DS 102: Data, Inference, and Decisions

Lecture 5

Michael Jordan
University of California, Berkeley

Two Kinds of Statistical Inference

- Bayesian and Frequentist
- Both inferential frameworks are useful
- It's akin to "waves" vs. "particles" in physics
  - they're both correct in some sense
  - they are complementary in many ways
  - but they also conflict in some serious ways
- Understanding Bayes/frequentist relationships can help you become a real problem solver, not just a person who runs downloads software and runs data analysis procedures

Frequentism

- We want to be able to say that a procedure works "on average"
  - or possibly "with high probability"
- Where does the randomness come from to be able to talk about an "average" or a "probability"?
- The frequentist idea (due to Neyman, Wald, and others) is to assume that we don't just have one dataset, but rather we repeatedly draw datasets independently from the population
  - and the randomness comes from this sampling process
  - for example, that's the meaning of the expectation in going from the FDP to the FDR

Bayesianism

- The idea is to condition on the data and consider the posterior distribution of various unknowns conditional on the data

$$P(\theta \mid \text{data}) \propto P(\text{data} \mid \theta) P(\theta)$$

- This updates the prior belief into a posterior belief
- A Bayesian doesn't talk about averages over multiple possible data sets; they want to condition on the observed data
- A Bayesian is happy to assign probabilities to things that can't be repeated

Frequentist Hypothesis Testing

- This is what one learns in classical statistics classes
- The basic idea is to specify, via a probability distribution, what data one expects to see under the null hypothesis
  - and similarly for the alternative hypothesis
- One then collects actual data and assesses, via some algorithm, how well the data fit that null distribution
- If the answer is “not so much,” then one rejects the null
- One then proves that such a decision-making algorithm will perform well on average
  - e.g., having a controlled probability of a Type I error
  - it’s that probability which is a frequentist concept

Bayesian Hypothesis Testing

- Has risen, fallen and risen again many times over history
- The basic idea is to specify, via a probability distribution, what data one expects to see under the null hypothesis and similarly for the alternative hypothesis
- One places a prior probability on the null and the alternative
- One now has all the ingredients to compute a conditional probability of the hypothesis given the data

Comparisons

- Bayesian perspective
  - conditional perspective--inferences should be made conditional on the actual observed data, not on possible data one could have observed
  - natural in the setting of a long-term project with a domain expert
  - the optimist---let’s make the best use possible of our sophisticated inferential tool
- Frequentist perspective
  - unconditional perspective---inferential procedures should give good answers in repeated use
  - natural in the setting of writing software that will be used by many people for many problems
  - the pessimist--let’s protect ourselves against bad decisions given that our inferential procedure is a simplification of reality

Comparisons

- Bayesian perspective
  - conditional perspective--inferences should be made conditional on the actual observed data, not on possible data one could have observed
  - natural in the setting of a long-term project with a domain expert
  - the optimist---let’s make the best use possible of our sophisticated inferential tool
- Frequentist perspective
  - unconditional perspective---inferential procedures should give good answers in repeated use
  - natural in the setting of writing software that will be used by many people for many problems
  - the pessimist--let’s protect ourselves against bad decisions
- Q: Are “bias” and “variance” frequentist or Bayesian?

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator
- Now suppose that someone tells you that the measuring device was broken, and anybody over 7 feet tall was recorded as 7 feet
  - but there actually was no one over 7 feet tall; everyone was actually less than 6.5 feet

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator
- Now suppose that someone tells you that the measuring device was broken, and anybody over 7 feet tall was recorded as 7 feet
  - but there actually was no one over 7 feet tall; everyone was actually less than 6.5 feet
- The right model for the truncated data is a truncated Gaussian, and the sample mean is no longer unbiased under the new model

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator
- Now suppose that someone tells you that the measuring device was broken, and anybody over 7 feet tall was recorded as 7 feet
  - but there actually was no one over 7 feet tall; everyone was actually less than 6.5 feet
- The right model for the truncated data is a truncated Gaussian, and the sample mean is no longer unbiased under the new model
- Should you alter your estimate?
  - consider this question from both a Bayesian and frequentist point of view

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\}$$

$$\delta(X) \in \{0, 1\}$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

		Decision	
		0	1
Reality	0		
	1		

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

Reality \ Decision	0	1
0	$l(0,0)$	$l(0,1)$
1	$l(1,0)$	$l(1,1)$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

Reality \ Decision	0	1
0	0	1
1	1	0

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: L2 loss

$$\theta \in \mathbb{R}$$

$$\delta(X) \in \mathbb{R}$$

$$l(\theta, \delta(X)) = (\delta(X) - \theta)^2$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

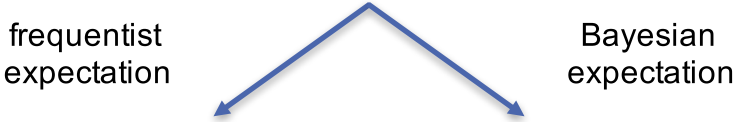
Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

$$l(\theta, \delta(X))$$



Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

$$l(\theta, \delta(X))$$

frequentist expectation

Bayesian expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

$$l(\theta, \delta(X))$$

frequentist expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

Bayesian expectation

$$\mathbb{E}[l(\theta, \delta(X)) \mid X]$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

$$l(\theta, \delta(X))$$

frequentist expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

Bayesian expectation

$$\mathbb{E}[l(\theta, \delta(X)) \mid X]$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

$$l(\theta, \delta(X))$$

frequentist expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

Bayesian expectation

$$\mathbb{E}[l(\theta, \delta(X)) \mid X]$$

Risk Functions

- The frequentist risk:

$$R(\theta) = \mathbb{E}_\theta l(\theta, \delta(X))$$

- The Bayesian posterior risk:

$$\rho(X) = \mathbb{E}[l(\theta, \delta(X)) \mid X]$$

Risk Functions

- The frequentist risk:

$$R(\theta) = \mathbb{E}_\theta l(\theta, \delta(X))$$

- The Bayesian posterior risk:

$$\rho(X) = \mathbb{E}[l(\theta, \delta(X)) \mid X]$$

- A fun bonus exercise: If we take an expectation of $R(\theta)$ with respect to θ, or an expectation of $\rho(X)$ with respect to X, we get a constant known as the "Bayes risk"

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$R(\theta) = \mathbb{E}_\theta[l(\theta, \delta(X))]$$

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \mathbb{E}_\theta[(\delta(X) - \theta)^2]
\end{aligned}
$$

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \mathbb{E}_\theta[(\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X) + \mathbb{E}_\theta\delta(X) - \theta)^2]
\end{aligned}
$$

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \mathbb{E}_\theta[(\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X) + \mathbb{E}_\theta\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))^2] + 2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)] + \mathbb{E}_\theta[(\mathbb{E}_\theta\delta(X) - \theta)^2]
\end{aligned}
$$

The Cross-Product Vanishes

$$2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)]$$

-

The Cross-Product Vanishes

$$2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)]$$

$$= 2(\mathbb{E}_\theta\delta(X) - \theta)\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))]$$

The Cross-Product Vanishes

$$
\begin{aligned}
&2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)] \\
&= 2(\mathbb{E}_\theta\delta(X) - \theta)\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))] \\
&= 2(\mathbb{E}_\theta\delta(X) - \theta)[\mathbb{E}_\theta\delta(X) - \mathbb{E}_\theta\delta(X)]
\end{aligned}
$$

The Cross-Product Vanishes

$$
\begin{aligned}
&2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)] \\
&= 2(\mathbb{E}_\theta\delta(X) - \theta)\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))] \\
&= 2(\mathbb{E}_\theta\delta(X) - \theta)[\mathbb{E}_\theta\delta(X) - \mathbb{E}_\theta\delta(X)] \\
&= 0
\end{aligned}
$$

- Essentially this is just orthogonality, and the risk decomposition on the previous page is the Pythagorean theorem…

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \mathbb{E}_\theta[(\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X) + \mathbb{E}_\theta\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))^2] + 2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)] + \mathbb{E}_\theta[(\mathbb{E}_\theta\delta(X) - \theta)^2]
\end{aligned}
$$

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \mathbb{E}_\theta[(\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X) + \mathbb{E}_\theta\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))^2] + 2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)] + \mathbb{E}_\theta[(\mathbb{E}_\theta\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))^2] + (\mathbb{E}_\theta\delta(X) - \theta)^2
\end{aligned}
$$

Example: Frequentist Risk Under L2 Loss

- The loss: $l(\theta, \delta(X)) = (\delta(X) - \theta)^2$
- Expanding out the frequentist risk:

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \mathbb{E}_\theta[(\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X) + \mathbb{E}_\theta\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))^2] + 2\mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))(\mathbb{E}_\theta\delta(X) - \theta)] + \mathbb{E}_\theta[(\mathbb{E}_\theta\delta(X) - \theta)^2] \\
&= \mathbb{E}_\theta[(\delta(X) - \mathbb{E}_\theta\delta(X))^2] + (\mathbb{E}_\theta\delta(X) - \theta)^2 \\
&= \text{variance} \ + \text{bias}^2
\end{aligned}
$$

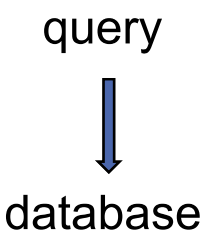
Consequences of this Decomposition

- Lots of frequentist statistics involves analyzing the bias and the variance of various procedures
- Generally speaking, the bias and the variance trade off
  - i.e., when one adjusts some tuning knob of the procedure to decrease the variance, the bias increases, and vice versa
- The classical statistical approach was again to formulate inference as a constrained optimization problem
  - e.g., consider only estimators that have zero bias and then minimize the variance
  - this approach has become less prominent over the years
  - e.g., Bayesian and empirical Bayesian procedures generally are biased
  - but they have lower variance
- So modern frequentist analysis usually tries to characterize this tradeoff, and it makes use of Bayesian ideas to find good trade offs
  - as you've hopefully understood, FDR is a great example of this!

Privacy and Data Analysis

- Individuals are not generally willing to allow their personal data to be used without control on how it will be used and now much privacy loss they will incur
- “Privacy loss” can be quantified via differential privacy
- We want to trade privacy loss against the value we obtain from data analysis
- The question becomes that of quantifying such value and juxtaposing it with privacy loss
- We’ll have an entire section on privacy later in the course, but let’s make some initial comments here

Privacy



Privacy

query

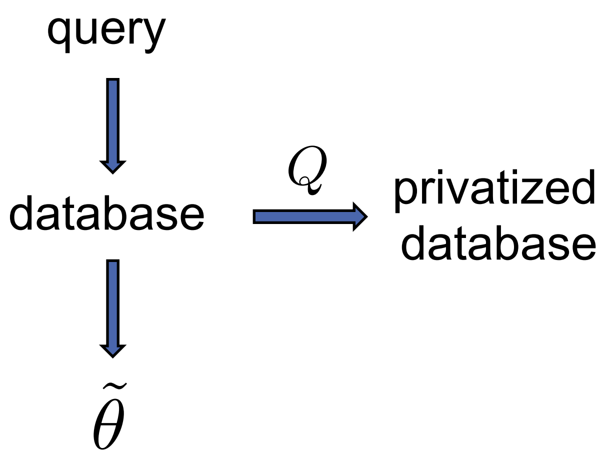
↓

database

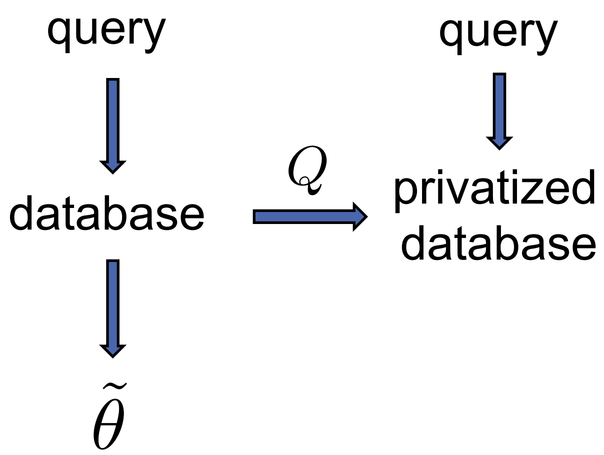
↓

$\tilde{\theta}$

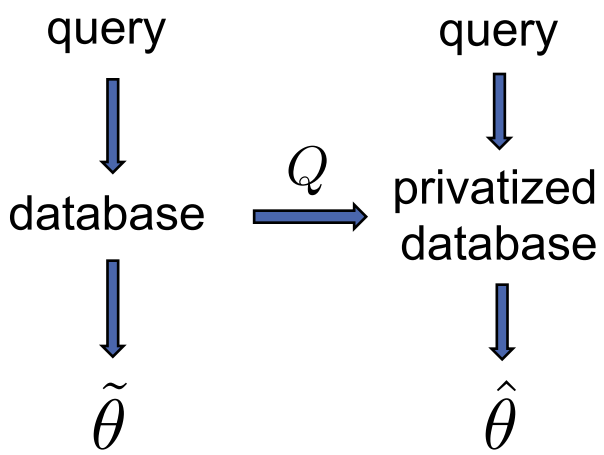
Privacy



Privacy

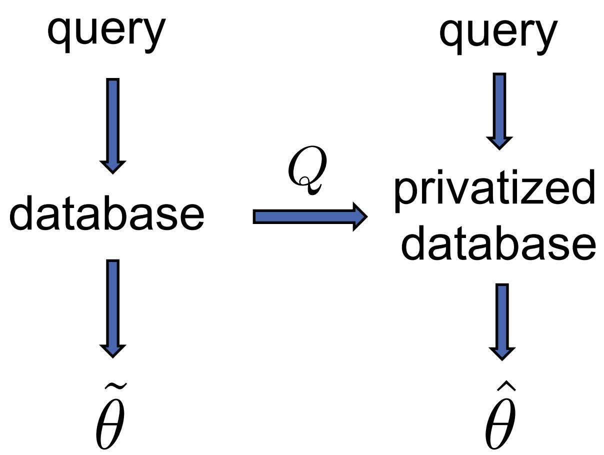


Privacy



query
query
Q
database
privatized database
$\tilde{\theta}$
$\hat{\theta}$

Privacy



Q is a “noisy channel”

Classical problem in differential privacy: show that $\hat{\theta}$ and $\tilde{\theta}$ are close under constraints on Q

Inference

query

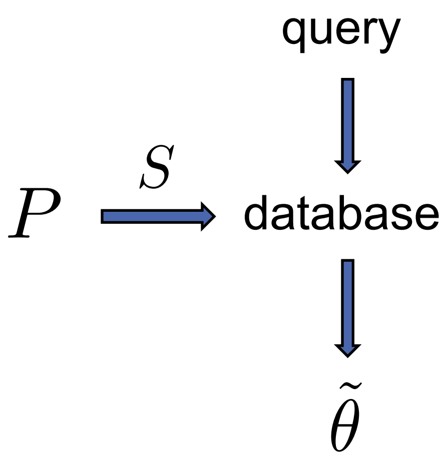
↓

database

↓

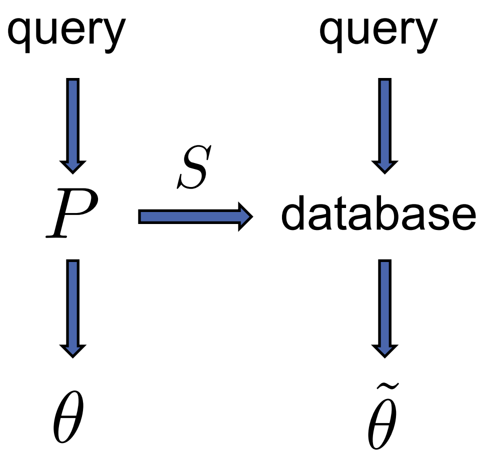
$\tilde{\theta}$

Inference

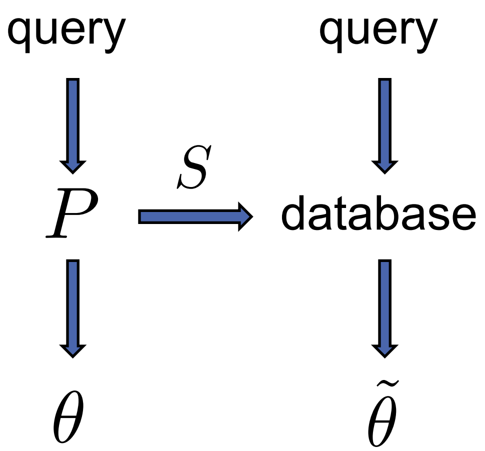


S is the sampling process

Inference

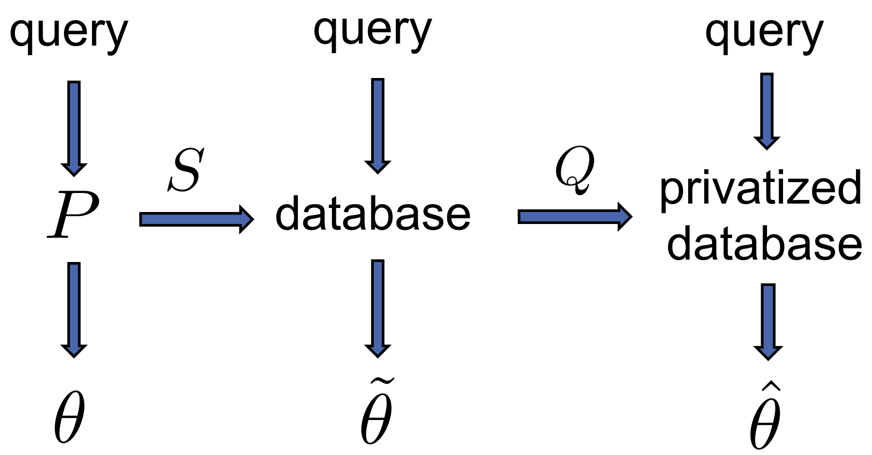


Inference



Classical problem in statistical theory: show that $\tilde{\theta}$ and θ are close under constraints on S

Privacy and Inference



The privacy-meets-inference problem: show that θ and $\hat{\theta}$ are close under constraints on Q and on S