STATS 300A: Theory of Statistics **Fall 2015**

Lecture 1 — September 22

Lecturer: Lester Mackey *Scribe: Amelia Lemionet, Ken Wang*

Warning: *These notes may contain factual and/or typographic errors.*

1.1 The Big Picture

Consider the following flowchart for statistics:

$$\left.\begin{array}{l}\text{natural sciences (biology, physics, climate, etc)}\\ \text{social sciences (economics, politics, etc)}\\ \text{engineered systems (networks, images, etc)}\end{array}\right\} \longrightarrow \text{data} \longrightarrow \text{statistics} \longrightarrow \text{inferences}$$

That is, as statisticians, we are tasked with turning the large amount of data generated by experiments and observations into inferences about the world. This simple directive gives rise to a number of core statistical questions:

1. *Modeling*: How do we capture the uncertainty in our data and the world that produced it?

2. *Methodology*: What are the right mathematical and computational tools that allow us to draw these statistical inferences?

3. *Analysis*: How do we compare and evaluate the statistical inferences we make and the procedures we use to make them? In particular, how do we do optimal inference?

Many of the classes in our department are focused on these questions but answer them in slightly different ways:

- Our department's introductory applied sequence, consisting of Stats 305, 306A, and 306B, explores many of the *empirical* and *applied* aspects of these questions, with a particular focus on *methodology* and *modeling*.

- Stats 300A and B focus on developing rigorous *mathematical* answers to these questions with a strong focus on notions of *optimality* of the statistical inference. In 300A we will develop finite sample answers to the core questions while 300B delivers asymptotic answers (letting the sample size $n \to \infty$).

1.2 Decision Theory

1.2.1 Framework

We will address our core questions within a framework for statistical inference developed by Abraham Wald in 1939 called **decision theory**. This decision theoretic framework will give us a way to answer all of the core questions. Hereafter, we will view our data as the realization of a random element X taking values in a sample space $\mathcal{X}$. Often $\mathcal{X}$ will be a subset of the Euclidean space, so X will be a vector (or matrix) $(X_1, \ldots, X_n)$ with i.i.d. (independent and identically distributed) entries (or columns).

Now, let us formalize the notion of inference as a **decision problem** consisting of three key ingredients:

1. A **statistical model** is a family of distributions $\mathcal{P}$, indexed by a parameter θ. We write

   $$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}.$$

   Here θ is the parameter, Ω is the parameter space, and each $\mathbb{P}_\theta$ is a distribution. Often $\Omega \subset \mathbb{R}^k$.

   $\mathcal{P}$ is the class of distributions to which we believe X belongs. In other words, we assume that the data X come from some $\mathbb{P}_\theta \in \mathcal{P}$ but that the true θ is unknown. The fact that we don't know θ captures our uncertainty about the problem.

   Example 1 (Weighted coin flips)**.** Observe a sequence of coin flips $X_1, \ldots, X_n \in \{0, 1\}$ where 0 encodes tails and 1 encodes heads. It's a weighted coin, so I don't know how often I expect heads to arrive. The goal is to estimate the probability of heads given the observations. Then we model this process as independent draws from a Bernoulli distribution: $\mathcal{P} = \{\text{Ber}(\theta) : \theta \in [0, 1] = \Omega\}$. In this case, $\mathbb{P}_\theta(X_i = 1) = \theta$.

2. A **decision procedure** δ is a map from $\mathcal{X}$ (the sample space) to the decision space $\mathcal{D}$.[1]

   Example 2 (Weighted coin flips)**.** Taking $\mathcal{P} = \{\text{Ber}(\theta)\}$ as before, we may be interested in estimating θ or testing hypotheses based on θ.

   (a) Estimating θ: the decision space is $\mathcal{D} = [0, 1]$, and the decision procedure might be $\delta(X) = \frac{1}{n}\sum_{i=1}^n X_i$. This procedure is an example of an **estimator.**

   (b) Accepting or rejecting the hypothesis $\theta > 1/2$: the decision space is $\mathcal{D} = \{\text{accept}, \text{reject}\}$, and one possible decision procedure is $\delta(X) =$ "reject if $\frac{1}{n}\sum_{i=1}^n X_i \le 1/2$, accept otherwise". This procedure is an example of a **hypothesis test**.

3. A **loss function** is a mapping $L : \Omega \times \mathcal{D} \to \mathbb{R}^+$. $L(\theta, d)$ represents the penalty for making the decision d when θ is in fact the true parameter for the distribution generating the data. The goal is to assign high penalties for bad decisions.

   Example 3 (Squared-error loss)**.** For estimating a real-valued parameter θ with decision $d \in \mathbb{R} = \mathcal{D}$, a common loss function is the squared-error loss $L(\theta, d) = (\theta - d)^2$.

[1] We will also have reason to consider randomized decision procedures, functions δ^* which map the data X and an independent random variable $U \sim \text{Unif}[0, 1]$ to a value $\delta^*(X, U) \in \mathcal{D}$.

1.2.2 Analyzing Procedures

Decision theory is useful because it allows us to analyze statistical procedures. Indeed, the three components of a decision problem together give rise to our primary basis for evaluation, the **risk function** $R(\theta, \delta) = E_\theta(L(\theta, \delta(X)))$.[2] The risk $R(\theta, \delta)$ is the average loss incurred when the decision procedure δ is used over many draws of the data from its generating distribution $\mathbb{P}_\theta$.

The risk function gives us a way to compare and rule out procedures. We say a procedure δ is **inadmissible** if another procedure never has greater risk than δ but sometimes has strictly lower risk. In other words, δ is inadmissible if there exists δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta', \delta') < R(\theta', \delta)$ for some θ'. So, δ' is always as good as δ and sometimes better. Decision theory rules out inadmissible procedures δ in favor of dominating procedures δ'.

This very bold statement should be taken with a grain of salt because there are cases when you might want to use an inadmissible procedure. For example, an explicit dominating procedure may be unknown or much more expensive to compute.

Example 4 (Weighted coin flips)**.** For estimating the probability of heads θ, let $\delta_n(X) = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean of the first n data points. Under the loss function $L(\theta, d) = (\theta - d)^2$, the risk of δ_n is

$$R(\theta, \delta_n) = E_\theta((\theta - \delta_n(X))^2) = \frac{\theta(1-\theta)}{n}.$$

This is computed by realizing that the expectation is just expressing the variance of a binomial distribution for the coin flips. Now, we can compare different decision procedures by considering holding out the data from various coin flips. In particular, when we have two flips, we can just use one of them: $R(\theta, \delta_1) = \theta(1-\theta)$, which is always higher than using both in our procedure $R(\theta, \delta_2) = \theta(1-\theta)/2$, so δ_1 is inadmissible.

Now consider the constant estimator $\delta_{\text{goofy}} = \frac{1}{2}$, which has risk $R(\theta, \delta_{\text{goofy}}) = (\frac{1}{2} - \theta)^2$. Since δ_{goofy} achieves a risk of 0 when $\theta = \frac{1}{2}$, it will be admissible. Nonetheless, it is intuitively unreasonable since it doesn't use the data at all; it also has unacceptably high risk for values of θ near 0 or 1.

The following figure plots the risk of the three estimators δ_1, δ_2, and δ_{goofy}.

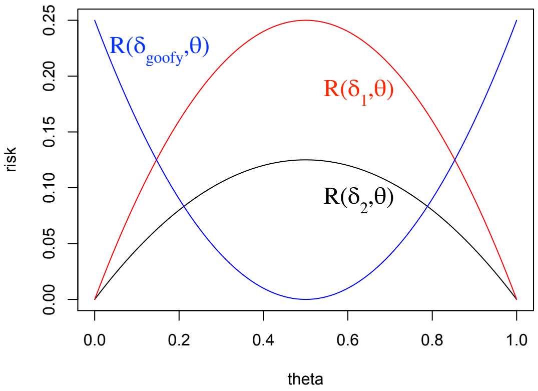
The lesson to take away from this is that there is typically no uniformly best procedure. Nonetheless, we can develop our theory of optimality by changing the requirements of our decision problem. Here are some common actions taken to induce an optimizable problem:

1. **Constrain** the set of decision procedures under consideration, by requiring our procedures to satisfy criteria like unbiasedness or invariance.

   (a) Unbiased estimators: we say that δ is unbiased for estimating $g(\theta)$ if $E_\theta(\delta(X)) = g(\theta)$.

   (b) Equivariance or invariance: enforce symmetries in the decision procedure. For example, location invariance requires an estimator to satisfy $\delta(X + c) = \delta(X) + c$.

[2] The expectation is taken over the data X with θ held fixed. In the case of randomized estimators $\delta^*(X, U)$, the expectation is also taken over the auxiliary randomness in U.



2. **Collapse** the risk function into a single numerical summary and minimize this overall summary of risk instead of requiring uniformly lower risk.

   (a) Bayes procedures minimize the average risk $\int R(\theta, \delta) d\Lambda(\theta)$ where Λ is a probability distribution (the prior distribution) over Ω.

   (b) Minimax procedures minimize the worst-case risk, $\sup_{\theta \in \Omega} R(\theta, \delta)$, and hence achieve the best worst-case performance.

In this course we will explore each of these principles for optimal inference, first in the context of point estimation and later in the context of hypothesis testing.

1.3 Data Reduction

Before constraining or collapsing, let's attend to a more basic fact that will aid us in the design of optimal procedures: **Not all data is relevant** to a particular decision problem. We will see that discarding irrelevant data can never hurt performance and results in a simpler inference procedure. To understand data reduction, let's introduce the following two definitions.

Definition 1 (Statistic)**.** A statistic $T : \mathcal{X} \to \mathcal{T}$ is a function of the data.

Definition 2 (Sufficient Statistic)**.** A statistic is sufficient for a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$ if for all t, the conditional distribution $X|T(x) = t$ does not depend on θ.

Let's take a look at an example of a sufficient statistic.

Example 5 (Weighted coin flips)**.** Let $X_1, X_2, ..., X_n$ be i.i.d. according to $Ber(\theta)$, is the number of heads, i.e. $\sum_{i=1}^n X_i$, sufficient? The answer is yes. To see that, let's show the conditional distribution does not depend on θ. First of all, we have

$$\mathbb{P}_\theta(X = (X_1, X_2, ..., X_n)) = \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_i X_i}(1-\theta)^{n-\sum_i X_i}$$

So the conditional distribution is

$$\begin{aligned}\mathbb{P}_\theta(X=x|T(x)=t) &= \frac{\mathbb{P}_\theta(X=x,T(x)=t)}{\mathbb{P}_\theta(T(x)=t)}\\ &= \frac{\mathbb{1}(t=\sum_{i=1}^n X_i)\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}\\ &= \frac{\mathbb{1}(t=\sum_{i=1}^n X_i)}{\binom{n}{t}}\end{aligned}$$

which does not depend on θ, so the sum of heads is a sufficient statistic.

Two other examples of sufficient statistic are:

Example 6 (Max of Uniform)**.** Let $X_1, X_2, ..., X_n$ be i.i.d. according to uniform distribution $U(0,\theta)$. Then $T(x) = \max(X_1, ..., X_n)$ is sufficient. The intuition behind this is the following: think of $X_1, X_2, \cdots, X_n$ as n numbers on the real line, then the remaining $n-1$ numbers, given the maximum is fixed at t, behave like $n-1$ i.i.d random samples drawn from $U(0,t)$. Since this conditional distribution is independent of θ, $T(x)$ is sufficient.

Example 7 (Order Statistics)**.** Let $X_1, X_2, ..., X_n$ be i.i.d. with any model. Then the *order statistics* $T = X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are sufficient. To see this, note that given T the possible values of X are the $n!$ permutations of T. By symmetry, we can see that each of these permutations has equal probability of $\frac{1}{n!}$. Thus the conditional distribution $X|T(x) = t$ is independent of θ. Therefore the order statistics are sufficient, regardless of the model.

From the viewpoint of decision theory, data reduction via a sufficient statistic represents *lossless* data compression: any risk curve that can be achieved by a decision procedure based on X can also be achieved by a (possibly randomized) decision procedure based on $T(X)$. This is made precise in the following theorem.

Theorem 1. (TPE 1.6, Theorem 6.1) If $X \sim \mathbb{P}_\theta \in \mathcal{P}$ and T is sufficient for $\mathcal{P}$, then, for any decision procedure δ, there is a (possibly randomized) decision procedure of equal risk that depends on X only through $T(X)$.

To see why the theorem is true, note that given an independent source of randomness U, we can always sample a new dataset $X' = f(T(X), U)$ from the conditional distribution $P(X \mid T(X))$ and define a randomized procedure

$$\delta^*(X,U) \triangleq \delta(f(T(X),U)) = \delta(X') \overset{d}{=} \delta(X).$$

The equality in distribution implies that $\hat{\hat{\delta}}$ and δ have equal risk, and this procedure is valid since, by sufficiency, $X \mid T(X)$ does not depend on θ.

In practice it is seldom necessary to regenerate a dataset from sufficient statistics to achieve accurate inference; rather, we will see in a future lecture that the risk of a decision procedure can often be matched or improved upon by a *non-randomized* decision procedure based on sufficient statistics alone.

1.3.1 The Neyman-Fisher Factorization Criterion

Checking the definition of sufficiency directly is often a tedious exercise. A much simpler characterization of sufficiency is available whenever our model distributions admit densities (w.r.t. a common σ-finite measure).

Theorem 2 (Neyman-Fisher Factorization Criterion (NFFC), TSH, p. 19)**.** Suppose each $\mathbb{P}_\theta \in \mathcal{P}$ has density $p(x;\theta)$ w.r.t. a common σ-finite measure μ, i.e., $\frac{d\mathbb{P}_\theta}{d\mu} = p(x;\theta)$. Then $T(X)$ is sufficient if and only if $p(x;\theta) = g_\theta(T(x))h(x)$ for some g_θ, h.

In other words, a necessary and sufficient condition for $T(X)$ to be sufficient is that the density $p(x;\theta)$ can be factorized into two factors where the first factor may depend on θ but depends on x only through $T(x)$ while the second factor is independent of θ.

Example 8 (i.i.d. Normal)**.** Let X_i be i.i.d. $N(\mu,\sigma^2)$ and $\theta = (\mu,\sigma^2)$. The joint distribution is

$$\begin{aligned} p(x;\theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{1}{2\sigma^2}(-\sum_{i=1}^n X_i^2 + 2\mu\sum_{i=1}^n X_i - n\mu^2)} \\ &= g_\theta(T(X)) \end{aligned}$$

where $T(X) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is sufficient.

Now, let's move on to the proof of NFFC. Here we will just prove the discrete case.

Proof. (Discrete Case)

Suppose $p(x;\theta) = g_\theta(T(x))h(x)$. Since $\mathbb{P}_\theta(X = x|T(X) = t) = 0$ whenever $t \neq T(x)$, so we may focus our attention on conditionals of the form $\mathbb{P}_\theta(X = x|T(X) = T(x))$. We have

$$\begin{aligned} \mathbb{P}_\theta(X = x|T(X) = T(x)) &= \frac{\mathbb{P}_\theta(X = x, T(X) = T(x))}{\mathbb{P}_\theta(T(X) = T(x))} = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T(X) = T(x))} \\ &= \frac{g_\theta(T(x))h(x)}{\sum_{x'\in\mathcal{X}} p(x';\theta)\mathbb{1}(T(x') = T(x))} \\ &= \frac{g_\theta(T(x))h(x)}{\sum_{x'\in\mathcal{X}} g_\theta(T(x))h(x')\mathbb{1}(T(x') = T(x))} \\ &= \frac{h(x)}{\sum_{x'\in\mathcal{X}} h(x')\mathbb{1}(T(x') = T(x))} \end{aligned}$$

which has no θ dependence, so T is sufficient.

Conversely, suppose $\mathbb{P}_\theta(X = x|T(X) = T(x))$ is independent of θ. Then, defining $h(x) \triangleq \mathbb{P}_\theta(X = x|T(X) = T(x))$, we have

$$\begin{aligned} p(x;\theta) = \mathbb{P}_\theta(X = x) &= \mathbb{P}_\theta(X = x, T(X) = T(x)) \\ &= \mathbb{P}_\theta(X = x|T(X) = T(x))\mathbb{P}_\theta(T(X) = T(x)) \\ &= h(x)g_\theta(T(x)), \end{aligned}$$

which establishes the factorization criterion. □

1.3.2 Summary

Let's wrap up this section by summarizing the benefits of data reduction:

1. Data reduction via sufficient statistics never impairs our risk, while (we will see that) irrelevant attributes can in fact lead to increased risk.

2. Data reduction can increase interpretability.

3. Data reduction generally reduces storage requirements and often reduces the subsequent computational costs of inference.

Note however that reduction via sufficiency can also increase the computational complexity of inference, in some instances even turning a computationally tractable inference problem into an intractable one. See *Montanari* (2014) for examples of this counterintuitive phenomenon.

References

1. Montanari, A. (2014). Computational implications of reducing data to sufficient statistics. arXiv preprint arXiv:1409.3821.