DS 102: Data, Inference, and Decisions

Lecture 2

Michael Jordan

University of California, Berkeley

Basics of Decision Making

- We'll start by considering the most simple of decision-making formulations
- Let's suppose that Reality is in one of two states, which we denote as 0 or 1
- We don't observe this state, but we do obtain Data that is drawn from a distribution that depends on whether the state is 0 or 1
- We make a Decision based on the Data, which we denote as 0 or 1
- We can think of the Decision as our best guess as to the state of Reality or, more generally, as an action we think is best given our guess of the state of Reality

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0		
	1		

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

The Basic Two-by-Two Table

Reality \ Decision	0	1
0	TN	FP
1	FN	TP

TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

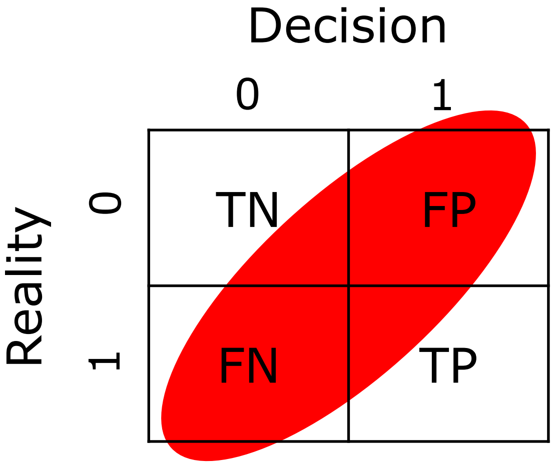
TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

The Basic Two-by-Two Table



TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

Rough goal: lots of green outcomes, few red outcomes!

Examples: How Serious are FP and FN (and How Desirable are TP and TN)?

- Medical: 0 = no disease, 1 = disease
- Commerce: 0 = no fraud, 1 = fraud
- Physics: 0 = no Higgs boson, 1 = Higgs boson
- Social network: 0 = no link, 1 = link
- Self-driving car: 0 = no pedestrian, 1 = pedestrian
- Search: 0 = not relevant, 1 = relevant
- Oil-Well Drilling: 0 = no oil, 1 = oil

- In real-world domains, there are many, many complications that arise

Towards a Statistical Framework

- Although the two-by-two table is useful conceptually, it’s not clear how to make use of it in a real problem, because we don’t know Reality
- We need to move towards a statistical framework, where we consider not just one decision, but a set of related decisions

Towards a Statistical Framework

- Let's now imagine that we not only make a decision, but we build a decision-making algorithm
- We want to evaluate the algorithm not just on one problem, but on a set of related problems

Towards a Statistical Framework

- Let's now imagine that we not only make a decision, but we build a decision-making algorithm
- We want to evaluate the algorithm not just on one problem, but on a set of related problems

- Concretely, we may have a collection of hypothesis-testing problems, where we repeatedly decide whether to accept the null or accept the alternative
- Or we may have a set of classification decisions, where we repeatedly classify data points into one of two classes

Towards a Statistical Framework

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$N = n_{00} + n_{01} + n_{10} + n_{11}$$

Towards a Statistical Framework

- Our language will start to involve rates and probabilities

- Indeed, the variables n_{00}, n_{01}, n_{10}, and n_{11} are random variables
- In just what sense they are random will need to be made clear (e.g., is the state of Reality random, is the Decision random, is N random?)

Some Row-Wise Rates

Decision

Reality		0	1
	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{sensitivity} = \frac{n_{11}}{n_{10}+n_{11}}$$

Some Row-Wise Rates

		Decision 0	Decision 1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{sensitivity} = \frac{n_{11}}{n_{10} + n_{11}}$$

aka, “true positive rate” or “recall” or “power”

Some Row-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{specificity} = \frac{n_{00}}{n_{00}+n_{01}}$$

Some Row-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{specificity} = \frac{n_{00}}{n_{00}+n_{01}}$$

aka, "true negative rate"
or "selectivity"

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
  - e.g., sensitivity approximates $P(\mathrm{Decision} = 1 \mid \mathrm{Reality} = 1)$

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
  - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$
- As such, they are not dependent on the prevalence (i.e., the probabilities of the two states of Reality in the population)

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
  - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$
- As such, they are not dependent on the prevalence (i.e., the probabilities of the two states of Reality in the population)
- They are the kinds of quantities that are the focus of Neyman-Pearson inferential theory, which we'll review later
  - specificity = 1 – Type I error rate
  - sensitivity = 1 – Type II error rate = power

Towards Inference

- We’d like to have have high sensitivity and high specificity
  - but in general there is a tradeoff (see whiteboard drawings)
  - we have to figure out how to manage the tradeoff

Towards Inference

- We'd like to have have high sensitivity and high specificity
  - but in general there is a tradeoff (see whiteboard drawings)
  - we have to figure out how to manage the tradeoff
- Neyman and Pearson (1932) formulated this problem as a constrained optimization problem:
  - maximize the power while constraining the false-positive rate to be under some fixed number (e.g., .05)
  - we're smudging over the distinction between probabilities and rates, which we'll clarify later
  - a very fruitful idea, and sometimes the right idea, but not to be viewed as written in stone

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

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, with 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

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
- One thresholds that probability to make the decision

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?

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

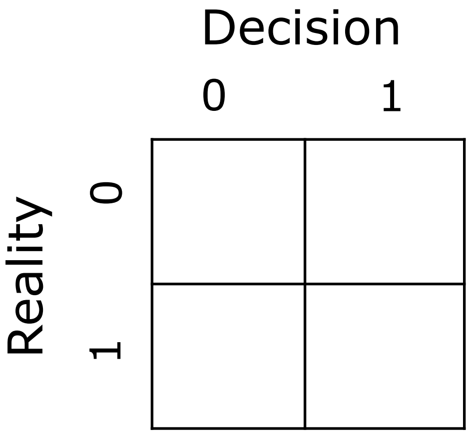
- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

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

- Example: 0/1 loss

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

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



Decision-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	$l(0,0)$	$l(0,1)$
	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

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

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

		Decision	
		0	1
Reality	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

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

frequentist
expectation

Bayesian
expectation

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

Bayesian expectation

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

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

$$\begin{tikzcd} & l(\theta, \delta(X)) \arrow[dl, "\text{frequentist expectation}"'] \arrow[dr, "\text{Bayesian expectation}"] & \\ \mathbb{E}_\theta l(\theta, \delta(X)) & & \mathbb{E}[l(\theta, \delta(X)) \mid X] \end{tikzcd}$$

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

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

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"

Examples (on the White Board)

- The risk under the 0/1 loss
- The risk under the L2 loss

Comparisons

- 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

Back to Hypothesis Testing

- Let's now consider a column-wise perspective

Back to Hypothesis Testing

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Back to Hypothesis Testing

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

- Let's now consider a column-wise perspective

Some Column-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

false discovery rate $= \dfrac{n_{01}}{n_{01}+n_{11}}$

Some Column-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{false omission rate} = \frac{n_{10}}{n_{00}+n_{10}}$$

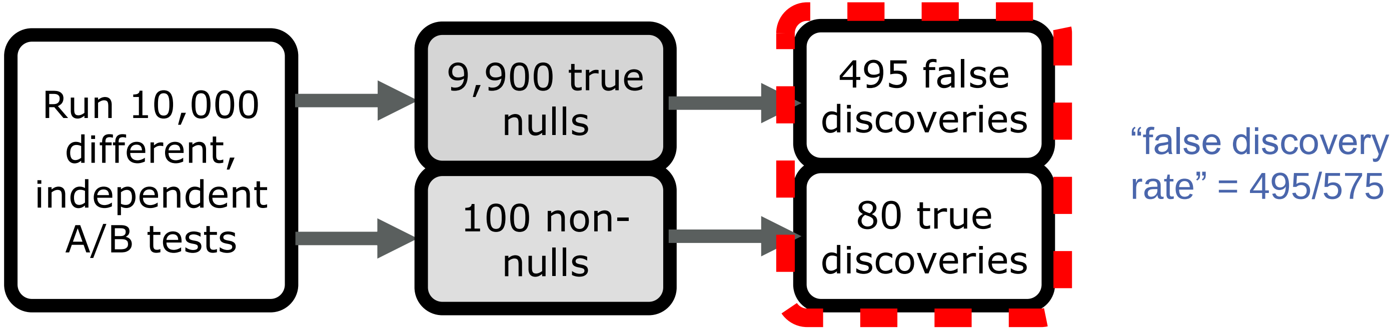
Comments on the Column-Wise Rates

- They can be thought of as estimates of conditional probabilities
  - e.g., false discovery rate approximates $P(\text{Reality} = 0 \,|\, \text{Decision} = 1)$
- They are dependent on the prevalence (i.e., the probabilities of the two states of Reality in the population), via Bayes' Theorem
  - as such, they are more Bayesian
- This is arguably a good thing, as we'll see on the next slide

A Bayesian Calculation

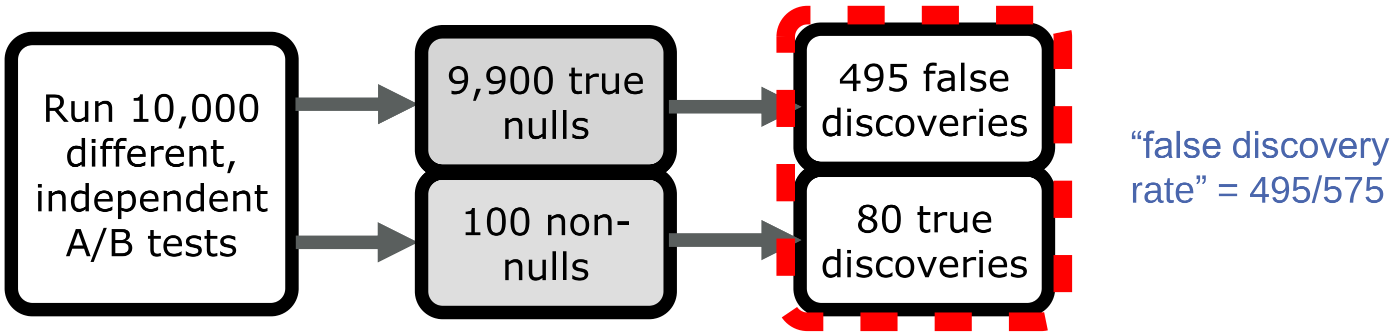
- Let’s calculate on the white board

Type I error rate (per test) = 0.05



Power (per test) = 0.80

Type I error rate (per test) = 0.05



Power (per test) = 0.80

(NB: We’re again not being rigorous at this point; FDR is actually an expectation of this proportion. We’ll do it right anon.)

Back to Inference

- Can we develop general frameworks that allow us to control column-wise quantities like the false-discovery rate (FDR)?
  - in a similar way as Neyman-Pearson controls the false-positive rate
- To be continued…