Statistics: Side B

Jason Preszler

10/12/2017

My path into Statistics

1. avoided stats in college, focused on Math and CS
2. first job after PhD had me teaching 4 sections of Math 125
3. enjoyed teaching stats
4. previous job allowed me do lots of statistical modeling and machine learning

My path into Statistics

1. avoided stats in college, focused on Math and CS
2. first job after PhD had me teaching 4 sections of Math 125
3. enjoyed teaching stats
4. previous job allowed me do lots of statistical modeling and machine learning

My path into Statistics

1. avoided stats in college, focused on Math and CS
2. first job after PhD had me teaching 4 sections of Math 125
3. enjoyed teaching stats
4. previous job allowed me do lots of statistical modeling and machine learning

My path into Statistics

1. avoided stats in college, focused on Math and CS
2. first job after PhD had me teaching 4 sections of Math 125
3. enjoyed teaching stats
4. previous job allowed me do lots of statistical modeling and machine learning

My path into Statistics

1. avoided stats in college, focused on Math and CS
2. first job after PhD had me teaching 4 sections of Math 125
3. enjoyed teaching stats
4. previous job allowed me do lots of statistical modeling and machine learning

Frequentist Perspective

- All probabilities are relative frequencies
- *How sample is extracted from population* is only source of uncertainty
- population parameters are fixed, random variables come from sample data
- **Goals:** parameter "estimation" or relationship description

EX: 95% confidence interval

Frequentist Perspective

- All probabilities are relative frequencies
- *How sample is extracted from population* is only source of uncertainty
- population parameters are fixed, random variables come from sample data
- **Goals:** parameter "estimation" or relationship description

EX: 95% confidence interval

Frequentist Perspective

- All probabilities are relative frequencies
- *How sample is extracted from population* is only source of uncertainty
- population parameters are fixed, random variables come from sample data
- **Goals:** parameter "estimation" or relationship description

EX: 95% confidence interval

Frequentist Perspective

- All probabilities are relative frequencies
- *How sample is extracted from population* is only source of uncertainty
- population parameters are fixed, random variables come from sample data
- **Goals:** parameter "estimation" or relationship description

EX: 95% confidence interval

Frequentist Perspective

- All probabilities are relative frequencies
- *How sample is extracted from population* is only source of uncertainty
- population parameters are fixed, random variables come from sample data
- **Goals:** parameter "estimation" or relationship description

EX: 95% confidence interval

Frequentist Perspective

- All probabilities are relative frequencies
- *How sample is extracted from population* is only source of uncertainty
- population parameters are fixed, random variables come from sample data
- **Goals:** parameter "estimation" or relationship description

EX: 95% confidence interval

Frequentist Success and Failure

Failure

- Not logical
- Usually getting samples is costly, why replicate over them?
- interpretation and prediction problems

Success

- good procedures are robust
- parameter estimation
- many successes for traditional experimental data with large effects

Frequentist Success and Failure

Failure

- Not logical
- Usually getting samples is costly, why replicate over them?
- interpretation and prediction problems

Success

- good procedures are robust
- parameter estimation
- many successes for traditional experimental data with large effects

Frequentist Success and Failure

Failure

- Not logical
- Usually getting samples is costly, why replicate over them?
- interpretation and prediction problems

Success

- good procedures are robust
- parameter estimation
- many successes for traditional experimental data with large effects

Frequentist History

- Galton created linear regression in the late 1800's
- K. Pearson added correlation and goodness of fit tests by 1900
- Fisher, E. Pearson, Neyman created hypothesis test and confidence interval procedures in 1920's and 30's.

Frequentist History

- Galton created linear regression in the late 1800's
- K. Pearson added correlation and goodness of fit tests by 1900
- Fisher, E. Pearson, Neyman created hypothesis test and confidence interval procedures in 1920's and 30's.

Frequentist History

- Galton created linear regression in the late 1800's
- K. Pearson added correlation and goodness of fit tests by 1900
- Fisher, E. Pearson, Neyman created hypothesis test and confidence interval procedures in 1920's and 30's.

Bayesian Origins

In 1763, Rev. Thomas Bayes (posthumously) pub. paper containing:

Theorem (Bayes)

$$P(\theta|A) = \frac{P(\theta)P(A|\theta)}{P(A)}$$

$$Posterior = \frac{Prior \cdot Likelihood}{Evidence}$$

Using this for **inference** deeply concerns some people (but not Laplace, Gauss, Turing, or others).

Bayesian Origins

In 1763, Rev. Thomas Bayes (posthumously) pub. paper containing:

Theorem (Bayes)

$$P(\theta|A) = \frac{P(\theta)P(A|\theta)}{P(A)}$$

$$Posterior = \frac{Prior \cdot Likelihood}{Evidence}$$

Using this for **inference** deeply concerns some people (but not Laplace, Gauss, Turing, or others).

Bayesian Origins

In 1763, Rev. Thomas Bayes (posthumously) pub. paper containing:

Theorem (Bayes)

$$P(\theta|A) = \frac{P(\theta)P(A|\theta)}{P(A)}$$

$$Posterior = \frac{Prior \cdot Likelihood}{Evidence}$$

Using this for **inference** deeply concerns some people (but not Laplace, Gauss, Turing, or others).

Bayesian Origins

In 1763, Rev. Thomas Bayes (posthumously) pub. paper containing:

Theorem (Bayes)

$$P(\theta|A) = \frac{P(\theta)P(A|\theta)}{P(A)}$$

$$Posterior = \frac{Prior \cdot Likelihood}{Evidence}$$

Using this for **inference** deeply concerns some people (but not Laplace, Gauss, Turing, or others).

Bayesian Origins

In 1763, Rev. Thomas Bayes (posthumously) pub. paper containing:

Theorem (Bayes)

$$P(\theta|A) = \frac{P(\theta)P(A|\theta)}{P(A)}$$

$$Posterior = \frac{Prior \cdot Likelihood}{Evidence}$$

Using this for **inference** deeply concerns some people (but not Laplace, Gauss, Turing, or others).

Bayesian Perspective

- probabilities represent degrees of belief (subjective)
- must account for uncertainty where it exists, not just sampling method
- data is fixed, parameters are random variables

Result Bayesian inference results in a probability distribution for all possible parameter values.

Bayesian Perspective

- probabilities represent degrees of belief (subjective)
- must account for uncertainty where it exists, not just sampling method
- data is fixed, parameters are random variables

Result Bayesian inference results in a probability distribution for all possible parameter values.

Bayesian Perspective

- probabilities represent degrees of belief (subjective)
- must account for uncertainty where it exists, not just sampling method
- data is fixed, parameters are random variables

Result Bayesian inference results in a probability distribution for all possible parameter values.

Bayesian Perspective

- probabilities represent degrees of belief (subjective)
- must account for uncertainty where it exists, not just sampling method
- data is fixed, parameters are random variables

Result Bayesian inference results in a probability distribution for all possible parameter values.

Bayesian Success and Failure

Failure

- Computationally infeasible until 1990 on most real problems
- How do we get priors?

Success

- Can easily update models as new data arrives
  Today's posterior is tomorrow's prior
- must articulate assumptions
- can answer direct questions, not just falsifications.
- can solve problems frequentists can't!

Bayesian Success and Failure

Failure

- Computationally infeasible until 1990 on most real problems
- How do we get priors?

Success

- Can easily update models as new data arrives
  Today's posterior is tomorrow's prior
- must articulate assumptions
- can answer direct questions, not just falsifications.
- can solve problems frequentists can't!

Bayesian Success and Failure

Failure

- Computationally infeasible until 1990 on most real problems
- How do we get priors?

Success

- Can easily update models as new data arrives
  Today's posterior is tomorrow's prior
- must articulate assumptions
- can answer direct questions, not just falsifications.
- can solve problems frequentists can't!

Bayesian Success and Failure

Failure

- Computationally infeasible until 1990 on most real problems
- How do we get priors?

Success

- Can easily update models as new data arrives
  Today's posterior is tomorrow's prior
- must articulate assumptions
- can answer direct questions, not just falsifications.
- can solve problems frequentists can't!

Bayesian Success and Failure

Failure

- Computationally infeasible until 1990 on most real problems
- How do we get priors?

Success

- Can easily update models as new data arrives
  Today's posterior is tomorrow's prior
- must articulate assumptions
- can answer direct questions, not just falsifications.
- can solve problems frequentists can't!

Overview of Bayesian Inference

All Bayesian methods follow the same procedure:

- develop prior probability distribution for our "hypothesis" or parameter values
- Find likelihood distribution of data given our hypothesis
- use Bayes theorem to construct posterior distribution

We then use the posterior to answer questions directly.

Overview of Bayesian Inference

All Bayesian methods follow the same procedure:

- develop prior probability distribution for our "hypothesis" or parameter values
- Find likelihood distribution of data given our hypothesis
- use Bayes theorem to construct posterior distribution

We then use the posterior to answer questions directly.

Overview of Bayesian Inference

All Bayesian methods follow the same procedure:

- develop prior probability distribution for our "hypothesis" or parameter values
- Find likelihood distribution of data given our hypothesis
- use Bayes theorem to construct posterior distribution

We then use the posterior to answer questions directly.

Overview of Bayesian Inference

All Bayesian methods follow the same procedure:

- develop prior probability distribution for our "hypothesis" or parameter values
- Find likelihood distribution of data given our hypothesis
- use Bayes theorem to construct posterior distribution

We then use the posterior to answer questions directly.

Overview of Bayesian Inference

All Bayesian methods follow the same procedure:

- develop prior probability distribution for our "hypothesis" or parameter values
- Find likelihood distribution of data given our hypothesis
- use Bayes theorem to construct posterior distribution

We then use the posterior to answer questions directly.

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)}$$ $$= 0.165$$

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)}$$ $$= 0.165$$

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)}$$
  $$= 0.165$$

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)} = 0.165$$

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$\begin{aligned} P(D|+) &= \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)} \\ &= 0.165 \end{aligned}$$

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)}$$
  $$= 0.165$$

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)}$$
  $$= 0.165$$

Frequentist use of Bayes

Typical example of drug or medical test

- Suppose everyone is either clean, C, or a drug user D.
- Let $+$ indicate a positive drug test and $-$ indicate a negative drug test
- It is believed that drug use is rare in the population, $P(D) = 0.01$
- Furthermore, the test has good sensitivity so $P(+|D) = 0.98$. The specificity of the test is also good, $P(-|C) = 0.95$.
- $$\begin{aligned} P(D|+) &= \frac{P(D)P(+|D)}{P(+|D)P(D) + P(+|C)P(C)} \\ &= 0.165 \end{aligned}$$

Drug Test Continued

- So using a good test means that there is a 16.5% chance of a being a drug user given a positive test result.
- This is ~ 16 times higher than without the test.
- If we were not looking at a random person, but someone suspected of drug use so $P(D) = 0.5$,then $P(D|+) = 0.95$.

Drug Test Continued

- So using a good test means that there is a 16.5% chance of a being a drug user given a positive test result.
- This is ~ 16 times higher than without the test.
- If we were not looking at a random person, but someone suspected of drug use so $P(D) = 0.5$,then $P(D|+) = 0.95$.

Drug Test Continued

- So using a good test means that there is a 16.5% chance of a being a drug user given a positive test result.
- This is ~ 16 times higher than without the test.
- If we were not looking at a random person, but someone suspected of drug use so $P(D) = 0.5$,then $P(D|+) = 0.95$.

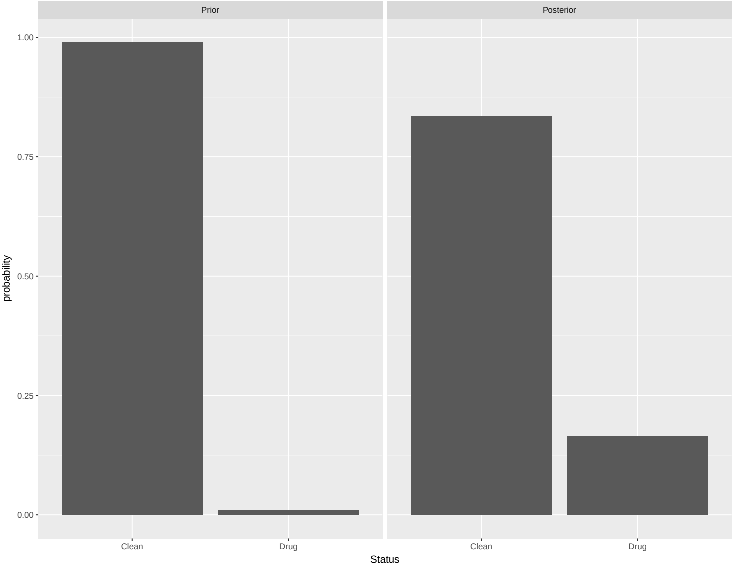
Drug Test Continued

- So using a good test means that there is a 16.5% chance of a being a drug user given a positive test result.
- This is ~ 16 times higher than without the test.
- If we were not looking at a random person, but someone suspected of drug use so $P(D) = 0.5$, then $P(D|+) = 0.95$.

Drug Test Continued

- So using a good test means that there is a 16.5% chance of a being a drug user given a positive test result.
- This is ~ 16 times higher than without the test.
- If we were not looking at a random person, but someone suspected of drug use so $P(D) = 0.5$,then $P(D|+) = 0.95$.

Drug Prior and Posterior



Bayesian Inference Example: Binomial

Situation:
A push rod eye is an essential part of a golf ball washer. We want to know the probability of push rod eyes being defective upon manufacture. The VP of Operations for the manufacturing company believes this probability is between 5% and 15%.

Data:
During a seven day period, 2340 push rod eyes were manufactured and 2211 passed quality control and were shipped.

Situation:

A push rod eye is an essential part of a golf ball washer. We want to know the probability of push rod eyes being defective upon manufacture. The VP of Operations for the manufacturing company believes this probability is between 5% *and* 15%*.*

Data:

During a seven day period, 2340 *push rod eyes were manufactured and* 2211 *passed quality control and were shipped.*

Bayesian Inference Example: Binomial

Situation:

A push rod eye is an essential part of a golf ball washer. We want to know the probability of push rod eyes being defective upon manufacture. The VP of Operations for the manufacturing company believes this probability is between 5% *and* 15%.

Data:

During a seven day period, 2340 *push rod eyes were manufactured and* 2211 *passed quality control and were shipped.*

Bayesian Binomial continued

- The VP's belief provides our prior: Suppose any value in $[.05, .15]$ is equally likely and there's a 100% chance that the VP is correct.
- The data should be binomially distributed with $n = 2340$, and the probability of failure, θ, the parameter to estimate. Let y be the number of failures.
- Write $P(y|\theta) \sim \text{Bin}(2340, \theta)$.

Bayesian Binomial continued

- The VP's belief provides our prior: Suppose any value in $[.05, .15]$ is equally likely and there's a 100% chance that the VP is correct.
- The data should be binomially distributed with $n = 2340$, and the probability of failure, θ, the parameter to estimate. Let y be the number of failures.
- Write $P(y|\theta) \sim \text{Bin}(2340, \theta)$.

Bayesian Binomial continued

- The VP's belief provides our prior: Suppose any value in $[.05, .15]$ is equally likely and there's a 100% chance that the VP is correct.
- The data should be binomially distributed with $n = 2340$, and the probability of failure, θ, the parameter to estimate. Let y be the number of failures.
- Write $P(y|\theta) \sim \text{Bin}(2340, \theta)$.

Binomial Posterior: Grid Approx.

- build "grid" of possible θ values (i.e. $seq(from = 0,\ to = 1,\ length.out = 1000)$)
- Likelihood is Binomial for each possible θ in grid.
- Let's switch to R...

Binomial Posterior: Conjugate Prior

- Our prior was overly simplistic: VP's can be wrong and $\theta = .2$ should be far more likely than $\theta = .8$.
- Better to use

$$\text{Beta}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{0,1}(\theta)$$

where $a = 12.06$ and $b = 116.06$ in this case.
- Then the posterior is $\theta|y \sim \text{Beta}(y+a, n-y+b)$.

Binomial Posterior: Conjugate Prior

- Our prior was overly simplistic: VP's can be wrong and $\theta = .2$ should be far more likely than $\theta = .8$.
- Better to use

$$\text{Beta}(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{0,1}(\theta)$$

where $a = 12.06$ and $b = 116.06$ in this case.

- Then the posterior is $\theta|y \sim \text{Beta}(y + a, n - y + b)$.

Binomial Posterior: Conjugate Prior

- Our prior was overly simplistic: VP's can be wrong and $\theta = .2$ should be far more likely than $\theta = .8$.
- Better to use

$$\mathrm{Beta}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{0,1}(\theta)$$

where $a = 12.06$ and $b = 116.06$ in this case.

- Then the posterior is $\theta|y \sim \mathrm{Beta}(y+a, n-y+b)$.

Methods to Work with Posterior

- Grid approximation is good for small numbers of variables
- Conjugate priors only work in very nice cases
- Best general method: Markov Chain Monte Carlo

Methods to Work with Posterior

- Grid approximation is good for small numbers of variables
- Conjugate priors only work in very nice cases
- Best general method: Markov Chain Monte Carlo

Methods to Work with Posterior

- Grid approximation is good for small numbers of variables
- Conjugate priors only work in very nice cases
- Best general method: Markov Chain Monte Carlo

Machine Learning

- In Machine Learning, need predictive models
- Often data wasn't gathered in nice sample
- Data usually not the result of experiment
- Models need to be updated as new data becomes available.
- Bayesian methods handle these "easily"

Machine Learning

- In Machine Learning, need predictive models
- Often data wasn't gathered in nice sample
- Data usually not the result of experiment
- Models need to be updated as new data becomes available.
- Bayesian methods handle these "easily"

Machine Learning

- In Machine Learning, need predictive models
- Often data wasn't gathered in nice sample
- Data usually not the result of experiment
- Models need to be updated as new data becomes available.
- Bayesian methods handle these "easily"

Machine Learning

- In Machine Learning, need predictive models
- Often data wasn't gathered in nice sample
- Data usually not the result of experiment
- Models need to be updated as new data becomes available.
- Bayesian methods handle these "easily"

Machine Learning

- In Machine Learning, need predictive models
- Often data wasn't gathered in nice sample
- Data usually not the result of experiment
- Models need to be updated as new data becomes available.
- Bayesian methods handle these "easily"

References

- "The Theory that Would not Die" by Sharon McGrayne
- "The Signal and the Noise" by Nate Silver
- ASA p-value Statement
- "Statistical Rethinking" by Richard McElreath
- "Bayesian Methods for Data Analysis" By Carlin and Louis
- "Bayesian Ideas and Data Analysis" By Christensen et al.
- "Bayesian Data Analysis" By Gelman et al.