Foundations

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

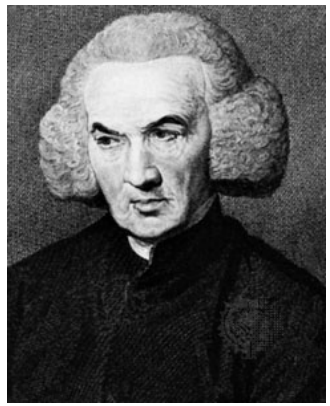
Outline

Outline

History

Bayes?

Price



Laplace

- Thomas Bayes (≈1701-1761) was an ordained minister and a talented mathematician.
- Bayes died before publishing his theorem, but Richard Price carried his work further and published it in 1764.
- Laplace rediscovered essentially the same idea in 1774, and developed it much further.

Bayes' theorem

- Let $x =$ observed data, let $\theta =$ unknown parameters.
- Suppose $p(x|\theta)$ is known for all x and θ.

- Bayes' insight was that if one assumes a *prior* distribution $p(\theta)$, then the conditional distribution $p(\theta|x)$ can be used to quantify uncertainty in θ.
- Bayes' theorem is simply a formula for this conditional:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

  where $p(x) = \int p(x|\theta)p(\theta)d\theta$.
- In words, we say "the posterior is proportional to the likelihood times the prior".

- From the modern perspective, Bayes' theorem is a trivial consequence of the definition of a conditional density—however, when Bayes wrote his paper, the idea of a conditional probability density did not yet exist!

Notation

- $f(x) \propto g(x)$ ("f is proportional to g") means there is a constant c such that $f(x) = cg(x)$ for all x.
- For functions of multiple variables, we write $f(x, y) \underset{x}{\propto} g(x, y)$ to indicate proportionality with respect to x. That is, for any y, there exists c_y such that $f(x, y) = c_y\, g(x, y)$ for all x.
- Proportionality is surprisingly useful in Bayesian statistics.

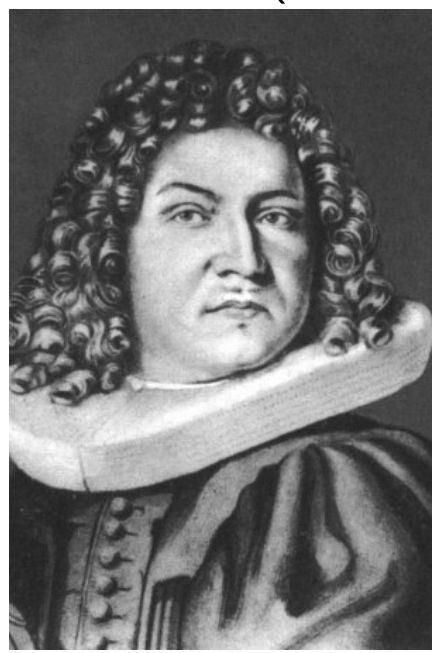
- Usually, we use capital letters to denote random variables (e.g., X) and lowercase for particular values (e.g., x).
- For θ and other greek letters, we use bold $\boldsymbol{\theta}$ to denote the r.v., and unbold θ for particular values. (Note: This is not a standard convention, but I like it.)

- We will usually use p for all p.d.f.s/p.m.f.s, following the usual convention that the symbol used (e.g., θ in the expression $p(\theta)$) indicates which random variable we are talking about.

Outline

Beta-Bernoulli example

Jacob Bernoulli (1655–1705)



(I'm not in a bad mood, everyone is just annoying)

Beta-Bernoulli example: Model/likelihood

- Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, i.e.,

$$p(x_i|\theta) = \mathbb{P}(X_i = x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}\text{I}(x_i \in \{0,1\}).$$

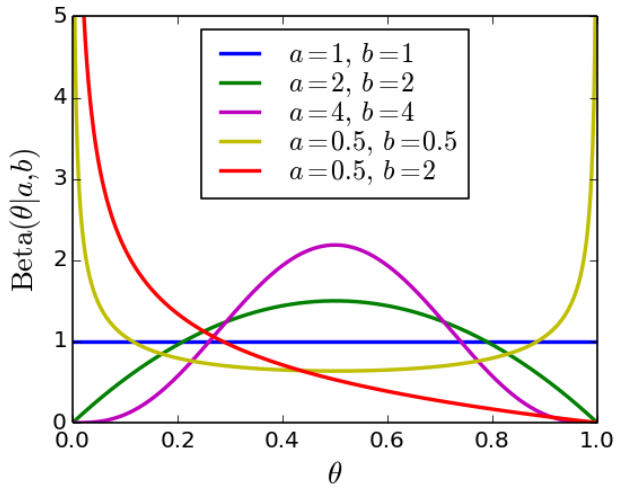
- Notation: The indicator function $\text{I}(E)$ equals 1 when E is true and 0 otherwise.

- For $x_1, \ldots, x_n \in \{0,1\}$, the likelihood function is

$$L(\theta; x_{1:n}) := p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

- Viewed as a function of θ, $p(x_{1:n}|\theta)$ is called the likelihood function. It is sometimes denoted $L(\theta; x_{1:n})$ to emphasize this.

Beta-Bernoulli example: Prior



- We write $\boldsymbol{\theta} \sim \mathrm{Beta}(a, b)$ to indicate that $\boldsymbol{\theta}$ has p.d.f.

$$p(\theta) = \mathrm{Beta}(\theta|a, b) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}\mathrm{I}(0 < \theta < 1).$$

  i.e., $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ on the interval from 0 to 1.

- Here, $B(a, b)$ is Euler's beta function and $a, b > 0$.
- The mean is $\mathrm{E}(\boldsymbol{\theta}) = \int \theta\, p(\theta) d\theta = a/(a+b)$.

Beta-Bernoulli example: Posterior

- Using Bayes' theorem, and plugging in the likelihood and prior, the posterior is

$$p(\theta|x_{1:n}) \propto ???$$

(Whiteboard exercise)

Beta-Bernoulli example: Posterior

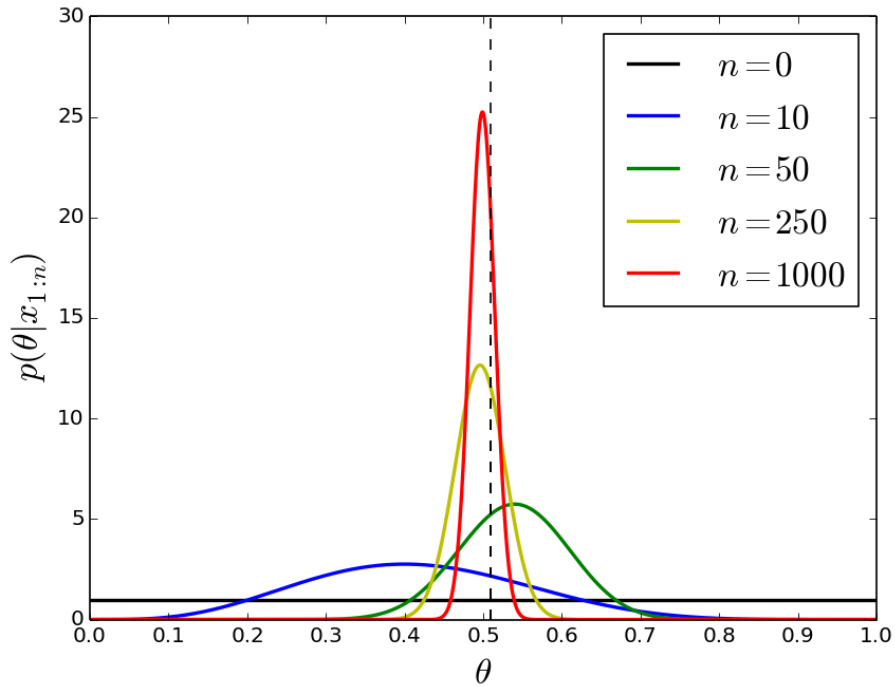
- Using Bayes' theorem, and plugging in the likelihood and prior, the posterior is

$$\begin{aligned} p(\theta|x_{1:n}) &\propto p(x_{1:n}|\theta)p(\theta) \\ &\propto \theta^{a+\sum x_i-1}(1-\theta)^{b+n-\sum x_i-1}\mathrm{I}(0<\theta<1) \\ &\propto \mathrm{Beta}\left(\theta \mid a+\sum x_i,\, b+n-\sum x_i\right). \end{aligned}$$

- So, the posterior has the same form (a Beta distribution) as the prior! When this happens, we say that the prior is *conjugate* (more on this later).

- Since the posterior has such a nice form, it is easy to work with—e.g., for computing certain integrals with respect to the posterior, sampling from the posterior, and computing the posterior p.d.f. and its derivatives.

Beta-Bernoulli example: Simulation



- Simulation: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_0)$ with $\theta_0 = 0.51$.
- Suppose $a = 1$ and $b = 1$, so that the prior is uniform.
- The figure shows the posterior p.d.f. for increasing amounts of data. The dotted line indicates the true value of θ.

Marginal likelihood and posterior predictive

- The *marginal likelihood* is

$$p(x) = \int p(x|\theta)p(\theta)\,d\theta$$

  i.e., the marginal p.d.f./p.m.f. of the observed data, obtained by integrating θ out of the joint density $p(x, \theta) = p(x|\theta)p(\theta)$.

- When the data is a sequence $x = (x_1, \ldots, x_n)$, the *posterior predictive* is the distribution of X_{n+1} given $X_{1:n} = x_{1:n}$.
- When $X_1, \ldots, X_n, X_{n+1}$ are independent given $\boldsymbol{\theta}$, the posterior predictive is

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta|x_{1:n})\,d\theta \\ &= \int p(x_{n+1}|\theta, x_{1:n})p(\theta|x_{1:n})\,d\theta \\ &= \int p(x_{n+1}|\theta)p(\theta|x_{1:n})\,d\theta. \end{aligned}$$

Beta-Bernoulli example: Marginal likelihood

- In the Beta-Bernoulli example, the marginal likelihood is

$$p(x_{1:n}) = ???$$

(Whiteboard exercise)

Beta-Bernoulli example: Marginal likelihood

- In the Beta-Bernoulli example, the marginal likelihood is

$$
\begin{aligned}
p(x_{1:n}) &= \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta \\
&= \frac{1}{B(a,b)} \int_0^1 \theta^{a_n-1}(1-\theta)^{b_n-1} d\theta \\
&= \frac{B(a_n,b_n)}{B(a,b)} \int_0^1 \mathrm{Beta}(\theta \mid a_n, b_n) d\theta \\
&= \frac{B(a_n,b_n)}{B(a,b)}
\end{aligned}
$$

where $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$.

Beta-Bernoulli example: Posterior predictive

- Letting $a_n = a + \sum_{i=1}^n x_i$ and $b_n = b + n - \sum_{i=1}^n x_i$ for brevity, and using the fact that $p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$,

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) &= \int \mathbb{P}(X_{n+1} = 1 \mid \theta) p(\theta|x_{1:n}) d\theta \\ &= \int \theta \ \text{Beta}(\theta|a_n, b_n) = \frac{a_n}{a_n + b_n}.\end{aligned}$$

- Hence, the posterior predictive is

$$p(x_{n+1}|x_{1:n}) = \text{Bernoulli}(x_{n+1} \mid a_n/(a_n + b_n)).$$

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/4uqYnrSQcjwYxVRPA

(Three people per room, randomly assigned. 15 minutes.)

Outline

Bayesian decision theory

- In decision theory, we start with the end in mind—how are we actually going to use our inferences and what consequences will this have?
- Basic goal: minimize loss (or equivalently, maximize utility/gain).
- Multiple ways of making this precise, e.g., minimax, Bayes.
- The standard Bayesian approach is to minimize posterior expected loss.

Bayesian decision theory

Abraham Wald (1902–1950)

"The father of statistical decision theory"

Bayesian decision theory

- General setup
  - $S =$ the state of nature (unknown)
  - $x =$ observation (known)
  - $a =$ action
  - $\ell(S, a) =$ loss incurred for action a when state is S

- Bayesian approach: Choose an action a that minimizes the *posterior expected loss*,

$$\rho(a, x) = \mathrm{E}(\ell(S, a)|x) = \int \ell(s, a) p(s|x) ds.$$

- Here, S is a r.v. and $p(s|x)$ is the posterior of S given x.

- A *decision procedure* δ is a map from x's to a's.
- A *Bayes procedure* δ satisfies $\delta(x) \in \operatorname{argmin}_a \rho(a, x)$.

Example 1: Estimating $\boldsymbol{\theta}$ with quadratic loss

- Setup:
  - State: $S = \boldsymbol{\theta}$
  - Observation: $x = x_{1:n}$
  - Action: $a = \hat{\theta}$
  - Loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (quadratic loss, a.k.a. square loss)

- Using quadratic loss here works out nicely, since the optimal decision is simply to estimate $\boldsymbol{\theta}$ by the posterior mean:

$$\hat{\theta} = \delta(x_{1:n}) = \mathrm{E}(\boldsymbol{\theta} | x_{1:n}).$$

- To see why, note that $\ell(\theta, \hat{\theta}) = \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2$, and thus

$$\rho(\hat{\theta}, x_{1:n}) = \mathrm{E}(\ell(\boldsymbol{\theta}, \hat{\theta}) | x_{1:n}) = ???.$$

  By calculus, the minimum occurs at ???.

(whiteboard)

Example 1: Estimating $\boldsymbol{\theta}$ with quadratic loss

- Setup:
  - State: $S = \boldsymbol{\theta}$
  - Observation: $x = x_{1:n}$
  - Action: $a = \hat{\theta}$
  - Loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (quadratic loss, a.k.a. square loss)

- Using quadratic loss here works out nicely, since the optimal decision is simply to estimate $\boldsymbol{\theta}$ by the posterior mean:

$$\hat{\theta} = \delta(x_{1:n}) = \mathrm{E}(\boldsymbol{\theta}|x_{1:n}).$$

- To see why, note that $\ell(\theta, \hat{\theta}) = \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2$, and thus

$$\rho(\hat{\theta}, x_{1:n}) = \mathrm{E}(\ell(\boldsymbol{\theta}, \hat{\theta})|x_{1:n}) = \mathrm{E}(\boldsymbol{\theta}^2|x_{1:n}) - 2\hat{\theta}\mathrm{E}(\boldsymbol{\theta}|x_{1:n}) + \hat{\theta}^2.$$

  By calculus, the minimum occurs at $\hat{\theta} = \mathrm{E}(\boldsymbol{\theta}|x_{1:n})$.

Example 2: Predicting the next outcome with 0-1 loss

- Assume X_{n+1} is a discrete random variable.

- Setup:
  - State: $S = X_{n+1}$
  - Observation: $x = x_{1:n}$
  - Action: $a = \hat{x}_{n+1}$
  - Loss: $\ell(s, a) = \mathrm{I}(s \neq a)$ (this is called the 0-1 loss)

- Optimal decision: Predict the most probable value according to the posterior predictive, i.e.,

$$\hat{x}_{n+1} = \delta(x_{1:n}) = \operatorname*{argmax}_{x_{n+1}} p(x_{n+1}|x_{1:n}).$$

Real-world decision problems

- Medical decision making
  - When to perform early cancer screening?
- Public health policy
  - How much and what type of flu vaccine to produce each year?
- Government regulations
  - What type of emissions regulations are most effective for improving health outcomes?
- Personal financial decisions
  - Should you buy life insurance?

- A word of caution: In the end, use good judgment!
  - A formal decision analysis is almost always oversimplified, and it's a bad idea to adhere strictly to such a procedure. Decision-theoretic analysis can help to understand a decision problem, but after all the analysis, decisions should be made based on your best judgment.

Outline

Example: Resource allocation for disease treatment

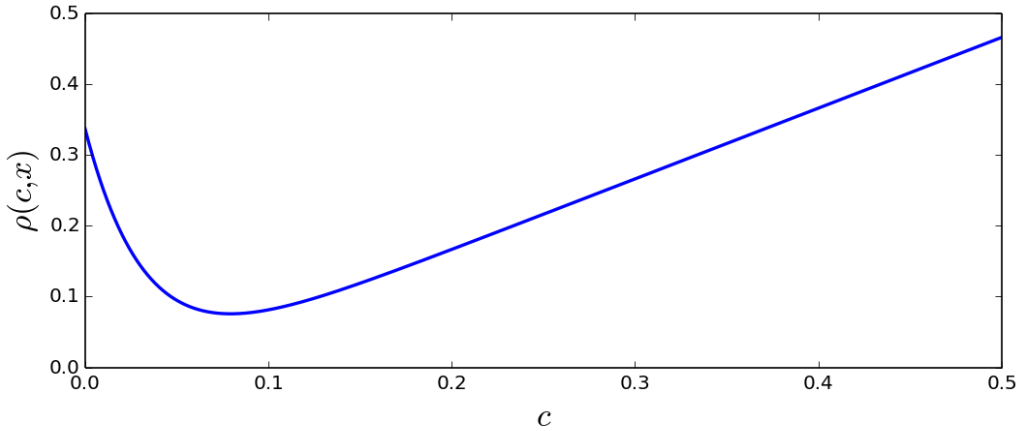
- City health officials need to decide how much money to use for prevention and treatment of a certain disease.
- The fraction θ of affected individuals in the city is unknown.
- Suppose they allocate enough resources to treat a fraction c of the population.
  - If c is too large, there will be wasted resources, while if it is too small, some individuals may go untreated.
- They tentatively adopt the following loss function:

$$\ell(\theta, c) = \begin{cases} |\theta - c| & \text{if } c \geq \theta \\ 10|\theta - c| & \text{if } c < \theta. \end{cases}$$

Example: Resource allocation for disease treatment

- Prior: Based on data from other similar cities, they select a $\text{Beta}(a, b)$ prior with $a = 0.05$ and $b = 1$.
- Data: They conduct a survey of the disease status of $n = 30$ individuals, $x_1, \ldots, x_n$.
- Likelihood/model: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$.
- Suppose that a single surveyed individual is affected, i.e., $\sum_{i=1}^{n} x_i = 1$.

Example: Resource allocation for disease treatment



- The Bayes procedure is to minimize the posterior expected loss

$$\rho(c, x) = \mathrm{E}(\ell(\boldsymbol{\theta}, c)|x) = \int \ell(\theta, c)p(\theta|x)d\theta$$

  where $x = x_{1:n}$. We can numerically compute this integral.
- The minimum of $\rho(c, x)$ occurs at $c \approx 0.08$, so under the assumptions above, this is the optimal amount to allocate.
- This makes more sense than choosing $c = \bar{x} = 1/30 \approx 0.03$, which does not account for the large loss that would result from possible under-resourcing.

Outline

Frequentist risk and Integrated risk

- Consider a decision problem in which $S = \boldsymbol{\theta}$.
- The *risk*, or *frequentist risk*, for a decision procedure δ is

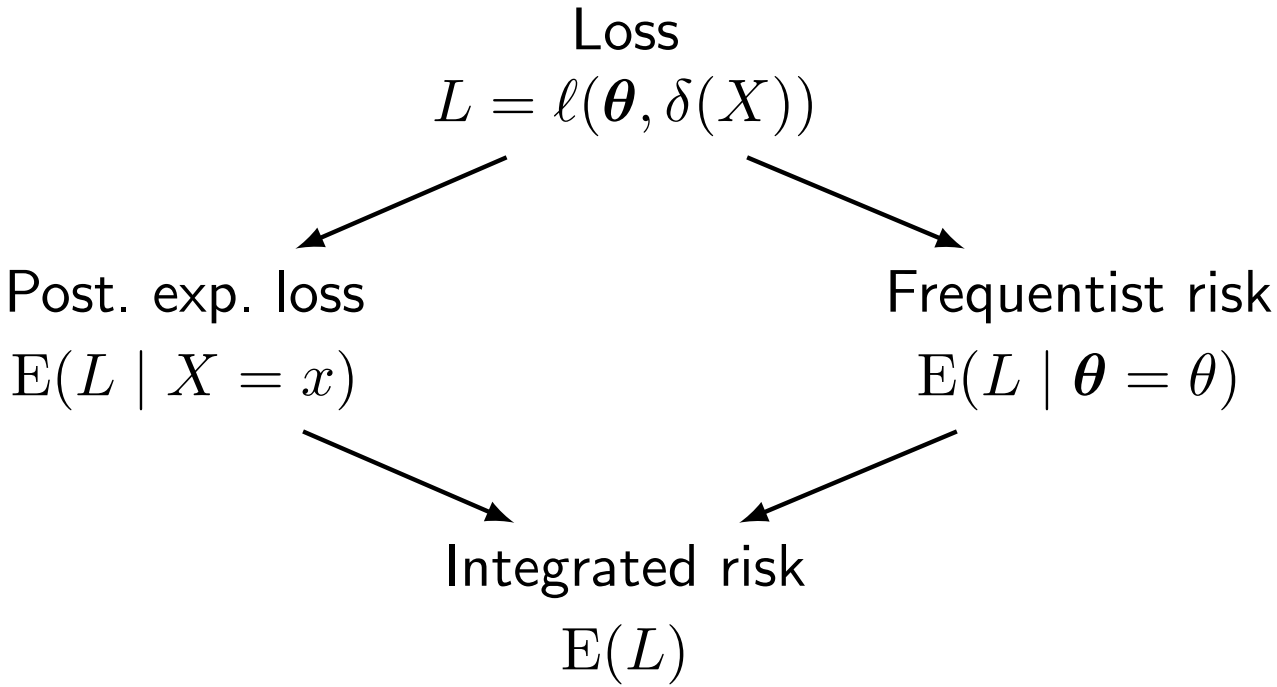
$$R(\theta, \delta) = \mathrm{E}\big(\ell(\boldsymbol{\theta}, \delta(X)) \mid \boldsymbol{\theta} = \theta\big) = \int \ell(\theta, \delta(x))\, p(x|\theta)\, dx.$$

- The *integrated risk* for δ is

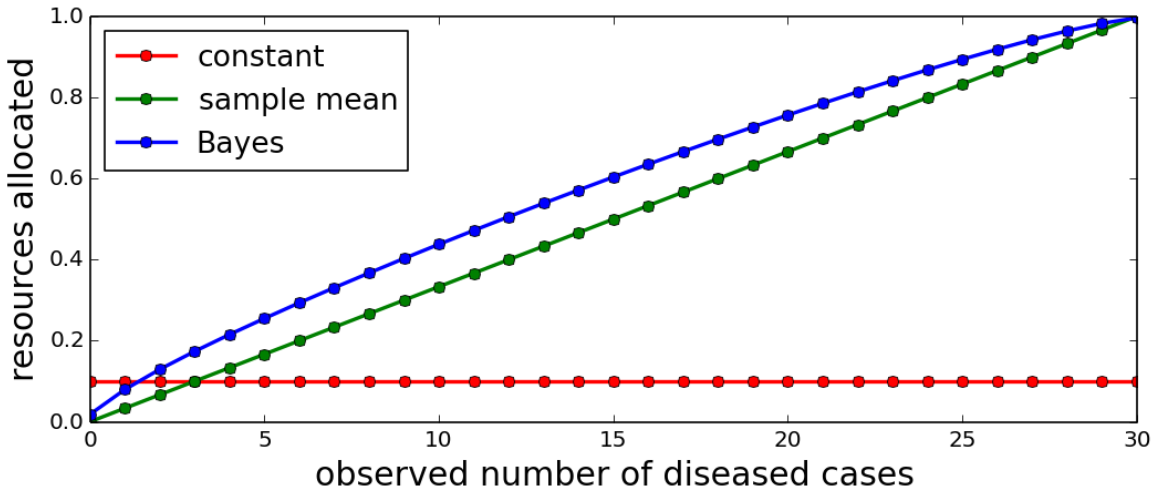
$$r(\delta) = \mathrm{E}(\ell(\boldsymbol{\theta}, \delta(X)) = \int R(\theta, \delta)\, p(\theta)\, d\theta.$$

Relationships between decision-theoretic objects

- Denoting $L = \ell(\boldsymbol{\theta}, \delta(X))$ for brevity, the diagram below visualizes the relationships between all of these concepts.

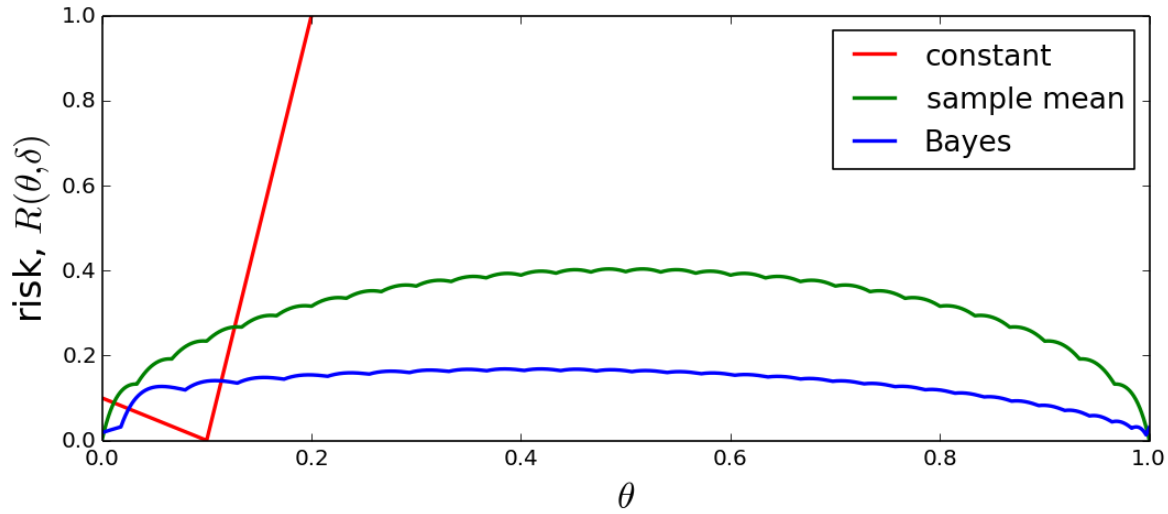


Resource allocation example: Comparing procedures



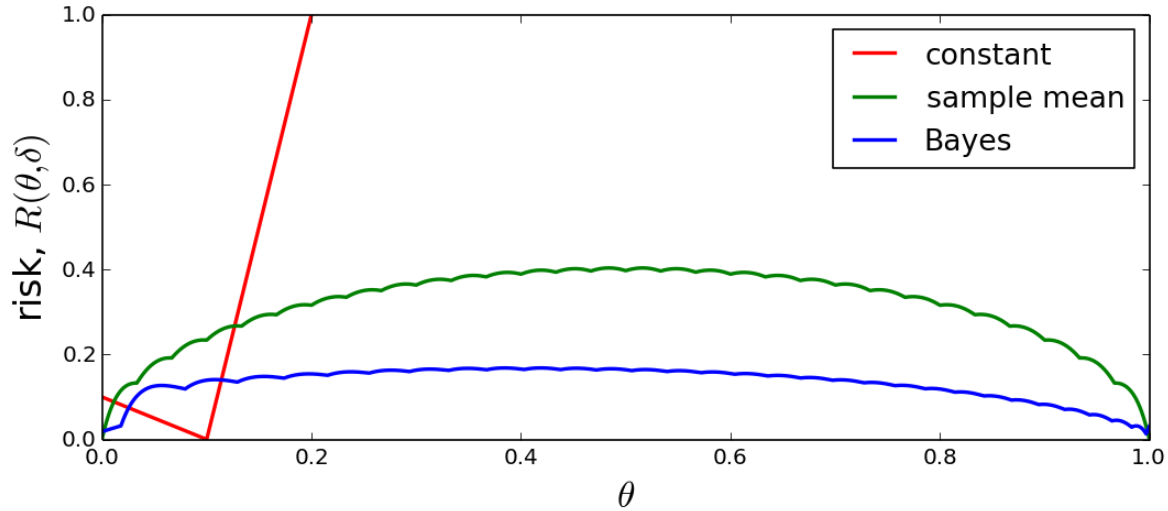
- Compare with two other procedures: choosing $c = \bar{x}$ (sample mean) or choosing $c = 0.1$ (constant).
- The figure shows each procedure as a function of $\sum x_i$, the observed number of affected individuals.
- The Bayes procedure picks c to be a little bigger than $\bar{x}$.

Resource allocation example: Comparing risk curves



- The frequentist risk provides a useful way to compare decision procedures in a prior-free way.
- The figure shows the risk $R(\theta, \delta)$ as a function of θ for each procedure. Smaller risk is better.
- Recall that for each θ, the risk is the expected loss, averaging over all possible data sets. The observed data doesn't factor into it at all.

Resource allocation example: Comparing risk curves



- The constant procedure is great when θ is near 0.1, but gets very bad very quickly for larger θ.
- The Bayes procedure is better than the sample mean for nearly all θ's.
- These curves reflect the usual situation—some procedures will work better for some θ's and some will work better for others.

Admissibility

- Suppose δ and δ' are two decision procedures.
- We say that δ' *dominates* δ if $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for at least one θ.
  - That is, δ' at least as good as δ for all θ and strictly better for some θ.
- A decision procedure is *admissible* if there is no other procedure that dominates it.
- Bayes procedures are admissible under very general conditions.
- Admissibility is nice to have, but it doesn't mean a procedure is necessarily good. E.g., in this example, the constant procedure $c = 0.1$ is admissible too!

References and supplements

- *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Robert, C. P. (2001). Springer Texts in Statistics.
- *Statistical Decision Theory and Bayesian Analysis*. Berger, J.O. (1985). Springer.
- *The History of Statistics: The Measurement of Uncertainty before 1900*. Stigler, S.M. (1986). Harvard University Press.

Individual activity: Exit ticket

Answer these questions individually:
https://forms.gle/faPhDsz5v72NKWk16