Introduction to Bayesian Inference

Yingbo Li

Clemson University

MATH 9810

Outline

Motivating Bayesian Inference

- Significance tests and confidence intervals are forms of "classical" or "frequentist" inference.
- When might classical inference be inadequate?
  - Suppose you flip a coin (with unknown probability of heads) three times and get tails all three times.
  - The sample percentage of heads equals zero. But, this can't be an accurate estimate of the true percentage of heads!
  - A priori of flipping the coin, we believe the true percentage is around 0.5, not 0.0.
- Bayesian inference provides a formal method for quantifying and incorporating our prior beliefs into inference.

Why Bayesian statistics?

- Long history: named after the 18^{th} century Presbyterian minister and mathematician Thomas Bayes (1701 - 1761).

- Modeling: incorporate prior belief or domain experts knowledge.
- Theoretical: doesn't need large sample assumption.
- Computational: Markov chain Monte Carlo (MCMC).

Bayesian approaches are largely popularized by revolutionary advance in computational technology during the last twenty years.

Celebrity statistician Nate Silver

He used Bayesian approaches to

- predict the result of 2008 presidential election and got 49 of the 50 states correct.
- predict the result of 2012 presidential election and got 50 of the 50 states correct.

Two schools of statistics

Frequentist (classical)

Probability: long run relative frequencies of repeatable events.

$$P(A) = \lim_{n\to\infty} \frac{\#(A)}{n}$$

- One time events?
- John Maynard Keynes (1883-1946) commented: In the long run, we are all dead.

Bayesian

Probability:
a subjective degree of belief.

- Two people could have differing probabilities $P(A)$.
- Probability changes as new information (data) arise according to Bayes rule.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayes Theorem Example: DNA Testing

- When two samples of DNA are from the same person, the chance of a true match is 0.999. (Actually, it is even higher.)
- When two samples of DNA are from different people, the chance of a false match is 0.01. (Actually, it is even lower.)
- A DNA test on a suspect is positive. What is the chance that the suspect owns the DNA found at the crime scene?

Setting Up the Problem

- Let $A =$ event that suspect owns DNA.
  $Pr(A)$ is our prior belief that the suspect owns the DNA. It is our belief before seeing the test results.
- Let $M =$ event that DNA test indicates a match. This is the data that we will collect.
- We want $Pr(A|M)$.
  $Pr(A|M)$ is our posterior belief, i.e., after seeing the data from the DNA test, that the suspect owns the DNA.

Formalizing a Model for Prior Beliefs

Suppose that the list of potential suspects comprises 5000 people. A reasonable model for the prior belief is:

$Pr(A) = 1/5000 = .0002$
$Pr(A^c) = 4999/5000 = .9998$

We could increase $Pr(A)$ by collecting more circumstantial evidence (other than DNA) that ties the suspect to the scene.

Formalizing a Model for the Data

From the given information, we have

$$Pr(M|A) = .999$$
$$Pr(M^c|A) = .001$$
$$Pr(M|A^c) = .01$$
$$Pr(M^c|A^c) = .99$$

These probabilities define the model for the data, M. Since we observe a match, we only need $Pr(M|A)$ and $Pr(M|A^c)$.

Combining Prior Beliefs with Data

Bayes Theorem allows us to compute $Pr(A|M)$.

$$P(A \mid M) = \frac{P(A \cap M)}{P(M)}$$
$$= \frac{P(M \mid A)P(A)}{P(M)}$$

Law of Total Probability:

$$P(M) = P(M \mid A)P(A) + Pr(M \mid A^c)Pr(A^c)$$

Result

Substituting the probabilities

$$P(M) = (.999)(.0002) + (.01)(.9998) = .010196$$
$$P(A \mid M) = \frac{(.999)(.0002)}{.010196} = .01942$$

The probability that the suspect owns the DNA, given that the test returns a positive match, is about 2%.

Reducing the list to 1000 people results in $P(A \mid M) = .09$.
Reducing the list to 100 people results in $P(A \mid M) = .50$.

Univariate random variable

Cumulative distribution function (cdf)

$$F_X(x) = P(X \leq x), \quad \text{for any } x \in \mathbb{R}$$

	Discrete	Continuous
pmf / pdf	$f_X(x) = P(X = x)$	$P(X \in B) = \int_B f(x)dx$
well-defined	$\sum_{i=1}^{\infty} f(x_i) = 1$	$\int_{-\infty}^{\infty} f(x)\ dx = 1$
cdf	$F(a) = \sum_{\text{all } x \leq a} f(x)$	$F(x) = \int_{-\infty}^{x} f(t)dt$ $f(x) = \frac{d}{dx}F(x)$
expectation	$E[X] = \sum_{\text{all } x} x \cdot f(x)$	$E[X] = \int_{-\infty}^{\infty} xf(x)dx$
	$E[g(X)] = \sum_{\text{all } x} g(x)\ f(x)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

Joint distribution

	Discrete	Continuous
pmf/pdf	$P(X = x, Y = y)$	$P[(X,Y) \in C] = \iint_{(x,y)\in C} f(x,y)\ dxdy$
marginal	$f_X(x) = \sum_y f(x,y)$,	$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\ dy$,
	$f_Y(y) = \sum_x f(x,y)$	$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\ dx$
cdf		$F(a,b) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\ dxdy$
		$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$

Conditional distribution

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

- Joint density $f(x,y) = f(x|y)f_Y(y) = f(y|x)f_X(x)$

Conditional expectation

- $E(X|Y=y) = \sum_{\text{all } x} xf(x|y)$ or $\int_{-\infty}^{\infty} xf(x|y)\ dx$
- Law of total expectation $E_Y\left[E(X|Y)\right] = E(X)$

What is fixed, Y or θ?

Frequentist:

$$p(Y|\theta)$$

- Parameter θ: fixed
- Data Y: random
- Take average on multiple datasets unconditionally.

Bayesian:

$$p(\theta|Y)$$

- Data Y: fixed
- Parameter θ: random
- Make inference conditional on the current data.

Extending to Statistical Analysis

Bayes Theorem extends naturally to parameters in statistical inference as well.

- "Characteristics" are akin to parameters θ in probability models, e.g., $\theta = p$ in the binomial distribution $Bin(n, p)$.
- "Data" are akin to measurements on sampled data subjects expressed numerically, say y.
- Before the sample is collected, both y and θ are unknown.
- "Model for data" is akin to a probability model for Y assuming we know θ, e.g., $Y \sim Bin(n, p)$.

Bayesian Inference

Bayesian inference provides a formal approach for updating prior beliefs with the observed data to quantify uncertainty *a posteriori* about θ

- Prior Distribution $p(\theta)$
- Sampling Model $p(y \mid \theta)$
- Posterior Distribution:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)} = \frac{p(y \mid \theta)\, p(\theta)}{\int_\Theta p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}$$

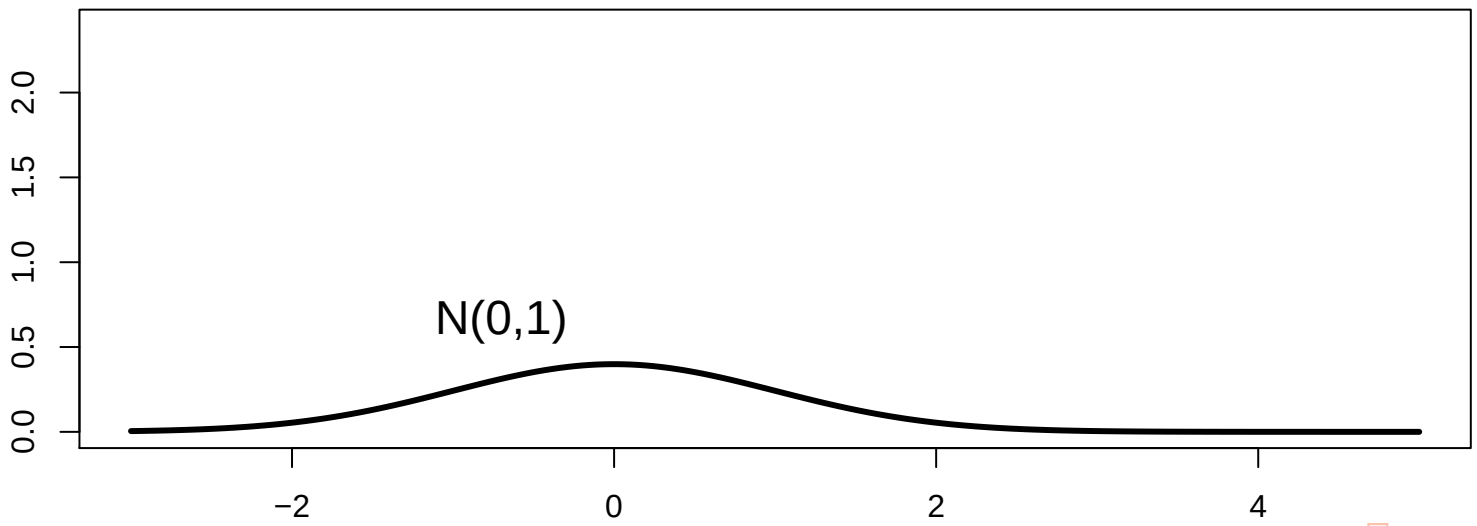
(for discrete support for θ replace integral with sum)

Bayesian inference

General steps

1. Specify sampling model $p(Y|\theta)$, e.g. $Y_i \overset{iid}{\sim} N(\theta, 1)$
2. Specify prior distribution $p(\theta)$, e.g. $\theta \sim N(0, 1)$
3. Observe data Y, e.g. $n = 5, \bar{Y} = 2$,
   update knowledge about θ, posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \sim N\left(\frac{n}{n+1}\bar{Y}, \frac{1}{n+1}\right)$$

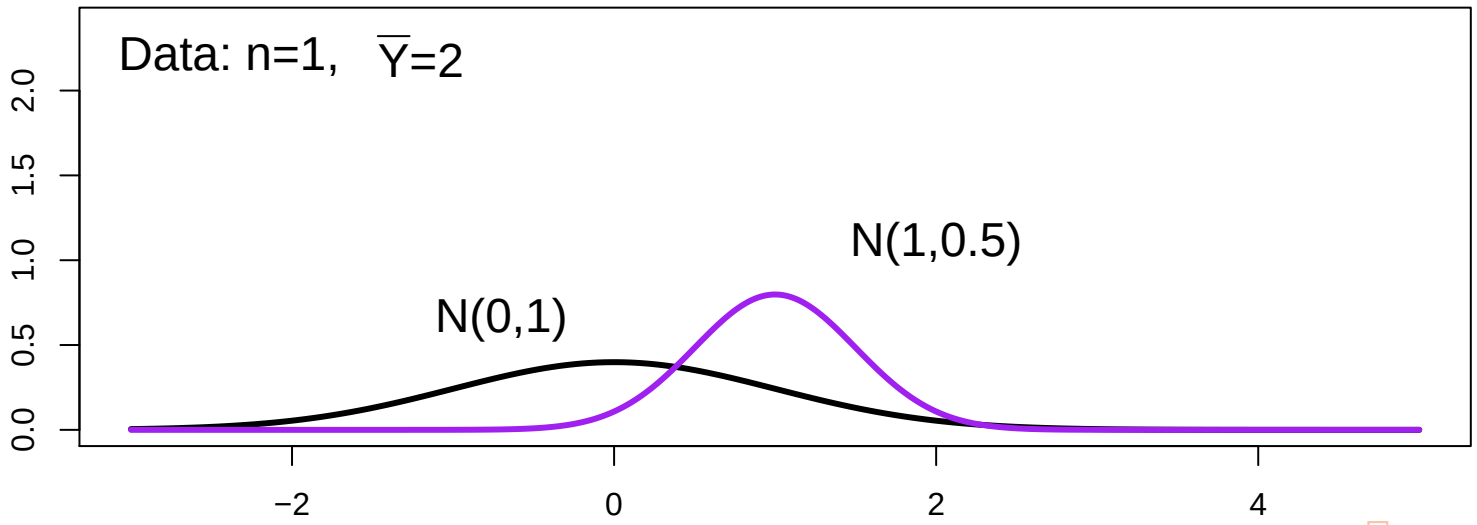


Bayesian inference

General steps

1. Specify sampling model $p(Y|\theta)$, e.g. $Y_i \overset{iid}{\sim} N(\theta, 1)$
2. Specify prior distribution $p(\theta)$, e.g. $\theta \sim N(0, 1)$
3. Observe data Y, e.g. $n = 5, \bar{Y} = 2$,
   update knowledge about θ, posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \sim N\left(\frac{n}{n+1}\bar{Y}, \frac{1}{n+1}\right)$$

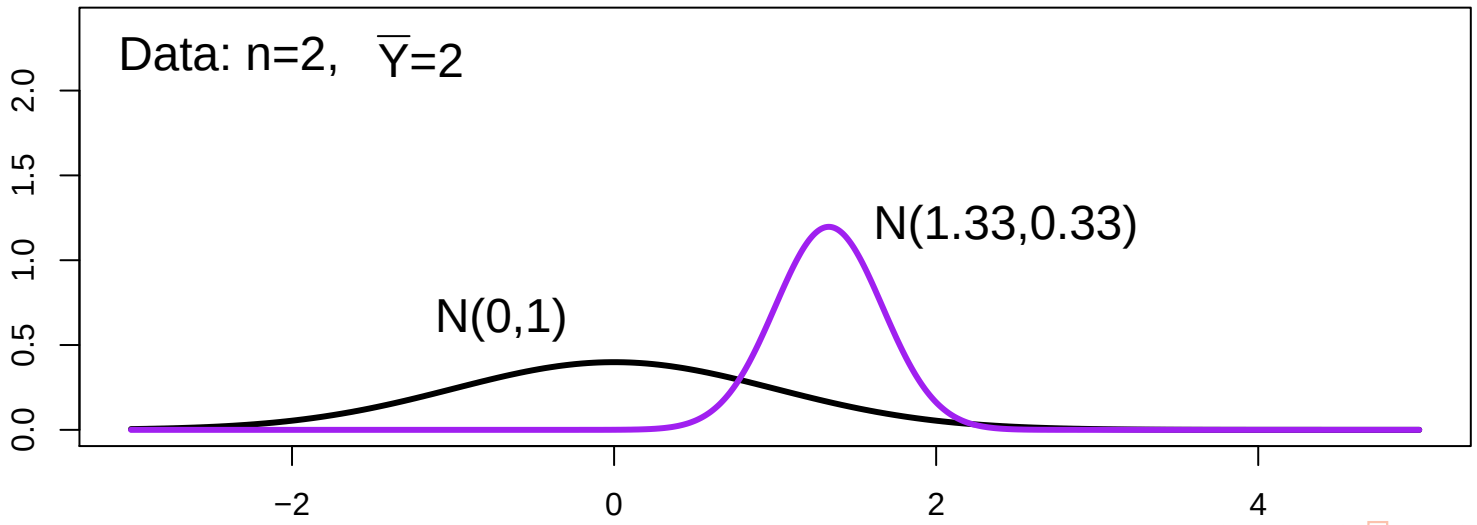


Bayesian inference

General steps

1. Specify sampling model $p(Y|\theta)$, e.g. $Y_i \overset{iid}{\sim} N(\theta, 1)$
2. Specify prior distribution $p(\theta)$, e.g. $\theta \sim N(0, 1)$
3. Observe data Y, e.g. $n = 5, \bar{Y} = 2$,
   update knowledge about θ, posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \sim N\left(\frac{n}{n+1}\bar{Y}, \frac{1}{n+1}\right)$$

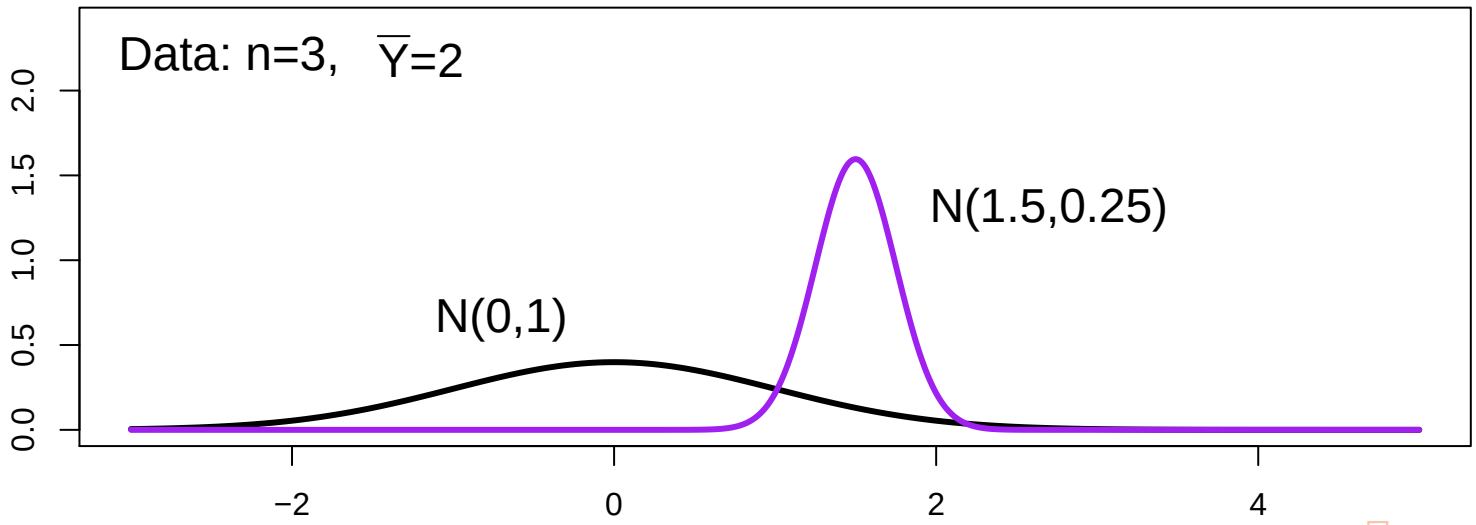


Bayesian inference

General steps

1. Specify sampling model $p(Y|\theta)$, e.g. $Y_i \overset{iid}{\sim} N(\theta, 1)$
2. Specify prior distribution $p(\theta)$, e.g. $\theta \sim N(0, 1)$
3. Observe data Y, e.g. $n = 5, \bar{Y} = 2$,
   update knowledge about θ, posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \sim N\left(\frac{n}{n+1}\bar{Y}, \frac{1}{n+1}\right)$$

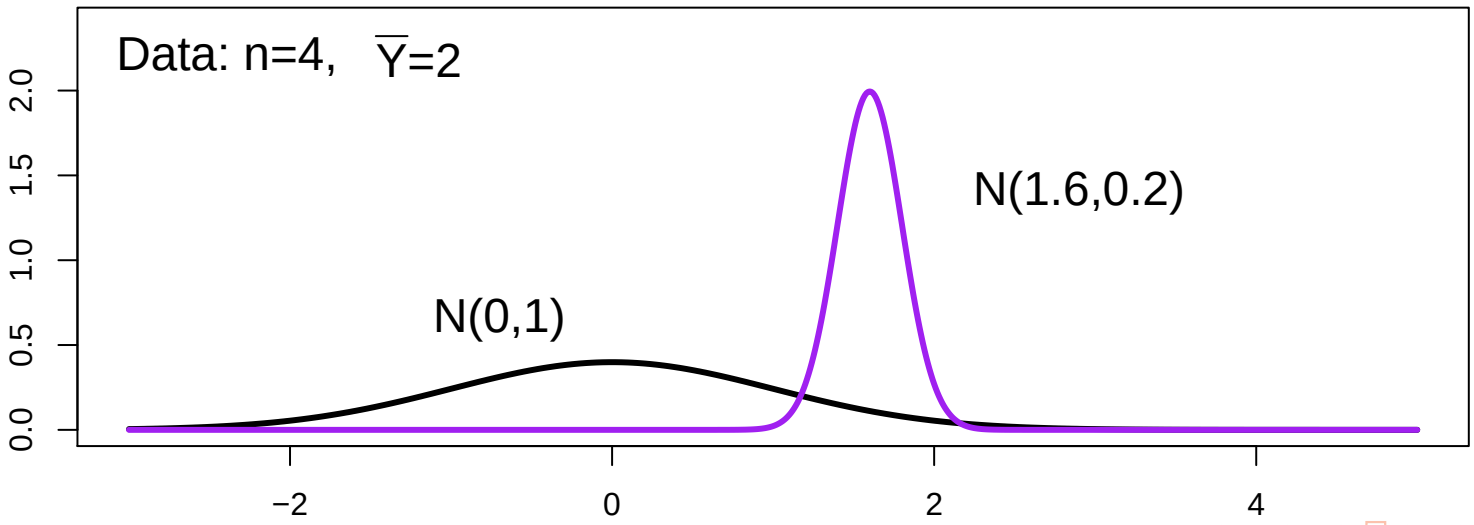


Bayesian inference

General steps

1. Specify sampling model $p(Y|\theta)$, e.g. $Y_i \overset{iid}{\sim} N(\theta, 1)$
2. Specify prior distribution $p(\theta)$, e.g. $\theta \sim N(0, 1)$
3. Observe data Y, e.g. $n = 5, \bar{Y} = 2$,
   update knowledge about θ, posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \sim N\left(\frac{n}{n+1}\bar{Y}, \frac{1}{n+1}\right)$$

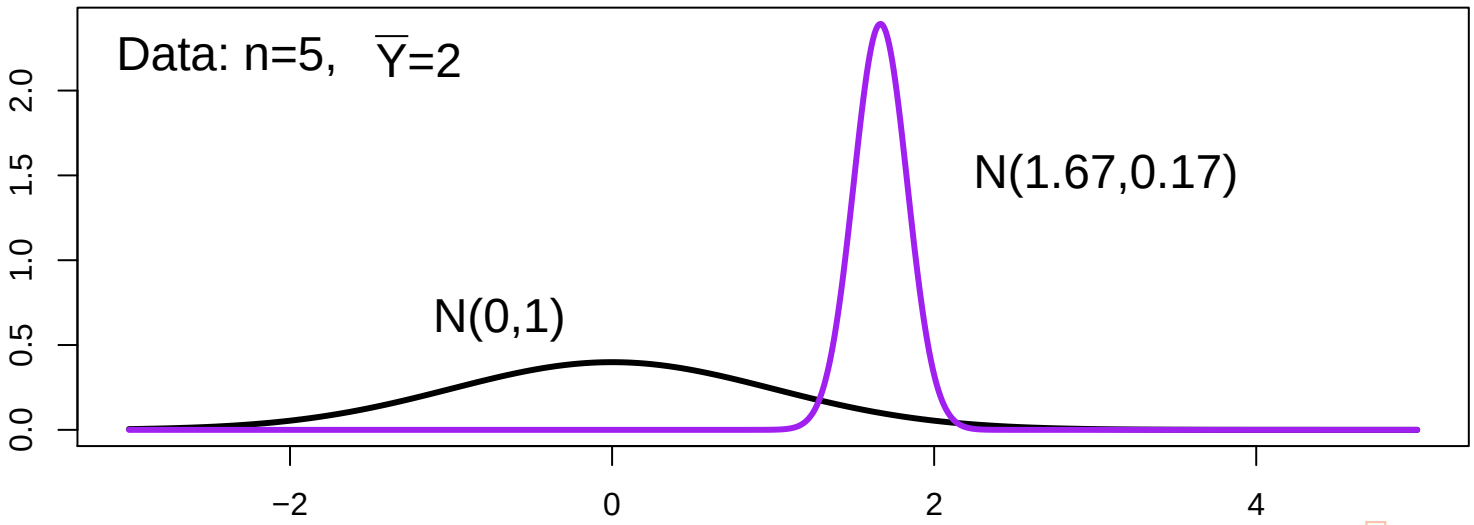


Bayesian inference

General steps

1. Specify sampling model $p(Y|\theta)$, e.g. $Y_i \overset{iid}{\sim} N(\theta, 1)$
2. Specify prior distribution $p(\theta)$, e.g. $\theta \sim N(0, 1)$
3. Observe data Y, e.g. $n = 5, \bar{Y} = 2$,
   update knowledge about θ, posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \sim N\left(\frac{n}{n+1}\bar{Y}, \frac{1}{n+1}\right)$$



Analysis Goals

Bayesian methods go beyond the formal updating of the prior distribution to obtain a posterior distribution

- Estimation of uncertain quantities (parameters) with good statistical properties
- Prediction of future events
- Tests of hypotheses
- Making decisions

Sampling Models

At least Two Physical Meanings for Sampling Models

1. Random Sampling Model
   - well-defined Population
   - individuals in sample are drawn from the population in a "random" way
   - could conceivably measure the entire population
   - probability model specifies properties of the full population

Infection rate

Interest is in the prevalence of a disease (say H1N1 flu) in a city. Rather than using a census of the population, take a random sample of 20 individuals.

- θ: fraction of infected individuals
- Y_i indicator that ith individuals in the sample of 20 is infected
- Model for Y_i given θ?

Sampling Models

2. Observations are made on a system subject to random fluctuations
   - probability model specifies what would happen if, hypothetically, observations were repeated again and again under the same conditions.
   - Population is hypothetical

Probability Distributions for the Random Outcomes

Parametric Probability Models:

$$Y_i \mid \theta \overset{iid}{\sim} f(y \mid \theta) \text{ for } i = 1, \ldots, n$$

- Bernoulli/Binomial
- Multinomial
- Poisson
- Normal
- Log-normal
- Exponential
- Gamma

Exchangeable

Let $p(y_1, \ldots, y_n)$ be the joint distribution of $Y_1, \ldots, Y_n$ and let $\pi_1, \ldots, \pi_n$ be a permutation of the indices $1, \ldots, n$.
If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations, then $Y_1, \ldots, Y_n$ are **exchangeable**.

de Finetti's Theorem $Y_1, Y_2, \ldots$ be a sequence of random variables. If for any n, $Y_1, \ldots, Y_n$ are exchangeable, then there exists a prior distribution $p(\theta)$ and sampling model $p(y \mid \theta)$ such that

$$p(y_1, \ldots, y_n) = \int_\Theta \left\{ \prod_1^n p(y_i \mid \theta) \right\} p(\theta)\, d\theta$$

Models

$$\left.\begin{array}{l} Y_1, \ldots Y_n \mid \theta \overset{\text{iid}}{\sim} p(y \mid \theta) \\ \theta \sim p(\theta) \end{array}\right\} \iff Y_1, \ldots Y_n \text{ are exchangeable for all } n$$

Applicable if $Y_1, \ldots, Y_n$ are

- outcomes of a repeatable experiment
- random sample from finite population with replacement
- sampled from an infinite population w/out replacement
- sampled from a finite population of size $N >> n$ w/out replacement (approximate)

Labels carry no information.

Infectious Disease Example

Does model of exchangeability make sense?