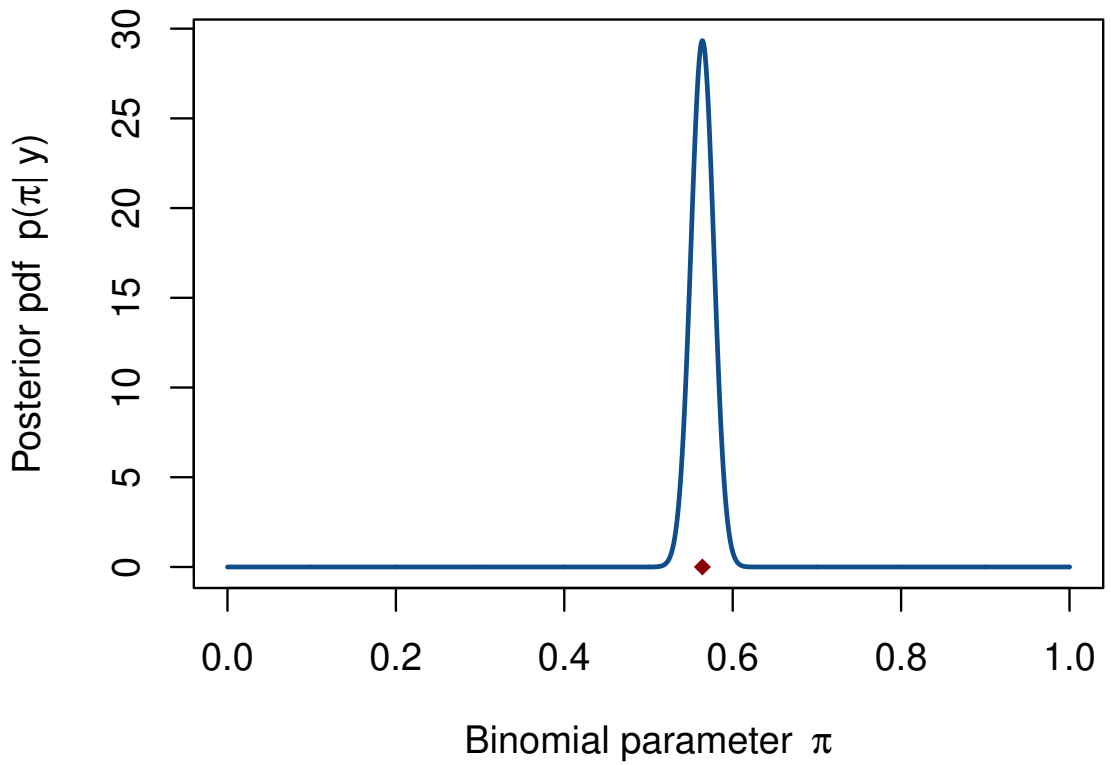
Mean $= (y + 1)/(n + 2) = 750/1330 = 0.5639$.

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) # beta quantile function
[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

From the `binom` package in `R`, for the beta posterior with hyperparameters $\alpha_1 = 750$ and $\alpha_2 = 580$, the 95% HPD interval is (0.537, 0.591).

```
------------------------------------------------------------------------------------------------
> 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)
  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% confidence interval

> binom.confint(749, 1328, conf.level=0.95, method="agresti-coull")
         method   x    n     mean     lower     upper
1 agresti-coull 749 1328 0.564006 0.5371881 0.5904547
------------------------------------------------------------------------------------------------
```

Frequentist 95% CI ("asymptotic," based on large-sample normal distribution of ML estimator $\hat{\pi} = y/n$) is

$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

This method has poor performance when n small or π near 0 or 1.

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 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.
- *Generalized linear models* permit us to assume response distributions for Y other than normality.
- Bias/variance tradeoff: In selecting a model, it's beneficial to 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.

Bayesian Approach to Normal Linear Model

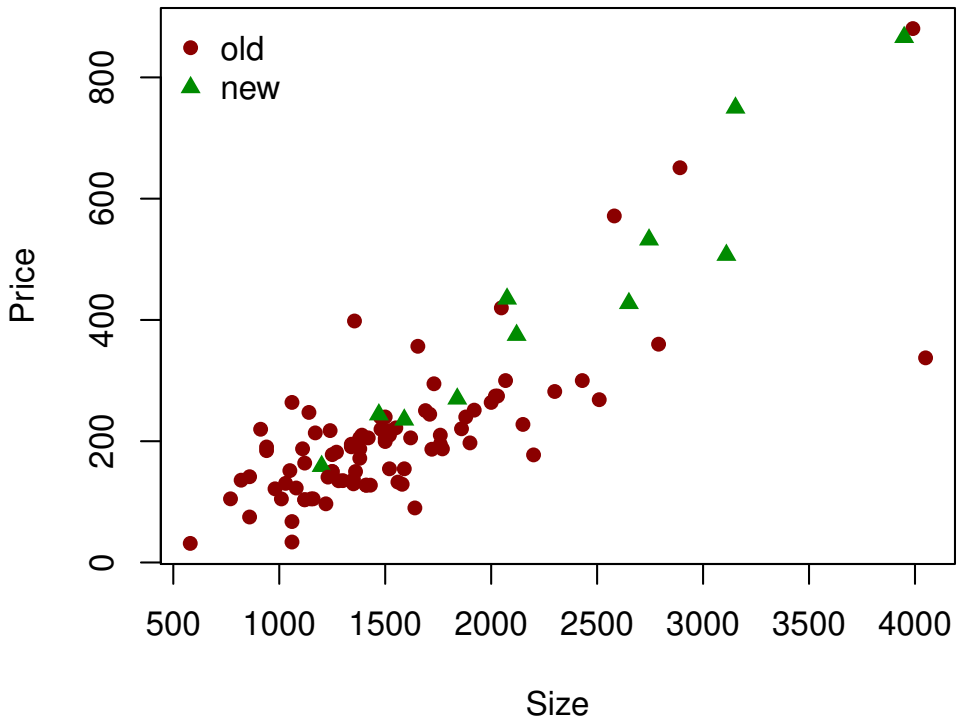
Model assumes independent observations such that for $i = 1, \ldots, n$,

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

- For prior, we can take $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$ where $\{\beta_j\}$ are independent, e.g., each with $\mathcal{N}(0, \sigma_0^2)$ distribution with large σ_0.
- A simple prior for σ is *exponential distribution*, $p(\sigma \mid \lambda) = \lambda e^{-\lambda\sigma}$, $\sigma \geq 0$, which has mean = standard deviation $= 1/\lambda$.
- With highly disperse prior distributions for $\{\beta_j\}$ and for σ, 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: Modeling Selling Price of a Home

For 100 home sales in Gainesville, Y = selling price (thousands of dollars), x_1 = size of house (square feet), and x_2 = whether the house is new (1 = yes, 0 = no).



Scatterplot suggests perhaps greater slope for new houses than old. We include an interaction term to allow effects of size to differ for new homes and older homes.

Example: Modeling Selling Price of a Home

Model $E(Y) = \beta_0 + \beta_1(size) + \beta_2(new) + \beta_3(size \times new)$; we center house size values around their mean, so β_2 compares $E(Y)$ of new homes ($new = 1$) and older homes ($new = 0$) at mean house size.

Here is the least squares fit, assuming linear model with constant response variance:

```
------------------------------------------------------------------------------------------
> Houses <- read.table("http://bayes4ds.rwth-aachen.de/data/Houses.dat", header=TRUE)
> Houses$size2 <- Houses$size - mean(Houses$size) # center size around its mean

> summary(lm(price ~ size2 + new + size2:new, data=Houses)) # least squares fit
             Estimate Std. Error t value Pr(>|t|)
Intercept      221.90       8.36  26.529  < 2e-16
size2            0.16       0.01  11.082  < 2e-16
new             33.53      32.78   1.023  0.30903  # estimated new effect = 33.53 at mean size
size2:new        0.09       0.03   2.855  0.00527
---
Residual standard error: 78 on 96 degrees of freedom # estimate of sigma parameter
------------------------------------------------------------------------------------------
```

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.

At the mean house size, the estimated selling price is \$33,530 higher for new homes than for older homes.

Example: Bayesian Modeling of Selling Price

For Bayesian fitting of the *normal GLM*, we use highly disperse prior distributions: $N(0, 1000^2)$ distribution for each model parameter and exponential distribution for σ with mean and standard deviation 1000.

Here are results using `brms`, Bayesian regression modeling using Stan (honoring Stanislaw Ulam and using "Markov chain Monte Carlo"):

```
-------------------------------------------------------------------------------------------
> library(brms) # Bayesian fit assuming normal response
> 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
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
-------------------------------------------------------------------------------------------
```

With highly disperse priors, similar results as least squares.

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 $E(Y) = \$100,000$ 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 $\alpha > 0$ and rate parameter $\beta > 0$ 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.
- With a *generalized linear model* (GLM), we can assume a gamma distribution for Y. Gamma GLMs usually assume that the shape parameter α is constant but unknown, like σ^2 in normal GLMs.
- Modeling the log of the mean ensures that the fitted values for the mean are all positive, with results similar to assuming *log-normal distribution* for Y.

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" to model the log of the mean.

```
-------------------------------------------------------------------------------
> 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))
> 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 (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 $\exp[1000(0.00059)] = 1.81$.

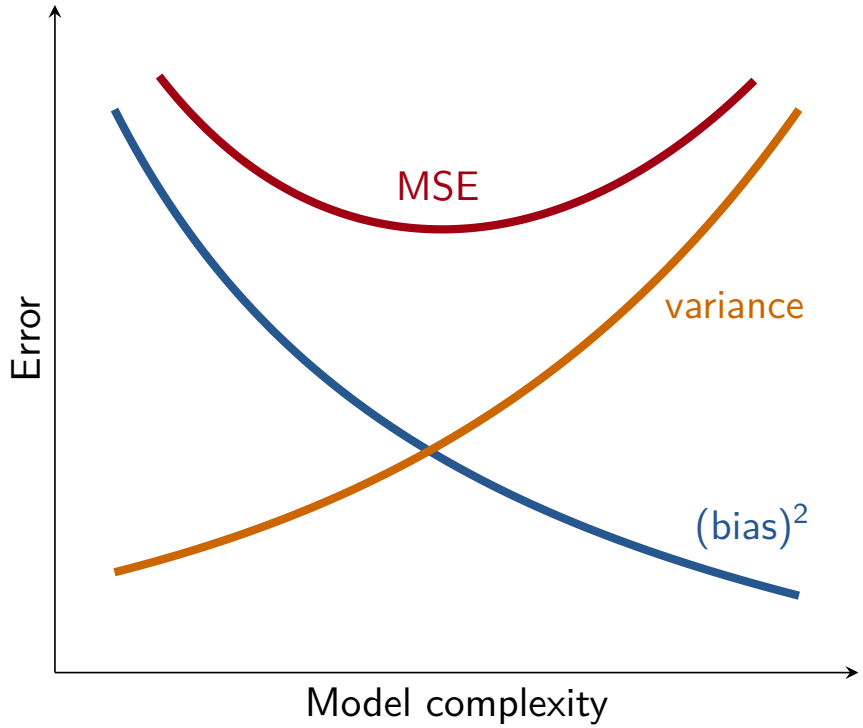
Model Comparison and the Bias/Variance Tradeoff

- 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* of an estimator $\hat{\theta}$ of a parameter θ,

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

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 (Akaike information criterion). 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,

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

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.

The end

This lecture was taken from a half-day short course that I presented this year at two Italian universities. You can access the pdf file for the entire course at the summary of my Bayesian book at

www.alanagresti.com

That short-course file also presents Bayesian methods for comparing means, comparing proportions, Bayesian fitting of other generalized linear models (such as logistic regression for binary responses, which is especially useful when ML estimates are infinite), and a summary of Markov chain Monte Carlo methods for approximating the posterior distribution.

I hope this presentation has been useful for giving you an introductory overview of the Bayesian approach.

Thanks very much for your time and attention!